JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 12 (2024) ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Viewpoints

Original Papers

Case Report

Implementation Reports

Corrigenda and Addendas

Research Letters

Electronic Health Record Data Quality and Performance Assessments: Scoping Review

Yordan P Penev[1,2], MTM; Timothy R Buchanan[1,2], BS; Matthew M Ruppert[1,2,3], MS; Michelle Liu[1], BS; Ramin Shekouhi[1], MD; Ziyuan Guan[1,2], MS; Jeremy Balch[4], MD; Tezcan Ozrazgat-Baslanti[1,2], PhD; Benjamin Shickel[1,2], PhD; Tyler J Loftus[2,4], MD, PhD; Azra Bihorac[2,5], MS, MD

1

2

3

4

5

Corresponding Author:
Azra Bihorac, MS, MD

Abstract

Background: Electronic health records (EHRs) have an enormous potential to advance medical research and practice through easily accessible and interpretable EHR-derived databases. Attainability of this potential is limited by issues with data quality (DQ) and performance assessment.

Objective: This review aims to streamline the current best practices on EHR DQ and performance assessments as a replicable standard for researchers in the field.

Methods: PubMed was systematically searched for original research articles assessing EHR DQ and performance from inception until May 7, 2023.

Results: Our search yielded 26 original research articles. Most articles had 1 or more significant limitations, including incomplete or inconsistent reporting (n=6, 30%), poor replicability (n=5, 25%), and limited generalizability of results (n=5, 25%). Completeness (n=21, 81%), conformance (n=18, 69%), and plausibility (n=16, 62%) were the most cited indicators of DQ, while correctness or accuracy (n=14, 54%) was most cited for data performance, with context-specific supplementation by recency (n=7, 27%), fairness (n=6, 23%), stability (n=4, 15%), and shareability (n=2, 8%) assessments. Artificial intelligence–based techniques, including natural language data extraction, data imputation, and fairness algorithms, were demonstrated to play a rising role in improving both dataset quality and performance.

Conclusions: This review highlights the need for incentivizing DQ and performance assessments and their standardization. The results suggest the usefulness of artificial intelligence–based techniques for enhancing DQ and performance to unlock the full potential of EHRs to improve medical research and practice.

(JMIR Med Inform 2024;12:e58130) doi:10.2196/58130

KEYWORDS

electronic health record; EHR; record; data quality; data performance; clinical informatics; performance; data science; synthesis; review methods; review methodology; search; scoping

Introduction

The adoption of electronic health records (EHRs) optimistically promises easily searchable databases as an accessible means for prospective and retrospective research applications [1]. EHRs often fall short of these promises due to limited local data and poor data quality (DQ) [2,3]. To overcome these limitations, several institutions have harmonized databases and model ontologies, including PCORnet (The National Patient-Centered Clinical Research Network), All of Us, MIRACUM (Medical Informatics in Research and Care in University Medicine), and the EHDEN Project [4-7]. These programs strive to offer high-quality data for research purposes [2]. However, EHR DQ remains highly variable, with some studies showing completeness in EHR parameter values ranging from 60% to 100% [8,9]. Similar inconsistencies present a significant limitation to the generalizability and applicability of lessons learned across these datasets for broader medical and research purposes.

Multiple initiatives have aimed to measure and improve EHR data [10,11]. Early efforts in DQ assessment (DQA) demonstrated inconsistent reporting and a need for universal

terminology standards in DQA efforts [11]. In response, attempts at a standardized ontology for DQA have been developed, such as through the efforts of the International Consortium for Health Outcomes Measurement, 3×3 DQA guidelines, and the terminologies proposed by Kahn et al [12] and Wang et al [8,12-15]. More recently, artificial intelligence (AI) and natural language processing techniques have automated quality initiatives, including data assessment and augmentation [16,17]. Nonetheless, these techniques introduce their own set of quality requirements, including fairness metrics, handling intolerable or lost data, and mitigating data drift [18]. Measuring the result of these techniques' application in real-world clinical contexts has given rise to another field that has become crucial for EHR improvement, namely, data performance assessment (DPA) [19].

In this review, we critically evaluate peer-reviewed literature on the intersection of DQA and DPA applications, as well as trends in their automation [10-13,20-22]. The purpose of this scoping review was to combine the 3 to formulate a more clear road map for evaluating EHR datasets for medical research and practice.

Methods

Overview

This scoping literature review was conducted according to the 2018 PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews), whose checklist is shown in Checklist 1 [23].

Literature Search

A search was performed for all full-text research articles published in English in PubMed from inception to May 7, 2023. A list of the exact search terms is included in Multimedia Appendix 1.

Article Selection

Four investigators (JB, RS, TRB, and YPP) reviewed the selected studies during the title and abstract screening. Further 4 investigators (ML, RS, TOB, and YPP) conducted the full-text review and final extraction of articles. Title or abstract screening, full-text review, and final extraction were based on the consensus opinion between 2 independent reviewers. Conflicts were resolved by a third reviewer. Article management and calculations of interrater reliability and Cohen κ were performed using Covidence systematic review software (Veritas Health Innovation).

Inclusion Criteria

Titles and abstracts were screened to include original research articles assessing the DQ and performance of all or part of a hospital's EHR system. We looked for studies reporting on 1 or more aspects of DQ (the assessment of EHR data without consideration of follow-up actions) and data performance (the assessment of EHR data applications) as defined (Table 1).

Table . Data quality and performance indicator definitions, mitigation strategies, and references.

	Definition	Mitigation strategies	Relevant studies
Data quality			
Completeness (or, conversely, missingness)	The absence of data points, without reference to data type or plausibility [12]	Automated data extraction; data imputation	[2-6,8,9,24-37]
Conformance	The compliance of data with expected formatting, relational, or absolute definitions [12]	Preemptively enforced data format standardization	[2-6,8,14,24-27,29-33,36,38]
Plausibility	The possibility that a value is true given the context of other variables or temporal sequences (ie, patient date of birth must precede date of treatment or diagnosis) [12]	Periodic realignment with logic rule sets or objective truth standards; thresholding	[4-6,8,14,25,27,28,30-33,35,37-39]
Uniqueness	The lack of duplicate data among other patient records [8]	Two-level encounter or visit data structure	[8]
Data performance			
Correctness or accuracy	Whether patient records are free from errors or inconsistencies when the information provided in them is true [10,13]	Periodic validation against internal and external gold standards	[2,7-9,14,23,24,28]
Currency or recency	Whether data were entered into the EHR[a] within a clinically relevant time frame and is representative of the patient state at a given time of interest [10,13]	Enforcing predetermined hard and soft rule sets for timeline of data entry	[2,4,9,27,32,34,36]
Fairness (or, conversely, bias)	The degree to which data collection, augmentation, and application are free from unwarranted over- or underrepresentation of individual data elements or characteristics	Periodic review against a predetermined internal gold standard or bias criterion	[3,19,22,24,27,35]
Stability (or, conversely, temporal variability)	Whether temporally dependent variables change according to predefined expectations [10,12]	Periodic measurement of data drift against a baseline standard of data distribution	[4,8,19,31]
Shareability	Whether data can be shared directly, easily, and with no information loss [3]	Preemptively enforced data standardization	[2,3]
Robustness	The percent of patient records with tolerable (eg, inaccurate, inconsistent, and outdated information) versus intolerable (eg, missing required information) data quality problems [24]	Timely identification of critical data quality issues	[24]

[a]EHR: electronic health record.

Data Quality

Conformance

Conformance refers to the compliance of data with expected formatting, relational, or absolute definitions [12].

Plausibility

Plausibility refers to the possibility that a value is true given the context of other variables or temporal sequences (ie, the patient's date of birth must precede the date of treatment or diagnosis) [12].

Uniqueness

Uniqueness refers to the lack of duplicated records [8].

Completeness (or Conversely, Missingness)

With regard to completeness, *missingness* is the absence of requested data points, without reference to conformance or plausibility as defined [12].

Data Performance

Correctness or Accuracy

Correctness or *accuracy* refers to whether patient records are free from errors or inconsistencies when the information provided in them is true [10,13].

Currency or Recency

Currency or *recency* refers to whether data were entered into the EHR within a clinically relevant time frame and are representative of the patient state at a given time of interest [10,13].

Fairness (or Conversely, Bias)

With regard to bias, *fairness* refers to the degree to which data collection, augmentation, and application are free from unwarranted over- or underrepresentation of individual data elements or characteristics.

Stability (or Conversely, Temporal Variability)

With regard to stability, *temporal variability* refers to whether temporally dependent variables change according to predefined expectations [10,12].

Shareability

Shareability refers to whether data can be shared directly, easily, and with no information loss [3].

Robustness

Robustness refers to the percent of patient records with tolerable (eg, inaccurate, inconsistent, and outdated information) versus intolerable (eg, missing required information) DQ problems [24].

We additionally included studies reporting on data imputation methods, defined as techniques used to fill in missing values in an EHR, such as through statistical approximation and the application of AI.

Exclusion Criteria

We excluded tangential analyses of DQ in articles focused primarily on clinical outcomes. As such, studies discussing data cleaning as part of quantifying clinical outcomes were excluded from our analysis. Proposals or study protocols with no results were also excluded during the screening process.

Article Quality Assessment

Full-text articles were additionally scored as having or missing the criteria for (1) data integrity: comprehensiveness for each main outcome, including attrition and exclusions from the analysis and reasons for them; (2) method clarity: a clear description of DQA data sources, analysis steps, and criteria; (3) outcome clarity: outcomes reporting in plain language, in their entirety, and without evidence for selective reporting; and (4) generalizability: applicability of DQ techniques described in the article to other clinical settings.

Results

Article Characteristics

The flow diagram for article selection is shown in Figure 1. A total of 154 records were identified using the search terms defined in Multimedia Appendix 1 using the PubMed library. After the removal of 31 duplicates and the 72 articles identified as irrelevant, 51 studies proceeded to full-text review. Full-text review excluded a further 25 articles owing to reasons listed in Figure 1, leaving a final total of 26 original research studies [2-6,8,9,14,19,22,24-39]. The Cohen κ between the different pairs of reviewers ranged from 0.28 to 0.54 during the screening process and from 0.54 to 1.00 during the full-text review.

Study characteristics are shown in Table 2 and Multimedia Appendix 2. Exactly half of the identified articles targeted general EHR DQ analysis [4-6,19,22,27-32,38,39], while the other half focused on a particular specialty or diagnosis (Table 2) [2,3,8,9,14,24-26,33-37]. The latter included primary care (n=3, 12%) [35-37], cardiovascular disease (n=3, 12%) [8,33,34], anesthesia or pain medicine (n=2, 8%) [14,26], intensive care units (n=2, 8%) [3,25], and pediatrics [24], oncology [2], and infectious disease (n=1 each, 4%) [9].

Article quality assessment conducted as part of our review process identified 14 (54%) of the articles [2-6,8,9,19,22,24-36,38,39] had at least 1 common study design or reporting limitation, with 5 of the articles having more than 1 [14,24,33,36,38]. Among these, 6 (30% of all errors) articles did not clearly state their methods [3,27,28,33,36,39], 5 (25%) had incomplete data [24,29,33,36,38], 5 were not generalizable to other settings [4,24-26,33], and 4 did not clearly state their outcomes (Table 2) [31,34,38,39].

Commonly referenced DQ and performance indicators are summarized in Table 3. Respective definitions, mitigation strategies, and references are listed in Table 1.

Figure 1. PRISMA 2020 flow diagram detailing study selection and reasons for exclusion for all articles considered for this scoping review. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

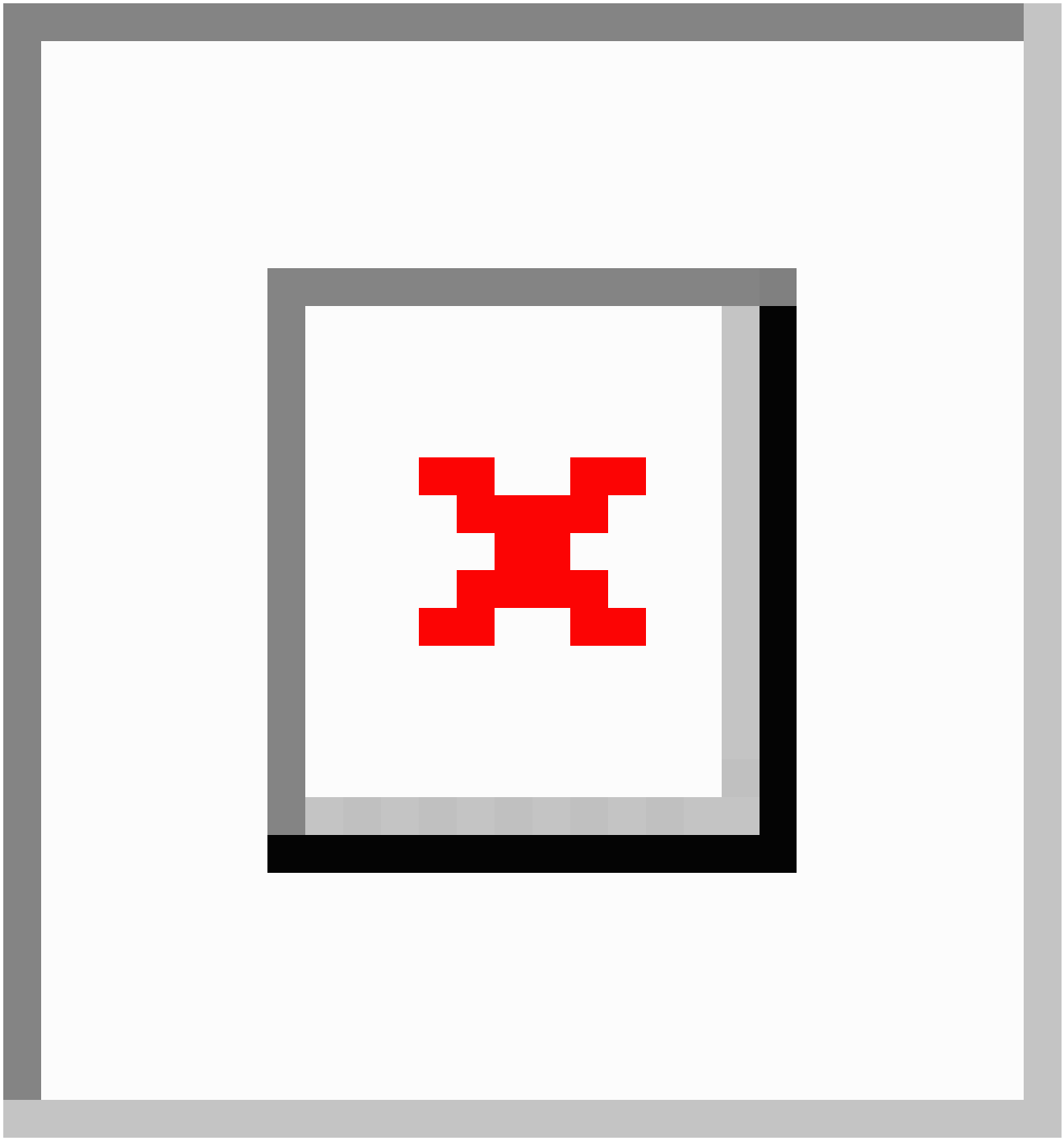


Table . Frequency of clinical specialties among all papers and study limitations among all limitations identified by reviewers in this analysis.

Setting		Values, n (%)
Specialty		
	ICU[a]	2 (8)
	Anesthesia or pain med	2 (8)
	General	13 (50)
	Cardiovascular	3 (12)
	Infectious disease	1 (4)
	Oncology	1 (4)
	Pain medicine	0 (0)
	Pediatrics	1 (4)
	Primary care	3 (12)
Limitations		
	Incomplete data	5 (25)
	Methods not clearly stated	6 (30)
	Outcomes not clearly stated	4 (20)
	Not generalizable to other settings	5 (25)

[a]ICU: intensive care unit.

Table . Elements of data quality and performance commonly referenced by papers included in this review.

Data Quality and Performance Element		Values, n (%)
Data quality		
	Completeness	21 (81)
	Conformance	18 (69)
	Plausibility	16 (62)
	Uniqueness	1 (4)
Data performance		
	Correctness or accuracy	14 (54)
	Currency	7 (27)
	Fairness or bias	6 (23)
	Stability	4 (15)
	Shareability	2 (8)
	Robustness	1 (4)

Data Quality Assessment

Completeness

Completeness was the most cited element of DQ analysis, with references in 21 (81%) of all articles [2-6,8,9,24-37]. Importantly, 19 (73%) studies integrated data from multiple clinical sites [2,4-6,9,19,22,24,26,30-39], which was associated with issues in data collection and missingness "across organizational structure, regulation, and data sourcing" [31]. Clinical domains reported to be prone to low data completeness included patient demographics, with Estiri et al [29] highlighting the issue for records of patient ethnicity and Thuraisingam et al [35] for mortality records (eg, missing year of death), and medication management, with Thuraisingam et al [35] highlighting the issue for dosage, strength, or frequency of prescriptions and Kiogou et al [34] for missing dates or reasons for discontinuation of medications.

To combat data missingness, Lee et al [22] used natural language processing algorithms to automatically extract data from patient records, while further 5 studies made use of data imputation techniques. Among the latter, 2 articles generated synthetic data, while another 3 supplemented datasets through information from external datasets. Fu et al [3] generated synthetic data by modeling providers' assessments of EHR data based on different information sources according to their individual characteristics (eg, tendency to ascertain delirium status based on Confusion

Assessment Method vs prior *International Statistical Classification of Diseases* coding or nursing flow sheet documentation), while Zhang et al [19] used a generative adversarial network (GAN) trained on real longitudinal EHR data to create single synthetic EHR episodes (eg, outpatient or inpatient visit). Meanwhile, Lee et al [33] supplemented existing EHR records on heart failure by aggregating data from open-source datasets of heart failure biomarkers (including the Database of Genotypes and Phenotypes and the Biologic Specimen and Data Repository Information Coordinating Center) and using literature guidelines to create a standard set of cardiovascular outcome measures, while Curtis et al [2] supplemented missing EHR mortality records with data from US Social Security Death Index and the National Death Index, and Mang et al [30] used a manually generated stand-alone synthetic dataset to test the development of a new software tool for DQ assessment.

Conformance

Conformance was the second most cited element of DQA, with references in 18 (69%) articles [2-6,8,14,24-27,29-33,36,38]. Similar to completeness, DQ checks on conformance were performed automatically across most studies. Mitigation strategies included enforcing strict formatting rules at the time of data entry, for example, by using *International Statistical Classification of Diseases* codes to define the cause of death or a diagnosis of delirium [2,3].

Plausibility

Plausibility was the third most cited element of DQA with references in 16 (62%) articles [4-6,8,14,25,27,28,30-33,35,37-39]. Clinical domains prone to issues with plausibility included patient baseline physical characteristics, medication, and laboratory records. Estiri et al [29] and Wang et al [39] reported significant rates of plausibility issues for baseline physical characteristics, with higher error rates for records of patient height as compared to weight, likely due to the multiple flow sheet fields for height, including "estimated," "reported," and "measured," which are generally averaged or selectively dropped. Pharmacologic data were prone to issues with plausibility due to timeliness (eg, antiretroviral therapy was dispensed before or more than 30 days after the visit date [9]) or discrepancies between diagnoses and drugs (eg, nonsteroidal anti-inflammatory drug prescription on the date of gastroduodenal ulcer diagnosis [6]). Finally, laboratory results were also prone to issues with plausibility due to value ranges, units, timing (eg, laboratory time was at an invalid time of day or in the future), and discrepancies between diagnoses and laboratory records (eg, drug was documented as present but there was no laboratory record) or drug prescriptions and laboratory records (eg, metformin was prescribed prior to a documented hemoglobin A_{1c} laboratory result, or warfarin was prescribed without a follow-up international normalized ratio laboratory result) [6]. Notably, this may reflect poorly integrated health care systems where laboratories are being drawn at disparate institutions.

A total of 18 (69%) studies used logic statements to assess plausibility [2,4-6,8,9,14,24,27,28,31-38], including rules to determine temporal plausibility (eg, laboratories drawn at an invalid time of day [eg, 10:65 AM] [6], extubation occurring prior to intubation [14], or death date occurring before birth date [32]), diagnostic or procedural plausibility (eg, a procedure marked as an outpatient when it is only performed on an inpatient basis [38] or an obstetric diagnosis given for a biologically male patient [6,9,38]), alignment with external standards or expectations (eg, laboratory result absent for diagnosis or drug [6] or demographic alignment of medication name and dose with expected value ranges [34]), and others. A total of 11 (42%) studies used thresholding to identify data of low or questionable quality [4,6,8,9,14,19,28,32,35,37,39], including clinical and physiological value ranges (eg, BMI between 12 and 90 kg/m^2 [35] or fraction of inspired oxygen between 10% and 100% [14]) and logical thresholds (eg, recorded date of arrival prior to the date of data collection initiation [8] or difference of >730 days when comparing age in years and date of birth fields [9]).

Uniqueness

Finally, 1 (4%) study reported on data uniqueness. Aerts et al [8] measured the frequency of patient record duplications (ie, when patient records were erroneously copied during data merging or reprocessing). To reduce the rate of record duplications, the researchers in the study suggest a 2-level data structure, with more general patient data being recorded at the encounter level (which can include multiple visits during a single clinical episode) and diagnosis or procedure-specific data at the level of the particular visit.

Data Performance Assessment

Correctness or Accuracy

Correctness or accuracy was the most cited element in data performance analysis, with references in 14 (54%) of all articles [2,8,9,14,19,25,26,32-37,39]. The metric was evaluated via manual review in 8 (57%) out of the 14 articles that reported the measure [2,8,14,25,26,34,36,39]. A total of 5 (36%) articles evaluated it in comparison to an external standard, including national registries [2,35], EHR case definitions based on billing codes [36], and literature guidelines with high research use [33], or, in the case of a newly proposed AI technique for synthetic data augmentation, comparison to a previously published GAN model performance [19]. A further 3 (21%) assessed correctness or accuracy against an internal standard by calculating the proportion of records satisfying internally predetermined rule sets [9,32,37]. Of note, Curtis et al [2] and Terry et al [36] used both manual review and comparison to an external gold standard for validation.

Currency or Recency

Recency was the second most cited data performance element, with references in 7 (27%) articles [2,4,9,27,32,34,36]. Among these, 5 (71%) studies evaluated the metric according to internally predetermined hard rule sets (eg, whether a patient who is obese had a weight recording within 1 year of the previous data point or whether data were entered into the EHR within 3 days of the clinical encounter [9,32,36]) or soft rule sets (eg, whether the data were entered into the EHR within a subjectively determined clinically actionable time limit [4,34]),

while 2 (29%) used external standards, including national registries and guidelines [2,27].

Fairness or Bias

The third most cited data performance element was fairness or bias, with references in 6 (23%) articles [3,19,22,24,27,35]. Among these, Lee et al [22], Thuraisingam et al [35], Tian et al [27], and García-de-León-Chocano et al [24] assessed fairness by manual review, while Fu et al [3] and Zhang et al [19] did so through automated review against a predetermined internal gold standard (ie, distribution of data characteristics within a real EHR dataset) or data bias criterion (ie, critic model measuring Jensen-Shannon divergence between real and synthetic data over time), respectively.

Stability

Data stability was the fourth most cited performance element, referenced in 4 (15%) articles [4,8,19,31]. All 4 articles that measured data stability did so via temporal statistical analyses of data drift according to a predetermined internal baseline standard of data distribution [8,9,32,37].

Shareability

Shareability was referenced in 2 (8%) articles from our analysis [2,3]. Both studies measured the performance metric by way of manual review in a pre- and posttest analysis of data standardization [2,3].

Robustness

Finally, García-de-León-Chocano et al [24] reported on information robustness by way of statistical estimation of critical (eg, missing or null required values) versus noncritical (all other) DQ issues that may obstruct subsequent data applications and performance measures.

Interventions for Improving DQ and Performance

Three articles included in our analysis reported effective interventions to improve DQ and performance [4,9,37]. In terms of DQ, Walker et al [37] reported an increase in compliance, with 155 completeness and plausibility data checks from 53% to 100% across 6 clinical sites after 3 rounds of DQA. In terms of DQ and performance, Puttkamer et al [9] reported both higher data completeness and recency following a continuous data reporting and feedback system implementation. Finally, Engel et al [4] reported increased shareability (concept success rate, ie, whether data partners converted information from their individual EHRs to the shared database)—an increase from 90% to 98.5%—and a notable reduction in the percentage of sites with over 3 DQ errors—a reduction from 67% to 35%—across 50+ clinical sites over 2 years.

Discussion

Principal Contributions and Comparison With Prior Work

This scoping review provides an overview of the most common and successful means of EHR DQ and performance analysis. The review adds to a growing body of literature on the subject, most recently supplemented by a systematic review by Lewis et al [40]. To our knowledge, ours is the first review of specialty-specific applications of DQ alongside performance assessments. We identified and analyzed a total of 26 original research articles recently published on the topic. The results serve to characterize the most common medical fields making use of such assessments, the methodologies they use for conducting them, and areas for specialty-specific, as well as generalizable, future improvement. Finally, the discussion proposes a set of 6 unique and practical recommendations for minimizing modifiable DQ and performance issues arising during data extraction and mapping.

Article Characteristics

Our review noted a paucity of DQ assessments within clinical specialties, where expert domain knowledge plays a key role in identifying logic inconsistencies. Half of all identified articles concerned general EHR data assessments, while the other half focused on medical fields such as primary care, cardiovascular diseases, or intensive care unit or anesthesia, with the notable absence of psychiatry, emergency medicine, and any of the surgical specialties. This points to a lack of peer-reviewed research and underuse of DQ and performance strategies across a wide spectrum of the medical field. There is a wide knowledge gap between how data are entered and acted upon clinically and how they appear in silico. Therefore, more efforts need to be directed toward supporting EHR data assessment initiatives in these specialties, with close collaboration between clinical users and data scientists.

More than half of the articles included in this scoping review had common limitations, including using or reporting incomplete data, methods, and outcomes. Among the articles scoring high for incomplete data, the chief issues include data attrition during extraction [24,29] and unclear or missing reporting [33,36,38], pointing to a need for higher information interoperability and reporting standards, such as those put forth by Kahn et al [12]. These standards recommend using a harmonized and inclusive framework for the reporting of DQ assessments, including standardized definitions for completeness, conformance, plausibility, and other measures as discussed previously.

Similar issues were observed with methods reporting, with several articles underreporting steps in their data extraction or analysis, thereby limiting the replicability and generalizability of their findings [3,27,28,33]. Unclear reporting or underreporting was a substantial issue for outcomes as well, with low-scoring articles reporting only partial or too high-level results suggesting selective reporting bias [14,31,34,38]. To align with the standards set forth by articles scoring high in reporting quality, we recommend stating all data sourcing, methods, and results according to predetermined definitions of DQ or performance (see above) in enough detail such that they would be easily replicated by researchers at an unrelated institution.

A final article quality pitfall concerned articles that were too specific to a particular health system or clinical context. The chief issues among original research articles that in house scored "low" in our generalizability assessment concerned their overreliance on internal DQ checks or measures that could only be implemented within their specific institutional EHR [4,24-26,33]. To increase generalizability, we recommend

relying on external DQ standards such as societal guidelines, previously published measures, or open-source databases, to the extent possible before resorting to the development of new in-house tools that impose limitations to generalizability outside the local clinical context [8,12-15].

Data Quality Assessment

The marked drop-off between the use of completeness, conformance, and plausibility versus other indicators (Table 3) demonstrates that the field has settled on these measures as the main components of EHR DQ analysis. Taking this into consideration, we recommend measuring all 3 for a general assessment of clinical DQ. Of note, there is a significant drop-off between 81% (n=21) of studies reporting on completeness versus 69% (n=18) on conformance and 62% (n=16) on plausibility, which indicates an opportunity for limited but quick DQ "checks" using completeness measures only. More specialized analyses may require further reporting, including uniqueness in the event of data merger with the possibility of duplicate results. These may be particularly important in the case of EHR DQ assessments following information reconciliation from the merger of multiple data sources, including patient demographics or baseline physical characteristics and laboratory or pharmacological data, which were shown to be particularly prone to errors in DQ.

Our review additionally demonstrates that issues with data completeness, conformance, and plausibility may be at least partially addressed with data imputation methods. While previously these methods were either too limited in scope (completeness only), crude (eg, augmenting missing data with the mean of the entire dataset or a value's k-nearest neighbor), or computationally expensive (eg, individual values calculated via regression models based on predetermined sets of correlated features), our review suggests that these tasks are being increasingly automated. Specifically, data attrition contributing to missingness and conformity at the extraction stage may be minimized with AI data extractor algorithms, such as the one described by Lee et al [22]. In cases where further extraction is no longer feasible, the dataset may be augmented by (1) using large language models for extracting structured data available in other formats (eg, laboratory values recorded in the text of media files from outside patient records); (2) incorporating or cross-referencing data from well-established outside data repositories (eg, the US Social Security Death Index for mortality records [2] or the Database of Genotypes and Phenotypes and the Biologic Specimen for biomarkers of heart failure and other conditions [33]); or (3) generating synthetic data, for example, by modeling providers' behaviors with respect to different information types or sources [3] and by using GANs to create synthetic care episodes based on longitudinal EHR observations [19].

Data Performance Assessment

Correctness or accuracy was by far the most reported measure among the data performance indicators examined in our review. While certainly integral to assessing a dataset's usability and potential for downstream clinical or research impact, correctness alone is insufficient to guarantee the success of said applications. A technically "correct" dataset may still be practically limited if it is outdated, biased, inconsistent, or entirely idiosyncratic. We, therefore, recommend that future data assessments consider including additional measures of recency, fairness, stability, and shareability, respectively, among their core set of performance indicators as they each contribute a unique measure of a dataset's applicability. Importantly, our review noted considerable heterogeneity in the definitions used for these additional measures (eg, by defining data recency in terms of whether the information was logged into the EHR within a set time or whether it represents a patient's state at a given time period [Table 1] [10,13]), suggesting that further efforts are needed to harmonize outcome definitions in the field of data performance analysis in particular. Nonetheless, the predominance of internal standard comparisons for measuring recency and stability in our review demonstrates that these indicators may be essential for individualized EHR DPAs and should, therefore, be considered on a case-by-case basis (eg, in epidemiology where the timing and consistency of reporting can be of essential importance, or quality improvement initiatives where a researcher might want to compare pre- vs postintervention results). Likewise, shareability ought to be considered in the case of assessing dataset performance for interoperability purposes (eg, with data integrations, sharing, and reporting).

As discussed previously, data fairness assessments can and should be considered for monitoring overall EHR bias, as well as the bias inherent to any data imputation methods as discussed above. Our review points to the fact that this is a rapidly developing field, with fairness assessments to date mostly requiring manual review against national guidelines or disease registries, or, in the case of synthetic data, real EHR datasets [41-43]. Nonetheless, such gold standards are not always readily available (eg, What is the standard distribution of age or race in the real world?), so tech-savvy researchers have more recently resorted to detecting fairness during the validation of machine learning models or algorithms instead of the data itself [41-43]. Several research articles from our analysis proposed ways of automating the process. Fu et al [3] present a straightforward way of measuring the agreement of AI-generated synthetic data against a gold standard dataset. Zhang et al [19] suggest that while such straightforward analysis may be valuable, it is insufficient to measure true fairness, and they go on to propose a method of measuring bias via Jansen-Shannon divergence, which can be calculated for comparisons of real-world and synthetic data. The latter article also suggests a way of preventing synthetic data drift through condition regularization (ie, minimizing contrastive loss by regularizing the synthetic dataset against a real dataset distribution) and fuzzying (ie, adding controlled noise to broaden the dataset distribution before the AI training phase). To our knowledge, this is the most recently proposed technique for fairness assessment in the field. More research is needed to validate and augment the technique. Whether through Jansen-Shannon divergence or alternative methods, we recommend that all future data assessment projects measure and report model performance and fairness for sensitive groups.

Finally, Garcia-a-de-Leon-Chocano et al [24] propose a way of calculating data robustness. The calculation draws on comparing

tolerable versus nontolerable issues with DQ, which may be particularly important prior to using the information. We highly suggest that DQ assessments conduct a robustness calculation immediately before calculating data performance measures for downstream applications, which will allow for timely intervention in the case of significant issues with data completeness, conformity, or plausibility that merit additional data collection, review, or imputation steps as discussed above. The above findings and recommendations are summarized in Table 4.

Table . Recommendations for future EHR[a] data quality and performance assessments.

Issue	Recommendation
Article characteristics	
Paucity of specialty-focused EHR data assessments	Incentivize (eg, through quality improvement initiatives and grants) more EHR data assessments, particularly in psychiatry, emergency medicine, and surgical specialties
Incomplete reporting	Use standardized frameworks for measuring and reporting data quality and performance assessments (eg, Table 1)
Poor replicability	Describe DQA[b] methods in enough details such that they could be replicated by a research team at a different institution
Limited generalizability	Use already available data quality tools and standards (eg, DQA Guidelines proposed by Weiskopf et al [21]) before developing proprietary methodologies
DQA	
Inconsistent methodologies	Analyze completeness, conformance, and plausibility at every DQA (completeness only may be applicable for quick data quality checks)
Data missingness and nonconformity	Use available AI-based data extraction algorithms (eg, Lee et al [22]), and augment data using external and synthetic datasets (eg, Zhang et al [19])
Data performance assessment	
Inconsistent methodologies	Augment correctness or accuracy measurement with recency, fairness, stability, and shareability performance metrics
EHR data bias	Automate data fairness assessments by measuring agreement of AI-extracted data against an gold standard dataset (eg, manually extracted data) and preventing drift via condition fuzzying and regularization (eg, Zhang et al [19])
Timeliness of analysis	Calculate dataset robustness prior to detailed data quality and performance analysis (eg, as described by García-de-León-Chocano et al [24])

[a]EHR: electronic health record.

[b]DQA: data quality assessment.

Further Recommendations

Based on the review and our team's experience with DQ improvement initiatives, we recommend that administrators minimize modifiable DQ and performance issues arising during extraction by first using Internet of Things devices (eg, "smart" patient beds and infusion pumps) that directly upload measurements or settings to the EHR instead of requiring manual data entry. Second, the EHR's interface should be anchored to a predefined data workflow and ontological structure agreed upon in collaboration with clinical and data administrators (eg, encounters start at the time of patient check-in instead of when a physician first sees the patient, and all encounter times are recorded in 1 location using standard units). Finally, the plausibility of automatically entered data should be periodically validated such that corrections can be made when necessary (eg, a minute-by-minute electrocardiogram plausibility check that can detect if an electrocardiography lead falls off a patient's chest and needs to be replaced to record accurate measurements). Wherever possible, a reference data format (eg, electrocardiogram voltage between 0.5 and 5 mV) for the validation should be provided.

To minimize modifiable issues arising during data mapping, we furthermore recommend first establishing rules for how to treat (1) "missing," (2) "modified," or (3) "overlapping" data, such as whether (1) fields with no value should be regarded as data points or artifacts; (2) data points that have been subsequently modified should be updated or retained; and (3) one data source should take precedence over another in case of duplicate records (eg, weight recordings measured by weighing scale should supersede those measured by a hospital bed). Finally, standards for parent-child encounters should be instituted (eg, if a postoperative outpatient clinic visit should be assigned as a unique encounter or as a child encounter of the parent surgery visit).

The provenance of outside facility records, which can be used to identify potential issues with externally collected data, should also be maintained (eg, keeping records of where and when outside laboratory measures were taken in order to identify

potential issues with more or less accurate laboratory techniques).

Limitations

While this scoping review provides valuable insight into the existing literature on EHR DQ analytics, it has several limitations. Foremost, it is important to acknowledge the limited sample size of 154 articles using our original search criteria, and consequently also the limited number of 26 original research articles which were included in our final analysis after full-text review. Among these articles, there was significant heterogeneity in settings and outcomes of interest, which may limit the validity of direct comparisons between the studies, as well as the generalizability of our findings. The review was furthermore restricted to articles available in the PubMed library, which may introduce a potential publication bias, as well as to articles available only in English, which may introduce a language bias to our study selection and subsequent analysis. Finally, while the review focused on EHR DQ and performance assessments, it did not include adjacent areas that may have a pronounced impact on clinical data recording and use such as EHR implementation or use. Future research should consider broader inclusion criteria and explore additional dimensions of EHR DQ to provide a more comprehensive understanding of this important topic.

Conclusions

The findings of this scoping review highlight the importance of EHR DQ analysis in ensuring the accuracy and reliability of clinical data. Our review identified a need for specialty-specific data assessment initiatives, particularly in the fields of psychiatry, emergency medicine, and surgery. We additionally identified a need for standardizing DQ reporting to enhance the replicability and generalizability of outcomes in the field. Based on our review of the existing literature, we recommend analyzing DQ in terms of completeness, conformance, and plausibility; data performance in terms of correctness; and use case–specific metrics such as recency, fairness, stability, and shareability. Notably, our review demonstrated several examples of DQ improvement with the use of AI-enhanced data extraction and supplementation techniques. Future efforts in augmenting DQ through AI should make use of data fairness assessments to prevent the introduction of synthetic data bias.

Acknowledgments

TOB was supported by the National Institutes of Health (NIH; OT2 OD032701); the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK; K01 DK120784 and R01 DK121730); the National Institute of General Medical Sciences (NIH/NIGMS; R01 GM110240 and R01 GM149657); the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB; R01 EB029699); the National Institute of Neurological Disorders and Stroke (NIH/NINDS; R01 NS120924); University of Florida (UF) Research (DRPD-ROSF2023 [00132783]); and the University of Florida Clinical and Translational Science Institute (AWD10247), which was supported in part by the NIH National Center for Advancing Translational Sciences (UL1TR001427). AB was supported by the NIH (OT2 OD032701), the National Institute of General Medical Sciences (NIH/NIGMS; R01 GM110240 and R01 GM149657), the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB; R01 EB029699), the National Institute of Neurological Disorders and Stroke (NIH/NINDS; R01 NS120924), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK; R01 DK121730). TJL was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01 GM149657). BS was supported by the NIH (OT2 OD032701), by the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK; R01 DK121730), and by the National Institute of General Medical Sciences (NIH/NIGMS; R01 GM110240 and R01 GM149657). JB was supported by the NIH (T32 GM008721). The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the paper; and decision to submit the paper for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and other funding sources.

Authors' Contributions

YPP performed the investigation, data curation, and writing—original draft, review, and editing. TRB contributed to investigation, data curation, and writing—original draft. MMR performed data curation, investigation, and writing—review and editing. ML performed investigation. RS contributed to investigation. ZG did the investigation, methodology, and writing—review and editing. JB did the data curation, methodology, writing—review and editing—and supervision. TOB performed data curation, methodology, and supervision. BS performed data curation, methodology, and supervision. TJL contributed to data curation, methodology, and supervision. AB performed data curation, methodology, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search terms.
[DOCX File, 11 KB - medinform_v12i1e58130_app1.docx]

Multimedia Appendix 2
Study characteristics.
[XLSX File, 12 KB - medinform_v12i1e58130_app2.xlsx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File, 54 KB - medinform_v12i1e58130_app3.docx]

References

1. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The "All of Us" research program. N Engl J Med 2019 Aug 15;381(7):668-676. [doi: 10.1056/NEJMsr1809937] [Medline: 31412182]
2. Curtis MD, Griffith SD, Tucker M, et al. Development and validation of a high-quality composite real-world mortality endpoint. Health Serv Res 2018 Dec;53(6):4460-4476. [doi: 10.1111/1475-6773.12872] [Medline: 29756355]
3. Fu S, Wen A, Pagali S, et al. The implication of latent information quality to the reproducibility of secondary use of electronic health records. Stud Health Technol Inform 2022 Jun 6;290:173-177. [doi: 10.3233/SHTI220055] [Medline: 35672994]
4. Engel N, Wang H, Jiang X, et al. EHR data quality assessment tools and issue reporting Workflows for the "All of Us" research program clinical data research network. AMIA Jt Summits Transl Sci Proc 2022 May;2022:186-195. [Medline: 35854725]
5. Kapsner LA, Mang JM, Mate S, et al. Linking a consortium-wide data quality assessment tool with the MIRACUM metadata repository. Appl Clin Inform 2021 Aug;12(4):826-835. [doi: 10.1055/s-0041-1733847] [Medline: 34433217]
6. Mohamed Y, Song X, McMahon TM, et al. Tailoring rule-based data quality assessment to the Patient-Centered Outcomes Research Network (PCORnet) Common Data Model (CDM). AMIA Annu Symp Proc 2023 Apr 29;2022:775-784. [Medline: 37128433]
7. Becoming the trusted open science community built with standardised health data via a European federated network. European Health Data & Evidence Network. URL: https://www.ehden.eu/ [accessed 2024-10-23]
8. Aerts H, Kalra D, Sáez C, et al. Quality of hospital electronic health record (EHR) data based on the International Consortium for Health Outcomes Measurement (ICHOM) in heart failure: pilot data quality assessment study. JMIR Med Inform 2021 Aug 4;9(8):e27842. [doi: 10.2196/27842] [Medline: 34346902]
9. Puttkammer N, Baseman JG, Devine EB, et al. An assessment of data quality in a multi-site electronic medical record system in Haiti. Int J Med Inform 2016 Feb;86:104-116. [doi: 10.1016/j.ijmedinf.2015.11.003]
10. Bian J, Lyu T, Loiacono A, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc 2020 Dec 9;27(12):1999-2010. [doi: 10.1093/jamia/ocaa245]
11. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010 Oct;67(5):503-527. [doi: 10.1177/1077558709359007] [Medline: 20150441]
12. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244. [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
13. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681]
14. Wang Z, Penning M, Zozus M. Analysis of anesthesia screens for rule-based data quality assessment opportunities. Stud Health Technol Inform 2019;257:473-478. [Medline: 30741242]
15. Kelley TA. International Consortium for Health Outcomes Measurement (ICHOM). Trials 2015 Dec;16(S3). [doi: 10.1186/1745-6215-16-S3-O4]
16. Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care 2019 Jul;8(7):2328-2331. [doi: 10.4103/jfmpc.jfmpc_440_19] [Medline: 31463251]
17. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
18. Gardner A, Smith AL, Steventon A, Coughlan E, Oldfield M. Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy AI practice. AI Ethics 2022 May;2(2):277-291. [doi: 10.1007/s43681-021-00069-w] [Medline: 34790951]
19. Zhang Z, Yan C, Malin BA. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. J Am Med Inform Assoc 2022 Oct 7;29(11):1890-1898. [doi: 10.1093/jamia/ocac131] [Medline: 35927974]
20. Ozonze O, Scott PJ, Hopgood AA. Automating electronic health record data quality assessment. J Med Syst 2023 Feb 13;47(1):23. [doi: 10.1007/s10916-022-01892-2] [Medline: 36781551]
21. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017 Sep 4;5(1):14. [doi: 10.5334/egems.218] [Medline: 29881734]

22. Lee RY, Kross EK, Torrence J, et al. Assessment of natural language processing of electronic health records to measure goals-of-care discussions as a clinical trial outcome. JAMA Netw Open 2023 Mar 1;6(3):e231204. [doi: 10.1001/jamanetworkopen.2023.1204] [Medline: 36862411]
23. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for Scoping Reviews (PRISMA-SCR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
24. García-de-León-Chocano R, Sáez C, Muñoz-Soler V, Oliver-Roig A, García-de-León-González R, García-Gómez JM. Robust estimation of infant feeding indicators by data quality assessment of longitudinal electronic health records from birth up to 18 months of life. Comput Methods Programs Biomed 2021 Aug;207:106147. [doi: 10.1016/j.cmpb.2021.106147] [Medline: 34020376]
25. Sirgo G, Esteban F, Gómez J, et al. Validation of the ICU-DaMa tool for automatically extracting variables for minimum dataset and quality indicators: the importance of data quality assessment. Int J Med Inform 2018 Apr;112:166-172. [doi: 10.1016/j.ijmedinf.2018.02.007] [Medline: 29500016]
26. Toftdahl AKS, Pape-Haugaard LB, Palsson TS, Villumsen M. Collect once - use many times: the research potential of low back pain patients' municipal electronic healthcare records. Stud Health Technol Inform 2018;247:211-215. [Medline: 29677953]
27. Tian Q, Han Z, Yu P, An J, Lu X, Duan H. Application of openEHR archetypes to automate data quality rules for electronic health records: a case study. BMC Med Inform Decis Mak 2021 Apr 3;21(1):113. [doi: 10.1186/s12911-021-01481-2] [Medline: 33812388]
28. Tian Q, Han Z, An J, Lu X, Duan H. Representing rules for clinical data quality assessment based on openEHR guideline definition language. Stud Health Technol Inform 2019 Aug 21;264:1606-1607. [doi: 10.3233/SHTI190557] [Medline: 31438254]
29. Estiri H, Stephens KA, Klann JG, Murphy SN. Exploring completeness in clinical data research networks with DQe-c. J Am Med Inform Assoc 2018 Jan 1;25(1):17-24. [doi: 10.1093/jamia/ocx109] [Medline: 29069394]
30. Mang JM, Seuchter SA, Gulden C, et al. DQAgui: a graphical user interface for the MIRACUM data quality assessment tool. BMC Med Inform Decis Mak 2022 Aug 11;22(1):213. [doi: 10.1186/s12911-022-01961-z] [Medline: 35953813]
31. Sengupta S, Bachman D, Laws R, et al. Data quality assessment and multi-organizational reporting: tools to enhance network knowledge. EGEMS (Wash DC) 2019 Mar 29;7(1):8. [doi: 10.5334/egems.280] [Medline: 30972357]
32. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of an ontology for characterizing data quality for a secondary use of EHR data. Appl Clin Inform 2016 Feb;7(1):69-88. [doi: 10.4338/ACI-2015-08-RA-0107] [Medline: 27081408]
33. Lee K, Weiskopf N, Pathak J. A framework for data quality assessment in clinical research datasets. AMIA Annu Symp Proc 2018 Apr;2017:1080-1089. [Medline: 29854176]
34. Kiogou SD, Chi CL, Zhang R, Ma S, Adam TJ. Clinical data cohort quality improvement: the case of the medication data in the University of Minnesota's clinical data repository. AMIA Jt Summits Transl Sci Proc 2022 May 23;2022:293-302. [Medline: 35854717]
35. Thuraisingam S, Chondros P, Dowsey MM, et al. Assessing the suitability of general practice electronic health records for clinical prediction model development: a data quality assessment. BMC Med Inform Decis Mak 2021 Oct 30;21(1):297. [doi: 10.1186/s12911-021-01669-6] [Medline: 34717599]
36. Terry AL, Stewart M, Cejic S, et al. A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak 2019 Feb 12;19(1):30. [doi: 10.1186/s12911-019-0740-0] [Medline: 30755205]
37. Walker KL, Kirillova O, Gillespie SE, et al. Using the CER Hub to ensure data quality in a multi-institution smoking cessation study. J Am Med Inform Assoc 2014;21(6):1129-1135. [doi: 10.1136/amiajnl-2013-002629] [Medline: 24993545]
38. Gadde MA, Wang Z, Zozus M, Talburt JB, Greer ML. Rules based data quality assessment on claims database. Stud Health Technol Inform 2020 Jun 26;272:350-353. [doi: 10.3233/SHTI200567] [Medline: 32604674]
39. Wang H, Belitskaya-Levy I, Wu F, et al. A statistical quality assessment method for longitudinal observations in electronic health record data with an application to the VA million veteran program. BMC Med Inform Decis Mak 2021 Oct 20;21(1):289. [doi: 10.1186/s12911-021-01643-2] [Medline: 34670548]
40. Lewis AE, Weiskopf N, Abrams ZB, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc 2023 Sep 25;30(10):1730-1740. [doi: 10.1093/jamia/ocad120] [Medline: 37390812]
41. IBM. AI Fairness 360 (AIF360). GitHub. 2023. URL: https://github.com/Trusted-AI/AIF360 [accessed 2023-09-21]
42. LinkedIn. The Linkedin Fairness Toolkit (LiFT). GitHub. 2023. URL: https://github.com/linkedin/LiFT [accessed 2023-09-21]
43. Microsoft. Responsible AI Toolbox. GitHub. 2023. URL: https://github.com/microsoft/responsible-ai-toolbox [accessed 2023-09-21]

Abbreviations

AI: artificial intelligence
DPA: data performance assessment
DQ: data quality

DQA: data quality assessment
EHR: electronic health record
GAN: generative adversarial network
MIRACUM: Medical Informatics in Research and Care in University Medicine
PCORnet: The National Patient-Centered Clinical Research Network
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by C Lovis; submitted 06.03.24; peer-reviewed by M Mun, Z Liu; revised version received 14.05.24; accepted 08.06.24; published 06.11.24.

Please cite as:
Penev YP, Buchanan TR, Ruppert MM, Liu M, Shekouhi R, Guan Z, Balch J, Ozrazgat-Baslanti T, Shickel B, Loftus TJ, Bihorac A
Electronic Health Record Data Quality and Performance Assessments: Scoping Review
JMIR Med Inform 2024;12:e58130
URL: https://medinform.jmir.org/2024/1/e58130
doi:10.2196/58130

Case Identification of Depression in Inpatient Electronic Medical Records: Scoping Review

Allison Grothman[1]; William J Ma[1]; Kendra G Tickner[1]; Elliot A Martin[1,2], PhD; Danielle A Southern[1,3], MSc; Hude Quan[1,3], MD, PhD

1

2

3

Corresponding Author:
Elliot A Martin, PhD

Abstract

Background: Electronic medical records (EMRs) contain large amounts of detailed clinical information. Using medical record review to identify conditions within large quantities of EMRs can be time-consuming and inefficient. EMR-based phenotyping using machine learning and natural language processing algorithms is a continually developing area of study that holds potential for numerous mental health disorders.

Objective: This review evaluates the current state of EMR-based case identification for depression and provides guidance on using current algorithms and constructing new ones.

Methods: A scoping review of EMR-based algorithms for phenotyping depression was completed. This research encompassed studies published from January 2000 to May 2023. The search involved 3 databases: Embase, MEDLINE, and APA PsycInfo. This was carried out using selected keywords that fell into 3 categories: terms connected with EMRs, terms connected to case identification, and terms pertaining to depression. This study adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines.

Results: A total of 20 papers were assessed and summarized in the review. Most of these studies were undertaken in the United States, accounting for 75% (15/20). The United Kingdom and Spain followed this, accounting for 15% (3/20) and 10% (2/20) of the studies, respectively. Both data-driven and clinical rule-based methodologies were identified. The development of EMR-based phenotypes and algorithms indicates the data accessibility permitted by each health system, which led to varying performance levels among different algorithms.

Conclusions: Better use of structured and unstructured EMR components through techniques such as machine learning and natural language processing has the potential to improve depression phenotyping. However, more validation must be carried out to have confidence in depression case identification algorithms in general.

(JMIR Med Inform 2024;12:e49781) doi:10.2196/49781

KEYWORDS

electronic medical records; EMR phenotyping; depression; algorithms; health services research; precision public health; inpatient; clinical information; phenotyping; data accessibility; scoping review; disparity; development; phenotype; PRISMA-ScR; Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Introduction

Background

Depression is a significant factor contributing to the global burden of disease. It contributes significantly to the cost of health care services, with depression treatment services costing an average of CAD $550 (US $406.12) per patient in Alberta, Canada, in the 2007/2008 fiscal year [1]. Depression also carries a significantly higher mortality rate [2]. Surveillance of depression in the population is necessary to understand the needs of patients and allocate limited resources where they are most needed. This surveillance will ultimately allow health care professionals to make more targeted decisions when implementing population-level interventions.

Electronic medical records (EMRs) are a digitized collection of patient records documented by medical professionals. They contain various types of patient information, including test results, demographic data, and information about medication orders, recorded in structured data fields and free-text data, such as discharge summaries and nurses' notes [3-5]. EMRs were designed to aid individual patient care but are increasingly used for other purposes, such as research and gathering data for precision public health efforts, as they are compiled in large data warehouses [6-9]. An area that will be instrumental in

applying EMRs to public health is case phenotyping, which is developing case definitions to identify positive cases of a disorder in EMR data.

Accurate case identification in EMRs is an area where more research needs to be conducted. This is especially true for case identification of psychiatric disorders. Previous reviews of phenotyping algorithms for psychiatric disorders only considered primary care databases as their setting [10,11]. However, these are very different from inpatient EMR systems. For one, hospital inpatients are more likely to identify errors and omissions than patients in outpatient care or primary care [12]. EMR data have been used in research for psychiatric patients in various specific inpatient use cases, including assessing patient safety events in psychiatric inpatient units [13]. Research has also shown that hospitals with electronic psychiatric EMRs had lower readmission rates for psychiatric patients compared to hospitals without electronic records. Similarly, hospitals where psychiatric records were accessible to nonpsychiatric physicians had lower 14- and 30-day readmission rates [14]. In 2015, patients with a mental health diagnosis made up over 11% of hospital separations and 25% of hospital days [15]. Accurate case identification for inpatient stays for this at-risk population can help to identify what treatments have been most successful more efficiently than traditional research methods and could work in personalizing care for a more successful treatment plan.

Objectives

This study aims to provide an overview of existing algorithms for depression case identification in inpatient EMRs. It examines the performance of the algorithms and how they were constructed to provide guidance to those wishing to use an existing algorithm or to construct new ones.

Methods

Identifying Relevant Literature

This review followed the methodology outlined in the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) 2018 statement [16]. First, we used the *ICD-9-CM* (*International Classification of Diseases, Ninth Revision, Clinical Modification*) codes for depression provided by Elixhauser et al [17] to identify relevant terms, then developed a Boolean algorithm using these terms, as well as terms related to EMRs and terms related to case identification (Multimedia Appendix 1). Finally, we searched the following 3 databases: Embase (1974 to May 2023), Ovid MEDLINE (1946 to May 2023), and APA PsycInfo (1806 to May 2023) for peer-reviewed papers and exported the results of the search to a reference manager program (Zotero; Corporation for Digital Scholarship and Roy Rosenzweig Center for History and New Media).

Selecting Studies

Identified papers were screened in 2 stages. First, titles and abstracts were screened by 2 reviewers working independently to determine whether they met our established eligibility criteria. Papers were included if they were retrieved by the Boolean search and presented a case definition, involved depression and EMRs, were published between January 2000 and May 2023, and were written in English. We excluded papers that only used administrative databases, as this study focused on case phenotyping using EMR data. Next, full papers were reviewed for all abstracts that both reviewers identified as eligible. This review was carried out by 2 reviewers working independently. To be included, studies had to use EMRs for phenotyping and use inpatient data, and the case definition developed had to be for depression. The inpatient data source requirement was added because of differences in coding standards between primary care and inpatient settings. Disagreements at either screening stage were resolved by consensus, and if necessary, a third reviewer was consulted. We searched the references of all included papers for additional eligible papers, which we then screened using the same criteria. The search was designed to include all papers that used an algorithm phenotyping for depression with an EMR. The 2 most common methods were natural language processing (NLP) and machine learning, which were included but were not limited to. The search terms used to identify this category were not specific to a type of algorithm or method of case identification, as the purpose was to include a broad range of variations in phenotypic methodology (Multimedia Appendix 1).

Extracting Data

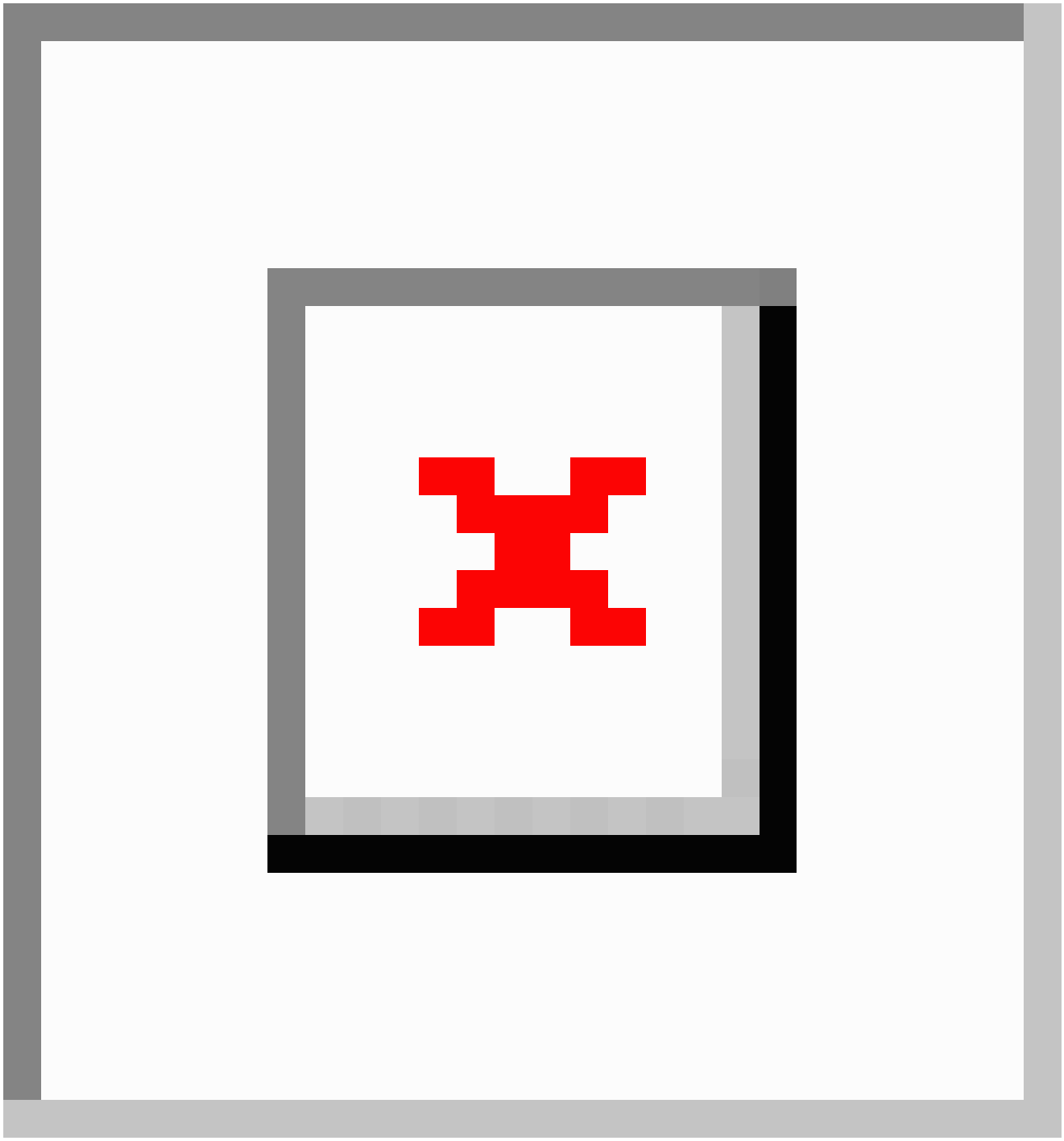
We adapted an existing data extraction form (Multimedia Appendix 2, Lee et al [18]) to collect the results of our review. Data were extracted by 1 reviewer and then confirmed by a second reviewer. Components we extracted included study characteristics (country, year, and inpatient or outpatient setting), the specific data source and details of the data, and the validation methodology (eg, medical record review), as well as detailed descriptions of the phenotype developed, the methods used, and the purpose for the case definition. We recorded the performance of the developed algorithms as reported in each study. We recorded the elements of EMRs used, whether other databases or diagnostic codes were used, and whether AI techniques (machine learning and NLP) were used as binary variables. Finally, based on this study's primary objective, we classified each study into 1 of 3 categories (algorithm development, outcome analysis, and comorbidity analysis).

Results

Paper Screening

The database search returned a total of 854 papers. After 257 duplicates were removed, 597 abstracts remained. Then, 522 abstracts were excluded in the title and abstract screening, leaving 75 papers for full-paper review. Of these, 20 papers could not be retrieved, and 36 were excluded based on the exclusion criteria. The 19 remaining papers met all eligibility criteria and were included in the review. Further, 1 additional paper was identified for inclusion from the references of the included papers, resulting in 20 papers for the review [6,19-37]. The PRISMA flow diagram illustrating these steps is shown in Figure 1.

Figure 1. PRISMA flow diagram. EMR: electronic medical record; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



Characterizing the Identified Literature

Of the 20 studies we identified, the majority occurred in the United States (15/20, 75%). The remaining studies were from the United Kingdom (3/20, 15%) and Spain (2/20, 10%). All the studies were published in 2005 or later.

Most studies looked at inpatient and outpatient data (16/20, 80%), while fewer focused solely on inpatient data (4/20, 20%). A few studies (4/20, 20%) linked EMR data to administrative databases. These studies used structured fields of EMRs and diagnostic codes found in administrative databases. They occurred in 3 countries (United States, United Kingdom, and Spain) and were all published in 2020 or later. Another 3 studies (3/20, 15%) linked EMRs to genomic data (the Partners HealthCare Biobank, United States; the Michigan Genomics Initiative, United States; and the pediatric biorepository database of the Center for Applied Genomics at the Children's Hospital of Philadelphia, United States). This linkage was conducted in an epidemiological analysis study to find genetic associations between conditions. These characteristics are shown in Table 1.

Table . Characteristics of included papers.

Paper reference	Country	EMR[a] setting	Additional data sources used
Dashti et al [19]	United States	Inpatient and outpatient	Genomic data
Dorr et al [20]	United States	Inpatient and outpatient	None
Edgcomb et al [21]	United States	Inpatient and outpatient	None
Estiri et al [22]	United States	Inpatient and outpatient	None
Fang et al [23]	United States	Inpatient and outpatient	Genomic data
Fernandes et al [24]	United Kingdom	Inpatient and outpatient	None
Goulet et al [25]	United States	Inpatient and outpatient	None
Hong et al [26]	United States	Inpatient and outpatient	Administrative data
Ingram et al [27]	United States	Inpatient and outpatient	None
Khapre et al [28]	United Kingdom	Inpatient and outpatient	Administrative data
Mar et al [29]	Spain	Inpatient and outpatient	Administrative data
Mason et al [30]	United Kingdom	Inpatient and outpatient	None
Mayer et al [31]	Spain	Inpatient and outpatient	None
McCoy et al [32]	United States	Inpatient	None
Parthipan et al [33]	United States	Inpatient	None
Perlis et al [6]	United States	Inpatient and outpatient	None
Slaby et al [34]	United States	Inpatient	Genomic data
Tvaryanas et al [35]	United States	Inpatient and outpatient	None
Yusufov et al [36]	United States	Inpatient and outpatient	Administrative data
Zhou et al [37]	United States	Inpatient	None

[a]EMR: electronic medical record.

Most of the identified studies (18/20, 90%) used diagnostic codes in their case definition for depression. The most common codes used were *ICD-9* (*International Classification of Diseases, Ninth Revision*), followed by *ICD-10* (*International Classification of Diseases, Tenth Revision*). In many studies, the diagnostic code case definitions were combined with structured data elements, such as patient demographics (sex, age, etc), laboratory results, medications, and procedures. For example, procedures were coded with Current Procedural Terminology codes and other types of classifications. Structured EMR data were used in 13/20 studies (65%). Fewer studies (8/20, 40%) incorporated unstructured data elements, such as clinical notes. To analyze these elements, some studies used standardized vocabularies, such as the Unified Medical Language System, to develop lists of keywords. Most studies using unstructured data used NLP techniques to analyze the free-text data in unstructured EMR fields (7/20, 35%). NLP is commonly used on free-text medical data to transform the data into a structured format that can be processed using statistical techniques and machine learning. A quarter of the identified studies (5/20, 25%) used machine learning to develop phenotyping algorithms. Machine learning models included logistic regression, random forest, and propositional rule learners. Table 2 contains details about the algorithms defined in each study.

Table . Summary of algorithms.

Paper reference	Diagnostic codes?	EMR[a] – structured?	EMR – unstructured?	ML[b]?	NLP[c]?	Validation methodology	Sensitivity	Specificity	PPV[d]	AUC[e]
Dashti et al [19]	No	Yes	Yes	Yes	Yes	Medical record review	0.81	—[f]	0.90	—
Dorr et al [20]	Yes	Yes	No	No	No	Not specified	—	—	—	—
Edgcomb et al [21]	Yes	No	No	No	No	Not specified	—	—	—	—
Estiri et al [22]	Yes	No	No	No	No	Not specified	—	—	—	—
Fang et al [23]	Yes	Yes	No	No	No	Not specified	—	—	—	—
Fernandes et al [24]	Yes	Yes	No	No	No	Not specified	—	—	—	—
Goulet et al [25]	Yes	No	No	No	No	Medical record review	0.45	0.90	—	—
Hong et al [26][g]	Yes	Yes	No	Yes	No	Medical record review	—	—	—	0.83
Ingram et al [27]	Yes	Yes	No	No	No	Convergent validity	—	—	—	—
Khapre et al [28]	Yes	Yes	Yes	No	Yes	Not specified	—	—	—	—
Mar et al [29]	Yes	Yes	Yes	Yes	No	Medical record review	—	—	—	0.80
Mason et al [30]	Yes	No	No	No	No	Not specified	—	—	—	—
Mayer et al [31]	Yes	Yes	No	No	No	Not specified	—	—	—	—
McCoy et al [32]	Yes	No	No	No	No	Not specified	—	—	—	—
Parthipan et al [33]	Yes	Yes	Yes	No	Yes	Medical record review	—	—	—	—
Perlis et al [6]	Yes	Yes	Yes	No	No	Medical record review	0.39	0.95	0.78	0.87
Slaby et al [34]	Yes	Yes	Yes	No	Yes	Medical record review	—	—	0.95	—
Tvaryanas et al [35]	Yes	No	No	No	No	Not specified	—	—	—	—
Yusufov et al [36]	Yes	Yes	Yes	No	Yes	Medical record review	0.85	0.95	—	—
Zhou et al [37]	No	No	Yes	Yes	Yes	Medical record review	0.87	0.92	—	—

[a]EMR: electronic medical record.

[b]ML: machine learning.

[c]NLP: natural language processing.

[d]PPV: positive predictive value.

[e]AUC: area under the receiver operating characteristic curve.

[f]Not available.

[g]Area under the precision-recall curve and F_1 -score were only available for Hong et al [26]. The best algorithm in that paper had an area under the precision-recall curve of 0.90 and an F_1 -score of 0.81.

Only 9 studies (45%) conducted a medical record review to produce a reference standard to which to compare phenotyping results. Since most of the identified studies (14/20, 70%) were conducted with a larger goal of which phenotyping depression was a small part, many did not provide much information on the methods of their phenotyping. Most studies did not report any metrics measuring the diagnostic accuracy of developed phenotyping algorithms; only 8 studies (40%) reported at least one performance metric. The 6 metrics reported were sensitivity, specificity, positive predictive value (PPV), area under the receiver operating characteristic curve, area under the precision-recall curve, and F_1 -score. No studies reported negative predictive value. These metrics are displayed in Table 2.

We classified each study into 1 of 3 general purposes: algorithm development, comorbidity analysis, and outcome analysis. A small percentage of the identified studies (6/20, 30%) were conducted for algorithm development. These studies did not look at applications of the phenotyping algorithms developed; instead, they focused on phenotyping methods and algorithm performance. The rest of the studies used a case definition for depression as a step toward a larger goal. For 9 of these studies (9/20, 45%), this goal was outcome analysis or analyzing the effect of depression on patient outcomes, such as mortality, suicide attempts, and psychotherapy receipt. For the remaining studies (5/20, 25%), the goal was comorbidity analysis, examining the prevalence of depression as a comorbidity of other conditions. The comorbidities studied included HIV, hepatitis C, and cancer. Outcome analysis studies have become more prevalent in recent years. Further, 6 were published between 2020 and 2022, up from 3 between 2000 and 2019. In addition, algorithms used for depression phenotyping in EMRs have become more prevalent since 2017.

Discussion

Principal Results

In this review, we found 20 papers describing phenotyping algorithms for depression in inpatient EMR data. Most of these algorithms were case definitions using diagnostic codes, specifically *ICD-9*. This reflects that *ICD* (*International Classification of Diseases*) codes are commonly used for billing purposes in the United States and are the most frequently used diagnostic codes in EMRs worldwide [38]. *ICD*-coded data are thus widely available, making them a practical choice when developing a case definition. However, case definitions using diagnostic codes achieved worse sensitivity than algorithms that only used other fields of EMRs. Many algorithms also used structured EMR data [6,19,20,23,24,26-29,31,33,34,36], but fewer used unstructured data [6,19,28,29,33,34,36,37]. NLP and machine learning techniques were used by a minority of algorithms (NLP [19,28,33,34,36,37] and machine learning [19,26,29,37]). These types of machine learning applications are relatively new and are receiving much attention from researchers [39]. The algorithms that used machine learning performed well on all the metrics they reported (sensitivity 0.81 - 0.87, specificity 0.82, PPV 0.90, and area under the receiver operating characteristic curve 0.80 - 0.83). This suggests that the information in free-text EMR data is valuable for developing accurate phenotyping algorithms. It also supports the effectiveness of machine learning techniques for phenotyping of depression. This is likely an area that will be explored further in future research.

Many of the papers we found did not include a medical record review. If algorithms are not validated against a reference standard such as a medical record review, their accuracy remains unknown. Most papers also did not report metrics measuring the validity of the algorithms developed. This limits the potential of these algorithms for application in precision health care. Conducting validation studies on the algorithms presented in these papers would make them more rigorous. Of the papers that did report metrics, few reported sensitivity, specificity, and PPV together. This could result in skewed interpretations of phenotype performance, as a high sensitivity may come at the cost of a low PPV (or vice versa) for instance.

Based on the validity reported in these papers, an EMR appears promising as a phenotyping tool for depression; however, few studies have reported metrics of diagnostic accuracy of EMR algorithms, especially comprehensive metrics to fully assess performance. Future validation studies conducted on existing case definitions would be valuable in establishing their validity and bringing these types of phenotyping algorithms to the attention of medical professionals and public health analysts. Machine learning and NLP are small but growing areas within phenotyping research. More work could be carried out using these techniques on the unstructured fields in EMRs, alone or in combination with other fields. Finally, as most of the studies we found were performed in the United States on US EMR data, it is to be determined how generalizable the identified case definitions are to data recorded in other jurisdictions. Both the standards of care and the methods of reporting diagnoses vary widely between health care systems, which could result in an algorithm only being valid in the region in which it was developed. There is a need for further research validating existing case definitions across health care regions or creating new case definitions specific to the EMR systems of other countries.

Limitations

Some relevant papers may have been missed, as we only searched 3 databases. It is also possible that our search terms were not sufficiently broad to return every pertinent paper. We also only considered peer-reviewed papers, not gray literature.

However, we developed our search strategy in consultation with librarians and experts in the field with experience performing scoping reviews. For these reasons, we believe our search was sufficient to find papers for the review.

Conclusions

We examined current algorithms for phenotyping depression in inpatient EMRs. This is an area in which more research needs to be performed. It is difficult to accurately identify cases of depression in EMR data because depression is inconsistently coded, as there is some subjectivity in its diagnosis. Diagnostic codes are primarily used in the algorithms we found, but machine learning on free-text data has recently achieved promising results. Most of the algorithms were developed in the United States; how well they will perform on data from other jurisdictions is yet to be known. In addition, many identified algorithms have yet to be validated against a reference standard, or their performance was not reported. To be useful for public health research, case definitions must be validated; this is an area in which future work is needed. From this study, we conclude that EMRs have the potential to provide valuable insight into the indicators of depression, as well as its prevalence, common comorbidities, and associated outcomes. Future research into applying machine learning and NLP techniques on unstructured EMR data and studies to ascertain the validity and generalizability of existing phenotyping algorithms will be valuable in establishing EMR-based case phenotyping as a reliable tool in precision public health.

Acknowledgments

We are grateful to Natalie Wiebe, MSc, for her help developing the search strategy; to Seungwon Lee, PhD, for creating the data extraction form; and to Oliver Slater-Kinghorn for helping to screen papers. This work is supported by a Foundation Grant, led by HQ, through the Canadian Institutes of Health Research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Developed search terms.
[DOC File, 38 KB - medinform_v12i1e49781_app1.doc]

Multimedia Appendix 2
Summary spreadsheet of identified papers.
[XLS File, 54 KB - medinform_v12i1e49781_app2.xls]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File, 531 KB - medinform_v12i1e49781_app3.pdf]

References

1. Slomp M, Jacobs P, Ohinmaa A, et al. The distribution of mental health service costs for depression in the Alberta population. Can J Psychiatry 2012 Sep;57(9):564-569. [doi: 10.1177/070674371205700907]
2. Chiu M, Vigod S, Rahman F, Wilton AS, Lebenbaum M, Kurdyak P. Mortality risk associated with psychological distress and major depression: a population-based cohort study. J Affect Disord 2018 Jul;234:117-123. [doi: 10.1016/j.jad.2018.02.075] [Medline: 29525352]
3. Offerman S, Rauchwerger A, Nishijima D, et al. Use of an electronic medical record “dotphrase” data template for a prospective head injury study. West JEM 2013 Mar 1;14(2):109-113. [doi: 10.5811/westjem.2012.11.13400]
4. Cohen S, Jannot AS, Iserin L, Bonnet D, Burgun A, Escudié JB. Accuracy of claim data in the identification and classification of adults with congenital heart diseases in electronic medical records. Arch Cardiovasc Dis 2019 Jan;112(1):31-43. [doi: 10.1016/j.acvd.2018.07.002] [Medline: 30612895]
5. Greiver M, Barnsley J, Glazier RH, Harvey BJ, Moineddin R. Measuring data reliability for preventive services in electronic medical records. BMC Health Serv Res 2012 May 14;12:116. [doi: 10.1186/1472-6963-12-116] [Medline: 22583552]
6. Perlis RH, Iosifescu DV, Castro VM, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2012 Jan;42(1):41-50. [doi: 10.1017/S0033291711000997] [Medline: 21682950]
7. LaFleur J, McAdam-Marx C, Alder SS, et al. Clinical risk factors for fracture among postmenopausal patients at risk for fracture: a historical cohort study using electronic medical record data. J Bone Miner Metab 2011 Mar;29(2):193-200. [doi: 10.1007/s00774-010-0207-y] [Medline: 20686803]

8. Patel RC, Amorim G, Jakait B, et al. Pregnancies among women living with HIV using contraceptives and antiretroviral therapy in Western Kenya: a retrospective, cohort study. BMC Med 2021 Aug 13;19(1):178. [doi: 10.1186/s12916-021-02043-z] [Medline: 34384443]
9. Canfell OJ, Kodiyattu Z, Eakin E, et al. Real-world data for precision public health of noncommunicable diseases: a scoping review. BMC Public Health 2022 Nov 24;22(1):2166. [doi: 10.1186/s12889-022-14452-7] [Medline: 36434553]
10. Carreira H, Williams R, Strongman H, Bhaskaran K. Identification of mental health and quality of life outcomes in primary care databases in the UK: A systematic review. BMJ Open 2019 Jul;9(7):e029227. [doi: 10.1136/bmjopen-2019-029227]
11. Larvin H, Peckham E, Prady SL. Case-finding for common mental disorders in primary care using routinely collected data: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2019 Oct;54(10):1161-1175. [doi: 10.1007/s00127-019-01744-4] [Medline: 31300893]
12. Wang B, Kristiansen E, Fagerlund AJ, et al. Users' experiences with online access to electronic health records in mental and somatic health care: cross-sectional study. J Med Internet Res 2023 Dec 25;25:e47840. [doi: 10.2196/47840] [Medline: 38145466]
13. Marcus SC, Hermann RC, Frankel MR, Cullen SW. Safety of psychiatric inpatients at the Veterans Health Administration. Psychiatr Serv 2018 Feb 1;69(2):204-210. [doi: 10.1176/appi.ps.201700224] [Medline: 29032707]
14. Kozubal DE, Samus QM, Bakare AA, et al. Separate may not be equal: a preliminary investigation of clinical correlates of electronic psychiatric record accessibility in academic medical centers. Int J Med Inform 2013 Apr;82(4):260-267. [doi: 10.1016/j.ijmedinf.2012.11.007] [Medline: 23266060]
15. Johansen H, Finès P. Acute care hospital days and mental diagnoses. H Rep 2012;23(4):1-7 [FREE Full text]
16. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLoS Med 2021 Mar;18(3):e1003583. [doi: 10.1371/journal.pmed.1003583] [Medline: 33780438]
17. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
18. Lee S, Doktorchik C, Martin EA, et al. Electronic medical record–based case phenotyping for the Charlson conditions: scoping review. JMIR Med Inform 2021 Feb 1;9(2):e23934. [doi: 10.2196/23934] [Medline: 33522976]
19. Dashti HS, Redline S, Saxena R. Polygenic risk score identifies associations between sleep duration and diseases determined from an electronic medical record biobank. Sleep 2019 Mar 1;42(3):zsy247. [doi: 10.1093/sleep/zsy247]
20. Dorr DA, Quiñones AR, King T, Wei MY, White K, Bejan CA. Prediction of future health care utilization through note-extracted psychosocial factors. Med Care 2022;60(8):570-578. [doi: 10.1097/MLR.0000000000001742]
21. Edgcomb JB, Thiruvalluru R, Pathak J, Brooks JO. Machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalization of women with mental illness. Med Care 2021 Feb 1;59:S58-S64. [doi: 10.1097/MLR.0000000000001467] [Medline: 33438884]
22. Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SN. Predicting COVID-19 mortality with electronic medical records. NPJ Digit Med 2021 Feb 4;4(1):15. [doi: 10.1038/s41746-021-00383-x] [Medline: 33542473]
23. Fang Y, Fritsche LG, Mukherjee B, Sen S, Richmond-Rakerd LS. Polygenic liability to depression is associated with multiple medical conditions in the electronic health record: phenome-wide association study of 46,782 individuals. Biol Psychiatry Cogn Neurosci Neuroimaging 2022 Dec;92(12):923-931. [doi: 10.1016/j.biopsych.2022.06.004]
24. Fernandes AC, Chandran D, Khondoker M, et al. Demographic and clinical factors associated with different antidepressant treatments: a retrospective cohort study design in a UK psychiatric healthcare setting. BMJ Open 2018 Sep;8(9):e022170. [doi: 10.1136/bmjopen-2018-022170]
25. Goulet JL, Fultz SL, McGinnis KA, Justice AC. Relative prevalence of comorbidities and treatment contraindications in HIV-mono-infected and HIV/HCV-co-infected veterans. AIDS 2005 Oct;19 Suppl 3:S99-105. [doi: 10.1097/01.aids.0000192077.11067.e5] [Medline: 16251836]
26. Hong C, Rush E, Liu M, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. NPJ Digit Med 2021 Oct 27;4(1):151. [doi: 10.1038/s41746-021-00519-z] [Medline: 34707226]
27. Ingram WM, Baker AM, Bauer CR, et al. Defining major depressive disorder cohorts using the EHR: multiple phenotypes based on ICD-9 codes and medication orders. Neurol Psychiatry Brain Res 2020 Jun;36:18-26. [doi: 10.1016/j.npbr.2020.02.002] [Medline: 32218644]
28. Khapre S, Stewart R, Taylor C. An evaluation of symptom domains in the 2 years before pregnancy as predictors of relapse in the perinatal period in women with severe mental illness. Eur Psychiatr 2021;64(1):e26. [doi: 10.1192/j.eurpsy.2021.18]
29. Mar J, Gorostiza A, Ibarrondo O, et al. Validation of random forest machine learning models to predict dementia-related neuropsychiatric symptoms in real-world data. J Alzheimers Dis 2020;77(2):855-864. [doi: 10.3233/JAD-200345] [Medline: 32741825]
30. Mason A, Irving J, Pritchard M, et al. Association between depressive symptoms and cognitive-behavioural therapy receipt within a psychosis sample: a cross-sectional study. BMJ Open 2022 May 10;12(5):e051873. [doi: 10.1136/bmjopen-2021-051873] [Medline: 35537795]

31. Mayer MA, Gutierrez-Sacristan A, Leis A, De La Peña S, Sanz F, Furlong LI. Using electronic health records to assess depression and cancer comorbidities. In: Informatics for Health: Connected Citizen-Led Wellness and Population Health: IOS Press; 2017:236-240. [doi: 10.3233/978-1-61499-753-5-236]
32. McCoy TH, Yu S, Hart KL, et al. High throughput phenotyping for dimensional psychopathology in electronic health records. Biol Psychiatry 2018 Jun 15;83(12):997-1004. [doi: 10.1016/j.biopsych.2018.01.011] [Medline: 29496195]
33. Parthipan A, Banerjee I, Humphreys K, et al. Predicting inadequate postoperative pain management in depressed patients: a machine learning approach. PLoS One 2019;14(2):e0210575. [doi: 10.1371/journal.pone.0210575] [Medline: 30726237]
34. Slaby I, Hain HS, Abrams D, et al. An electronic health record (EHR) phenotype algorithm to identify patients with attention deficit hyperactivity disorders (ADHD) and psychiatric comorbidities. J Neurodev Disord 2022 Jun 11;14(1):37. [doi: 10.1186/s11689-022-09447-9] [Medline: 35690720]
35. Tvaryanas AP, Maupin GM. Risk of incident mental health conditions among critical care air transport team members. Aviat Space Environ Med 2014 Jan;85(1):30-38. [doi: 10.3357/asem.3782.2014] [Medline: 24479256]
36. Yusufov M, Pirl WF, Braun I, Tulsky JA, Lindvall C. Natural language processing for computer-assisted chart review to assess documentation of substance use and psychopathology in heart failure patients awaiting cardiac resynchronization therapy. J Pain Symptom Manage 2022 Oct;64(4):400-409. [doi: 10.1016/j.jpainsymman.2022.06.007]
37. Zhou L, Baughman AW, Lei VJ, et al. Identifying patients with depression using free-text clinical documents. Stud Health Technol Inform 2015;216:629-633. [Medline: 26262127]
38. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005 Oct;40(5 Pt 2):1620-1639. [doi: 10.1111/j.1475-6773.2005.00444.x] [Medline: 16178999]
39. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 4;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]

Abbreviations

EMR: electronic medical record
***ICD*:** *International Classification of Diseases*
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
***ICD-9-CM*:** *International Classification of Diseases, Ninth Revision, Clinical Modification*
NLP: natural language processing
PPV: positive predictive value
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by C Lovis; submitted 14.06.23; peer-reviewed by K Allen, L Herrle; revised version received 05.07.24; accepted 07.07.24; published 14.10.24.

Please cite as:
Grothman A, Ma WJ, Tickner KG, Martin EA, Southern DA, Quan H
Case Identification of Depression in Inpatient Electronic Medical Records: Scoping Review
JMIR Med Inform 2024;12:e49781
URL: https://medinform.jmir.org/2024/1/e49781
doi:10.2196/49781

Diagnostic Accuracy of Artificial Intelligence in Endoscopy: Umbrella Review

Bowen Zha[*], BMed; Angshu Cai[*], BMed; Guiqi Wang, MD, PhD

Department of Endoscopy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Guiqi Wang, MD, PhD

Abstract

Background: Some research has already reported the diagnostic value of artificial intelligence (AI) in different endoscopy outcomes. However, the evidence is confusing and of varying quality.

Objective: This review aimed to comprehensively evaluate the credibility of the evidence of AI's diagnostic accuracy in endoscopy.

Methods: Before the study began, the protocol was registered on PROSPERO (CRD42023483073). First, 2 researchers searched PubMed, Web of Science, Embase, and Cochrane Library using comprehensive search terms. Then, researchers screened the articles and extracted information. We used A Measurement Tool to Assess Systematic Reviews 2 (AMSTAR2) to evaluate the quality of the articles. When there were multiple studies aiming at the same result, we chose the study with higher-quality evaluations for further analysis. To ensure the reliability of the conclusions, we recalculated each outcome. Finally, the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) was used to evaluate the credibility of the outcomes.

Results: A total of 21 studies were included for analysis. Through AMSTAR2, it was found that 8 research methodologies were of moderate quality, while other studies were regarded as having low or critically low quality. The sensitivity and specificity of 17 different outcomes were analyzed. There were 4 studies on esophagus, 4 studies on stomach, and 4 studies on colorectal regions. Two studies were associated with capsule endoscopy, two were related to laryngoscopy, and one was related to ultrasonic endoscopy. In terms of sensitivity, gastroesophageal reflux disease had the highest accuracy rate, reaching 97%, while the invasion depth of colon neoplasia, with 71%, had the lowest accuracy rate. On the other hand, the specificity of colorectal cancer was the highest, reaching 98%, while the gastrointestinal stromal tumor, with only 80%, had the lowest specificity. The GRADE evaluation suggested that the reliability of most outcomes was low or very low.

Conclusions: AI proved valuabe in endoscopic diagnoses, especially in esophageal and colorectal diseases. These findings provide a theoretical basis for developing and evaluating AI-assisted systems, which are aimed at assisting endoscopists in carrying out examinations, leading to improved patient health outcomes. However, further high-quality research is needed in the future to fully validate AI's effectiveness.

(JMIR Med Inform 2024;12:e56361) doi:10.2196/56361

KEYWORDS

endoscopy; artificial intelligence; umbrella review; meta-analyses; AI; diagnostic; researchers; researcher; tools; tool; assessment

Introduction

Gastrointestinal diseases impose a serious burden on health care systems worldwide. The data show that gastrointestinal diseases cause millions of deaths worldwide every year [1]. Endoscopy, as an efficient and convenient method, can effectively diagnose various gastrointestinal diseases [2]. Endoscopic intervention can also effectively treat early gastrointestinal cancers [3].

In recent years, with the rise of artificial intelligence (AI), numerous studies have been conducted to explore its application in the field of endoscopy, aiming to assist medical professionals in lesion identification and endoscopy quality control [4,5].

At present, some meta-analyses have reported the diagnostic value of AI in endoscopy [6-9]. Although AI has high sensitivity and specificity in identifying lesions in some studies, due to merger heterogeneity and sample size variations, the reliability of merger analysis outcomes needs further discussion [10-12].

In this study, an umbrella review methodology was used to elucidate current research directions and identify potential future research ideas by evaluating existing meta-analyses on AI in endoscopy. The meta-analyses of current studies were screened and extracted, and the quality of outcomes was assessed.

Methods

Registration

The protocol was registered on PROSPERO (CRD42023483073) before the study began. PROSPERO is an open access database of systematic reviews. Registration before the start of the study effectively reduced selective reporting [13,14]. This umbrella review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The details can be seen in Checklist 1.

Search Strategy

Two researchers searched PubMed, Web of Science, Embase, and Cochrane Library with a comprehensive search strategy up to November 2023. In addition, we searched "Google Scholar" to identify gray literature and searched for references of eligible articles. Two researchers independently screened the titles and abstracts and reviewed the full texts to identify eligible studies. Any discrepancies were resolved through consultation with a third researcher until a consensus was reached. The search strategy details are available in Table S1 in Multimedia Appendix 1.

Inclusion and Exclusion Criteria

The inclusion criteria were as follows: (1) studies evaluating the diagnostic value of AI in endoscopy; (2) studies that provided at least one outcome data—sensitivity or specificity; (3) articles that had meta-analyses and were conducted by systematic methods; and (4) articles published in English.

We excluded studies that met the following criteria: (1) experiments not on humans, (2) unavailable full text, (3) duplicate studies, and (4) studies lacking critical information.

Data Extraction

Two researchers independently extracted data. The third researcher would extract data if there were any discrepancies. The following basic information was included: the first author, year of publication, country, kind of endoscopy, detection, followed guidelines, registered number, number of included studies in the meta-analyses, outcomes, included study types in the meta-analyses, and tools for assessing the risk of the Bias. Then, we collected outcome information, including sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, diagnostic odds ratio, and area under the curve. We searched for missed information in primary studies if necessary.

Evaluation of Article Quality

Two reviewers independently evaluated the quality of the articles using A Measurement Tool to Assess Systematic Reviews 2 (AMSTAR2). AMSTAR is a tool for evaluating the systematic reviews of randomized trials [15,16]. In 2015, researchers introduced AMSTAR2, which expanded the application scope of AMSTAR to include the evaluation of systematic reviews of nonrandomized trials [17]. AMSTAR2 consists of a 16-item questionnaire prompting reviewers to respond with "yes," "partly yes," or "no" to each item. We viewed 2 "partly yes" answers as 1 "yes." In total, 7 items were considered important. If all the items were in conformity or only 1 unimportant item was out of conformity, the study was evaluated as having high quality. If more than 1 unimportant item did not fit, the study was rated as having moderate quality. If 1 important item did not conform, the study was rated as having low quality; the study was regarded as having critically low quality if more than 1 important item did not conform.

Data Analysis

We collected the outcome indicators of applying AI technology in different scenarios. This study evaluated the application of AI diagnostic techniques in different endoscopes. Considering that there are several studies analyzing the same issues, if there were multiple meta-analyses, we selected high-quality studies according to the AMSTAR2 criteria. If the quality of different studies was consistent, we chose the latest published study among them. After that, the most recent meta-analysis was collected and performed again to ensure that the most recent results were obtained. To make the results more reliable, we chose a more conservative method. Moreover, we used the random effect model to ensure the reliability of the result.

We calculated the effect quantity and 95% CI of each meta-analysis. In each meta-analysis, the *P* value of the Cochran Q test and the I^2 metric were used to evaluate the heterogeneity caused by the threshold effect. The Deek test was used to test publication bias. We used forest figures to show the diagnostic value of AI in endoscopy. We also used the bar accumulation charts to show the conformity of the included articles. In this study, we used R (version 4.3.2; R Foundation for Statistical Computing) for calculation. If the *P* value was more than .05, we considered that there was no statistically significant difference.

Grading of the Evidence

Using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) principle, 2 reviewers evaluated the credibility of evidence independently. GRADE proposes 5 factors for downgrading certainty in the evidence (the risk of bias, inconsistency, indirectness, imprecision, and publication bias) and 2 factors for upgrading certainty in the evidence (large effect and dose-response). These factors were used to evaluate outcomes as being of high, moderate, low, or very low quality. The body of evidence for diagnostic test accuracy studies begins with high quality. There was no guidance on the up factors in the diagnostic test accuracy study; we only downgraded the evidence using the 5 downgrading factors. For the comparative study, we defined its initial reliability according to the results of AMSTAR2 and then adjusted it according to the above factors.

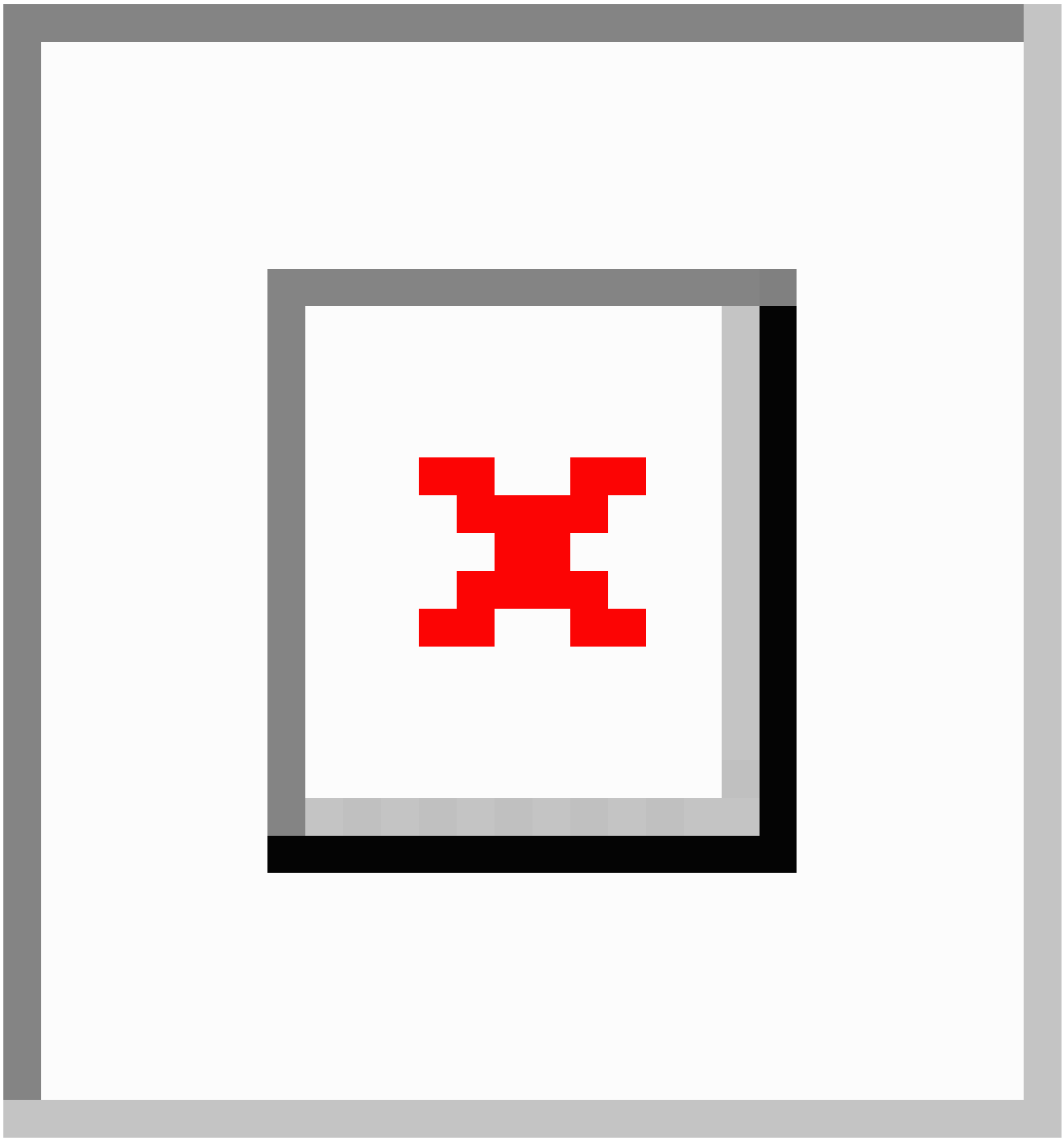
Results

Study Selection

We initially identified 3230 studies through the database and 208 studies through manual retrieval. After eliminating duplicates, we had 3013 studies. Then, the researchers eliminated 2897 studies that did not meet the criteria based on their titles and abstracts. After reading the full text of 116 studies, 80 irrelevant studies and 15 studies without

meta-analyses were excluded, and finally, 21 studies were included for statistical analysis and evaluation. These included 10 studies pertaining to upper gastrointestinal endoscopy[9,18-26], 5 studies focusing on colonoscopy [27-31], and 4 studies on capsule endoscopy [32-35]. Additionally, there was 1 study about endoscopic ultrasound (EUS) and 1 study about laryngoscopy [36,37]. Detail can be seen in Figure 1.

Figure 1. Search strategy and study screening.



Included Study Characteristics

A total of 10 studies reported the diagnostic value of AI technoloyg in upper gastrointestinal endoscopy. These studies encompassed various original research papers, ranging from 7 to 39 studies per investigation. These studies analyzed the diagnostic value of AI in various diseases, including esophageal and gastric neoplasia, Barrett esophagus, and *Helicobacter pylori* infection. In terms of research strategies, 9 research reports followed PRISMA guidelines, and 5 studies were registered on PROSPERO. With regard to evaluating bias, 8 studies used Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2), 1 study used QUADAS, and 1 study was not evaluated. QUADAS evaluates the diagnostic accuracy of the research system, including patient selection, index test, reference standard, flow, and timing. In 2011, researchers developed QUADAS-2 for better evaluation [38]. All studies were included in observational studies for diagnostic evaluation.

A total of 5 research studies on the diagnostic value of AI in colonoscopy were included. Among them, 1 study focused on ulcerative colitis, one focused on colon polyps and tumors, one used Prediction Model Risk of Bias Assessment Tool (PROBAST) to evaluate bias, 3 used QUADRAS-2, and one was not evaluated.

Of the remaining 6 included studies, 4 studies reported the value of AI in capsule endoscopy for diagnosing bleeding and ulcers; 2 studies reported AI's diagnostic value of laryngoscopes in examining normal or diseased throat structures and in EUS for diagnosing gastrointestinal stromal tumors separately; 5 studies were conducted according to the PRISMA guidelines; and 3 studies were registered in advance. All 6 studies were included in the observational study, and 5 of them used QUADRAS-2. The details can be seen in Table 1 and Table S2 in Multimedia Appendix 1.

Table . Basic information of included studies.

Study	Year	Country	Kind of endoscopy	Aim	Included studies, n	Followed guidelines
Tan et al [24]	2022	Australia	Upper endoscopy	Detection of Barrett esophagus	12	PRISMA[a]
Ma et al [22]	2022	China	Upper endoscopy	Detection of esophagus cancer	7	PRISMA
Bang et al [32]	2020	Korea	Upper endoscopy	Detection of *Helicobacter pylori* infection	8	PRISMA
Shi et al [23]	2022	China	Upper endoscopy	Detection of chronic atrophic gastritis	8	PRISMA
Guidozzi et al [20]	2023	South Africa	Upper endoscopy	Detection of Barrett esophagus and cancer	14	PRISMA
Jahagirdar et al [29]	2023	America	Colonoscopy	Detection of ulcerative colitis	12	PRISMA
Keshtkar et al [30]	2023	Iran	Colonoscopy	Detection of colorectal polyp and cancer	24	NR[b]
Bang et al [32]	2022	Korea	Wireless capsule Endoscopy	Detection of ulcers, polyps, celiac disease, bleeding, and hookworm	39	PRISMA
Soffer et al [35]	2020	Israel	Wireless capsule Endoscopy	Detection of ulcers, polyps, celiac disease, bleeding, and hookworm	19	PRISMA
Gomes et al [36]	2023	America	Endoscopic ultrasonography	Detection of gastrointestinal stromal tumor	8	PRISMA
Zurek et al [37]	2022	Poland	Laryngeal endoscopy	Detection of lesions in the larynx	11	PRISMA
Bai et al [27]	2023	China	Colonoscopy	Prediction of invasion depth of colorectal cancer or neoplasms	10	PRISMA
Qin et al [34]	2021	China	Wireless capsule endoscopy	Detection of erosion/ulcer, gastrointestinal bleeding, and polyps/cancer	16	PRISMA
Mohan et al [33]	2021	America	Wireless capsule endoscopy	Detection of gastrointestinal ulcers	9	NR
Bang et al [28]	2021	Korea	Colonoscopy	Detection of diminutive colorectal polyps	13	PRISMA
Lui et al [31]	2020	China	Colonoscopy	Detection of colorectal polyp and cancer	18	PRISMA
Lui et al [21]	2020	China	Upper endoscopy	Detection of gastric and esophageal neoplastic lesions and *Helicobacter pylori*	23	PRISMA
Visaggi et al [9]	2021	Italy	Upper endoscopy	Detection of Barrett neoplasia	19	NR

Study	Year	Country	Kind of endoscopy	Aim	Included studies, n	Followed guidelines
Zhang et al [26]	2021	China	Upper endoscopy	Detection of esophageal cancer and neoplasm	16	PRISMA
Xie et al [25]	2022	China	Upper endoscopy	Detection of gastric cancer and prediction invasion depth	17	PRISMA
Chen et al [19]	2022	China	Upper endoscopy	Detection of early gastric cancer	12	PRISMA

[a]PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses.

[b]NR: not reported.

Methodological Quality of Included Studies

In all the included studies, methodological quality ranged from very low to moderate. Results show that the methodology was rated as moderate for 6 studies, low for 2 studies, and very critically low for the remaining 13 studies. Among the articles about upper endoscopy, it was found that 5 studies exhibited a moderate level of methodological quality. In comparison, 2 studies were deemed to have low quality, and 3 studies had were very low quality. The critical problems were the need for advanced registration and an incomplete retrieval strategy. The noncritical problem was that the original literature funding had not been reported. Besides, studies of moderate methodological quality were conducted on both the stomach and esophagus of the upper gastrointestinal tract. Three studies on colonoscopy were of moderate quality, 2 were of low quality, and the remaining 8 were of very low quality. The main problems were the meta-merging method and the evaluation of publication bias.

Regarding the application of AI in capsule endoscopy, 1 study was of moderate quality, and the other 3 had critically low quality. In addition, the research on applying EUS to identify gastrointestinal stromal tumors and laryngoscope to identify normal and pathological structures of the throat had critically low quality. The details can be seen in Table S3 and Table S4 in Multimedia Appendix 1.

Meta-Analyses

There were 4 outcomes for the esophagus. The sensitivity was 0.89 (95% CI 0.84-0.93) for esophageal neoplasia, 0.95 (95% CI 0.91-0.98) for esophageal squamous cell carcinoma, 0.94 (95% CI 0.67-0.99) for abnormal intrapapillary loops, and 0.97 (95% CI 0.67-1.00) for gastroesophageal reflux disease. Their specificity was 0.86 (95% CI 0.83-0.93) for esophageal neoplasia, 0.92 (95% CI 0.82-0.97) for esophageal squamous cell carcinoma, 0.94 (95% CI 0.84-0.98) for abnormal intrapapillary loops, and 0.97 (95% CI 0.75-1.00) for gastroesophageal reflux disease. The sensitivity of gastric cancer and chronic atrophic gastritis was 0.89 (95% CI 0.85-0.93) and 0.94 (95% CI 0.88-0.97), respectively. At the same time, their specificity was 0.93 (95% CI 0.88-0.97) and 0.96 (95% CI 0.88-0.98), respectively. The sensitivity and specificity of judging the invasion depth of gastric cancer were 0.82 (95% CI 0.78-0.85) and 0.90 (95% CI 0.82-0.95), respectively. The sensitivity and specificity of *Helicobacter pylori* infection were 0.87 (95% CI 0.72-0.94) and 0.86 (95% CI 0.72-0.96).

In colonoscopy, the sensitivity and specificity of colon polyps were 0.93 (95% CI 0.91-0.95) and 0.87 (95% CI 0.76-0.93), respectively. The sensitivity and specificity of colon neoplasia were 0.94 (95% CI 0.85-0.98) and 0.98 (95% CI 0.94-0.99), respectively. The sensitivity and specificity of ulcerative colitis were 0.83 (95% CI 0.78-0.87) and 0.92 (95% CI 0.89-0.95), respectively. For invasion depth of colon neoplasia, the sensitivity and specificity were 0.71 (95% CI 0.58-0.81) and 0.95 (95% CI 0.91-0.97), respectively.

For wireless capsule endoscopy, we got 2 results. The sensitivity and specificity of the diagnosis of gastrointestinal ulcer were 0.93 (95% CI 0.89-0.95) and 0.92 (95% CI 0.89-0.95), respectively. The sensitivity and specificity of the diagnosis of gastrointestinal bleeding were 0.96 (95% CI 0.94-0.97) and 0.97 (95% CI 0.95-0.99), respectively. The sensitivity and specificity of EUS in diagnosing gastrointestinal stromal tumors were 0.92 (95 % CI 0.89-0.95) and 0.80 (95 % CI 0.75-0.85), respectively. The sensitivity of healthy and diseased tissues in AI-identified laryngoscope was 0.91 (95 % CI 0.83-0.98) and 0.91 (95 % CI 0.86-0.96), respectively, and the specificity was 0.97 (95 % CI 0.96-0.99) and 0.95 (95 % CI 0.90-0.99), respectively. The details can be seen in Figure 2 and Table 2.

Figure 2. Diagnostic value of artificial intelligence in different endoscopic outcomes.

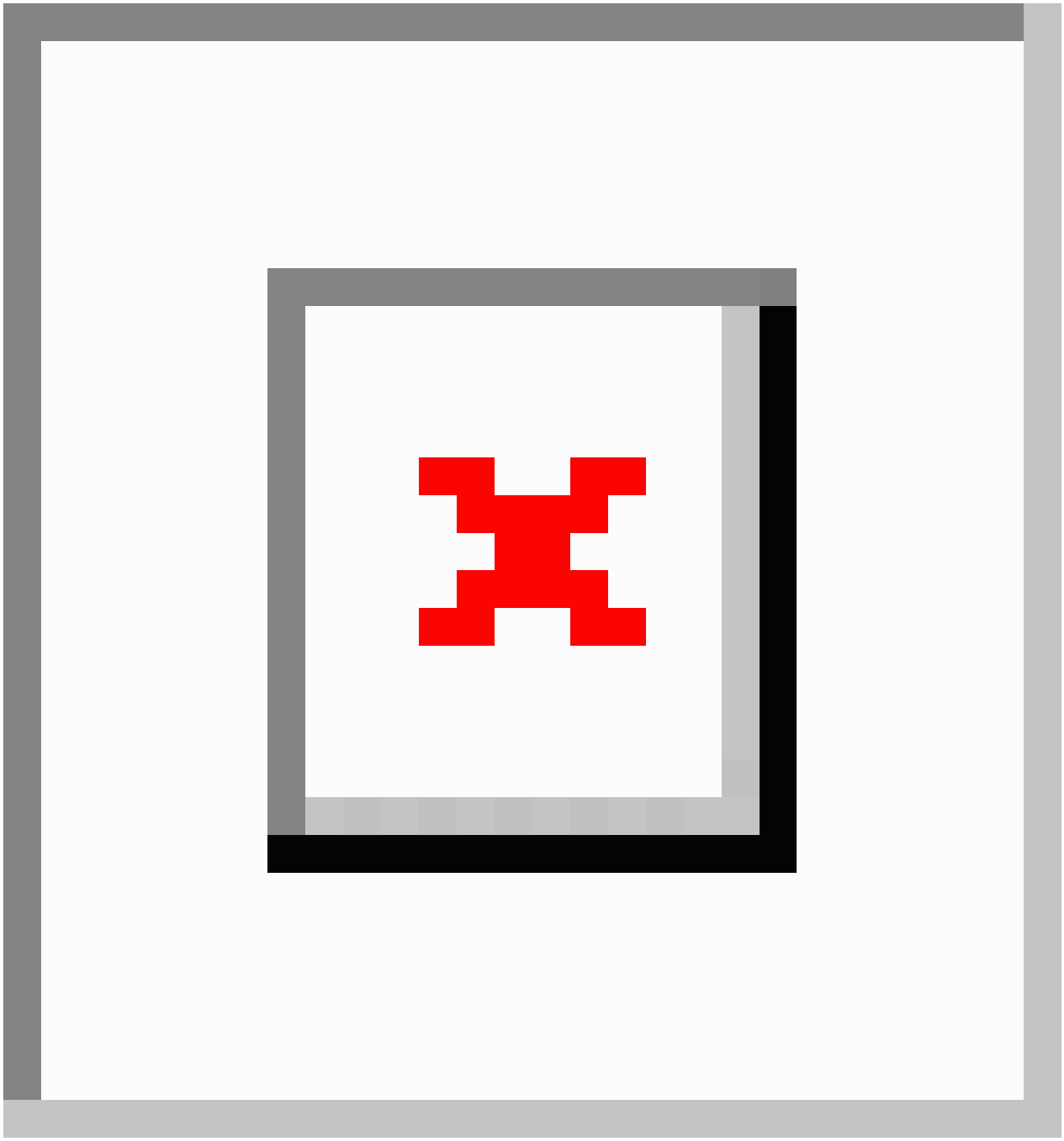


Table . Outcomes of artificial intelligence in endoscopy diagnosis.

Study	Detection	Sensitivity (95% CI)	Specificity (95% CI)	PLR[a] (95% CI)	NLR[b] (95% CI)	DOR[c] (95% CI)	AUC[d] (95% CI)	Model
Tan et al [24]	Early Barrett esophagus	0.90 (0.87-0.93)	0.84 (0.80-0.88)	NR[e]	NR	0.90 (0.87-0.93)	NR	Random
Ma et al et al [22]	Early esophageal cancer	0.90 (0.82-0.94)	0.91 (0.79-0.96)	9.8 (3.8-24.8)	0.11 (0.06-0.21)	NR	0.95[f]	NR
Bang et al [18]	*Helicobacter pylori* Infection	0.87 (0.72-0.94)	0.86 (0.72-0.96)	6.2 (3.8-10.1)	0.15 (0.07-0.34)	40 (15 - 112)	0.92 (0.90-0.94)	NR
Guidozzi et al[20]	Esophageal squamous cell carcinoma	0.91 (0.84-0.95)	0.80 (0.63-0.90)	NR	NR	NR	NR	Random
Guidozzi et al [20]	Esophageal adenocarcinoma	0.91 (0.87-0.94)	0.87 (0.82-0.91)	NR	NR	NR	NR	NR
Shi et al [23]	Chronic atrophic gastritis	0.94 (0.88-0.97)	0.96 (0.88-0.98)	21.58 (7.91-58.85)	0.07 (0.04-0.13)	320.19 (128.5-797.84)	0.98 (0.96-0.99)	NR
Jahagirdar et al [29]	Ulcerative colitis	0.83 (0.78-0.87)	0.92 (0.89-0.95)	NR	NR	NR	0.92 (0.88-0.94)	NR
Keshtkar et al [30]	Colorectal polyp	0.92 (0.85-0.96)	0.94 (0.89-0.96)	14.5 (8.4-25.2)	0.09 (0.05-0.16)	162 (59.44-5)	0.97 (0.96-0.99)	NR
Keshtkar et al [30]	Colorectal cancer	0.94 (0.85-0.98)	0.98 (0.94-0.99)	41.2 (13.7-124.2)	0.06 (0.02-0.16)	677 (108-4240)	0.99 (0.98-1.00)	NR
Bang et al [32]	Gastrointestinal ulcer	0.93 (0.89-0.95)	0.92 (0.89-0.94)	NR	NR	138 (79-243)	0.97 (0.95-0.98)	NR
Bang et al [32]	Gastrointestinal hemorrhage	0.96 (0.94-0.97)	0.97 (0.95-0.99)	NR	NR	888 (343-2303)	0.99 (0.98-0.99)	NR
Soffer et al [35]	Mucosal ulcers	0.95 (0.89-0.98)	0.94 (0.90-0.96)	NR	NR	NR	NR	Random
Soffer et al [35]	Bleeding	0.98 (0.96-0.99)	0.99 (0.97-0.99)	NR	NR	NR	NR	Random
Gomes et al [36]	Gastrointestinal stromal tumors	0.92 (0.89-0.95)	0.80 (0.75-0.85)	4.26 (2.7-6.7)	0.09 (0.14-0.18)	71.74 (22.43-229.46)	0.949[f]	NR
Zurek et al [37]	Healthy laryngeal tissue	0.91 (0.83-0.98)	0.97 (0.96-0.99)	NR	NR	NR	0.945[f]	Random
Zurek et al [37]	Benign and malignant lesions	0.91 (0.86-0.96)	0.95 (0.90-0.99)	NR	NR	NR	0.924[f]	Random
Bai et al [27]	Invasion depth of early colorectal cancer	0.71 (0.58-0.81)	0.95 (0.91-0.97)	NR	NR	NR	0.93 (0.90-0.95)	NR
Qin et al [34]	Erosion or ulcers	0.96 (0.91-0.98)	0.97 (0.93-0.99)	36.8 (12.3-110.1)	0.04 (0.02-0.09)	893 (103-5834)	0.99 (0.98-1.00)	NR
Qin et al [34]	Gastrointestinal bleeding	0.97 (0.93-0.99)	1.00 (0.99-1.00)	289.4 (80.3-1043.0)	0.03 (0.01-0.08)	10,291 (1539-68,791)	1.00 (0.99-1.00)	NR
Qin et al [34]	Polyps and cancer	0.97 (0.82-0.99)	0.98 (0.92-0.99)	42.7 (11.3-161.8)	0.03 (0.01-0.21)	1291 (60-27-808)	0.99 (0.98-1.00)	NR
Mohan et al [33]	Gastrointestinal ulcers or hemorrhage	0.96 (0.94-0.97)	0.96 (0.95-0.97)	NR	NR	NR	95.4 (94.3-96.3)	NR

Study	Detection	Sensitivity (95% CI)	Specificity (95% CI)	PLR[a] (95% CI)	NLR[b] (95% CI)	DOR[c] (95% CI)	AUC[d] (95% CI)	Model
Bang et al [28]	Colorectal polyps	0.93 (0.91-0.95)	0.87 (0.76-0.93)	7.1 (3.8-13.3)	0.08 (0.06-0.11)	87 (38-201)	0.96 (0.93-0.97)	NR
Lui et al [31]	Colorectal polyps	0.92 (0.89-0.95)	0.90 (0.85-0.93)	NR	NR	NR	0.96 (0.95-0.98)	Random
Lui et al [21]	Neoplastic lesions in the stomach	0.92 (0.88-0.95)	0.88 (0.78-0.95)	NR	NR	NR	0.96 (0.94-0.99)	NR
Lui et al [21]	Barrett esophagus	0.88 (0.83-0.92)	0.90 (0.86-0.95)	NR	NR	NR	0.96 (0.93-0.99)	NR
Lui et al [21]	Neoplastic lesions in squamous esophagus	0.76 (0.48-0.93)	0.92 (0.67-0.99)	NR	NR	NR	0.88 (0.82-0.96)	NR
Lui et al [21]	Helicobacter pylori status	0.84 (0.71-0.93)	0.90 (0.79-0.96)	NR	NR	NR	0.92 (0.88-0.97)	NR
Visaggi et al [9]	Barrett neoplasia	0.89 (0.84-0.93)	0.86 (0.83-0.93)	6.50 (1.59-2.15)	0.13 (0.20-0.08)	50.53 (24.74-103.22)	0.90 (0.85-0.94)	Random
Visaggi et al [9]	Esophageal squamous cell carcinoma	0.95 (0.91-0.98)	0.92 (0.82-0.97)	12.65 (1.61-3.51)	0.05 (0.11-0.02)	258.36 (44.18-1510.7)	0.97 (0.92-0.98)	Random
Visaggi et al [9]	Abnormal intrapapillary capillary loops	0.94 (0.67-0.99)	0.94 (0.84-0.98)	14.75 (1.46-3.70)	0.07 (0.39-0.01)	225.83 (11.05-4613.93)	0.98 (0.86-0.99)	Random
Visaggi et al [9]	Gastroesophageal reflux disease	0.97 (0.67-1.00)	0.97 (0.75-1.00)	38.26 (0.98-6.22)	0.03 (0.44-0.00)	1159.6 (6.12-219711.69)	0.99 (0.80-0.99)	Random
Zhang et al [26]	Esophageal neoplasms	0.94 (0.92-0.96)	0.85 (0.73-0.92)	6.40 (3.38-12.11)	0.06 (0.04-0.10)	98.88 (39.45-247.87)	0.97 (0.95-0.98)	Random
Xie et al [25]	Gastric cancer	0.89 (0.85-0.93)	0.93 (0.88-0.97)	13.4 (7.3-25.5)	0.11 (0.07-0.17)	NR	0.94 (0.91-0.98)	Random
Xie et al [25]	Invasion depth of gastric cancer	0.82 (0.78-0.85)	0.90 (0.82-0.95)	8.4 (4.2-16.8)	0.20 (0.16-0.26)	NR	0.90 (0.87-0.93)	Random
Chen et al [19]	Gastric cancer	0.86 (0.75-0.92)	0.90 (0.84-0.93)	NR	NR	NR	0.94[f]	NR

[a]PLR: positive likelihood ratio.

[b]NLR: negative likelihood ratio.

[c]DOR: diagnostic odds ratio.

[d]AUC: area under the curve.

[e]NR: not reported.

[f]95% CIs were not reported.

Grading of Evidence

We evaluated the reliability of each outcome through GRADE. Results showed that the quality was evaluated as very low for 44.1% of the outcomes and low for 55.9% of the outcomes. Our research found that the sensitivity and specificity of Barrett neoplasia, esophageal squamous cell carcinoma, *Helicobacter pylori* infection, chronological gastritis, colorectal polyp, gastrointestinal ulcer, and gastrointestinal hemorrhage had low credibility. The other outcomes had very low credibility. Generally speaking, the primary defects were indirectness and imprecision. These problems were caused by the different AI models and training methods used in the original literature, and there were also differences in the selection of recognition samples. Endoscopists in different regions used different samples and chose different AI algorithms to train and test the models, making the synthesized results less credible. Detail can be seen in Table S4 in Multimedia Appendix 1.

Discussion

Principal Findings

In this study, we conducted a systematic review of the current use of AI in endoscopic diagnosis, assessing the quality of

research and meta-analyses conducted in this field. AI has been studied and applied in upper gastrointestinal endoscopy, colorectal endoscopy, capsule endoscopy, and laryngoscopy. The meta-analysis results showed that AI has high sensitivity and specificity for these types of endoscopy. However, the overall evidence level of the outcomes was low.

In previous studies, AI could effectively assist in sedation and training in the operation process of upper digestive tract examination [39,40]. The earliest research we examined was conducted in 2007, when computers were trained to identify esophageal cancer [41]. At that time, the research only distinguished malignant and nonmalignant esophageal tissues in vitro.

With the rise of AI and the continuous upgrading of training methods, the application of AI in gastrointestinal endoscopy, including esophageal cancer, gastric cancer, and *Helicobacter pylori* infection, has been widely studied. In addition to the ordinary white light examination, computer-aided systems have shown a certain diagnostic value in stained and magnifying endoscopic imaging [42,43]. Moreover, some studies have found that trained models have research value in diagnosing gastric cancer's infiltration depth [44].

A study in 2022 compared the diagnostic value of computer-aided systems and professional endoscopists in gastric cancer images through retrospective data and found no significant difference in the diagnostic rate between the two groups [45]. This shows that AI aid is not inferior to endoscopists in image diagnosis. Wu conducted a single-center randomized controlled trial and found that the missed diagnosis rate of gastric adenoma could be significantly reduced using AI [46]. Multi-center randomized controlled studies are still needed for further analysis in the future.

AI has been widely studied in colorectal endoscopy. A meta-analysis showed that AI could effectively improve adenoma detection rate [7]. However, another meta-analysis based on real-world research reached the opposite conclusion [47]. The findings of our study proposed that AI has a noticeable effect in identifying intestinal lesions. However, many problems still need to be effectively addressed, particularly in terms of clinical implementation and practical translation.

In November 2023, the team at West China Hospital led a 12-center study with more than 10,000 patients [48]. This randomized controlled trial compared the relationship between AI-assisted and routine examinations in the missed diagnosis rate of esophageal lesions. The results showed that AI could not significantly improve the missed diagnosis rate of esophageal lesions. Many teams are constantly developing, improving and trying to use AI models in clinics. As mentioned above, although AI has been shown to have a significant effect in many studies, there has been an increase in research regarding the failure of AI to significantly improve the effectiveness of endoscopy in the clinical situation. In the application process, we found that the recognition threshold of AI greatly affected its application value. We are explored the possiblity of classifying patients according to some baseline information or endoscopic mucosal background images and then continuously optimized the AI recognition threshold according to the risk stratification of different patients. This approach aimed to achieve individualized endoscopic examinations and improve overall identification accuracy [49-51].

Moreover, the economic impact of a large-scale rollout of AI systems in clinical work on patients and health care institutions must be further studied. In addition, the differences in validation sets make it difficult to truly achieve accurate side-by-side comparisons when evaluating the capabilities of different AI models, which may lead to biased results. We believe it would be beneficial to produce an open platform that includes test data sets from different parts and different lesions of the gastrointestinal tract so that researchers can test the effectiveness of AI recognition in the future.

This study has several strengths. According to our preliminary understanding, the umbrella evaluation of using AI in endoscopic applications must be revised. To a certain extent, we have filled this blank. Second, we conducted a strict analysis and discussion following the PRISMA guidelines. Third, two researchers conducted all analyses, and the results were reliable.

There are also some limitations to this study. First, various computer-aid models have certain heterogeneity, and this could not be avoided in the analysis. Therefore, our results are a general summary of the current technology. Second, we could not gather the data of some unpublished studies. Third, the limited number of studies made it difficult to do further subgroup analyses. Fourth, we only included studies reported in English, which might have introduced some biases to our study.

Conclusions

This study found that AI has high diagnostic value in endoscopy. These findings provide a theoretical basis for the development and evaluation of AI-assisted systems, aimed at assisting endoscopists in conducting examinations, thereby improving patient health outcomes. However, it is worth noting that there is no convincing high-quality evidence in the existing research and further research is needed in the future.

Acknowledgments

This research was supported by the following grants: (1) Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS; grants 2021-I2M-1-015, 2021-I2M-1-061, 2021-I2M-1-013, and 2022-I2M-C&T-B-054); (2) Sanming Project of Medicine in Shenzhen (SZSM201911008); (3) Capital's Funds for Health Improvement and Research (grant CRF2020-2-4025); and (4) Beijing Hope Run Special Fund of Cancer Foundation of China (grant LC2021A03).

Data Availability

The data supporting this study's findings are available on request from the corresponding author.

Authors' Contributions

BZ was responsible for data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, writing the original draft, as well as reviewing and editing the final draft. AC was responsible for data curation, formal analysis, investigation, and resources. GW was responsible for conceptualization, funding acquisition, investigation, and resources.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional statistics.
[DOCX File, 41 KB - medinform_v12i1e56361_app1.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) checklist.
[DOCX File, 21 KB - medinform_v12i1e56361_app2.docx]

References

1. Milivojevic V, Milosavljevic T. Burden of gastroduodenal diseases from the global perspective. Curr Treat Options Gastroenterol 2020 Jan;18(1):148. [doi: 10.1007/s11938-020-00277-z] [Medline: 31993967]
2. Wu S, Zhang R, Yan J, et al. High-speed and accurate diagnosis of gastrointestinal disease: learning on endoscopy images using lightweight transformer with local feature attention. Bioengineering (Basel) 2023 Dec 13;10(12):1416. [doi: 10.3390/bioengineering10121416] [Medline: 38136007]
3. Brand M, Fuchs KH, Troya J, Hann A, Meining A. The role of specialized instruments for advanced endoscopic resections in gastrointestinal disease. Life (Basel) 2023 Nov 7;13(11):2177. [doi: 10.3390/life13112177] [Medline: 38004317]
4. Chino A, Ide D, Abe S, et al. Performance evaluation of a computer-aided polyp detection system with artificial intelligence for colonoscopy. Dig Endosc 2024 Feb;36(2):185-194. [doi: 10.1111/den.14578] [Medline: 37099623]
5. Karsenti D, Tharsis G, Perrot B, et al. Effect of real-time computer-aided detection of colorectal adenoma in routine colonoscopy (COLO-GENIUS): a single-centre randomised controlled trial. Lancet Gastroenterol Hepatol 2023 Aug;8(8):726-734. [doi: 10.1016/S2468-1253(23)00104-8] [Medline: 37269872]
6. Lou S, Du F, Song W, et al. Artificial intelligence for colorectal neoplasia detection during colonoscopy: a systematic review and meta-analysis of randomized clinical trials. EClinicalMedicine 2023 Dec;66:102341. [doi: 10.1016/j.eclinm.2023.102341] [Medline: 38078195]
7. Wei MT, Fay S, Yung D, Ladabaum U, Kopylov U. Artificial intelligence-assisted colonoscopy in real-world clinical practice: a systematic review and meta-analysis. Clin Transl Gastroenterol 2024 Mar 1;15(3):e00671. [doi: 10.14309/ctg.0000000000000671] [Medline: 38146871]
8. Liu Y, Ai YQ, Yang XJ, Zhang P, Zhong C. A meta-analysis of artificial intelligence in detection of colonoscopy adenoma and polyp. Jiangxi Medical Journal 2023;58(5):543-548. [doi: 10.3969/j.issn.1006-2238.2023.05.007]
9. Visaggi P, Barberio B, Gregori D, et al. Systematic review with meta-analysis: artificial intelligence in the diagnosis of oesophageal diseases. Aliment Pharmacol Ther 2022 Mar;55(5):528-540. [doi: 10.1111/apt.16778] [Medline: 35098562]
10. Gimeno-García AZ, Hernández-Pérez A, Nicolás-Pérez D, Hernández-Guerra M. Artificial intelligence applied to colonoscopy: is it time to take a step forward? Cancers (Basel) 2023 Apr 7;15(8):2193. [doi: 10.3390/cancers15082193] [Medline: 37190122]
11. Maida M, Marasco G, Facciorusso A, et al. Effectiveness and application of artificial intelligence for endoscopic screening of colorectal cancer: the future is now. Expert Rev Anticancer Ther 2023 Jul;23(7):719-729. [doi: 10.1080/14737140.2023.2215436] [Medline: 37194308]
12. Farid AB, Irdza TA, Li XJ, ZQ Z. Meta-analysis of the diagnostic value of artificial intelligence technology based on deep learning for early gastric cancer under endoscope. Modern Interv Diagnos Treatment Gastroenterol 2023;28(1):63-67. [doi: 10.3969 /j.issn.1672-2159.2023.01.013]
13. Chien PFW, Khan KS, Siassakos D. Registration of systematic reviews: PROSPERO. BJOG 2012 Jul;119(8):903-905. [doi: 10.1111/j.1471-0528.2011.03242.x] [Medline: 22703418]
14. Farrah K, Young K, Tunis MC, Zhao L. Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. Syst Rev 2019 Nov 15;8(1):280. [doi: 10.1186/s13643-019-1172-8] [Medline: 31730014]

15. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: A Measurement Tool to Assess the Methodological Quality of Systematic Reviews. BMC Med Res Methodol 2007 Feb 15;7:10. [doi: 10.1186/1471-2288-7-10] [Medline: 17302989]
16. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009 Oct;62(10):1013-1020. [doi: 10.1016/j.jclinepi.2008.10.009] [Medline: 19230606]
17. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017 Sep 21;358:j4008. [doi: 10.1136/bmj.j4008] [Medline: 28935701]
18. Bang CS, Lee JJ, Baik GH. Artificial intelligence for the prediction of Helicobacter pylori infection in endoscopic images: systematic review and meta-analysis of diagnostic test accuracy. J Med Internet Res 2020 Sep 16;22(9):e21983. [doi: 10.2196/21983] [Medline: 32936088]
19. Chen PC, Lu YR, Kang YN, Chang CC. The accuracy of artificial intelligence in the endoscopic diagnosis of early gastric cancer: pooled analysis study. J Med Internet Res 2022 May 16;24(5):e27694. [doi: 10.2196/27694] [Medline: 35576561]
20. Guidozzi N, Menon N, Chidambaram S, Markar SR. The role of artificial intelligence in the endoscopic diagnosis of esophageal cancer: a systematic review and meta-analysis. Dis Esophagus 2023 Nov 30;36(12):doad048. [doi: 10.1093/dote/doad048] [Medline: 37480192]
21. Lui TKL, Tsui VWM, Leung WK. Accuracy of artificial intelligence-assisted detection of upper GI lesions: a systematic review and meta-analysis. Gastrointest Endosc 2020 Oct;92(4):821-830. [doi: 10.1016/j.gie.2020.06.034] [Medline: 32562608]
22. Ma H, Wang L, Chen Y, Tian L. Convolutional neural network-based artificial intelligence for the diagnosis of early esophageal cancer based on endoscopic images: a meta-analysis. Saudi J Gastroenterol 2022;28(5):332-340. [doi: 10.4103/sjg.sjg_178_22] [Medline: 35848703]
23. Shi Y, Wei N, Wang K, Tao T, Yu F, Lv B. Diagnostic value of artificial intelligence-assisted endoscopy for chronic atrophic gastritis: a systematic review and meta-analysis. Front Med (Lausanne) 2023;10:1134980. [doi: 10.3389/fmed.2023.1134980] [Medline: 37200961]
24. Tan JL, Chinnaratha MA, Woodman R, et al. Diagnostic accuracy of artificial intelligence (AI) to detect early neoplasia in Barrett's esophagus: a non-comparative systematic review and meta-analysis. Front Med (Lausanne) 2022;9:890720. [doi: 10.3389/fmed.2022.890720] [Medline: 35814747]
25. Xie F, Zhang K, Li F, et al. Diagnostic accuracy of convolutional neural network-based endoscopic image analysis in diagnosing gastric cancer and predicting its invasion depth: a systematic review and meta-analysis. Gastrointest Endosc 2022 Apr;95(4):599-609. [doi: 10.1016/j.gie.2021.12.021] [Medline: 34979114]
26. Zhang SM, Wang YJ, Zhang ST. Accuracy of artificial intelligence-assisted detection of esophageal cancer and neoplasms on endoscopic images: a systematic review and meta-analysis. J Dig Dis 2021 Jun;22(6):318-328. [doi: 10.1111/1751-2980.12992] [Medline: 33871932]
27. Bai J, Liu K, Gao L, et al. Computer-aided diagnosis in predicting the invasion depth of early colorectal cancer: a systematic review and meta-analysis of diagnostic test accuracy. Surg Endosc 2023 Sep;37(9):6627-6639. [doi: 10.1007/s00464-023-10223-6] [Medline: 37430125]
28. Bang CS, Lee JJ, Baik GH. Computer-aided diagnosis of diminutive colorectal polyps in endoscopic images: systematic review and meta-analysis of diagnostic test accuracy. J Med Internet Res 2021 Aug 25;23(8):e29682. [doi: 10.2196/29682] [Medline: 34432643]
29. Jahagirdar V, Bapaye J, Chandan S, et al. Diagnostic accuracy of convolutional neural network-based machine learning algorithms in endoscopic severity prediction of ulcerative colitis: a systematic review and meta-analysis. Gastrointest Endosc 2023 Aug;98(2):145-154. [doi: 10.1016/j.gie.2023.04.2074] [Medline: 37094691]
30. Keshtkar K, Safarpour AR, Heshmat R, Sotoudehmanesh R, Keshtkar A. A systematic review and meta-analysis of convolutional neural network in the diagnosis of colorectal polyps and cancer. Turk J Gastroenterol 2023 Oct;34(10):985-997. [doi: 10.5152/tjg.2023.22491] [Medline: 37681266]
31. Lui TKL, Guo CG, Leung WK. Accuracy of artificial intelligence on histology prediction and detection of colorectal polyps: a systematic review and meta-analysis. Gastrointest Endosc 2020 Jul;92(1):11-22. [doi: 10.1016/j.gie.2020.02.033] [Medline: 32119938]
32. Bang CS, Lee JJ, Baik GH. Correction: computer-aided diagnosis of gastrointestinal ulcer and hemorrhage using wireless capsule endoscopy: systematic review and diagnostic test accuracy meta-analysis. J Med Internet Res 2022 Jan 11;24(1):e36170. [doi: 10.2196/36170] [Medline: 35015660]
33. Mohan BP, Khan SR, Kassab LL, et al. High pooled performance of convolutional neural networks in computer-aided diagnosis of GI ulcers and/or hemorrhage on wireless capsule endoscopy images: a systematic review and meta-analysis. Gastrointest Endosc 2021 Feb;93(2):356-364. [doi: 10.1016/j.gie.2020.07.038] [Medline: 32721487]
34. Qin K, Li J, Fang Y, et al. Convolution neural network for the diagnosis of wireless capsule endoscopy: a systematic review and meta-analysis. Surg Endosc 2022 Jan;36(1):16-31. [doi: 10.1007/s00464-021-08689-3] [Medline: 34426876]

35. Soffer S, Klang E, Shimon O, et al. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. Gastrointest Endosc 2020 Oct;92(4):831-839. [doi: 10.1016/j.gie.2020.04.039] [Medline: 32334015]
36. Gomes RSA, de Oliveira GHP, de Moura DTH, et al. Endoscopic ultrasound artificial intelligence-assisted for prediction of gastrointestinal stromal tumors diagnosis: a systematic review and meta-analysis. World J Gastrointest Endosc 2023 Aug 16;15(8):528-539. [doi: 10.4253/wjge.v15.i8.528] [Medline: 37663113]
37. Żurek M, Jasak K, Niemczyk K, Rzepakowska A. Artificial intelligence in laryngeal endoscopy: systematic review and meta-analysis. J Clin Med 2022 May 12;11(10):2752. [doi: 10.3390/jcm11102752] [Medline: 35628878]
38. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
39. Di Giulio E, Fregonese D, Casetti T, et al. Training with a computer-based simulator achieves basic manual skills required for upper endoscopy: a randomized controlled trial. Gastrointest Endosc 2004 Aug;60(2):196-200. [doi: 10.1016/s0016-5107(04)01566-4] [Medline: 15278044]
40. Pambianco DJ, Vargo JJ, Pruitt RE, Hardi R, Martin JF. Computer-assisted personalized sedation for upper endoscopy and colonoscopy: a comparative, multicenter randomized study. Gastrointest Endosc 2011 Apr;73(4):765-772. [doi: 10.1016/j.gie.2010.10.031] [Medline: 21168841]
41. Kodashima S, Fujishiro M, Takubo K, et al. Ex vivo pilot study using computed analysis of ENDO-cytoscopic images to differentiate normal and malignant squamous cell epithelia in the oesophagus. Dig Liver Dis 2007 Aug;39(8):762-766. [doi: 10.1016/j.dld.2007.03.004] [Medline: 17611178]
42. Kanesaka T, Lee TC, Uedo N, et al. Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging. Gastrointest Endosc 2018 May;87(5):1339-1344. [doi: 10.1016/j.gie.2017.11.029] [Medline: 29225083]
43. Nagao S, Tsuji Y, Sakaguchi Y, et al. Highly accurate artificial intelligence systems to predict the invasion depth of gastric cancer: efficacy of conventional white-light imaging, nonmagnifying narrow-band imaging, and Indigo-carmine dye contrast imaging. Gastrointest Endosc 2020 Oct;92(4):866-873. [doi: 10.1016/j.gie.2020.06.047] [Medline: 32592776]
44. Zhu Y, Wang QC, Xu MD, et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest Endosc 2019 Apr;89(4):806-815. [doi: 10.1016/j.gie.2018.11.011] [Medline: 30452913]
45. Niikura R, Aoki T, Shichijo S, et al. Artificial intelligence versus expert endoscopists for diagnosis of gastric cancer in patients who have undergone upper gastrointestinal endoscopy. Endoscopy 2022 Aug;54(8):780-784. [doi: 10.1055/a-1660-6500] [Medline: 34607377]
46. Wu L, Shang R, Sharma P, et al. Effect of a deep learning-based system on the miss rate of gastric neoplasms during upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial. Lancet Gastroenterol Hepatol 2021 Sep;6(9):700-708. [doi: 10.1016/S2468-1253(21)00216-8] [Medline: 34297944]
47. Patel HK, Mori Y, Hassan C, et al. Lack of effectiveness of computer aided detection for colorectal neoplasia: a systematic review and meta-analysis of nonrandomized studies. Clin Gastroenterol Hepatol 2024 May;22(5):971-980. [doi: 10.1016/j.cgh.2023.11.029] [Medline: 38056803]
48. Wei R, Wei P, Yuan H, et al. Inflammation in metal-induced neurological disorders and neurodegenerative diseases. Biol Trace Elem Res 2024 Jan 11. [doi: 10.1007/s12011-023-04041-z] [Medline: 38206494]
49. Cai C, Chen C, Lin X, et al. An analysis of the relationship of triglyceride glucose index with gastric cancer prognosis: a retrospective study. Cancer Med 2024 Feb;13(3):e6837. [doi: 10.1002/cam4.6837] [Medline: 38204361]
50. Jia K, Kundrot S, Palchuk MB, et al. A Pancreatic cancer risk prediction model (Prism) developed and validated on large-scale US clinical data. EBioMedicine 2023 Dec;98:104888. [doi: 10.1016/j.ebiom.2023.104888] [Medline: 38007948]
51. Liu Y, Chen S, Shen W, Qu X, Li S, Shi Y. Construction and validation of a gastric cancer diagnostic model based on blood groups and tumor markers. J Cancer 2024;15(3):729-736. [doi: 10.7150/jca.88190] [Medline: 38213731]

Abbreviations

AI: artificial intelligence
AMSTAR: A Measurement Tool to Assess Systematic Reviews
EUS: endoscopic ultrasound
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
QUADAS: Quality Assessment of Diagnostic Accuracy Studies

Edited by C Lovis; submitted 15.01.24; peer-reviewed by F Yu, K Liu; revised version received 25.05.24; accepted 26.05.24; published 15.07.24.

Please cite as:
Zha B, Cai A, Wang G
Diagnostic Accuracy of Artificial Intelligence in Endoscopy: Umbrella Review
JMIR Med Inform 2024;12:e56361
URL: https://medinform.jmir.org/2024/1/e56361
doi:10.2196/56361

Machine Learning Models for Parkinson Disease: Systematic Review

Thasina Tabashum[1], MSc; Robert Cooper Snyder[1], MSc; Megan K O'Brien[2,3], PhD; Mark V Albert[1,4], PhD

1

2

3

4

Corresponding Author:
Thasina Tabashum, MSc

Abstract

Background: With the increasing availability of data, computing resources, and easier-to-use software libraries, machine learning (ML) is increasingly used in disease detection and prediction, including for Parkinson disease (PD). Despite the large number of studies published every year, very few ML systems have been adopted for real-world use. In particular, a lack of external validity may result in poor performance of these systems in clinical practice. Additional methodological issues in ML design and reporting can also hinder clinical adoption, even for applications that would benefit from such data-driven systems.

Objective: To sample the current ML practices in PD applications, we conducted a systematic review of studies published in 2020 and 2021 that used ML models to diagnose PD or track PD progression.

Methods: We conducted a systematic literature review in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines in PubMed between January 2020 and April 2021, using the following exact string: “Parkinson’s” AND (“ML” OR “prediction” OR “classification” OR “detection” or “artificial intelligence” OR “AI”). The search resulted in 1085 publications. After a search query and review, we found 113 publications that used ML for the classification or regression-based prediction of PD or PD-related symptoms.

Results: Only 65.5% (74/113) of studies used a holdout test set to avoid potentially inflated accuracies, and approximately half (25/46, 54%) of the studies without a holdout test set did not state this as a potential concern. Surprisingly, 38.9% (44/113) of studies did not report on how or if models were tuned, and an additional 27.4% (31/113) used ad hoc model tuning, which is generally frowned upon in ML model optimization. Only 15% (17/113) of studies performed direct comparisons of results with other models, severely limiting the interpretation of results.

Conclusions: This review highlights the notable limitations of current ML systems and techniques that may contribute to a gap between reported performance in research and the real-life applicability of ML models aiming to detect and predict diseases such as PD.

(JMIR Med Inform 2024;12:e50117) doi:10.2196/50117

KEYWORDS

Parkinson disease; machine learning; systematic review; deep learning; clinical adoption; validation techniques; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Parkinson disease (PD) is a progressive neurodegenerative disease that results in a loss of motor function with muscle weakness, tremors, and rigidity. Secondary symptoms include speech difficulties, sleep disorders, and cognitive changes. Research suggests that pathophysiological symptoms can be used to detect PD before the onset of the motor features [1]. For these reasons, multiple clinical assessments and analyses are required to diagnose PD and allow for early detection. However, clinical diagnosis of PD is an error-prone process [2]. A UK autopsy study found that the misdiagnosis rate of PD is 24% [3]. Early detection is especially important for PD since early neuroprotective treatment slows down the progression of the disease and lessens the symptoms, which improves the patient’s quality of life [4]. From diagnosis to treatment, each case of PD is unique [5,6]. Precision medicine using machine learning (ML) has the potential to better use the varied data of individuals. Therefore, ML-based solutions can play an important role in PD diagnosis [7,8].

Here, ML refers to the branch of artificial intelligence that uses computational methods to perform a specific task without being explicitly programmed, by learning from previous examples of data and making predictions about new data [9]. ML includes

a broad range of standard learning algorithms, such as decision trees, support vector machines, and linear or logistic regression, as well as the subfield of deep learning that uses sophisticated, biologically inspired learning algorithms called neural networks. Generally, supervised algorithms learn from labeled data (eg, classification or regression), whereas unsupervised algorithms learn from hidden patterns in the unlabeled data (eg, clustering).

In the medical field, ML is becoming an increasingly central technique. For example, ML-based prediction models are being developed to detect early signs of diseases, improve decision-making processes, and track rehabilitation efficacy. Fueled by advances in data-recording technology, the increasing availability of patient data, and more accessible databases and code libraries, these models can generate more accurate insights about patients from large, existing health data sets. Contreras and Vehi [10] showed that within a decade, the number of articles proposing artificial intelligence models in diabetes research grew by 500%. Despite the large number of promising studies reported in the literature, the adoption of ML models in real-life clinical practice is low [11]. A wide range of ML models have been proposed for the automatic detection of PD [12]. Searching with only 1 query related to ML and PD results in over 1000 publications in 1 year alone. Despite the rising popularity of ML in PD research, models are rarely deployed in the field due to their irreproducibility and are limited for research purposes [13]. Although there may be many explanations, one possibility is a disconnect between the models developed in research and real-life implementation.

In contrast to previous systematic reviews that primarily explored data types and model variations, the emphasis of this review lies in the critical context of model validation approaches to provide a comprehensive understanding of the strengths and limitations of ML models in the PD field. Previous reviews emphasized data types; for instance, Ramdhani et al [14] reviewed sensor-based ML algorithms for PD predictions, and Mei et al [15] provided a comprehensive overview of outcomes associated with the type and source of data for 209 studies that applied ML models for PD diagnosis. Mei et al [15] also noted concerns about insufficient descriptions of methods, results, and validation techniques. We focused on the critical evaluation of validation techniques that are instrumental for the clinical integration of ML.

In this review, we examined a cross-section of recent ML prediction models related to PD detection and progression. Our goal was to summarize the different ML practices in PD research and identify areas for improvement related to model design, training, validation, tuning, and evaluation. Implementing best ML practices would help researchers develop PD prediction models that are more reproducible and generalizable, which in turn would improve their impact on the entire landscape of patient care and outcomes.

Methods

Search Strategy

We conducted a systematic literature review in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Checklist 1) guidelines in PubMed between January 2020 and April 2021, using the following exact string: “Parkinson’s” AND (“ML” OR “prediction” OR “classification” OR “detection” or “artificial intelligence” OR “AI”). The search resulted in 1085 publications.

Inclusion and Exclusion

Inclusion criteria were studies (1) on ML applied for predicting PD, PD subscores or PD severity, and PD symptoms; (2) published between January 2020 and April 2021; (3) written in English; and (4) with an available title and abstract.

Questionnaire Design

We designed a customized questionnaire to easily parse the literature and extract characteristics of the different ML approaches. Textbox 1 summarizes the model details extracted from the questionnaire, and the exact questionnaire is provided in Multimedia Appendix 1. This questionnaire was not intended to extract exhaustive details about these models, but rather to target specific concepts that seem to be inconsistently reported in the PD modeling literature. Our rationale for each question, and how they were designed specifically for PD, is provided below.

PD is a progressive neurological disorder, and symptoms can vary widely for each individual. To categorize PD progression and assess patient status, clinicians use standardized metrics such as the Unified Parkinson’s Disease Rating Scale [16] and Hoehn and Yahr (H&Y) scores [17]. The first question is related to clearly defining the research objectives or target outcomes of a particular study. The challenge of classifying PD versus non-PD may depend on symptom severity, which can be more readily assessed when severity metrics are available. In certain stages of PD, symptoms can be controlled or lessened through careful medication regimens, such as levodopa. This medication’s *on* and *off* periods are essential components for clinicians and researchers to consider. *On* and *off* episodes can create a substantially different effect on symptoms [18,19], and these symptoms are being used in ML algorithms to classify or assess PD. For example, Jahanshahi et al [20] investigated the levodopa medication’s effect on PD probabilistic classification learning and demonstrated that learning is associated with the patient with PD being in an *on* or *off* state. Warmerdam et al [21] showed that the patient’s state relative to dopaminergic medication correlated with the arm-swing task during PD walking. PD characteristics are important while researching PD, and the application of the models might play different roles depending on the data. As a result, the questions regarding the severity and medication state of patients can play a crucial role. In addition, class imbalance, cross-validation techniques, and hyperparameter tuning are critical concepts in ML. Class imbalance can lead to biased models or misinterpretation of results. Cross-validation and hyperparameter tuning allow systematic exploring of models and are essential for assessing models’ generalization performance. Lastly, comparing model performance to benchmark data can be valuable for research goals, but this process is not always applicable or possible.

Textbox 1. Model details obtained during data extraction (n=113).

1. What have the authors classified using machine learning?
2. Was there any information about the participants being on or off medication prior to the experiment?
3. Of the study participants, how many were (1) individuals with Parkinson disease, (2) controls, and (3) individuals with other diseases?
4. Did the study mention the distribution of the Unified Parkinson's Disease Rating Scale and Hoehn and Yahr scores?
5. What class imbalance mitigation techniques did the authors perform?
6. How did the authors split or cross-validate the data set while training the model? If cross-validation was applied, which particular strategies were applied?
7. If applicable, have the authors made the reader aware of the potential overinflated performance results (eg, the model overfitting the training data)? If so, how?
8. How was the hyperparameter tuning done?
9. Did the authors analyze and discuss the models' errors or misclassifications?
10. How did they compare their model to other modeling approaches by themselves or other authors, directly or indirectly?
11. Did the authors use multiple evaluation metrics to measure the performance of the model(s)?

Data Extraction

Two authors assessed the inclusion criteria of 1085 studies based on the title and abstract. During the initial manual screening of the title and abstract, 155 studies that met the initial inclusion criteria were identified. A total of 42 studies were excluded after assessing the full text for eligibility. These authors also extracted data from the studies using the questionnaire described above. Ultimately, 113 studies and the corresponding questionnaire responses were rechecked independently by both reviewers, and disagreements were resolved through discussion to reach a consensus. Questionnaire data from each study are provided in in Multimedia Appendix 2.

For the multiple-choice and checkbox questions (ie, questions 1, 7, 8, 9, 10, 11, 13, 14, and 15), we counted the number of times each response occurred in the results.

Results

First, we provide a general overview of the study characteristics in each publication. Then, we examine specific results evaluating the ML modeling practices using the following categories: PD characteristics, class imbalance, data set splitting, overfitting, hyperparameter tuning, and model comparisons.

General Overview of Studies

Methods Applied

The most prevalent ML classification algorithms were support vector machines (53/113, 46.9%), boosting ensemble learning (48/113, 42.5%; eg, gradient boosting, extreme gradient boosting, and random forest), naive Bayes (4/113, 3.5%), decision tree (13/113, 11.5%), and *k*-nearest neighbor (22/113, 19.5%). In regression models, the most prevalent methods included multiple linear or logistic regression (32/113, 28.3%), regression trees, *k*-means clustering, and Bayesian regression (3/113, 2.6%). Deep learning methods included convolutional neural networks (10/113, 8.8%), variants of recurrent neural networks (4/113, 3.5%; eg, long short-term memory [LSTM] and bidirectional-LSTM), and fully connected neural networks (22/113, 19.5%).

Data Modalities and Sources

More than half of the studies (65/113, 57.5%) used data collected by the authors, whereas 38.9% (44/113) used a public data set and 3.6% (4/113) used a mixture of public and private data sets. The most common data modalities were magnetic resonance imaging, single-photon emission computerized tomography imaging, voice recordings or features, gait movements, handwriting movements, surveys, and cerebrospinal fluid features.

ML Modeling Practices

PD Prediction Target

We categorized the studies based on 5 ML outcomes for PD models: *PD versus non-PD classification*, *PD severity prediction*, *PD versus non-PD versus other diseases classification*, *PD symptoms quantification*, and *PD progression prediction*. A total of 10 studies fell into more than 1 category; among them, 8 (80%) studies examined both *PD versus non-PD classification* and *PD severity regression*, and 2 (20%) studies examined *PD versus non-PD classification* and *PD symptoms quantification*.

1. *PD versus non-PD classification* (59/113, 52.2%): studies that proposed ML methods to distinguish between individuals with PD from controls without PD
2. *PD severity prediction* (30/113, 26.5%): studies that proposed ML methods to predict the stages of Unified Parkinson's Disease Rating Scale scores or H&Y scores of PD
3. *PD versus non-PD versus other diseases classification* (24/113, 21.2%): studies that proposed ML methods to distinguish between PD, non-PD, and other diseases (eg, Alzheimer disease)
4. *PD symptoms quantification* (9/113, 8%): studies that proposed ML methods to distinguish between PD symptoms (eg, tremor and bradykinesia) from no symptoms or non-PD symptoms

5. *PD progression prediction* (1/113, 0.9%): studies that proposed ML methods to predict PD progression

PD versus non-PD classification and *PD versus non-PD versus other diseases classification* have target settings that are binary variable predictions, as these targets are mostly for predicting the presence or absence of PD. *PD severity prediction* can be categorical (multilabel classification) or continuous (regression), such as predicting the H&Y score. *PD symptoms quantification* can also be categorical, such as predicting the presence of resting tremors, rigidity, and bradykinesia, or continuous, such as predicting the degree of tremor intensity. *PD progression prediction* measures the changes in overall disease severity at multiple time points. We found that most studies (107/113, 94.6%) indicated PD severity. However, fewer than half (53/113, 46.9%) of the studies reported the patient medication status directly, with 38.9% (44/113) using public data sets.

Class Imbalance

Class imbalance occurs when 1 training class contains significantly fewer samples than another class. In this case, the learners tend to focus on the better performance of the majority group, making it difficult to interpret the evaluation metrics, such as accuracy, for groups with less representation. Prediction models can be significantly affected by the imbalance problem. ML models can be highly unstable with different imbalance ratios [22]. On predicting *PD versus non-PD classification*, performance can suffer significantly from an imbalanced data set and generate impaired results [23]. Class imbalance can impact model external validity, and either mitigating or at least reporting the potential concerns in the interpretability of outcomes due to imbalances would help the reader interpret the model's power for predicting each class.

There are multiple ways to handle a class imbalance in the training phase, such as using resampling techniques or weighted evaluation metrics. Resampling creates a more balanced training data set, such as by oversampling the minority class or undersampling the majority class [24,25]. Moreover, there are alternative evaluation metrics, for example, balanced accuracy and *F*-measure, but these improvements on the standard evaluation metrics are also affected by class imbalance [26]. We observed that among the studies that attempted to mitigate class imbalance, many of them adopted under- or oversampling methods and then applied class weights to the evaluation metrics. Other techniques were data augmentation and grouping data to use the same ratio of minority and majority classes. In the case of extreme class imbalance, Megahed et al [27] were not able to mitigate overfitting. Overall, there is no perfect solution to tackle this critical issue in ML; however, recognizing that the problem exists and investigating appropriate mitigation strategies should be standard practice. Our results found at least moderate class imbalance in more than two-thirds (77/113, 68.1%) of the studies, and only 18% (5/27), 31% (5/16), 27% (8/30), and 25% (1/4) of studies for the *PD versus non-PD classification*, *PD versus non-PD versus other diseases classification*, *PD severity prediction*, and *PD symptoms quantification and progression prediction* target categories applied strategies to mitigate the effects of class imbalance, respectively. In Figure 1, we illustrate the number of studies with more than 30% class imbalance and how many of them applied imbalance mitigation strategies.

In some cases, authors applied class imbalance strategies but found no significant improvement in their model performance. Reporting these cases still provides valuable perspectives. For instance, van den Goorbergh et al [28] illustrated that correcting for imbalance resulted in the model exhibiting strong miscalibration and did not improve the model's capability to distinguish between patients and controls. A total of 4 studies compared results when using imbalanced data compared to imbalance-mitigated data. Details of these studies are provided in Table 1.

Figure 1. Number of studies with more than 30% class imbalance and the percentage of studies that applied the class imbalance strategies, separated by PD prediction target. In the *PD versus non-PD classification*, *PD versus non-PD versus other diseases classification*, *PD severity prediction*, and *PD symptoms quantification and progression prediction* categories, 46% (27/59), 67% (16/24), 100% (30/30), and 40% (4/10) had class imbalance, but only 8% (5/59), 21% (8/30), 27% (8/30), and 10% (1/10) applied mitigation strategies, respectively. PD: Parkinson disease.

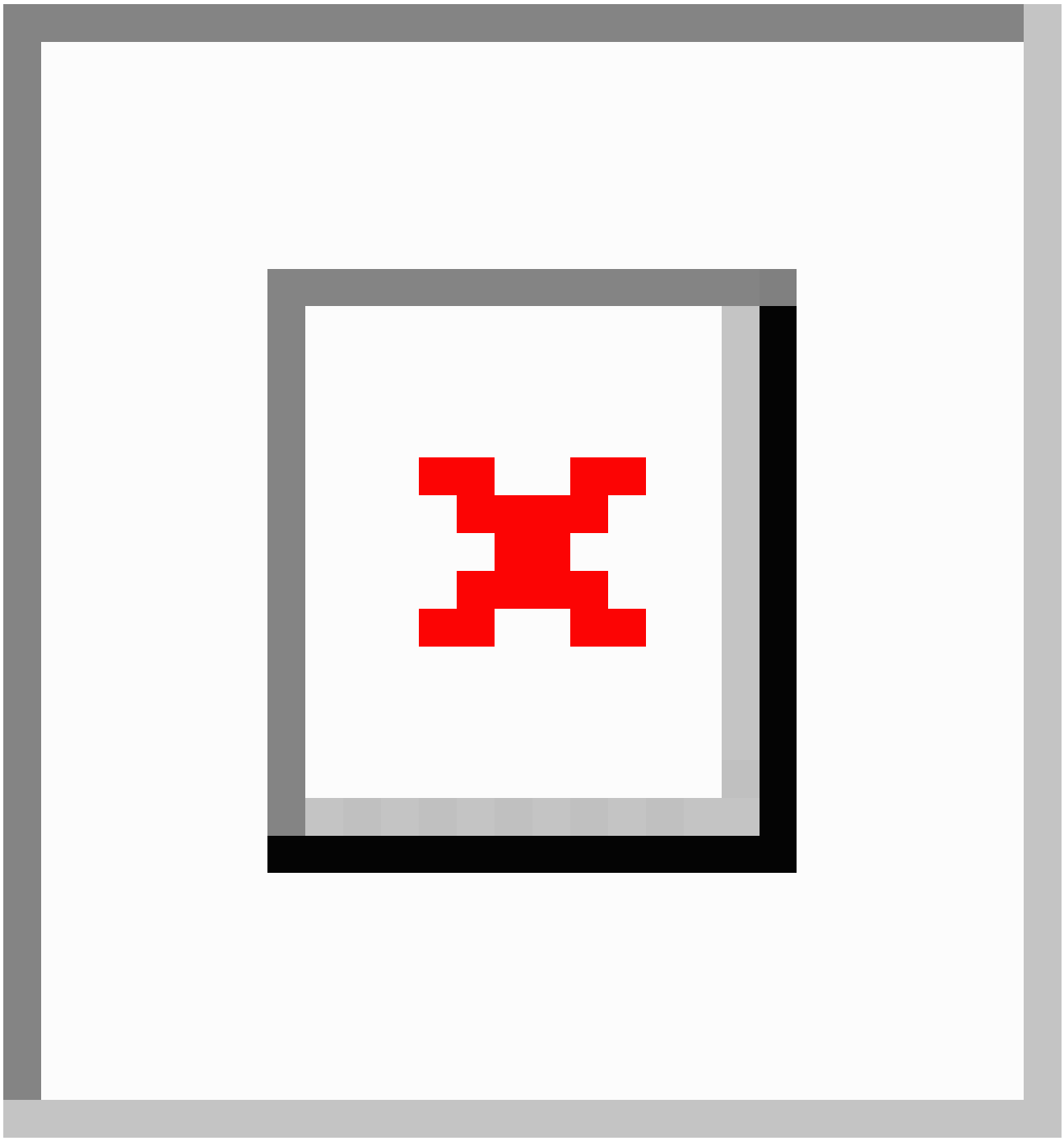


Table . Comparison between imbalanced data versus imbalance mitigation strategies.

Studies	Participant distribution	Techniques	Conclusion
Moon et al [29]	524 patients with PD[a] and 43 patients with essential tremor	• SMOTE[b]	• F_1-score improved
Veeraragavan et al [30]	93 patients with idiopathic PD and 73 controls; 10 patients with H&Y[c] 3; 28 patients with H&Y 2.5; and 55 patients with H&Y 2	• SMOTE	• Test accuracy improved
Falchetti et al [31]	388 patients with idiopathic PD and 323 controls	• Oversampling • Undersampling • Combination of oversampling and undersampling	• Without any sampling, the combination of oversampling and undersampling methods is comparable
Jeancolas et al [32]	115 patients with PD and 152 controls	• Data augmentation	• Performed better for free speech task • No consistent improvement in the sentence repetition task

[a]PD: Parkinson disease.

[b]SMOTE: synthetic minority oversampling technique.

[c]H&Y: Hoehn and Yahr.

Data Set Splitting

It is universally acknowledged that ML models can perform arbitrarily well on data that were used to create the model—that is, the training data set. This is why standard procedure in training models uses separate data sets to try different model variations and select the better variants. The confusion that sometimes occurs is when these separate data sets are used to select from a large number of model variants (validation set) or only used for the evaluation of selected variants (test set). The distinction in these 2 use cases of separate data is sometimes not clear and depends on the number of model variants tested. Critically, with modern ML practice, many model variants are often tested on provided data, which readily leads to overfitting on both the original training data and validation set used for evaluation. A separate holdout test set would be needed to properly evaluate model performance [33]. A single split can be error prone in estimating performance [34]. It is critical to have a holdout test set to provide better performance estimation. Additionally, cross-validation is a technique largely used to estimate and compare model performance or to optimize the hyperparameters [35]. Cross-validation divides the data into folds and iterates on these folds to test and train the models using different partitions of the data set. We found that 78.8% (89/113) of the studies used cross-validation; however, 5.3% (6/113) of the studies either did not mention the details of the validation procedure or did not do any splitting. A total of 9.7% (11/113) of the studies split the data set into only 2 sets, but it was not clear if the separate set was a validation set or a test set. Only 19.5% (22/113) of the studies applied cross-validation without a holdout test set (Table 2 and Figure S1 in Multimedia Appendix 3).

Table . Distribution of studies according to data set splitting techniques.

Data set splitting techniques	Studies (n=113), n (%)
Not mentioned	6 (5.3)
Split into 2 sets (training, test, or validation sets)	11 (9.7)
Only cross-validation	22 (19.5)
Split into 3 sets	7 (6.2)
Cross-validation and holdout test set	67 (59.3)

Cross-Validation

There are multiple types of cross-validation techniques. In *k*-fold cross-validation, the data set is divided into *k* equal folds randomly, and the model is trained and evaluated *k* times. Each time, the model is trained using *k*–1 folds and evaluated in the remaining fold. When the observations are independent and identically distributed, *k*-fold cross-validation works well. When the data are not identically distributed, *k*-fold cross-validation makes the model prone to overfitting and not generalize well [36]. For instance, multiple data samples from the same patient should generally not be present in both training and testing data sets. Subject-wise cross-validation separates folds according to the subject. Although Saeb et al [37] concluded that subject-wise methods are more clinically relevant compared to record-wise methods, Little et al [38] argued that subject-wise methods might not be the best in all use cases. However, Westerhuis et al [39] demonstrated that cross-validation can be overoptimistic and suggested that it is good practice to include a separate test set at the end to properly evaluate a model. To reduce bias in

model evaluation, nested cross-validation is another technique that involves 2 cross-validation loops [40]. The outer loop generates *k*-folds and iterates through them, so each fold is eventually used as a holdout test fold for a model developed using the remaining data. The inner loop uses a similar *k*-fold procedure to create a holdout validation fold that is used to select the best model during model tuning. Nested cross-validation is a more robust way to evaluate models than *k*-fold cross-validation alone, since using all available data to select the model architecture can lead to biased, overfitted results [40]. However, nested cross-validation is more computationally intensive, and these models can be difficult to interpret or implement (since they actually result in *k*-best models, so performance is usually averaged over all *k*-best models). In our analysis, we found that the most common cross-validation technique is *k*-fold cross-validation (68/113, 60.2%), whereas only 4.4% (5/113) of the studies adopted nested cross-validation (Table 3 and Figure S2 in Multimedia Appendix 3). Of the 113 studies, 20 (17.7%) adopted 2 types of cross-validation techniques, and 5 (4.4%) adopted 3 types of techniques.

Table . Distribution of studies that adopted cross-validation techniques.

Cross-validation techniques	Studies (n=113), n (%)
k-fold cross-validation	68 (60.2)
Leave-p-out cross-validation	25 (22.1)
Stratified or subject-wise cross-validation	21 (18.6)
Nested cross-validation	5 (4.4)
No cross-validation	24 (21.2)

Overfitting

We selected publications that did not evaluate their models with a holdout test set and then we analyzed if they mentioned that the proposed models could possibly be overfitting. Models can be overfitted for multiple reasons, such as an imbalanced data set or the lack of proper model selection and validation technique. Even with cross-validation, if a separate holdout set is not used, then the results can be inflated. Rao et al [41] demonstrated that leave-one-out cross-validation can achieve 100% sensitivity, but performance on a holdout test set can be significantly lower. Cross-validation alone is not sufficient model validation when the dimensionality of the data is high [41]. However, there are multiple ways to address or prevent overfitting, such as the examples provided by Ying [42]. Making the reader aware of overfitting concerns in the interpretability of results should be standard practice. Therefore, we searched to see if the authors mentioned that their model can suffer from overfitting. For this analysis, we excluded studies that applied the cross-validation technique with a holdout test set. We found that just over 54% (25/46) of the studies that likely suffer from overfitting did not mention it as a concern. Although 45% (21/46) of studies mentioned overfitting as a potential limitation, many of them did not have any detailed discussion about this.

Hyperparameters

While training a model, hyperparameters are selected to define the architecture of the model. These hyperparameters are often tuned so that the model gives the best performance. A common method of finding the best hyperparameters is by defining a range of parameters to test, then applying a grid search or random search on the fixed search space, and finally selecting parameters to minimize the model error [43]. These methods can be extremely computationally expensive and time-consuming depending on data complexity and available computation power [44]. Regardless of the method applied, it is considered good practice to make clear statements about the tuning process of hyperparameters to improve reproducibility [45]. This practice ensures parameters are properly selected and models are ready for direct comparison. Our results demonstrated that 38.9% (44/113) of studies did not report on hyperparameter tuning (Table 4 and Figure S3 in Multimedia Appendix 3). Of these, 2 adopted least absolute shrinkage and selection operator logistic regression, and 3 used a variant of logistic regression or linear regression, which typically have few or no hyperparameters to adjust.

Table . Distribution of studies according to hyperparameter tuning methods.

Hyperparameter tuning methods	Studies (n=113), n (%)
Not reported	44 (38.9)
Ad hoc	31 (27.4)
Random search	1 (0.9)
Grid search	27 (23.9)
Others	10 (8.8)

For many other models, there are inherently only a few hyperparameters that are usually adjusted; for instance, the major hyperparameter for the neighbor model is the number of neighbors, *k*. On the other hand, more complex models such as convolutional neural networks and LSTM require thorough tuning to achieve meaningful performance. Regardless of the

number of hyperparameters in a model, proper tuning would likely still contribute to achieving optimal performance. The choice of hyperparameters will impact model generalization, so it is worthwhile to examine changes in performance with different settings [46].

Model Comparison

In research domains that require complex deep learning models to achieve state-of-the-art performance, such as computer vision and natural language processing, it has become a regular practice to compare models with numeric benchmark data sets to contextualize their proposed model and provide insight into the model's relative performance to peers. Although such rigorous benchmarking and comparison is not possible given the heterogeneous data sets in PD research, it is important to contextualize a model's performance relative to other models, strategies, and data sets. We found that 66.4% (75/113) of studies compared results from multiple alternative models in their work, and 15% (17/113) of studies compared their results with previously published models. However, 18.6% (21/113) of studies only reported their single model performance and made no comparison to any other models or benchmarks (Table 5 and Figure S4 in Multimedia Appendix 3).

Table . Distribution of studies according to model comparison methods; 18.6% (21/113) of studies did not compare their model results to any alternative models or previously published models or benchmarks.

Model comparison methods	Studies (n=113), n (%)
Compared with their own multiple models	75 (66.4)
Compared with previous models or benchmarks	4 (3.5)
Compared with previous models and their own multiple models	13 (11.5)
No comparisons	21 (18.6)

Discussion

Principal Findings

In summary, we have comprehensively reviewed the general practices of ML research applied to PD in a recent cross-section of publications. We have identified several important areas of improvement for model building to reduce the disparity between in-the-lab research and real-world clinical applications. Standardizing the model reporting techniques and implementing best ML practices would increase the acceptability and reliability of these models to improve patient evaluation and care [47].

For the interoperability and usability of the models, clinicians need detailed information about the patients included in the model's training data, such as their medication state and PD progression stage. This information determines the predictive validity of a model to new patients and settings. We found that 94.7% (107/113) of the studies explained the PD severity of their patients, whereas only 46.9% (53/113) of studies reported the medication state of the patients. To incorporate data-driven algorithms in real life, the description of medication is significantly relevant to PD [48,49]. The overall representation of demographic samples in the training set should be accounted for as well. Our results show that 68.1% (77/113) of the studies had a class imbalance greater than 30% difference in their data set, and less than one-third (from 5/27, 18% to 5/16, 31%) of the studies addressed imbalance as a potential issue or considered its impact on the model results.

Another major finding is the lack of a standard reporting framework for a model's hyperparameter search and tuning. Hyperparameter tuning has a major impact on the model configuration and, by extension, its performance [50]. For example, Wong et al [51] demonstrated that a model using tuned (grid-searched) hyperparameters outperformed a model using default hyperparameters. Addressing hyperparameters is also essential for reproducibility, including a report on the final model configuration and how the authors made the decision. Although this is a considerably important aspect of ML model reporting, our study showed that 44 (38.9%) of the 113 studies did not report the hyperparameter tuning approach. Of these, 5 studies adopted logistic regression or linear regression. Traditional regression models are not expected to undergo significant hyperparameter tuning; however, variants that involve hyperparameters would likely still benefit from tuning. Consistent reporting of hyperparameter tuning practices will enhance the robustness and reliability of these models.

Moreover, to provide context to the results of model performance, comparisons of different models or with previously published models give a general idea of the quality of the proposed models. We found that 18.6% (21/113) of the studies only reported their proposed models; on the contrary, the reporting standard of proposed models in the computer vision and natural language processing fields is extensive. For instance, Wang et al [52] and Liu et al [53] proposed methods for visual recognition, and they reported large-scale experiment results with different data sets and compared their results with more than 10 previously proposed methods. Similarly, in natural language processing, to propose a task such as emotion cause extraction, Xia and Ding [54] compared around 8 methods with different evaluation metrics. These are a few cases to demonstrate that such comparisons are widely executed in the computer vision and natural language processing communities to propose a method. This systematic practice of comparison with previously published approaches results in reproducibility. Unfortunately, we found that only 15% (17/113) of the studies compared with previously proposed methods. However, in the medical field, due to the challenges of data availability, proper comparisons might not be possible.

There are several factors in ML and deep learning research that can create misleading results. One major factor is proper model validation, particularly in how the training and test data are separated. We found that 5.3% (6/113) of studies either did not

provide the details about data set splitting or did not do any splitting, and 15.9% (18/113) of studies performed static training, validation, and test set separation, which provides limited stability of scores. Cross-validation is a more stable validation method conducted while training the model and reduces the risk of overfitting [55]. The majority (89/113, 78.8%) of studies adopted some form of cross-validation, and the most common cross-validation technique adopted was *k*-fold (68/113, 60.2%). Nevertheless, the use case of different validation techniques depends on the data set and is problem specific. As powerful as cross-validation is in creating reliable models, applying simple cross-validation does not guarantee that the model is not overfitted [41]. For the studies that did not evaluate their results with a holdout test set in a cross-validation manner, we extracted information from their discussion sections. To be precise, we checked if they made their reader aware of how the study results might be overfitting. We found that 46% (21/46) of the studies that are potentially reporting overfitted scores did not mention this concern. The developed models should be reported with their limitations for transparency to allow for further improvement and real-world adoption.

In this systematic review, we sampled 113 recent studies on PD to summarize the standard ML practices and addressed broader concerns on reporting strategies. It is challenging for authors to always implement the best practices considering the practical realities of health care data, including limited sample sizes, noisy data, medical data privacy, etc. However, whenever possible, authors should consider these reporting practices, especially to acknowledge limitations in their data, model design, and performance. This will help to determine reasonable use cases for these models or to identify areas of improvement before they are ready for clinical translation. These considerations can also extend to other health care applications of ML.

Conclusion

Despite the increasing number of studies, our results demonstrate there are still many opportunities for improvement in reporting and implementing ML for applications in PD detection and progression. Studies should report detailed, standardized patient characteristics; use robust validation techniques to ensure the model's reliability; and justify choices of evaluation as well as hyperparameters. We found that 75% (58/77) of the studies sampled from 2020 to 2021 did not address class imbalance, and one-third (44/113, 38.9%) of studies did not report hyperparameter tuning. Reporting is the first step to understanding the usability and interpretation of models. By shifting the focus to the critical evaluation of these methods, we aim to improve the reporting and review of ML to strengthen the connection between research and real-world clinical applications. Ideally, the processes can be standardized, and clinical measurements can be leveraged more effectively for prediction models to improve the real-world impact on individuals with PD or other health conditions.

Data Availability

All data generated or analyzed during this study are included in this paper.

Authors' Contributions

TT, MVA, and MKO conceptualized the study. TT and RCS conducted the review, extracted the data, and conducted the analysis. TT wrote the paper. MKO and MVA revised the paper and supervised the study. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Customized questionnaire.
[PDF File, 74 KB - medinform_v12i1e50117_app1.pdf]

Multimedia Appendix 2
List of included studies.
[PDF File, 171 KB - medinform_v12i1e50117_app2.pdf]

Multimedia Appendix 3
Graphical representations of data.
[DOCX File, 15 KB - medinform_v12i1e50117_app3.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File, 83 KB - medinform_v12i1e50117_app4.pdf]

References

1. Garrote JAD, Cervantes CE, Díaz MS. Prediagnostic presentations of Parkinson's disease in primary care: a case-control study [Article in Spanish]. Semergen 2015;41(5):284-286. [doi: 10.1016/j.semerg.2015.01.007] [Medline: 25752864]
2. Rizzo G, Copetti M, Arcuti S, Martino D, Fontana A, Logroscino G. Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis. Neurology 2016 Feb 9;86(6):566-576. [doi: 10.1212/WNL.0000000000002350] [Medline: 26764028]
3. Pagan FL. Improving outcomes through early diagnosis of Parkinson's disease. Am J Manag Care 2012 Sep;18(7 Suppl):S176-S182. [Medline: 23039866]
4. Postuma RB, Berg D. Advances in markers of prodromal Parkinson disease. Nat Rev Neurol 2016 Oct 27;12(11):622-634. [doi: 10.1038/nrneurol.2016.152] [Medline: 27786242]
5. Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008 Apr;79(4):368-376. [doi: 10.1136/jnnp.2007.131045] [Medline: 18344392]
6. Massano J, Bhatia KP. Clinical approach to Parkinson's disease: features, diagnosis, and principles of management. Cold Spring Harb Perspect Med 2012 Jun;2(6):a008870. [doi: 10.1101/cshperspect.a008870] [Medline: 22675666]
7. Zhang J. Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease. NPJ Parkinsons Dis 2022 Jan 21;8(1):13. [doi: 10.1038/s41531-021-00266-8] [Medline: 35064123]
8. Miljkovic D, Aleksovski D, Podpečan V, Lavrač N, Malle B, Holzinger A. Machine learning and data mining methods for managing Parkinson's disease. In: Holzinger A, editor. Machine Learning for Health Informatics. Lecture Notes in Computer Science: Springer; 2016, Vol. 9605:209-220. [doi: 10.1007/978-3-319-50478-0_10]
9. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach: Prentice Hall/Pearson Education; 2003.
10. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018 May 30;20(5):e10775. [doi: 10.2196/10775] [Medline: 29848472]
11. Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509. [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
12. Bind S, Tiwari AK, Sahani AK, et al. A survey of machine learning based approaches for Parkinson disease prediction. International Journal of Computer Science and Information Technologies 2015;6(2):1648-1655 [FREE Full text]
13. Salari N, Kazeminia M, Sagha H, Daneshkhah A, Ahmadi A, Mohammadi M. The performance of various machine learning methods for Parkinson's disease recognition: a systematic review. Curr Psychol 2023 Jul;42(20):16637-16660. [doi: 10.1007/s12144-022-02949-8]
14. Ramdhani RA, Khojandi A, Shylo O, Kopell BH. Optimizing clinical assessments in Parkinson's disease through the use of wearable sensors and data driven modeling. Front Comput Neurosci 2018 Sep 11;12:72. [doi: 10.3389/fncom.2018.00072] [Medline: 30254580]
15. Mei J, Desrosiers C, Frasnelli J. Machine learning for the diagnosis of Parkinson's disease: a review of literature. Front Aging Neurosci 2021 May 6;13:633752. [doi: 10.3389/fnagi.2021.633752] [Medline: 34025389]
16. Martínez-Martín P, Gil-Nagel A, Gracia LM, Gómez JB, Martínez-Sarriés J, Bermejo F. Unified Parkinson's Disease Rating Scale characteristics and structure. Mov Disord 1994 Jan;9(1):76-83. [doi: 10.1002/mds.870090112] [Medline: 8139608]
17. Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. Neurology 1967 May;17(5):427-442. [doi: 10.1212/wnl.17.5.427] [Medline: 6067254]
18. Verbaan D, van Rooden SM, van Hilten JJ, Rijsman RM. Prevalence and clinical profile of restless legs syndrome in Parkinson's disease. Mov Disord 2010 Oct 15;25(13):2142-2147. [doi: 10.1002/mds.23241] [Medline: 20737549]
19. Martínez-Fernández R, Schmitt E, Martinez-Martin P, Krack P. The hidden sister of motor fluctuations in Parkinson's disease: a review on nonmotor fluctuations. Mov Disord 2016 Aug;31(8):1080-1094. [doi: 10.1002/mds.26731] [Medline: 27431515]
20. Jahanshahi M, Wilkinson L, Gahir H, Dharmaindra A, Lagnado DA. Medication impairs probabilistic classification learning in Parkinson's disease. Neuropsychologia 2010 Mar;48(4):1096-1103. [doi: 10.1016/j.neuropsychologia.2009.12.010] [Medline: 20006629]
21. Warmerdam E, Romijnders R, Hansen C, et al. Arm swing responsiveness to dopaminergic medication in Parkinson's disease depends on task complexity. NPJ Parkinsons Dis 2021 Oct 5;7(1):89. [doi: 10.1038/s41531-021-00235-1] [Medline: 34611152]
22. Yu Q, Jiang S, Zhang Y. The performance stability of defect prediction models with class imbalance: an empirical study. IEICE Trans Inf Syst 2017;E100.D(2):265-272. [doi: 10.1587/transinf.2016EDP7204]
23. Dinov ID, Heavner B, Tang M, et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PLoS One 2016 Aug 5;11(8):e0157077. [doi: 10.1371/journal.pone.0157077] [Medline: 27494614]
24. Brownlee J. Imbalanced Classification with Python: Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning: Machine Learning Mastery; 2020.
25. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from Imbalanced Data Sets: Springer; 2018. [doi: 10.1007/978-3-319-98074-4]

26. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
27. Megahed FM, Chen YJ, Megahed A, Ong Y, Altman N, Krzywinski M. The class imbalance problem. Nat Methods 2021 Nov;18(11):1270-1272. [doi: 10.1038/s41592-021-01302-4] [Medline: 34654918]
28. van den Goorbergh R, van Smeden M, Timmerman D, van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc 2022 Aug 16;29(9):1525-1534. [doi: 10.1093/jamia/ocac093] [Medline: 35686364]
29. Moon S, Song HJ, Sharma VD, et al. Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. J Neuroeng Rehabil 2020 Sep 11;17(1):125. [doi: 10.1186/s12984-020-00756-5] [Medline: 32917244]
30. Veeraragavan S, Gopalai AA, Gouwanda D, Ahmad SA. Parkinson's disease diagnosis and severity assessment using ground reaction forces and neural networks. Front Physiol 2020 Nov 9;11:587057. [doi: 10.3389/fphys.2020.587057] [Medline: 33240106]
31. Falchetti M, Prediger RD, Zanotto-Filho A. Classification algorithms applied to blood-based transcriptome meta-analysis to predict idiopathic Parkinson's disease. Comput Biol Med 2020 Sep;124:103925. [doi: 10.1016/j.compbiomed.2020.103925] [Medline: 32889300]
32. Jeancolas L, Petrovska-Delacrétaz D, Mangone G, et al. X-vectors: new quantitative biomarkers for early Parkinson's disease detection from speech. Front Neuroinform 2021 Feb 19;15:578369. [doi: 10.3389/fninf.2021.578369] [Medline: 33679361]
33. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods 2016 Sep;13(9):703-704. [doi: 10.1038/nmeth.3968]
34. Harrington P. Multiple versus single set validation of multivariate models to avoid mistakes. Crit Rev Anal Chem 2018 Jan 2;48(1):33-46. [doi: 10.1080/10408347.2017.1361314] [Medline: 28777019]
35. Refaeilzadeh P, Tang L, Liu H. Cross-validation. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems: Springer; 2009:532-538. [doi: 10.1007/978-0-387-39940-9_565]
36. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998 Sep 15;10(7):1895-1923. [doi: 10.1162/089976698300017197] [Medline: 9744903]
37. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. Gigascience 2017 May 1;6(5):1-9. [doi: 10.1093/gigascience/gix019] [Medline: 28327985]
38. Little MA, Varoquaux G, Saeb S, et al. Using and understanding cross-validation strategies. perspectives on Saeb et al. Gigascience 2017 May 1;6(5):1-6. [doi: 10.1093/gigascience/gix020] [Medline: 28327989]
39. Westerhuis JA, Hoefsloot HCJ, Smit S, et al. Assessment of PLSDA cross validation. Metabolomics 2008 Mar;4(1):81-89. [doi: 10.1007/s11306-007-0099-6]
40. Cawley GC, Talbo NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010 Oct 7;11:2079-2107 [FREE Full text]
41. Rao RB, Fung G, Rosales R. On the dangers of cross-validation. an experimental evaluation. In: Apte C, Park H, Wang K, et al, editors. Proceedings of the 2008 SIAM International Conference on Data Mining: Society for Industrial and Applied Mathematics; 2008:588-596. [doi: 10.1137/1.9781611972788.54]
42. Ying X. An overview of overfitting and its solutions. J Phys Conf Ser 2019;1168(2):022022. [doi: 10.1088/1742-6596/1168/2/022022]
43. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012 Dec 2;13:281-305 [FREE Full text]
44. Claesen M, de Moor B. Hyperparameter search in machine learning. arXiv. Preprint posted online on Apr 6, 2015. [doi: 10.48550/arXiv.1502.02127]
45. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020 Oct;13(10):e006556. [doi: 10.1161/CIRCOUTCOMES.120.006556] [Medline: 33079589]
46. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 2020 Nov 20;415:295-316. [doi: 10.1016/j.neucom.2020.07.061]
47. Bin Rafiq R, Modave F, Guha S, Albert MV. Validation methods to promote real-world applicability of machine learning in medicine. In: DMIP '20: 2020 3rd International Conference on Digital Medicine and Image Processing: Association for Computing Machinery; 2020:13-19. [doi: 10.1145/3441369.3441372]
48. Goberman A, Coelho C, Robb M. Phonatory characteristics of Parkinsonian speech before and after morning medication: the on and off states. J Commun Disord 2002;35(3):217-239. [doi: 10.1016/s0021-9924(01)00072-7] [Medline: 12064785]
49. Adamson MB, Gilmore G, Stratton TW, Baktash N, Jog MS. Medication status and dual-tasking on turning strategies in Parkinson disease. J Neurol Sci 2019 Jan 15;396:206-212. [doi: 10.1016/j.jns.2018.11.028] [Medline: 30504066]
50. Liao L, Li H, Shang W, Ma L. An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. ACM Trans Softw Eng Methodol 2022 Apr 9;31(3):1-40. [doi: 10.1145/3506695]

51. Wong J, Manderson T, Abrahamowicz M, Buckeridge DL, Tamblyn R. Can hyperparameter tuning improve the performance of a super learner? a case study. Epidemiology 2019 Jul;30(4):521-531. [doi: 10.1097/EDE.0000000000001027] [Medline: 30985529]
52. Wang P, Han K, Wei XS, Zhang L, Wang L. Contrastive learning based hybrid networks for long-tailed image classification. Presented at: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 20 to 25, 2021; Nashville, TN. [doi: 10.1109/CVPR46437.2021.00100]
53. Liu J, Li W, Sun Y. Memory-based jitter: improving visual recognition on long-tailed data with diversity in memory. Proc AAAI Conf Artif Intell 2022 Jun 28;36(2):1720-1728. [doi: 10.1609/aaai.v36i2.20064]
54. Xia R, Ding Z. Emotion-cause pair extraction: a new task to emotion analysis in texts. In: Korhonen A, Traum D, Márquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2019:1003-1012. [doi: 10.18653/v1/P19-1096]
55. King RD, Orhobor OI, Taylor CC. Cross-validation is safe to use. Nat Mach Intell 2021 Apr 20;3(4):276. [doi: 10.1038/s42256-021-00332-z]

Abbreviations

H&Y: Hoehn and Yahr
LSTM: long short-term memory
ML: machine learning
PD: Parkinson disease
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Benis; submitted 19.06.23; peer-reviewed by J Wong, S Marceglia, U Kanike; revised version received 12.02.24; accepted 01.04.24; published 17.05.24.

Please cite as:
Tabashum T, Snyder RC, O'Brien MK, Albert MV
Machine Learning Models for Parkinson Disease: Systematic Review
JMIR Med Inform 2024;12:e50117
URL: https://medinform.jmir.org/2024/1/e50117
doi:10.2196/50117

Ventilator-Associated Pneumonia Prediction Models Based on AI: Scoping Review

Jinbo Zhang[1,2], BA; Pingping Yang[1,2], BA; Lu Zeng[1,2], BA; Shan Li[1,2], BA; Jiamei Zhou[1,2], BA, MA

[1]

[2]

Corresponding Author:
Jiamei Zhou, BA, MA

Abstract

Background: Ventilator-associated pneumonia (VAP) is a serious complication of mechanical ventilation therapy that affects patients' treatments and prognoses. Owing to its excellent data mining capabilities, artificial intelligence (AI) has been increasingly used to predict VAP.

Objective: This paper reviews VAP prediction models that are based on AI, providing a reference for the early identification of high-risk groups in future clinical practice.

Methods: A scoping review was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. The Wanfang database, the Chinese Biomedical Literature Database, Cochrane Library, Web of Science, PubMed, MEDLINE, and Embase were searched to identify relevant articles. Study selection and data extraction were independently conducted by 2 reviewers. The data extracted from the included studies were synthesized narratively.

Results: Of the 137 publications retrieved, 11 were included in this scoping review. The included studies reported the use of AI for predicting VAP. All 11 studies predicted VAP occurrence, and studies on VAP prognosis were excluded. Further, these studies used text data, and none of them involved imaging data. Public databases were the primary sources of data for model building (studies: 6/11, 55%), and 5 studies had sample sizes of <1000. Machine learning was the primary algorithm for studying the VAP prediction models. However, deep learning and large language models were not used to construct VAP prediction models. The random forest model was the most commonly used model (studies: 5/11, 45%). All studies only performed internal validations, and none of them addressed how to implement and apply the final model in real-life clinical settings.

Conclusions: This review presents an overview of studies that used AI to predict and diagnose VAP. AI models have better predictive performance than traditional methods and are expected to provide indispensable tools for VAP risk prediction in the future. However, the current research is in the model construction and validation stage, and the implementation of and guidance for clinical VAP prediction require further research.

(JMIR Med Inform 2024;12:e57026) doi:10.2196/57026

KEYWORDS

artificial intelligence; machine learning; ventilator-associated pneumonia; prediction; scoping; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Background

Ventilator-associated pneumonia (VAP) is a pulmonary infectious disease that occurs in patients who receive mechanical ventilation for more than 48 hours and is primarily caused by pathogens that are present in the hospital environment. VAP is one of the most common complications in patients who undergo invasive mechanical ventilation. The incidence of VAP among patients who undergo mechanical ventilation ranges from 5% to 40%, depending on the setting and diagnostic criteria. The estimated attributable mortality rate of VAP is approximately 10%, with higher mortality rates among surgical intensive care unit (ICU) patients and those with moderate severity scores at admission [1]. VAP seriously affects the treatments and prognoses of patients, resulting in prolonged hospital stays, increased medical costs, and increased mortality rates. The early identification of groups at high risk for VAP is important for reducing VAP incidence and mortality [2].

Artificial intelligence (AI) can contribute to significant developments in the medical field. With the popularity of electronic health records, advancements in hardware computing power, and the development of big data, AI has become the optimal tool [3]. Among predictive models, AI models perform better than traditional models in various ways [4]. Data mining of patient cases via AI technology is conducted to create tools

that can predict groups at high risk for VAP to help medical staff initiate preventive interventions early, which is critical for reducing VAP incidence and mortality. Therefore, we aimed to explore the application of AI technology in predicting VAP and report our findings to provide a reference for the future development of VAP prevention.

Research Problem and Objective

Many studies have been conducted on the application of AI to VAP prediction. However, there is a lack of integrated evidence describing the AI techniques and model features that have been used in existing research. Therefore, this review aims to explore the characteristics of AI models for VAP prediction to assist the scientific community in advancing research within this field by identifying gaps and planning for the future.

Methods

Overview

We conducted a scoping review of studies that used AI to predict and diagnose VAP. For a transparent review, the guidelines of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [5] were followed.

Search Strategy

The following seven literature databases were searched for this study: the Wanfang database, the Chinese Biomedical Literature Database, Cochrane Library, Web of Science, PubMed, MEDLINE, and Embase. Databases were searched by using terms related to the target technology, population, and outcomes of interest. The search queries used for each database are listed in Table 1. In addition to searching the databases, backward citation screening was performed on the included studies to identify additional relevant studies. The search was conducted from January 12 to January 16, 2024.

Table . Search terms used to find studies.

Database	Hits, n	Search terms
Wanfang database	3	*(“Ventilator-associated pneumonia” OR “ventilator-associated pneumonia” OR “ventilator-associated pneumonia”) AND (“Prediction” OR “predictive models” OR “risk prediction” OR “assessment” OR “risk assessment tools”) AND (“Artificial intelligence” OR “machine learning” OR “artificial learning” OR “deep learning” OR “Bayesian learning” OR “neural networks” OR “support vector machines” OR “statistical learning” OR “decision trees” OR “random forests”)* (in Chinese)
Chinese Biomedical Literature Database	1	*(“Ventilator-associated pneumonia” OR “ventilator-associated pneumonia” OR “ventilator-associated pneumonia”) AND (“Prediction” OR “predictive models” OR “risk prediction” OR “assessment” OR “risk assessment tools”) AND (“Artificial intelligence” OR “machine learning” OR “artificial learning” OR “deep learning” OR “Bayesian learning” OR “neural networks” OR “support vector machines” OR “statistical learning” OR “decision trees” OR “random forests”)* (in Chinese)
Cochrane Library	10	*(“vap” OR “Pneumonia Ventilator-Associated” OR “Ventilator-Associated Pneumonia”) AND (“Prediction” OR “prediction model” OR “risk prediction” OR “assessment” OR” risk assessment” OR “assessment tool”) AND (“artificial intelligence” OR “machine learning” OR “Artificial Learning” OR “deep learning” OR “Bayesian Learning” OR “Neural Network” OR “Support vector machine” OR “Statistical Learning” OR “Decision tree*” OR “Random Forest”)*
Web of Science	29	*(“vap” OR “Pneumonia Ventilator-Associated” OR “Ventilator-Associated Pneumonia”) AND (“Prediction” OR “prediction model” OR “risk prediction” OR “assessment” OR” risk assessment” OR “assessment tool”) AND (“artificial intelligence” OR “machine learning” OR “Artificial Learning” OR “deep learning” OR “Bayesian Learning” OR “Neural Network” OR “Support vector machine” OR “Statistical Learning” OR “Decision tree*” OR “Random Forest”)*
PubMed	45	*(“vap” OR “Pneumonia Ventilator-Associated” OR “Ventilator-Associated Pneumonia”) AND (“Prediction” OR “prediction model” OR “risk prediction” OR “assessment” OR” risk assessment” OR “assessment tool”) AND (“artificial intelligence” OR “machine learning” OR “Artificial Learning” OR “deep learning” OR “Bayesian Learning” OR “Neural Network” OR “Support vector machine” OR “Statistical Learning” OR “Decision tree*” OR “Random Forest”)*

Database	Hits, n	Search terms
MEDLINE	21	*("vap" OR "Pneumonia Ventilator-Associated" OR "Ventilator-Associated Pneumonia") AND ("Prediction" OR "prediction model" OR "risk prediction" OR "assessment" OR" risk assessment" OR "assessment tool") AND ("artificial intelligence" OR "machine learning" OR "Artificial Learning" OR "deep learning" OR "Bayesian Learning" OR "Neural Network" OR "Support vector machine" OR "Statistical Learning" OR "Decision tree*" OR "Random Forest")*
Embase	28	*("vap" OR "Pneumonia Ventilator-Associated" OR "Ventilator-Associated Pneumonia") AND ("Prediction" OR "prediction model" OR "risk prediction" OR "assessment" OR" risk assessment" OR "assessment tool") AND ("artificial intelligence" OR "machine learning" OR "Artificial Learning" OR "deep learning" OR "Bayesian Learning" OR "Neural Network" OR "Support vector machine" OR "Statistical Learning" OR "Decision tree*" OR "Random Forest")*

Eligibility Criteria

This review included studies on AI technology for VAP diagnosis and risk prediction. However, this review excluded literature reviews and other articles that only summarized AI approaches to VAP analysis and studies that were based solely on clinical trials and experimental studies. We included only journal articles and conference papers and excluded case reports, reviews, white papers, conference abstracts, editorials, and gray literature. Studies that used non-AI techniques to predict VAP were excluded. Moreover, this review considered only studies that were written in English and Chinese and were published between the date of the establishment of the repository and January 2024. There were no constraints with regard to the study settings, study designs, study outcomes, publication months, or publication countries.

Study Selection

The screening process was performed by 2 researchers. First, we imported document titles into EndNote (Clarivate) software to eliminate duplicates. As per the inclusion criteria, irrelevant articles were further excluded by reading the titles and abstracts. Subsequently, the full texts were read to determine the final included articles. Any objections during screening were discussed with a third investigator.

Data Extraction and Synthesis

Two reviewers independently extracted the data from the included literature and discussed them with a third reviewer in cases of any objections. The extracted information included the authors; year of publication; study design; country; sample source; study population; sample size; positive outcomes; tool type; construction method; main evaluation content; model presentation form; verification method; and indicators related to reliability, validity, and predictive power.

Narrative synthesis was used to analyze the extracted data. The results included in this study were categorized as technical characteristics of the included studies (eg, AI models and algorithms used), AI model data (eg, data sources), and predictive performance indices.

Ethical Considerations

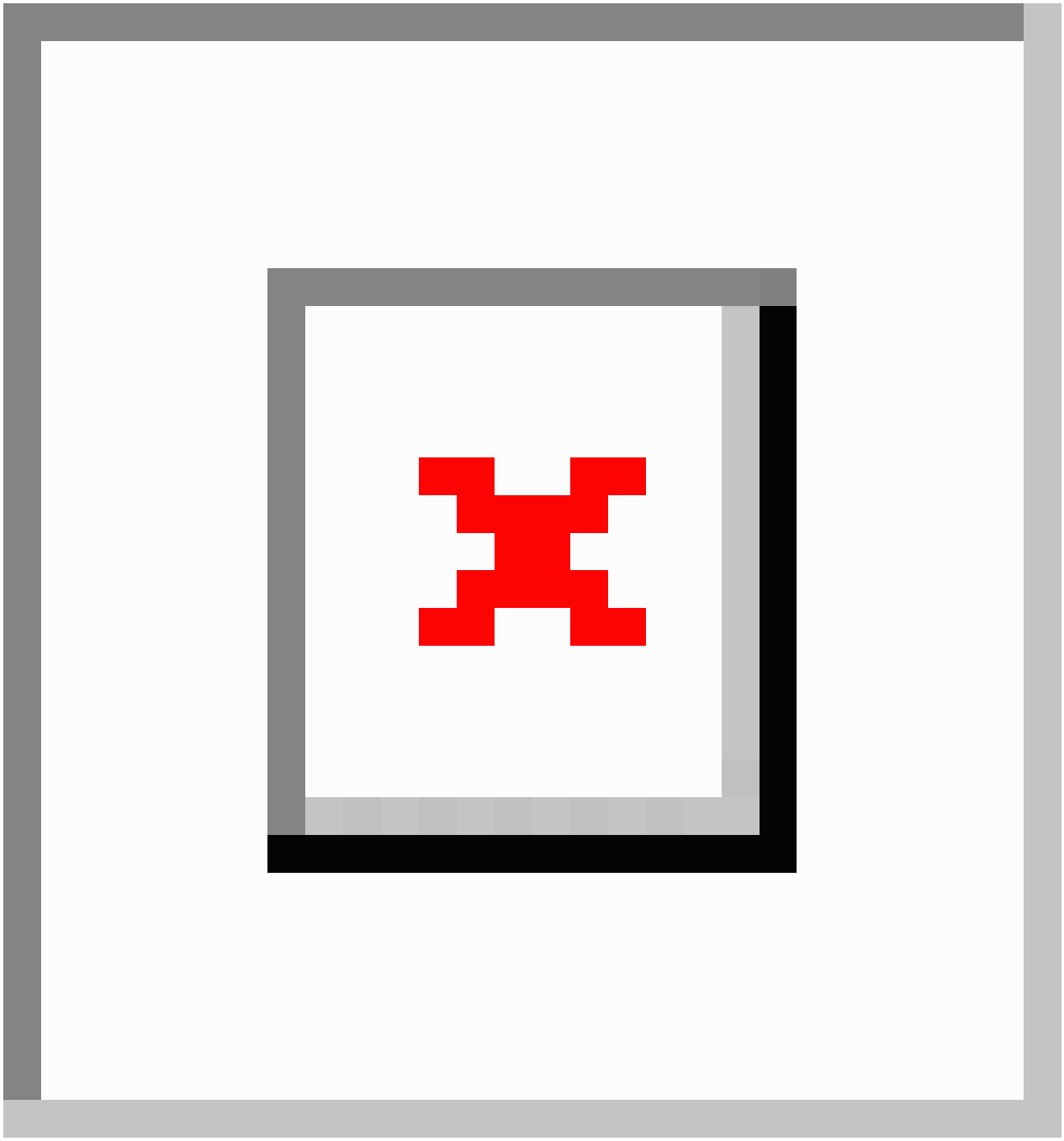
This study did not require ethical approval because we did not study any human or animal subjects and did not collect any personal information or sensitive data.

Results

Search Results

As shown in Figure 1, 137 studies were retrieved from the search, and 59 were duplicates. A total of 78 study titles and abstracts were screened, and 66 were excluded. Figure 1 presents the reasons for exclusion. Because the full text of 1 study could not be found, 11 studies were screened for eligibility; all of them met the criteria and were included in this review.

Figure 1. Flow diagram of the review process and the identification of studies via databases. VAP: ventilator-associated pneumonia.



Characteristics of Included Studies

All included studies (11/11, 100%) were published in peer-reviewed journals. The studies were published between 2007 and 2023 (Table 1), with most (3/11, 27%) published in 2023. The included studies were from 4 countries but were predominantly from the United States (5/11, 45%), followed by China (4/11, 36%). In addition, ICU patients were the most frequently studied population (studies: 6/11, 55%), 2 studies involved neurosurgical ICU patients, 1 study involved patients with traumatic brain injury, 1 study involved pediatric ICU patients, and 1 study involved older patients (age≥65 y). Public databases were the most common sources of samples (studies: 6/11, 55%), with 4 studies using the MIMIC-III (Medical Information Mart for Intensive Care III) data set. The detailed characteristics of the included studies are summarized in Table 2.

Table . Characteristics of the included studies (N=11).

Author, year	Publication type	Study design	Country	Sample source	Study population
Schurink et al [6], 2007	Journal article	Prospective cohort study	Netherlands	Recruit volunteers	Medical ICU[a] and neurosurgical ICU patients
Rambaud et al [7], 2023	Journal article	Retrospective cohort study	France	Electronic medical records	PICU[b] patients
Pearl and Bar-Or [8], 2012	Journal article	Retrospective cohort study	United States	NTDB[c] data set 6.2	ICU patients
Chen et al [9], 2020	Journal article	Prospective case-control study	China	Recruit volunteers	ICU patients
Liang et al [10], 2022	Journal article	Retrospective cohort study	China	MIMIC-III[d] data set	ICU patients
Faucher et al [11], 2022	Preprint article	Retrospective cohort study	United States	MIMIC-III data set	ICU patients
Liao et al [12], 2019	Journal article	Prospective case-control study	China	Recruit volunteers	Neurosurgical ICU patients
Abujaber et al [13], 2021	Journal article	Retrospective cohort study	United States	Electronic medical records	Patients with traumatic brain injury
Giang et al [14], 2021	Journal article	Retrospective cohort study	United States	MIMIC-III data set	ICU patients
Samadani et al [15], 2023	Journal article	Retrospective case-control study	United States	Philips eRI[e] data set	ICU patients
Mingwei et al [16], 2023	Journal article	Retrospective cohort study	China	MIMIC-III data set	Older patients (aged ≥65 y)

[a]ICU: intensive care unit.

[b]PICU: pediatric intensive care unit.

[c]NTDB: National Trauma Data Bank.

[d]MIMIC-III: Medical Information Mart for Intensive Care III.

[e]eRI: eICU Research Institute.

AI Technical Characteristics of Included Studies

All 11 included studies used only machine learning algorithms, and none of them involved deep learning algorithms or large language models. The random forest model was the most commonly used model (studies: 5/11, 45%), followed by the XGBoost (extreme gradient boost) model (studies: 4/11, 36%) and neural networks (studies: 3/11, 27%). Only 4 studies mentioned the programming languages for model building (Python: 3/11, 27%; R: 1/11, 9%). Further, 3 studies used model-building software to develop predictive models (ie, Hugin, Tiberius, and SPSS Modeler 18.2). Further details are presented in Table 3.

Table . Basic characteristics, predictors, and performance of artificial intelligence models for ventilator-associated pneumonia prediction (studies: N=11).

Author	Model	Development methodology	Sample size, n	Positive outcome, n	Predictors	Application	Verification method	Prediction performance
Schurink et al [6]	BDSS[a]	Hugin	872	157	Body temperature: <36.5 °C or >38.5 °C; ICU[b] daily sputum score: none=+0, rarely=+1, moderate=+2, severe=+3; sputum score: >14; sputum color: yellow or green; PaO_2[c]/FiO_2[d]: ≤205 mm Hg or decrease of >35 mm Hg from the previous day; use of acetaminophen, nonsteroidal anti-inflammatory drugs, or steroid antipyretics; chest x-ray showing localized or diffuse infiltration of the lungs; WBC[e] count: <4×10^9/L or >11×10^9/L; MV[f] time: >48 h	—[g]	Not reported	AUC[h]: 0.846 (95% CI 0.794-0.899); sensitivity: 0.79; specificity: 0.79; positive predictive value: 0.87; negative predictive value: 0.66

Author	Model	Development methodology	Sample size, n	Positive outcome, n	Predictors	Application	Verification method	Prediction performance
Rambaud et al [7]	IRF[i]	R	827	77	Body weight (kg); WBC count (per mm^3); neutrophil count (per mm^3); PaO_2 (mm Hg); FiO_2 (%); PEEP[j] (cmH_2O); PIP[k] (cmH_2O); MAwP[l] (cmH_2O); respiratory rate (respirations per min); tidal volume (mL); subjective volume of respiratory secretions (0, +, ++, and +++); lung dynamic compliance calculated by the oxygenation index and oxygen saturation index (in barometric mode: tidal volume/[PIP – PEEP]; in volumetric mode: tidal volume/[peak pressure – PEEP]); PIM[m] 2 score; PELOD-2[n] score	—	k-fold cross-validation	AUC: 0.82 (95% CI 0.71-0.93); sensitivity: 0.797; specificity: 0.727; positive predictive value: 0.09; negative predictive rate: 0.99; accuracy: 0.795
Pearl and Bar-Or [8]	ANN[o]	Tiberius	1,438,035	598,066	ICU length of stay; trauma score (ISS[p]); no ventilation; gender; systolic blood pressure: <40 mm Hg; age: ≤16 y; respiratory rate: <10 respirations per minute; respiratory rate: >29 respirations per minute; full model; age: >55 y	—	Not reported	Gini coefficient: 0.80435

Author	Model	Development methodology	Sample size, n	Positive outcome, n	Predictors	Application	Verification method	Prediction performance
Chen et al [9]	KNN[q], NBM[r], DT[s], NN[t], SVM[u], and RF[v]	Python	59	26	Electronic nose sensor data	—	Not reported	Best model—AUC: 0.94 (95% CI 0.74-1.00); accuracy: 0.77 (95% CI 0.46-0.95); sensitivity: 0.71; specificity: 0.83; positive predictive value: 0.93; negative predictive rate: 0.71
Liang et al [10]	RF	Python	10,431	212	Internal intensive care (control: other intensive care); emergency admission; hypertension; liver failure; PaO_2/FiO_2; APACHE[w] III score; temperature; respiratory rate; A-aDO_2[x]/PaO_2; urinary output; blood sodium; bilirubin; GCS[y]; SOFA[z]; pulmonary function; coagulation function; liver function; cardiovascular disease; central nervous system disease; aspiration admission; trauma admission	—	Not reported	AUC: mean 0.84 (SD 0.02); sensitivity: mean 0.74 (SD 0.03); specificity: mean 0.71 (SD 0.01)

Author	Model	Development methodology	Sample size, n	Positive outcome, n	Predictors	Application	Verification method	Prediction performance
Faucher et al [11]	LR[aa], fEBM[ab], and XGBoost[ac]	—	18,671	470	WBC first; WBC mean; MV hours (value); WBC median; WBC last; GCS last; WBC max; WBC min; GCS median; GCS mean; GCS max; RespRate[ad] first; Dias ABP[ae] max; blood (count) × MV hours (value); MV hours (value) × WBC last; MV hours (value) × WBC first; weight; weight × MV hours (value); SpO_2[af] first; MV hours (value) × WBC median	—	Not reported	Best model (fEBM)—AUC: 0.893
Liao et al [12]	ENN[ag] and SVM	—	12	12	Electronic nose sensor data	—	Not reported	ENN—accuracy: mean 0.9479 (SD 0.0135); sensitivity: mean 0.9714 (SD 0.0131); positive predictive value: mean 0.9288 (SD 0.0306); AUC: mean 0.9842 (SD 0.0058). SVM—accuracy: mean 0.8686 (SD 0.0422); sensitivity: mean 0.9250 (SD 0.0423); positive predictive value: mean 0.8639 (SD 0.0276); AUC: mean 0.9410 (SD 0.0301)

Author	Model	Development methodology	Sample size, n	Positive outcome, n	Predictors	Application	Verification method	Prediction performance
Abujaber et al [13]	DT	SPSS Modeler 18.2	772	169	Time to emergency department; blood transfusion; ISS[p]; pneumothorax; comorbidity	—	Not reported	Accuracy: 0.835; AUC: 0.805; precision: 0.71; negative predicted value: 0.86; sensitivity: 0.43; specificity: 0.95; *F*-score: 0.54
Giang et al [14]	LR, MLP[ah], RF, and XGBoost	—	6126	524	MV hours; biotics indicator; sputum indicator; sputum count; GCS_LAST; Platelets_MIN; Platelets_MAX; Platelets_AVERAGE; blood culture count; Temp_FIRST; GCS_AVERAGE; Platelets_FIRST; GCS_MAX; Platelets_MEDIAN; WBC_LAST	—	Not reported	Best model—AUC: 0.854
Samadani et al [15]	XGBoost	—	14,923	6811	Body temperature; FiO_2; age; MV times; total CO_2[ai]; chloride; SpO_2; heart rate; respiratory rate; gender; $PaCO_2$[aj]; creatinine; BUN[ak]; mean blood pressure; hematocrit	—	Hold-out cross-validation	AUC: 0.76; AUPRC[al]: 0.75

Author	Model	Development methodology	Sample size, n	Positive outcome, n	Predictors	Application	Verification method	Prediction performance
Mingwei et al [16]	LR, RF, XGBoost, and LightGBM[am]	Python	1523	336	SOFA; maximum WBC count; maximum respiratory rate; maximum base remaining; age; maximum creatinine; minimum $PaCO_2$; minimum oxygenation index; diabetes; ICU admission, paraplegia, gender, COPD[an]	—	10-fold cross-validation	Best models: LightGBM—AUC: 0.85 (95% CI 0.82-0.88); accuracy: 0.77; precision: 0.80; recall: 0.72; specificity: 0.82; F_1: 0.75. XGBoost—AUC: 0.84 (95% CI 0.81-0.87); accuracy: 0.76; precision: 0.78; recall: 0.73; specificity: 0.79; F_1:0.75

[a]BDSS: Bayesian decision support system.
[b]ICU: intensive care unit.
[c]PaO_2: partial pressure of oxygen.
[d]FiO_2: fraction of inspired oxygen.
[e]WBC: white blood cell.
[f]MV: mechanical ventilation.
[g]Not applicable.
[h]AUC: area under the curve.
[i]IRF: imbalanced random forest model.
[j]PEEP: positive end-expiratory pressure.
[k]PIP: peak inspiratory pressure.
[l]MAwP: mean airway pressure.
[m]PIM: pediatric index of mortality.
[n]PELOD-2: Pediatric Logistic Organ Dysfunction-2.
[o]ANN: artificial neural network.
[p]ISS: Injury Severity Score.
[q]KNN: k-nearest neighbor.
[r]NBM: naive Bayes model.
[s]DT: decision tree.
[t]NN: neural network.
[u]SVM: support vector machine.
[v]RF: random forest.
[w]APACHE: Acute Physiology and Chronic Health Evaluation.
[x]A-aDO2: alveolar-arterial oxygen difference.
[y]GCS: Glasgow Coma Scale.
[z]SOFA: Sequential Organ Failure Assessment.
[aa]LR: logistic regression.
[ab]fEBM: full feature explainable boosting machine.
[ac]XGBoost: extreme gradient boost.
[ad]RespRate: respiratory rate of the ventilator.
[ae]Dias ABP: diastolic blood pressure.
[af]SpO_2: peripheral blood oxygen saturation.
[ag]ENN: ensemble neural network.
[ah]MLP: multilayer perceptron.
[ai]CO_2: carbon dioxide.

[aj]$PaCO_2$: carbon dioxide partial pressure.

[ak]BUN: blood urea nitrogen.

[al]AUPRC: area under the precision-recall curve.

[am]LightGBM: light gradient boosting machine.

[an]COPD: chronic obstructive pulmonary disease.

Different types of data were used in the included studies, including laboratory data (eg, white blood cell count, neutrophil count, and bilirubin level), clinical data (including temperature, sputum volume, and ventilator parameters), and demographic data (eg, age, weight, and sex). Of note, 2 studies used sensor data to build predictive models, and the remaining 9 studies used clinical data. In addition, 67% (6/9) of these studies used laboratory data, with white blood cell count being the most commonly used laboratory data (studies: 4/9, 44%), followed by neutrophil count (studies: 1/9, 11%), bilirubin level (studies: 1/9, 11%), and blood urea nitrogen level (studies: 1/9, 11%). Demographic data were used in 56% (5/9) of the studies; age was used as a predictor in 4 studies, and weight and age were both included in only 1 study.

In terms of data set size, of the 11 studies, 6 (55%) had sample sizes of >1000; however, with regard to the data from the electronic nose sensors that were used in 2 studies, multiple sensors were placed on the electronic nose, and each sensor collected data more than once. Therefore, the actual sample sizes for these two studies were 1888 [9] and 3360 [12]. Nevertheless, because the data were collected by the same electronic nose sensor and came from the same patient, we did not include these two studies in the number of studies with sample sizes of >1000. Further, 3 studies used data sets with <1000 samples, and 4 studies had data sets with >10,000 samples. The AI performance index was mentioned in all 11 studies. The area under the curve (AUC) was the most commonly used predictive performance index (studies: 10/11, 90%), followed by sensitivity (studies: 6/11, 55%) and specificity (studies: 6/11, 55%). The AUC values, which were reported in 10 studies, averaged to 0.86 (SD 0.07) and ranged from 0.76 to 0.98. The sensitivity, which was reported in 6 studies, averaged to 0.74 (SD 0.18) and ranged from 0.43 to 0.97. The specificity, which was reported in 6 studies, averaged to 0.80 (SD 0.09) and ranged from 0.71 to 0.95. Additionally, 5 studies reported accuracy (mean 0.82, SD 0.07, range 0.77-0.95).

Discussion

Principal Findings

In this review, we explored AI techniques for the prediction of VAP. Of the 11 included studies, 9 (82%) were published in the past 5 years, and the number of studies has increased annually with the evolution of AI technology (1 in 2019, 1 in 2020, 2 in 2021, 2 in 2022, and 3 in 2023). Most (9/11, 82%) of the AI-based prediction model studies were published in the United States (5/11, 45%) and China (4/11, 36%). To explore the application of AI in predicting VAP, the results were divided into 3 categories, and each of them classified the included studies from a different perspective.

The first category included the technical characteristics of the studies. All studies used only machine learning algorithms, with the random forest model being the most commonly used model (studies: 5/11, 45%), followed by neural networks (studies: 4/11, 36%) and the XGBoost model (studies: 4/11, 36%). The second category focused on AI model data, in which we explored the data types, data sources, and data set sizes. Different types of data, including laboratory, clinical, and demographic data, were used in the included studies. In terms of data set size, apart from 2 studies that used electronic noses, 6 (55%) had sample sizes of >1000. Public databases were the most common sources of data (studies: 6/11, 55%). The third category focused on the predictive performance of AI models, including studies that used different performance validation indices, such as the AUC, accuracy, sensitivity, and specificity.

Implications for Practice and Research

This review highlights the most common AI models that have been used to predict VAP. Based on our findings, AI models can predict VAP by using various data types. In our review, no studies that used deep learning and large language models were found. A possible reason for this is that chest computed tomography data are not available in most public databases, and in clinical practice, patients who do not exhibit pneumonia symptoms do not undergo chest computed tomography examinations; therefore, such data are not available for research. The random forest and XGBoost models are the most frequently used machine learning–based VAP prediction models, probably because ensemble learning models exhibit better prediction performance and robustness when dealing with multiple types of data compared to other models [17].

Based on the data sources of the prediction models, the use of more data types for comprehensive predictions may be the main focus of future research. Current research may be constrained to using structured data, owing to the limitations of algorithms and data collection workloads, while electronic health records contain unstructured clinical text, such as admission records and progress notes. Furthermore, much data remain to be mined. Tsai et al [18] found that information extracted from unstructured clinical text could make predictive models more comprehensive and improve their predictive performance. In addition to unstructured clinical text, lung radiography and computed tomography can be used to predict the occurrence of pneumonia.

In terms of predictive tools, natural language processing and deep learning may be the direction of future research, and the development of large language models, such as ChatGPT, that are based on natural language processing is sufficient to prove the ability of natural language processing algorithms to process unstructured clinical text [19]. Traditional machine learning algorithms are not competent in the image recognition domain, while deep learning algorithms can analyze and process clinical

imaging data effectively. Lee et al [20] found that deep learning–based predictive models that used preoperative imaging data from patients could effectively predict the occurrence of postoperative pneumonia; however, no studies have used deep learning algorithms to construct VAP prediction models.

Of further note, the studies reviewed herein rarely mentioned nurse-related data, and it has been suggested that nursing is important for VAP prevention [21,22]. The potential of various data types in predicting VAP should be explored in future studies. Additionally, none of the studies included in this review considered the application of the final model. The deployment of feasible predictive models in clinical settings needs to be explored.

Strengths

This review discusses all of the AI techniques and study populations that have been used to date to predict VAP, with no major restrictions on paper status, research environment, and geographic location. In addition, the characteristics of each AI model and the data sets that were used to build the models were discussed in depth.

Based on our findings, Frondelius et al [4,23] explored diagnostic and prognostic models for VAP and performed a meta-analysis of the performance of machine learning–based predictive models for VAP. However, to the best of our knowledge, ours is the first review of all AI VAP prediction models that have been explored thus far, filling research gaps to improve understanding of prediction techniques rather than focusing solely on the final predictive performance of models. Moreover, in the literature search, we did not place any limitations on types of technology and included all branches of AI to gain insight into the research on different AI technologies for VAP prediction.

Finally, study selection and data extraction were performed independently by 2 evaluators to ensure minimal bias.

Limitations

This review has certain limitations. Reviews, conference abstracts, case reports, white papers, proposals, editorials, and gray literature were excluded to reduce the complexity of the results. We also included Chinese databases in our search but did not explore articles in languages other than English or Chinese, which might have reduced the comprehensiveness of our study.

Conclusions

This paper reviews the application of AI technology in VAP prediction and provides new evidence on the role of AI technology. We believe that the findings will help researchers better understand the application of AI technology in VAP prediction and provide a reference for future research on VAP prediction models. Lastly, we believe that advances in AI technology will provide further possibilities for predicting VAP and that interdisciplinary developments will improve the health care industry.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

J Zhang implemented the research and drafted the manuscript. PY and LZ collected the data. SL made important revisions to the manuscript. J Zhou approved the final paper.

Conflicts of Interest

None declared.

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File, 86 KB - medinform_v12i1e57026_app1.docx]

References

1. Papazian L, Klompas M, Luyt CE. Ventilator-associated pneumonia in adults: a narrative review. Intensive Care Med 2020 May;46(5):888-906. [doi: 10.1007/s00134-020-05980-0] [Medline: 32157357]
2. Modi AR, Kovacs CS. Hospital-acquired and ventilator-associated pneumonia: diagnosis, management, and prevention. Cleve Clin J Med 2020 Oct 1;87(10):633-639. [doi: 10.3949/ccjm.87a.19117] [Medline: 33004324]
3. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. Br Med Bull 2021 Sep 10;139(1):4-15. [doi: 10.1093/bmb/ldab016] [Medline: 34405854]
4. Frondelius T, Atkova I, Miettunen J, Rello J, Jansson MM. Diagnostic and prognostic prediction models in ventilator-associated pneumonia: systematic review and meta-analysis of prediction modelling studies. J Crit Care 2022 Feb;67:44-56. [doi: 10.1016/j.jcrc.2021.10.001] [Medline: 34673331]

5. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
6. Schurink CAM, Visscher S, Lucas PJF, et al. A Bayesian decision-support system for diagnosing ventilator-associated pneumonia. Intensive Care Med 2007 Aug;33(8):1379-1386. [doi: 10.1007/s00134-007-0728-6] [Medline: 17572880]
7. Rambaud J, Sajedi M, Al Omar S, et al. Clinical decision support system to detect the occurrence of ventilator-associated pneumonia in pediatric intensive care. Diagnostics (Basel) 2023 Sep 18;13(18):2983. [doi: 10.3390/diagnostics13182983] [Medline: 37761350]
8. Pearl A, Bar-Or D. Decision support in trauma management: predicting potential cases of ventilator associated pneumonia. Stud Health Technol Inform 2012;180:305-309. [Medline: 22874201]
9. Chen CY, Lin WC, Yang HY. Diagnosis of ventilator-associated pneumonia using electronic nose sensor array signals: solutions to improve the application of machine learning in respiratory research. Respir Res 2020 Feb 7;21(1):45. [doi: 10.1186/s12931-020-1285-6] [Medline: 32033607]
10. Liang Y, Zhu C, Tian C, et al. Early prediction of ventilator-associated pneumonia in critical care patients: a machine learning model. BMC Pulm Med 2022 Jun 25;22(1):250. [doi: 10.1186/s12890-022-02031-w] [Medline: 35752818]
11. Faucher M, Chetty SS, Shokouhi S, et al. Early prediction of ventilator-associated pneumonia in ICU patients using an interpretable machine learning algorithm. Preprints.org. Preprint posted online on Jun 10, 2022. [doi: 10.20944/preprints202206.0149.v1]
12. Liao YH, Wang ZC, Zhang FG, Abbod MF, Shih CH, Shieh JS. Machine learning methods applied to predict ventilator-associated pneumonia with pseudomonas aeruginosa infection via sensor array of electronic nose in intensive care unit. Sensors (Basel) 2019 Apr 18;19(8):1866. [doi: 10.3390/s19081866] [Medline: 31003541]
13. Abujaber A, Fadlalla A, Gammoh D, Al-Thani H, El-Menyar A. Machine learning model to predict ventilator associated pneumonia in patients with traumatic brain injury: the C.5 decision tree approach. Brain Inj 2021 Jul 29;35(9):1095-1102. [doi: 10.1080/02699052.2021.1959060] [Medline: 34357830]
14. Giang C, Calvert J, Rahmani K, et al. Predicting ventilator-associated pneumonia with machine learning. Medicine (Baltimore) 2021 Jun 11;100(23):e26246. [doi: 10.1097/MD.0000000000026246] [Medline: 34115013]
15. Samadani A, Wang T, van Zon K, Celi LA. VAP risk index: early prediction and hospital phenotyping of ventilator-associated pneumonia using machine learning. Artif Intell Med 2023 Dec;146:102715. [doi: 10.1016/j.artmed.2023.102715] [Medline: 38042602]
16. Mingwei S, Jun L, Chunping S, Xinmin L. Construction of early warning model for ventilator-associated pneumonia in the elderly based on machine learning algorithm [Article in Chinese]. Chinese Journal of Geriatrics 2023;42(6):670-675. [doi: 10.3760/cma.j.issn.0254-9026.2023.06.009]
17. Zheng H, Sherazi SWA, Lee JY. A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data. IEEE Access 2021;9:113692-113704. [doi: 10.1109/ACCESS.2021.3099795]
18. Tsai HC, Hsieh CY, Sung SF. Application of machine learning and natural language processing for predicting stroke-associated pneumonia. Front Public Health 2022 Sep 29;10:1009164. [doi: 10.3389/fpubh.2022.1009164] [Medline: 36249261]
19. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785. [doi: 10.2196/48785] [Medline: 37862079]
20. Lee T, Hwang EJ, Park CM, Goo JM. Deep learning-based computer-aided detection system for preoperative chest radiographs to predict postoperative pneumonia. Acad Radiol 2023 Dec;30(12):2844-2855. [doi: 10.1016/j.acra.2023.02.016] [Medline: 36931951]
21. Collins T, Plowright C, Gibson V, et al. British Association of Critical Care Nurses: evidence-based consensus paper for oral care within adult critical care units. Nurs Crit Care 2021 Jul;26(4):224-233. [doi: 10.1111/nicc.12570] [Medline: 33124119]
22. Wang Y, Lan Y, Jia T, Ma M, Liu C, Tang H. Construction and application of a training program for ICU nurses to manage artificial airway gasbags to prevent ventilator-associated pneumonia. J Multidiscip Healthc 2023 Dec 2;16:3737-3748. [doi: 10.2147/JMDH.S438316] [Medline: 38076591]
23. Frondelius T, Atkova I, Miettunen J, et al. Early prediction of ventilator-associated pneumonia with machine learning models: a systematic review and meta-analysis of prediction model performance. Eur J Intern Med 2024 Mar;121:76-87. [doi: 10.1016/j.ejim.2023.11.009] [Medline: 37981529]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
ICU: intensive care unit
MIMIC-III: Medical Information Mart for Intensive Care III

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
VAP: ventilator-associated pneumonia
XGBoost: extreme gradient boost

Edited by C Lovis; submitted 02.02.24; peer-reviewed by A Hassan, R Bidkar; revised version received 08.04.24; accepted 11.04.24; published 14.05.24.

Please cite as:
Zhang J, Yang P, Zeng L, Li S, Zhou J
Ventilator-Associated Pneumonia Prediction Models Based on AI: Scoping Review
JMIR Med Inform 2024;12:e57026
URL: https://medinform.jmir.org/2024/1/e57026
doi:10.2196/57026

Semantic Interoperability of Electronic Health Records: Systematic Review of Alternative Approaches for Enhancing Patient Information Availability

Sari Palojoki[1,*], PhD; Lasse Lehtonen[2,*], MD, PhD; Riikka Vuokko[1,*], PhD

[1]

[2]

[*]all authors contributed equally

Corresponding Author:
Sari Palojoki, PhD

Abstract

Background: Semantic interoperability facilitates the exchange of and access to health data that are being documented in electronic health records (EHRs) with various semantic features. The main goals of semantic interoperability development entail patient data availability and use in diverse EHRs without a loss of meaning. Internationally, current initiatives aim to enhance semantic development of EHR data and, consequently, the availability of patient data. Interoperability between health information systems is among the core goals of the European Health Data Space regulation proposal and the World Health Organization's *Global Strategy on Digital Health 2020-2025*.

Objective: To achieve integrated health data ecosystems, stakeholders need to overcome challenges of implementing semantic interoperability elements. To research the available scientific evidence on semantic interoperability development, we defined the following research questions: What are the key elements of and approaches for building semantic interoperability integrated in EHRs? What kinds of goals are driving the development? and What kinds of clinical benefits are perceived following this development?

Methods: Our research questions focused on key aspects and approaches for semantic interoperability and on possible clinical and semantic benefits of these choices in the context of EHRs. Therefore, we performed a systematic literature review in PubMed by defining our study framework based on previous research.

Results: Our analysis consisted of 14 studies where data models, ontologies, terminologies, classifications, and standards were applied for building interoperability. All articles reported clinical benefits of the selected approach to enhancing semantic interoperability. We identified 3 main categories: increasing the availability of data for clinicians (n=6, 43%), increasing the quality of care (n=4, 29%), and enhancing clinical data use and reuse for varied purposes (n=4, 29%). Regarding semantic development goals, data harmonization and developing semantic interoperability between different EHRs was the largest category (n=8, 57%). Enhancing health data quality through standardization (n=5, 36%) and developing EHR-integrated tools based on interoperable data (n=1, 7%) were the other identified categories. The results were closely coupled with the need to build usable and computable data out of heterogeneous medical information that is accessible through various EHRs and databases (eg, registers).

Conclusions: When heading toward semantic harmonization of clinical data, more experiences and analyses are needed to assess how applicable the chosen solutions are for semantic interoperability of health care data. Instead of promoting a single approach, semantic interoperability should be assessed through several levels of semantic requirements A dual model or multimodel approach is possibly usable to address different semantic interoperability issues during development. The objectives of semantic interoperability are to be achieved in diffuse and disconnected clinical care environments. Therefore, approaches for enhancing clinical data availability should be well prepared, thought out, and justified to meet economically sustainable and long-term outcomes.

(JMIR Med Inform 2024;12:e53535) doi:10.2196/53535

KEYWORDS

electronic health record; health records; EHR; EHRs; semantic; health care data; semantic interoperability; interoperability; standardize; standardized; standardization; cross-border data exchange; systematic review; synthesis; syntheses; review methods; review methodology; search; searches; searching; systematic; data exchange; information sharing; ontology; ontologies; terminology; terminologies; standard; standards; classification; PRISMA; data sharing; Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Introduction

Over the past 2 decades, there has been growing interest in digital technologies and eHealth integration into national health care systems to promote health [1]. The World Health Organization (WHO) has launched the *Global Strategy on Digital Health 2020-2025* [2]. To implement digital health strategy objectives, a toolkit was set up to help countries to integrate eHealth into their health care systems [3]. The objectives of the WHO strategy include standards for interoperability. Another current large-scale international initiative is the European Health Data Space (EHDS) regulation proposal. EHDS is a health-specific ecosystem comprised of rules, common standards and practices, infrastructures, and a governance framework. It supports the use of health data for better health care delivery, research, innovation, and policy making. Moreover, it aims at empowering patients through increased digital access to and control of their personal health data [3-6].

Interoperability ensures health data availability and use. It is the ability of different organizations and professionals to interact and share information according to standards of data transfer and common protocols that support data exchange [4-8]. In clinical context, interoperable electronic health records (EHRs) help health care practitioners gather, store, and communicate essential health information reliably and securely across care settings. This aims to guarantee coordinated and patient-centered care while creating many efficiencies in the delivery of health care [9]. EHRs use health-related information pertinent to an individual patient, whereas registries are mainly focused on population management and are designed to obtain information on predefined health outcomes data and data for public health surveillance, for example. Although technological possibilities for using various types of data grow, new demands are placed on data quality and usability and, consequently, on interoperability [5,10,11].

Moreover, semantic interoperability enhances the unambiguous representation of clinical concepts, supported by the use of international standard reference systems and ontologies. Since there are different types of health information, such as data from EHRs, patient registries, genomics data, and data from health applications, the development of international data standardization, common guidelines, and recommendations are needed [4-8]. Without applying appropriate semantic standards, such as domain-relevant terminologies, interoperability will be limited. This may diminish the availability and potential value of data. The various parties involved have to address the importance of shared digital health standards and especially semantic interoperability features [12-15]. In the clinical context, interoperability is required to enhance the quality, efficiency, and effectiveness of the health care system by providing information in the appropriate format whenever and wherever it is needed by eliminating unnecessary replication [16].

Therefore, our study aims to provide readers with up-to-date information about the different types of approaches to resolve semantic interoperability in EHRs specifically and to summarize the benefits of these choices. We aimed to research the topic with an emphasis on patient data availability and use. Our research questions were as follows: What are the key elements of and approaches for building semantic interoperability integrated in EHRs? What kinds of goals are driving the development? and What kinds of clinical benefits are perceived following this development?

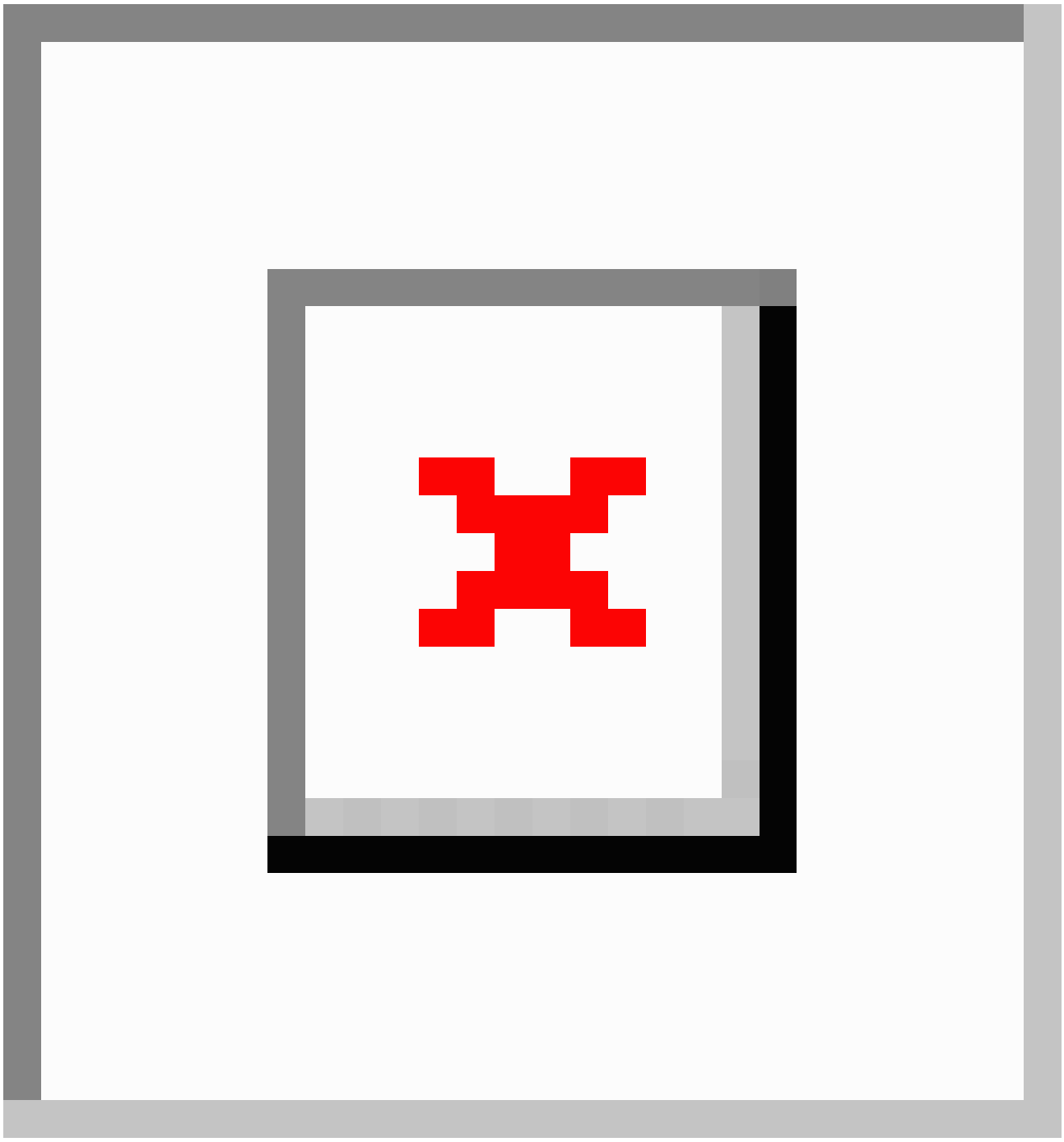
Methods

Methodological Framework

With our research questions as a starting point, we set out to perform a systematic literature review of semantic interoperability. Regarding different layers of interoperability, legal interoperability ensures overcoming potential barriers for data exchange. Interoperability agreements are made binding via international- or national-level legislation and via bilateral and multilateral agreements. Organizational interoperability defines, for example, business goals and processes. Semantic interoperability ensures that the precise meaning of exchanged information is understandable by any other application. It enables systems to combine received information with other information resources and process it in a meaningful manner. Technical interoperability covers various issues of linking computer systems and services, such as open interfaces, data integration, data presentation and exchange, accessibility, and security services [6,7].

For the study design, we first defined our core concepts to refine the literature search strategy. The scope of the review was semantic interoperability, that is, organizational, legal, and technical interoperability were excluded [7]. Semantic interoperability was apprehended based on the European Interoperability Framework (EIF) that provides a common set of principles and guidance for the design and development of interoperable digital services. In the EIF, semantic interoperability covers both semantic and syntactic aspects. The semantic aspect refers to the meaning of data elements and their relationships, whereas the syntactic aspect refers to the format of the information to be exchanged. With semantic interoperability, it is ensured that data can be shared in such a way that the meaning of data does not change [7,15,17,18]. There are also other models for analyzing interoperability layers [18]. For example, in comparison to the European approach [7], the Healthcare Information and Management Systems Society defines 4 levels of interoperability for health care technology: foundational, structural, semantic, and organizational [19,20]. Since the EIF is a well-established and largely applied framework [6], we chose the EIF definitions to primarily guide our review framework, as illustrated in Figure 1. Our review deals with semantic interoperability, which is highlighted in gray in the figure. Thus, we did not analyze, for example, standards that are related to processes or information quality.

Figure 1. Our framework for defining semantic interoperability elements for conducting the literature search and guiding our study design. ATC: Anatomical Therapeutic Chemical; CDA: Clinical Document Architecture; EHR: electronic health record; EMR: electronic medical record; FHIR: Fast Health Interoperability Resources; HL7: Health Level 7; *ICD-10: International Classification of Diseases, Tenth Revision*; *ICD-11: International Classification of Diseases, 11th Revision*; LOINC: Logical Observation Identifiers Names and Codes; RIM: reference information model; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terminology.



As shown in Figure 1, processing, storing, and exchanging health care data in EHRs and between EHRs or other clinical applications is, for example, governed and regulated at the legal layer. To continue, processes and workflows regarding information exchange are arranged at the organizational interoperability layer and resolved in the technical layer, for example, according to the principles of data protection and information security. To illustrate the point, for example, the EHDS proposal suggests that compliance with essential requirements on interoperability and data security may be demonstrated by the manufacturers of EHR systems through the implementation of common specifications. To that end, implementation can be grounded on common specifications, such as data sets, coding systems, technical specifications, standards, and profiles for data exchange, as well as requirements and principles related to security, confidentiality, integrity, patient safety, and he protection of personal data and so on [6].

The semantic interoperability layer in Figure 1 covers various approaches to resolve interoperability issues, such as more established international or domain-specific health care

classifications, clinical terminologies, and ontologies and applications of international standards for EHRs. In Figure 1, we provided some examples to illustrate various semantic aspects, but this is not an exhaustive list. Similarly, for other interoperability levels, real-world examples were given. Based on the EIF, semantic interoperability also covers syntactic features, such as data format and, for example, structured data content. We identified these key features of semantic interoperability based on previous research [8,16-19,21]. In our framework, a data model is a generic concept that describes various applications of data models from a reference information model (RIM) to a clinical information model. Data models define structures and semantics for storing, exchanging, querying, and processing health care data. Clinical information models can be implemented in an EHR, for example, as archetypes and templates, whereas RIMs refer to standards-based approaches to enable health care documentation and messages, such as the Health Level 7 (HL7) RIM or the International Organization for Standards' EN/ISO 13606 standard for EHR communication [19,22]. When designing EHRs, for semantic interoperability, a dual-level method can be applied to represent both information and knowledge levels of interoperability requirements, properties, and structures for data. This approach is used, for example, for representing the dual levels of knowledge by an archetype model and information structures by the chosen RIM [16,21,22].

Study Design

In the design of the review, we applied the Cochrane review protocol [23] to ensure the scientific reliability and validity of our review (Checklist 1). The search strategy (see Textbox 1) was defined based on the framework for semantic interoperability presented in Figure 1. We performed the search in the PubMed database in December 2022. To conduct a systematic literature review, PubMed is regarded as a comprehensive database [24]. Therefore, no further data searches were performed. We documented the search so that it can be reproduced (see Textbox 1). The search resulted in 131 unique articles. One article was removed because it did not include an abstract, and 1 was removed because it was not in English. In total, the authors screened 129 articles.

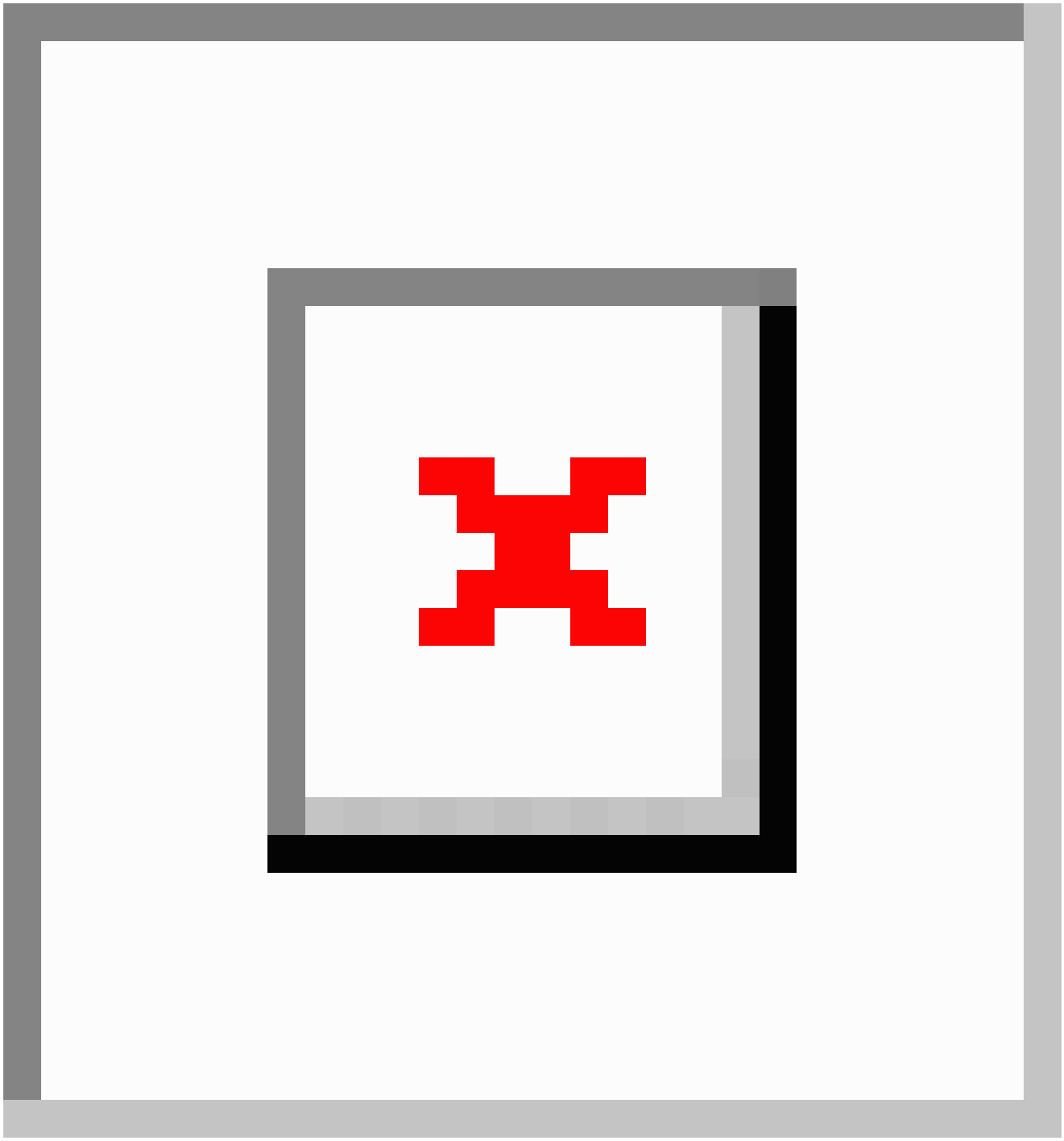
Textbox 1. Search strategy and filters used.

- Search terms: (((((EHR) OR (EMR)) OR ("Electronic Health Record")) OR ("Electronic Medical Record")) AND (((((“Semantic interoperability”) OR ((“data model”) AND (“Semantic interoperability”))) OR ((((“classification”) OR (ontology)) OR (terminology)) AND (“Semantic interoperability”))) OR (((“data content”) OR (“data format”)) AND (“Semantic interoperability”))) OR ((“Semantic interoperability”) AND (standard)))
- Filters used: abstract, full text, and English

The research team first screened all the remaining articles by title and abstract from January to March 2023. After the first test reading, the researchers discussed the inclusion and exclusion criteria and coherence of the understanding. Researchers were blinded and performed the analysis independently based on the inclusion and exclusion criteria and then compared the results. Selecting the same alternative created a match. Choosing a different alternative or failing to recognize the category at all was considered a nonmatch. In data-model cases, discussion was needed for alignment, but no complex situations developed. During the first screening, after discussion by the research team, 71 articles were excluded from the review for the following four reasons: (1) EHR was not a key factor but a contextual factor in the original research setting; (2) the original research did not focus on semantic interoperability but on another level of interoperability; (3) the original study did not entail practical implementation goals, but the focus was predominantly theoretical or methodological; and (4) the original research was not a research article but, for example, a poster. The remaining 58 articles were sought for retrieval. For 4 articles, the full text was not available. To evaluate eligibility, full texts of the 54 remaining articles were read by the research team. At this point, 17 articles were excluded because the original research was out of scope, that is, semantic interoperability was not developed with practical goals for advancing the availability and use of interoperable patient data. In addition, 15 articles were excluded as the semantic interoperability case did not involve EHR use or development, 3 articles were excluded due to the absence of semantic interoperability altogether, and 5 more were excluded because they were not research articles. After agreeing upon the final exclusion within our research team, 14 articles were analyzed for semantic interoperability in EHRs. Our final inclusion criteria were grounded on our research questions: the research article should explore an EHR use or development case with the focus on semantic interoperability of clinical data. Preferably, the case would document the stage of interoperability development or use, expected or realized clinical benefits, semantic development goals, and aspects of interoperability to be implemented, as well as the method of application.

The extraction and documentation of the information from the research articles was informed by our research questions, the review framework (Figure 1), and by previous research literature. At this stage, previous reviews [16-19,21] were especially used in compiling our study framework (see Figure 1). Based on our framework, the documentation of the review analysis included elements of interoperability already identified in the search strategy. Consequently, it was necessary to investigate which documented elements are typically examined in research and with what methods they are applied in EHRs [8,16-18]. Moreover, we deemed it important to document how semantic interoperability is described in the clinical use context, consisting of various EHRs, clinical applications, registers, and other data resources. Lastly, the information documentation had to include not only the semantic implementation, use goals, or intended benefits but also practical goals or benefits in the clinical use context (see Figure 2). We defined and agreed upon the information documentation categories within our research team to conduct a well-grounded analysis for the review.

Figure 2. Flowchart of article identification, screening, and final inclusion. EHR: electronic health record.



Results

Contextual Results

We identified 14 articles describing semantic interoperability in EHRs, published between 2011 and 2022, as shown in Multimedia Appendix 1 [24-37]. The results revealed predominantly European advances in the study topic. Most (n=11, 79%) of the research cases were affiliated with different types of institutions in the European Union member states or in European Economic Area countries. One of the publications was coproduced by authors from Columbia and Germany, and the authors of another article represented organizations from the United States, South Korea, China, and Egypt. We decided not to limit the included studies to a certain geographical area but to analyze any potential use case for enabling the interoperability of EHRs.

Two of the reported research cases focused on patients with heart failure [24,30], 1 focused on patients with neurosurgical tumors [28], 2 focused on patients in cancer care [33,37], and 1 focused on patients with type 1 diabetes [31]. Other clinical use domains described were a prehospital unit at the site of an incident or during transfer to the emergency department and a hospital emergency care unit where prehospital patient documentation must be reassessed. A primary care–related case documented experimental laboratory test results of a population of 230,000 patients. Examples of older adult medication care and multiprofessional health care were part of our sample. Two articles described multipurpose clinical use of physician's notes

and tertiary care data. One article concerned the domain of clinical research using data from different EHR systems, and another described semantic aspects for retrieval of medication, laboratory test, and diagnosis-related data.

Although all studies concerned data from the EHRs, some studies included more detailed descriptions on data sources. Heart failure summaries containing clinical situation data and diagnoses (severity and certainty), as well as heart failure summaries covering clinical situations and symptoms data (a symptom's presence, absence, and severity), were represented in the sample. One study regarded clinical history, observations, and findings during tumor control. One study focused on histories of patients with diabetes and diabetes care plans (eg, insulin regimen, diet, and exercise plans) and patients' self-monitoring of vital signs, and 1 study used self-monitoring data on daily activities, side effects, and patient-reported outcomes. One article reported results around diagnosis and laboratory data; 1 article reported on medication, laboratory, and diagnosis data; and another article reported on neurosurgical imaging and laboratory data, although the starting point in the paper was diagnosis and medication data. The remaining 4 studies generally applied prehospital patient case data, emergency care–related EHR data, laboratory data, and diagnosis data.

Interoperability Results

In our sample, data were transferred and shared between different EHRs and clinical applications with no loss of data or changes in their meaning (Multimedia Appendix 2 [24-37]). Half (7/14, 50%) of the studies were aimed at developing semantic interoperability between different EHRs or within different EHR modules, such as a medication module in 1 EHR system. One case concentrated specifically on an EHR and a clinical application. Two articles reported results about the interoperability between EHRs and personal health records. Interoperability with the laboratory system and the EHR was the focus of study in 2 cases. Two studies reported advances in interoperability development between EHR and clinical research resources or a clinical registry. Regarding the state of development, the largest number of studies were categorized as "in development" (n=5, 36%) and "in use" (n=6, 43%). Two articles reported results regarding the testing phase, and the remaining study was in an implementation stage.

All articles reported clinical benefits of the selected approach to enhancing semantic interoperability. We identified 3 main categories of clinical benefits within the articles: increasing the availability of data for clinicians (n=6, 43%), increasing the quality of care (n=4, 29%), and enhancing clinical data use and reuse for varied purposes (n=4, 29%). The first category describes use cases where patient care would benefit from better availability of data. This was to be achieved by enhancing interoperable data and its transfer from clinical applications (eg, a laboratory system) to a central EHR and between EHRs to increase accessible data for making informed clinical decisions. These advances were in implementation to enhance the quality and effectiveness of care. Moreover, developing better access to health data and providing homogeneous access to heterogeneous data sets may facilitate resource effectiveness; patient management; and overall, the optimization of data for different purposes. The second category included benefits for the quality of care. The category had largely been implemented in EHRs already. Benefits entail better resource effectiveness and optimization of patient care planning and monitoring and better patient management, as well as the continuity of care based on interoperable and accessible health data that facilitates informed decision-making by clinicians. One of these cases documented improved patient safety based on interoperable health data across EHRs. The third category, enhancing clinical data use and reuse, included 2 use cases where data were used across EHRs. One use case described data transfer between an EHR and a national oncology registry, where interoperability enhanced data integration and redesign of the systems in use. The other 2 cases documented the evidence of data use, where better availability of data provided a means for developing new EHR integrated tools, such as clinical alerts, dynamic patient lists, and clinical follow-up dashboards. In summary, semantic development goals emphasized better access to data regardless of underlying standards and data structures or EHRs in use. The underlying assumption is that with better access to data, it is possible to facilitate better communication between professionals and the continuity of care.

In our analysis, semantic development goals were divided in 3 categories. All of these were closely coupled with the need to build usable and available data based on heterogeneous medical information that is accessible through various EHRs and databases, such as registers. Data harmonization and developing semantic interoperability between different EHRs or between EHRs and clinical application was the largest category (n=8, 57%). Enhancing health data quality through standardization (n=5, 36%) and developing EHR-integrated tools based on interoperable data (n=1, 7%) were the other identified categories. Semantic development goals were described as harmonizing data or otherwise processing semantically equivalent data across different medical domains and among different clinical data sources including EHRs and applications, thus facilitating clinicians' availability of health data. One case included the formalization of data with a semantic converter to increase the interoperability of data. In 2 research cases, the main semantic development goals concentrated on advancing the interoperability of EHR data and patient-generated data or sensor data to monitor the situation of patients who are chronically ill. Regarding data standardization, 1 research case reported increasing data quality as the semantic interoperability development goal. Standardized data content decreased information overload of clinicians. Through data standardization, it was possible to increase conceptualization and, thus, access to data within an EHR regardless of the underlying standards and data structures, by providing a semantic standardized layer to facilitate clinicians' data use, or by otherwise ensuring complete and coherent information with no errors due to the loss of meaning or context. One of these research cases documented improvements for system-level efficiency for EHR functions and integrated tools based on advances of semantic interoperability.

Features of semantic interoperability were described in all 14 articles. Most (9/14, 64%) of the analyzed cases incorporated

1 or more semantic aspects. In more detail, the aspects of semantic interoperability were described as follows: ontologies were the chosen aspect in 3 research cases, terminologies in 6 cases, classifications in 4 cases, various clinical documentation standards in 8 cases, and different data models in 10 cases. In this categorization, data model refers to various semantic model layers, namely, the use of various types of data models that include, for example, data content specifications, RIMs, and clinical information models depending on the development context. A dual model was discussed in 2 of the cases for the application of data models.

Closely related to the aspects of interoperability, several interoperability standard solutions were named. Named ontology solutions included a top-domain ontology for the life sciences (BioTopLite) in 2 cases, a HL7 Fast Health Interoperability Resources (FHIR) and semantic sensor network–based type 1 diabetes ontology for type 1 diabetes data, and a system of several ontologies to be used for building EHR interoperability. Systematized Nomenclature of Medicine Clinical Terminology was the common terminology application in 7 cases, whereas classification systems were applied in more heterogeneous ways. The following international classifications were named: *International Classification of Diseases, Tenth Revision*; *International Classification of Diseases, Ninth Revision, Clinical Modification*; The Anatomical Therapeutic Chemical Classification System; and Logical Observation Identifiers Names and Codes. One article documented national classification use. Applied health care–specific standards included the open standard specification in health informatics (openEHR; n=6), Digital Imaging and Communications in Medicine (n=1), HL7 FHIR (n=5), and the HL7 Clinical Document Architecture (n=2). Regarding data models or reference information models, several types were applied for distinct use environments. These included the Observational Medical Outcomes Partnership common data model, an EHR-specific data component model, the i2b2 common data model for data warehouse development, the HL7 FHIR RIM, and the EN/ISO 13606 standard–based model. Moreover, 1 case reported using openEHR as a data model reference.

The method for applying an interoperability framework or approach is related to the overall design of the data use purposes and the needs driving the semantic development. The chosen methodology for semantic development was based on ontology development or the application of an ontology framework in 4 research cases, data model–based development in 5 cases, archetype development in 1 case, and clinical data warehouse development to enhance access and processing of data in 1 case. In data model–based approaches, use cases document a method's capability in separating different semantic levels of development, that is, system level, application level, clinical user interface level, or patient information level. The reusability of data model–based semantic approaches and related methods were assessed for resource savings in time and cost in development projects and, thus, to justify the choice of the approach. For example, clinical knowledge model–based development may allow recycling archetypes that further promote semantic interoperability.

Discussion

Principal Findings

Our results are related to the main goals of semantic interoperability development, such as enabling patient data use regardless of which EHR the data originated from and by which terminologies, classifications, or other semantic features they are supported [16-19,21]. Regarding key elements of semantic interoperability, of the documented terminologies, Systematized Nomenclature of Medicine Clinical Terminology seemed to prevail as the dominant choice for clinical terminology [24-30]. For international classifications that are typically integrated into EHRs, a selection of well-established classifications was documented [25,26,31,32]. Likewise, several health care specific standards [24-26,28,31,33], ontologies [21,24,32,33], and data models [25,27,28,30-36] were presented, albeit in a relatively small sample in this study. One possible factor affecting the selection of interoperability features such as international standards may be open availability and the level of cost of the standard-specific resources and their deployment. Consequently, shared implementation experiences and recommendations from previous projects or from collaboration in international communities may promote and facilitate decision-making concerning future implementations.

Our review illustrates several approaches for building sematic interoperability. For ontologies and data models, based on the review, several layers may be deployed to address semantic interoperability development needs. For ontologies, deploying a system of ontologies seeks to bridge, for example, domain-specific ontologies and application-specific ontologies. In our sample, a case with a data model–based development approach enhanced the communication of clinical information with the application. The application was used by the patients in self-monitoring, and the EHR served as a clinical data repository to avoid the loss of meaningful information. In general, when applying data model–based approaches, a dual model or multimodel approach may be needed to address different semantic interoperability issues during development—from the clinician as an EHR user to the system transaction level.

Our review highlights several clinical benefits of semantic interoperability. Primarily semantic interoperability fulfills the need to support the implementation of applications that enhance the continuity of care and ensure access to safe and high-quality health care. The reported clinical benefits of developing semantic interoperability reflect well common international goals [2,3,5]. The results in our sample show that an evident goal driving the development in these studies is the following assumption: through increased access to patient information, better quality and outcomes in care can be achieved [24,26,27,33,37]. Better communication based on easily accessible data across EHRs is facilitated not only between clinicians but also between professionals and patients [28,34,35]. Further advances are related to efficiency and subsequent economic factors, for example, reducing the clinicians' workload for documenting and evaluating extensive patient data, to avoid information overload and support multiprofessional care

[26,31-33,35]. In addition, interoperable patient data provide opportunities for a wide range of EHR-related clinical development, for example, regarding decision-making support, other EHR integrated tools, clinical research, or other types of secondary use [25,28-31,33,36]. Essentially, the interoperability cases in our review demonstrated a well-documented selection of development goals in EHRs, including considerations of patient-generated, self-monitoring data and related interoperability features.

Finally, when reflecting on the goal-related semantic interoperability results, there is evidently not one universal approach available to tackle all interoperability-related needs and challenges. One reason for this is that interoperability is to be achieved in diffuse and disconnected clinical care settings and in registry data use across borders. However, regulations and international recommendations can support the choosing of common tools and standards for building interoperability for patient data generated in various EHRs and clinical applications. This may be the strongest selling point for evolving international frameworks, such as the EHDS regulation proposal. If adopted, unified toolkits of the most crucial means can be achieved for building international eHealth interoperability. Through these mechanisms, common solutions and standards can be agreed upon to remedy existing inconsistencies and avoid possible future imparities that hinder the realization of the common goals. It is noteworthy that all member states have steps to take to meet the international requirements with a country-specific road map to achieve the common goal [3,5]. Moreover, it would require cooperation to align on which level of interoperability should be reached when the operating environment consists of a diverse set of clinical practices and related data needs, such as between public and private care or between primary and specialized care. Additionally, it may be worthwhile to consider whether instead of promoting a single approach, semantic interoperability requirements should be assessed through several levels of semantic requirements, such as standards, data models, classifications, and terminologies. Moreover, developing the necessary skills and increasing capabilities is an essential component of this development.

Specifically, regarding European development, one of the main goals is to support the use of health data for better health care delivery and better research. The comprehensive and timely availability of EHR data is known to improve the quality of care and patient safety [26,38]. Concurrently, the lack of not only technical or organizational but also semantic interoperability has been recognized as one of the barriers for the cross-border exchange of health data [2-8]. Therefore, commonly recognized interoperability approaches and standards for the harmonization of semantic interoperability are needed.

Limitations

Our goal was to ensure that we did not overlook any important studies and to minimize any potential biases by conducting a thorough and comprehensive search of the available literature. However, it is worth noting that our search was limited to a single database, PubMed. Nevertheless, recent literature suggests that PubMed can serve as a primary search tool. It possesses the necessary capabilities for systematic reviews, including the ability to formulate and interpret queries accurately, as well as ensuring search reproducibility. It is important to acknowledge that even a well-performing system such as PubMed might not always yield the desired results in different scenarios [23]. Our data set was limited by a small sample size of 14 articles. Therefore, findings can only be regarded as descriptive in nature. Relatively large heterogeneity in study environments and selected research approaches limit us from drawing strong conclusions. Despite these limitations, this review demonstrates potentially feasible approaches for promoting semantic interoperability toward harmonized approaches. Additional real-world studies accounting for semantic interoperability are needed to reinforce understanding of the most promising, scalable examples such as international reference models (eg, HL7 RIM). Moreover, it was challenging to determine the "development status" category for certain studies. This was due to varying levels of details in the study reports, where some of the studies provided a wealth of detail, whereas some were more restricted in their scope.

Suggestions for Future Research

Future research directions are 2-fold from the current development perspective. First, evidence-based recommendations on semantic interoperability features, for example, data models and terminologies, are needed. Initially, the applicability of international data models and standards such as HL7 V2 might be evaluated. Second, more experiences of interoperability development should be reported in the peer-reviewed research literature to contribute evidence around successful and not so successful experiences instead of leaning solely on individual expert opinions. Presumably, due to the evolving implementation status of semantic interoperability cases illustrated in the research literature, systematic research–based evaluation of benefits and outcomes is still scarce.

Conclusions

We conclude that based on our review, the research literature highlights valuable aspects in promoting semantic interoperability in terms of the efficiency and feasibility of solutions integrated in EHRs and possibly for enhancing care. However, when heading toward semantic harmonization, more data, pilot experiences, and analyses are needed to assess how applicable the chosen specific solutions are for the standardization and semantic interoperability of patient data. Instead of promoting a single approach, semantic interoperability could be assessed through several levels of semantic approaches. A dual model or multimodel approach is usable to address different semantic interoperability issues during development—from the clinician as an EHR user to the system transaction level. Since interoperability is being implemented in complex and disconnected clinical care environments, choices should be well prepared and justified to meet sustainable and long-term outcomes. From that point of view, it is possible to outline future directions in selecting semantic interoperability approaches for the realization of the international patient data–related goals.

Acknowledgments

The study was supported by Finnish governmental study grant TYH2021319.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of study and sample characteristics.
[DOCX File, 25 KB - medinform_v12i1e53535_app1.docx]

Multimedia Appendix 2
Summary of results on semantic interoperability in electronic health records.
[DOCX File, 27 KB - medinform_v12i1e53535_app2.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File, 439 KB - medinform_v12i1e53535_app3.pdf]

References

1. Iyamu I, Gómez-Ramírez O, Xu AXT, et al. Defining the scope of digital public health and its implications for policy, practice, and research: protocol for a scoping review. JMIR Res Protoc 2021 Jun 30;10(6):e27686. [doi: 10.2196/27686] [Medline: 34255717]
2. Global strategy on digital health 2020-2025. : World Health Organization; 2021 URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2023-09-25]
3. Godinho MA, Ansari S, Guo GN, Liaw ST. Toolkits for implementing and evaluating digital health: a systematic review of rigor and reporting. J Am Med Inform Assoc 2021 Jun 12;28(6):1298-1307. [doi: 10.1093/jamia/ocab010] [Medline: 33619519]
4. Abboud L, Cosgrove S, Kesisoglou I, et al. TEHDAS Deliverable 4.1 Country factsheets: Mapping health data management systems through country visits: development, needs and expectations of the EHDS. : TEHDAS Consortium Partners; 2023 Apr 28 URL: https://tehdas.eu/app/uploads/2023/04/tehdas-mapping-health-data-management-systems-through-country-visits.pdf [accessed 2024-03-26]
5. Hussein R, Scherdel L, Nicolet F, Martin-Sanchez F. Towards the European Health Data Space (EHDS) ecosystem: a survey research on future health data scenarios. Int J Med Inform 2023 Feb;170:104949. [doi: 10.1016/j.ijmedinf.2022.104949] [Medline: 36521422]
6. Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space: COM/2022/197 final. European Union. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197 [accessed 2023-09-25]
7. Kouroubali A, Katehakis DG. The new European interoperability framework as a facilitator of digital transformation for citizen empowerment. J Biomed Inform 2019 Jun;94:103166. [doi: 10.1016/j.jbi.2019.103166] [Medline: 30978512]
8. Stellmach C, Muzoora MR, Thun S. Digitalization of health data: interoperability of the proposed European Health Data Space. Stud Health Technol Inform 2022 Aug 31;298:132-136. [doi: 10.3233/SHTI220922] [Medline: 36073471]
9. Gottumukkala M. Development, and evaluation of an automated solution for electronic information exchange between acute and long-term postacute care facilities: design science research. JMIR Form Res 2023 Feb 17;7:e43758. [doi: 10.2196/43758] [Medline: 36800213]
10. Carlson J, Laryea J. Electronic health record-based registries: clinical research using registries in colon and rectal surgery. Clin Colon Rectal Surg 2019 Jan;32(1):82-90. [doi: 10.1055/s-0038-1673358] [Medline: 30647550]
11. Hohman KH, Martinez AK, Klompas M, et al. Leveraging electronic health record data for timely chronic disease surveillance: the multi-state EHR-based network for disease surveillance. J Public Health Manag Pract 2023;29(2):162-173. [doi: 10.1097/PHH.0000000000001693] [Medline: 36715594]
12. 2015 Edition health information technology (health IT) certification criteria, 2015 Edition base electronic health record (EHR) definition, and ONC health IT certification program modifications. Federal Register. 2015 Oct 16. URL: https://www.federalregister.gov/documents/2015/10/16/2015-25597/2015-edition-health-information-technology-health-it-certification-criteria-2015-edition-base [accessed 2023-02-21]
13. Kush RD, Warzel D, Kush MA, et al. FAIR data sharing: the roles of common data elements and harmonization. J Biomed Inform 2020 Jul;107:103421. [doi: 10.1016/j.jbi.2020.103421] [Medline: 32407878]

14. Horgan D, Hajduch M, Vrana M, et al. European Health Data Space-an opportunity now to grasp the future of data-driven healthcare. Healthcare (Basel) 2022 Aug 26;10(9):1629. [doi: 10.3390/healthcare10091629] [Medline: 36141241]
15. Palojoki S, Vakkuri A, Vuokko R. The European cross-border health data exchange: focus on clinically relevant data. Stud Health Technol Inform 2021 May 27;281:442-446. [doi: 10.3233/SHTI210197] [Medline: 34042782]
16. Gamal A, Barakat S, Rezk A. Standardized electronic health record data modeling and persistence: a comparative review. J Biomed Inform 2021 Feb;114:103670. [doi: 10.1016/j.jbi.2020.103670] [Medline: 33359548]
17. Lee AR, Kim IK, Lee E. Developing a transnational health record framework with level-specific interoperability guidelines based on a related literature review. Healthcare (Basel) 2021 Jan 13;9(1):67. [doi: 10.3390/healthcare9010067] [Medline: 33450811]
18. de Mello BH, Rigo SJ, da Costa CA, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272. [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
19. Hwang KH, Chung KI, Chung MA, Choi D. Review of semantically interoperable electronic health records for ubiquitous healthcare. Healthc Inform Res 2010 Mar;16(1):1-5. [doi: 10.4258/hir.2010.16.1.1] [Medline: 21818417]
20. Interoperability in healthcare. Healthcare Information and Management Systems Society (HIMSS). URL: https://www.himss.org/resources/interoperability-healthcare [accessed 2023-12-21]
21. Moreno-Conde A, Moner D, Cruz WD, et al. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inform Assoc 2015 Jul;22(4):925-934. [doi: 10.1093/jamia/ocv008] [Medline: 25796595]
22. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.2: Cochrane; 2023. URL: http://www.training.cochrane.org/handbook [accessed 2023-01-27]
23. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of Google scholar, PubMed, and 26 other resources. Res Synth Methods 2020 Mar;11(2):181-217. [doi: 10.1002/jrsm.1378] [Medline: 31614060]
24. Martínez-Costa C, Schulz S. Ontology content patterns as bridge for the semantic representation of clinical information. Appl Clin Inform 2014 Jul 23;5(3):660-669. [doi: 10.4338/ACI-2014-04-RA-0031] [Medline: 25298807]
25. Sun H, Depraetere K, de Roo J, et al. Semantic processing of EHR data for clinical research. J Biomed Inform 2015 Dec;58:247-259. [doi: 10.1016/j.jbi.2015.10.009] [Medline: 26515501]
26. Andersen SNL, Brandsborg CM, Pape-Haugaard L. Use of semantic interoperability to improve the urgent continuity of care in Danish ERs. Stud Health Technol Inform 2021 May 27;281:203-207. [doi: 10.3233/SHTI210149] [Medline: 34042734]
27. Martínez-Costa C, Cornet R, Karlsson D, Schulz S, Kalra D. Semantic enrichment of clinical models towards semantic interoperability. the heart failure summary use case. J Am Med Inform Assoc 2015 May;22(3):565-576. [doi: 10.1093/jamia/ocu013] [Medline: 25670758]
28. Frid S, Fuentes Expósito MA, Grau-Corral I, et al. Successful integration of EN/ISO 13606-standardized extracts from a patient mobile app into an electronic health record: description of a methodology. JMIR Med Inform 2022 Oct 12;10(10):e40344. [doi: 10.2196/40344] [Medline: 36222792]
29. Højen AR, Brønnum D, Gøeg KR, Elberg PB. Applying the SNOMED CT concept model to represent value sets for head and neck cancer documentation. Stud Health Technol Inform 2016;228:436-440. [Medline: 27577420]
30. Pedrera M, Garcia N, Blanco A, et al. Use of EHRs in a tertiary hospital during COVID-19 pandemic: a multi-purpose approach based on standards. Stud Health Technol Inform 2021 May 27;281:28-32. [doi: 10.3233/SHTI210114] [Medline: 34042699]
31. Boussadi A, Zapletal E. A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. BMC Med Inform Decis Mak 2017 Aug 14;17(1):120. [doi: 10.1186/s12911-017-0513-6] [Medline: 28806953]
32. González C, Blobel B, López DM. Ontology-based framework for electronic health records interoperability. Stud Health Technol Inform 2011;169:694-698. [doi: 10.3233/978-1-60750-806-9-694] [Medline: 21893836]
33. Kropf S, Chalopin C, Lindner D, Denecke K. Domain modeling and application development of an archetype- and XML-based EHRs. practical experiences and lessons learnt. Appl Clin Inform 2017 Jul 28;8(2):660-679. [doi: 10.4338/ACI-2017-01-RA-0009] [Medline: 28657637]
34. Marco-Ruiz L, Moner D, Maldonado JA, Kolstrup N, Bellika JG. Archetype-based data warehouse environment to enable the reuse of electronic health record data. Int J Med Inform 2015 Sep;84(9):702-714. [doi: 10.1016/j.ijmedinf.2015.05.016] [Medline: 26094821]
35. El-Sappagh S, Ali F, Hendawi A, Jang JH, Kwak KS. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the Hl7 FHIR standard. BMC Med Inform Decis Mak 2019 May 10;19(1):97. [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
36. Yang L, Huang X, Li J. Discovering clinical information models online to promote interoperability of electronic health records: a feasibility study of OpenEHR. J Med Internet Res 2019 May 28;21(5):e13504. [doi: 10.2196/13504] [Medline: 31140433]
37. Terner A, Lindstedt H, Sonnander K. Predefined headings in a multiprofessional electronic health record system. J Am Med Inform Assoc 2012;19(6):1032-1038. [doi: 10.1136/amiajnl-2012-000855] [Medline: 22744962]

38. Vuokko R, Vakkuri A, Palojoki S. Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT) clinical use cases in the context of electronic health record systems: systematic literature review. JMIR Med Inform 2023 Feb 6;11:e43750. [doi: 10.2196/43750] [Medline: 36745498]

Abbreviations

EHDS: European Health Data Space
EHR: electronic health record
EIF: European Interoperability Framework
FHIR: Fast Health Interoperability Resources
HL7: Health Level 7
RIM: reference information model
WHO: World Health Organization

Edited by C Lovis; submitted 10.10.23; peer-reviewed by H Ulrich, X Huang; revised version received 21.02.24; accepted 24.02.24; published 25.04.24.

Please cite as:
Palojoki S, Lehtonen L, Vuokko R
Semantic Interoperability of Electronic Health Records: Systematic Review of Alternative Approaches for Enhancing Patient Information Availability
JMIR Med Inform 2024;12:e53535
URL: https://medinform.jmir.org/2024/1/e53535
doi:10.2196/53535

Predicting Hypoxia Using Machine Learning: Systematic Review

Lena Pigat[1], MPH; Benjamin P Geisler[1], MD, MPH; Seyedmostafa Sheikhalishahi[1], PhD; Julia Sander[1], PhD; Mathias Kaspar[1], PhD; Maximilian Schmutz[1,2], MD; Sven Olaf Rohr[1], MD; Carl Mathis Wild[1,3], MD; Sebastian Goss[1], MD; Sarra Zaghdoudi[1], MSc; Ludwig Christian Hinske[1,4], Prof Dr

1

2

3

4

Corresponding Author:
Lena Pigat, MPH

Abstract

Background: Hypoxia is an important risk factor and indicator for the declining health of inpatients. Predicting future hypoxic events using machine learning is a prospective area of study to facilitate time-critical interventions to counter patient health deterioration.

Objective: This systematic review aims to summarize and compare previous efforts to predict hypoxic events in the hospital setting using machine learning with respect to their methodology, predictive performance, and assessed population.

Methods: A systematic literature search was performed using Web of Science, Ovid with Embase and MEDLINE, and Google Scholar. Studies that investigated hypoxia or hypoxemia of hospitalized patients using machine learning models were considered. Risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool.

Results: After screening, a total of 12 papers were eligible for analysis, from which 32 models were extracted. The included studies showed a variety of population, methodology, and outcome definition. Comparability was further limited due to unclear or high risk of bias for most studies (10/12, 83%). The overall predictive performance ranged from moderate to high. Based on classification metrics, deep learning models performed similar to or outperformed conventional machine learning models within the same studies. Models using only prior peripheral oxygen saturation as a clinical variable showed better performance than models based on multiple variables, with most of these studies (2/3, 67%) using a long short-term memory algorithm.

Conclusions: Machine learning models provide the potential to accurately predict the occurrence of hypoxic events based on retrospective data. The heterogeneity of the studies and limited generalizability of their results highlight the need for further validation studies to assess their predictive performance.

Trial Registration: PROSPERO CRD42023381710; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=381710

(JMIR Med Inform 2024;12:e50642) doi:10.2196/50642

KEYWORDS

artificial intelligence; machine learning; hypoxia; hypoxemia; anoxia; hypoxic; deterioration; oxygen; prediction; systematic review; review methods; review methodology; systematic; hospital; predict; prediction; predictive

Introduction

A key factor in risk assessment for sequelae and mortality in hospitalized patients is hypoxia. It describes the decreased availability of oxygen in specific body regions (tissue hypoxia) or in the body as a whole (general hypoxia) [1-3]. To prevent general hypoxia and to detect deterioration quickly, hypoxemia monitoring is commonly performed using pulse oximetry as a continuous and noninvasive assessment, especially in the intensive care unit (ICU) and operating room (OR) [4]. Hypoxemia is defined as an abnormally low level of blood oxygen. In addition to pulse oximetry, it can be assessed through an arterial blood gas analysis or imaging techniques, which can additionally serve as reliable indicators of subsequent tissue damage [3]. A multinational, multicenter study including 117 ICUs found a hypoxemia prevalence of more than 50% among all ICU patients [5]. The severity of hypoxemia was shown to be a direct risk factor for mortality in patients with hypoxemia. Being able to validly assess the individual risk of future hypoxemic and ultimately hypoxic events is therefore highly relevant.

To determine the risk or stage of a disease, artificial intelligence (AI) has been increasingly introduced into clinical routine in recent years to exploit underlying causal mechanisms that may not be accessible to humans. As a prime example, machine learning (ML) as a discipline of AI is being successfully used

for cancer tissue classification in medical imaging [6,7]. ML is also already being applied for prognostic purposes, for example, in the examination of patient characteristics to identify an increased risk of deterioration tendencies such as atrial fibrillation and of developing sequelae of diabetes mellitus or hereditary diseases [8-10].

Efforts to date of using ML to predict hypoxic events are being conducted in a variety of settings and demonstrate diverse approaches and methodologies. Studies differ significantly in terms of the patient population assessed, definition of prediction outcome, features used to predict hypoxia, and ML algorithms used, thus increasing the difficulty to generalize the conclusions of individual studies. It is therefore challenging to compare and evaluate these studies comprehensively.

This review aimed to provide a systematic and structured overview of the existing approaches to predict hypoxic events in the hospital setting. Our specific objectives were to summarize the different populations, model details, and prediction performance to capture the current state of available models; identify gaps and limitations; highlight promising approaches and methodologies; and provide guidance for future research in this area.

Methods

Protocol

This review was reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Checklist 1) [11]. The protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO) prior to data extraction (reference CRD42023381710).

Search Strategy

Relevant literature was searched for using Ovid with Embase and MEDLINE, Web of Science, and Google Scholar. Although the prior 2 databases were searched via their web query interface, Google Scholar was searched using the software Publish or Perish, as it allows for more complex queries [12].

Publications on the topic of hypoxia prediction using ML were searched by creating 2 sets of search terms, with the first set addressing hypoxia (including hypoxemia) and the second set addressing ML. With the identified search engines, the intersection of these 2 groups was then searched for, adjusting the syntax according to the search logic of the respective search engine. If Medical Subject Headings or thesaurus entries were available, the selected terms were included in the search logic accordingly. For the searches using Ovid and Web of Science, the search results were filtered to only include studies that did not use wearables for data collection and that were published in the English and German languages. Those filters were not applicable for the search of Google Scholar using Publish or Perish.

The selection and deduplication process was performed using Covidence (Veritas Health Innovation Ltd), with undetected duplicates removed by hand [13]. The search results of all databases were included, and duplicates were removed. The abstracts of the remaining results were independently screened by 2 reviewers. Results that met the selection criteria were reviewed in their entirety for the assessment of eligibility by 2 reviewers. In addition, references of the included studies were also screened for studies that meet the inclusion criteria and were subsequently included where appropriate. The search strategy was developed by 1 team member and reviewed by another with expertise in conducting systematic reviews. The detailed search strategy can be found in Multimedia Appendix 1.

Selection Criteria

Primary outcomes were model features, definition of the prediction end point, and predictive performance. Studies developing ML models to predict hypoxia or hypoxemia in continuously monitored human inpatients were included. Both studies of patients who were mechanically ventilated and spontaneously breathing were included. Hypoxia could be a main outcome or an auxiliary goal.

Studies that assessed hypoxia only in specific tissues were excluded, as this review addresses the prediction of general hypoxia as an important indicator of critical illness for risk stratification and early detection of patients at risk of acute health deterioration. Additionally, studies focusing on a population <18 years of age were not included, since the distinct etiologies, risk factors, and clinical presentations of hypoxia in pediatric patients may limit the generalizability of the findings to the population of adult inpatients.

The definition of the end point of hypoxia prediction (eg, specific oximetry thresholds or time frames of prediction) was left unspecified due to the expected heterogeneity in the approaches. The patient population of the included studies was not limited to a specific hospital setting or ward.

Data Extraction and Risk of Bias

Data extracted included the data source; sample size and setting; model variables; prediction end point and time frame; type of model; and the predictive performance of each model, usually expressed as classification measures such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), or area under the receiver operating characteristics (AUROC). Missing values of performance measures and summary data influenced the risk-of-bias assessment.

A qualitative synthesis of the included studies was conducted. For this purpose, an overview of all studies was provided in a narrative summary by categorizing them into subgroups based on the population, model features, model types, and setting. For each study, the model with the highest performance according to performance metrics was selected to summarize AUROC, sensitivity, specificity, PPV, and NPV as the most reported performance measures. In the case of studies that examined multiple prediction outcomes, the outcome definition that is the most similar to those of the other studies was chosen for reporting. For studies reporting 1 performance value per patient, a mean value was calculated for each measure. Because of the heterogeneous study designs and characteristics of the data used, as well as missing summary data of model performances, conducting a meta-analysis was not feasible.

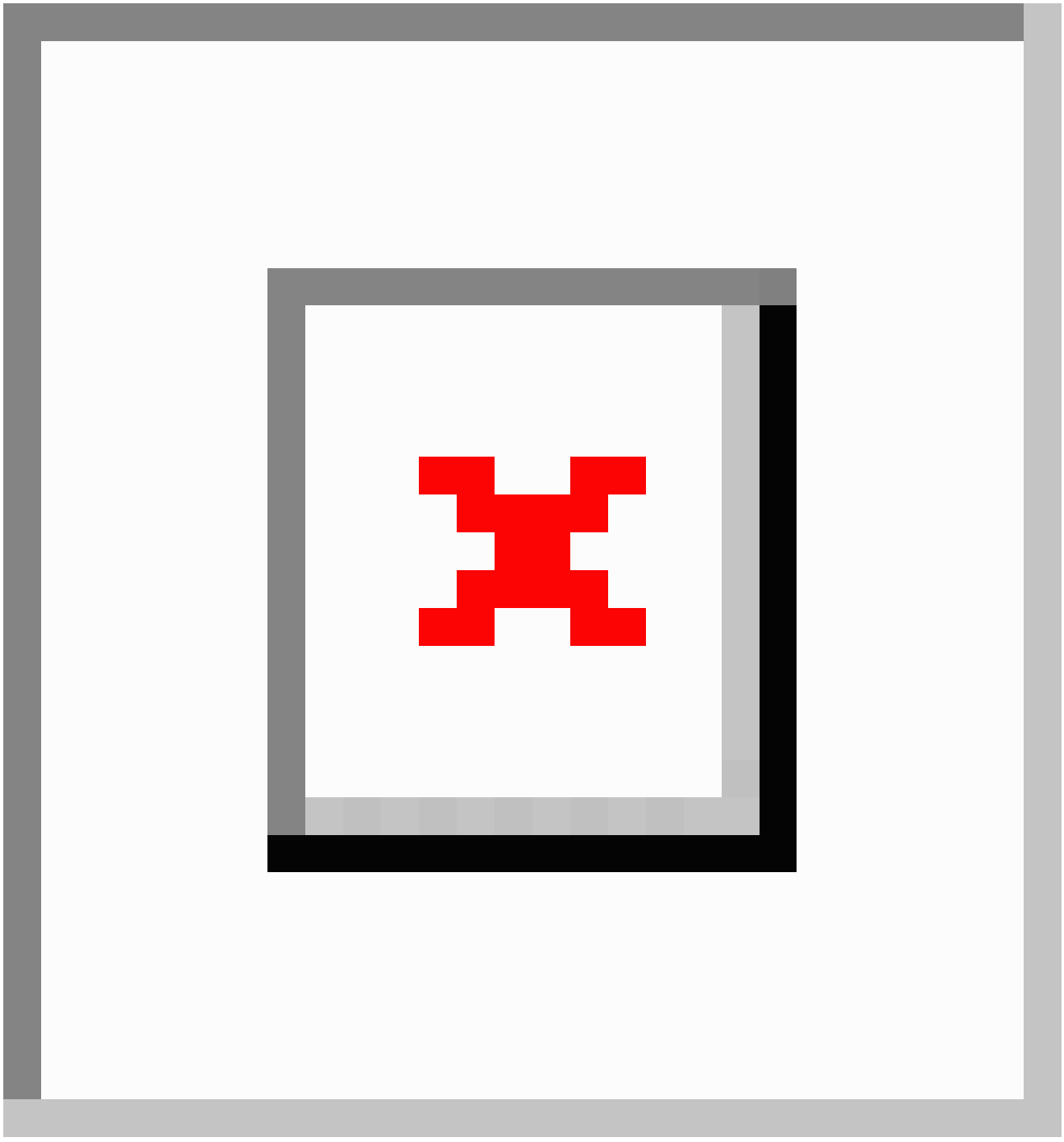
To assess the risk of bias, quality, and applicability of the studies included, Prediction Model Risk of Bias Assessment Tool (PROBAST) was used [14]. This tool is specifically designed to investigate the quality of prediction models and has become increasingly prevalent in systematic reviews in recent years. Assessment outcomes were evaluated based on 4 segments—participants, predictors, outcome, and analysis—and were determined by a comprehensive questionnaire. Risk of bias was rated as high, low, or unclear. If 1 domain suggested a high risk of bias, the overall risk of bias for that study was considered high. The assessment was conducted by a single researcher, with a second researcher reviewing the process independently.

Results

Literature Search

The initial search retrieved a total of 3734 studies (Figure 1). After removing a total of 700 duplicates, title and abstract screening identified the full texts of 31 studies for the assessment of eligibility. Of these, 19 studies were excluded due to not being a full study (n=6), not assessing a hypoxia outcome (n=4), not using machine learning (n=3), inability to obtain the full text (n=2), having an outpatient setting (n=2), having a pediatric patient population (n=1), and being in the Chinese language (n=1). The remaining 12 studies were included in the review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.



Study Characteristics

Overview

Table 1 presents the characteristics of all included studies and gives an overview of the best-performing model in each study, divided into conventional ML and deep learning models for studies including both. The studies were conducted in the United States [15-22], China [23,24], Germany [25], and the United Arab Emirates [26]. Half (6/12, 50%) of them were published after 2020 [15,16,19,21,22,26]. In 3 (25%) of the 12 studies, the prediction of hypoxia was a side or auxiliary goal [17,19,21], whereas it was the main study aim for the other studies.

Table . Study characteristics of the reviewed studies (n=12). The model with the highest performance in each study is reported. For studies using both conventional machine learning and deep learning models, each best-performing model is reported. For studies examining multiple prediction outcomes, the outcome definition that is the most similar to those of other studies was chosen for reporting. For studies reporting 1 performance value per patient, a mean value was calculated.

Reference	Sample size n	Clinical variables, n	Prediction end point	Model	Performance	External validation
Annapragada et al [15] (2021)	2435	1	SpO_2[a] <92% within the next 5 and 30 min (occurrence and magnitude of hypoxemic events)	• LSTM[b]	• PPV[c]: 0.94 • Sensitivity: 0.80 • Specificity: 0.99	Yes
Chen et al [16] (2021)	57,171	21	SaO_2[d] <93% within the next 5 min	• GBT[e]	• AUROC[f]: 0.89	Yes
ElMoaqet et al [17] (2014)	119	1	SpO_2 ≤89% within the next 20 and 60 s	• Lin[g]	• AUROC: 0.93	No
Erion et al [18] (2017)	57,173	1	SpO_2 ≤92% within the next 5 min	• LSTM • GBT	• LSTM AUROC: 0.87 • GBT AUROC: 0.86	No
Geng et al [23] (2018)	308	3	SpO_2 <90% for any duration during the endoscopic procedure	• LR[h]	• AUROC: 0.76	No
Geng et al [24] (2019)	220	3	SpO_2 <90% for any duration during the endoscopy procedure	• ANN[i]	• AUROC: 0.80	No
Lam et al [19] (2022)	39,630	26	SpO_2 <91% and <96% after algorithm evaluation and any time during hospitalization	• XGB[j] • RNN[k]	• XGB AUROC: 0.64 • RNN AUROC: 0.64	Yes
Lundberg et al [20] (2018)	36,232	>65	SpO_2 ≤92% initial status and within the next 5 min	• GBM[l]	• AUROC: 0.90	No
Ren et al [21] (2022)	17,818	3	PaO_2[m]/FiO_2[n] ≤150 at any time during ventilation	• NN[o] • LR	• NN AUROC: 0.83 • LR AUROC: 0.81	Yes
Sippl et al [25] (2017)	620	17, RF[p] and NN used subsets of 6 and 7	Presence and severity of temporary oxygen desaturation during anesthesia induction and intubation based on expert annotations	• NN • RF	• NN sensitivity: 0.74 • NN specificity: 0.93 • RF sensitivity: 0.35 • RF specificity: 0.99	No
Statsenko et al [26] (2022)	605	2D and 3D diagnostic images of the chest	Markers of systemic oxygenation: functional (HR[q], BR[r], SBP[s], and DBP[t]) and biochemical findings (SpO_2, serum potassium level, and AG[u])	• CNN[v]	• MAE[w]: mean 7.941% (SD 4.131%)	No

Reference	Sample size n	Clinical variables, n	Prediction end point	Model	Performance	External validation
Xia et al [22] (2022)	14,777	29	PaO_2 <60 mm Hg after extubating	• RF	• AUROC: 0.792	No

[a]SpO_2: peripheral oxygen saturation.

[b]LSTM: long short-term memory.

[c]PPV: positive predictive value.

[d]SaO_2: arterial oxygen saturation.

[e]GBT: gradient boosted tree.

[f]AUROC: area under the receiver operating characteristics.

[g]Lin: linear regression.

[h]LR: logistic regression.

[i]ANN: artificial neural network.

[j]XGB: extreme gradient boosting.

[k]RNN: recurrent neural network.

[l]GBM: gradient boosting machine.

[m]PaO_2: partial pressure of oxygen.

[n]FiO_2: fraction of inspired oxygen.

[o]NN: neural network.

[p]RF: random forest.

[q]HR: heart rate.

[r]BR: breath rate.

[s]SBP: systolic blood pressure.

[t]DBP: diastolic blood pressure.

[u]AG: anion gap.

[v]CNN: convolutional neural network.

[w]MAE: mean averaged error to the range of values.

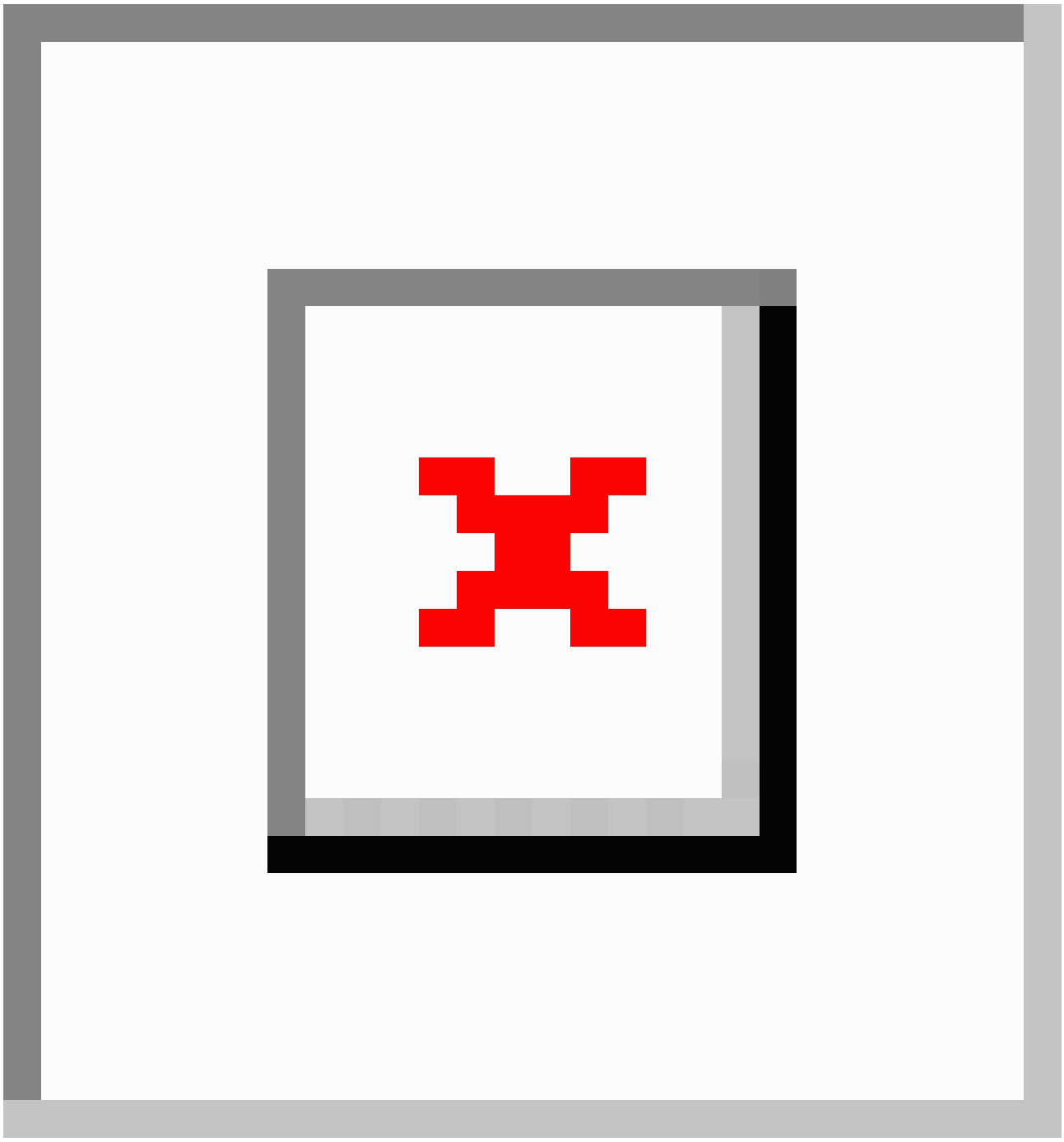
Data Sources and Population

Most studies (9/12, 75%) analyzed a large sample size of 500 or more patients [15,16,18-22,25,26]. Data from the publicly available databases Medical Information Mart for Intensive Care and eICU Collaborative Research Database were used in 4 of the studies [15,16,21,22], whereas 3 studies relied on data collected via an anesthesia information management system (AIMS) [16,18,20]. AIMSs are widely adopted hardware and software solutions that are integrated into a hospital's electronic health record system and are used to manage and document a patient's perioperative measurements [27,28]. The studies were set in the OR (n=5) [16,18,20,23,24], the ICU (n=3) [15,21,22], and mixed or general care units (n=4) [17,19,25,26]. Of the 12 studies analyzed, 10 (83%) did not include patients with COVID-19 [16-25], whereas the remaining 2 (17%) studies either were performed only on patients who tested positive for COVID-19 or were externally validated on a COVID-19 cohort [15,26].

ML Model Specifics

Figure 2 [15-26] gives an overview of the models and the number of variables used in each study. Exclusively conventional ML algorithms were applied in 5 of the identified studies [16,17,20,22,23], whereas 7 studies included deep learning algorithms [15,18,19,21,24-26]. Models based on logistic regression were used most often (n=4) [18,21-23], followed by artificial neural networks (n=3) [21,24,25].

Figure 2. Machine learning (ML) methods used by each study. ML methods (upper half) in gray: conventional ML; ML methods in black: deep learning. Studies are sorted by the number of clinical variables used. Studies in blue: 1 clinical variable; studies in green: 2-5 clinical variables; studies in yellow to red: >5 clinical variables. ANN: artificial neural network; Autoreg: autoregressive model; CNN: convolutional neural network; DTW: dynamic time warping; GBM: gradient boosting machine; GBT: gradient boosted tree; kNN: k-nearest neighbor; Lin: linear regression; LR: logistic regression; LSTM: long short-term memory; RF: random forest; RNN: recurrent neural network; SVM: support vector machine; XGB: extreme gradient boosting.



The number of clinical variables included ranged from 1 to over 65 different variables. The prediction of hypoxic events was based solely on prior peripheral oxygen saturation (SpO_2) values in 3 studies [15,17,18], whereas 4 studies used 2 or 3 clinical variables as input [21,23,24,26]. The remaining 5 studies relied on at least 6 variables [16,19,20,22,25]. The most frequently used variable sources were oximetry measurements (9/12, 75%) [15-22,25] and static patient characteristics such as age (5/12, 42%) [16,19,20,23,25]. Additionally, a single study relied on diagnostic images of the chest to make predictions [26].

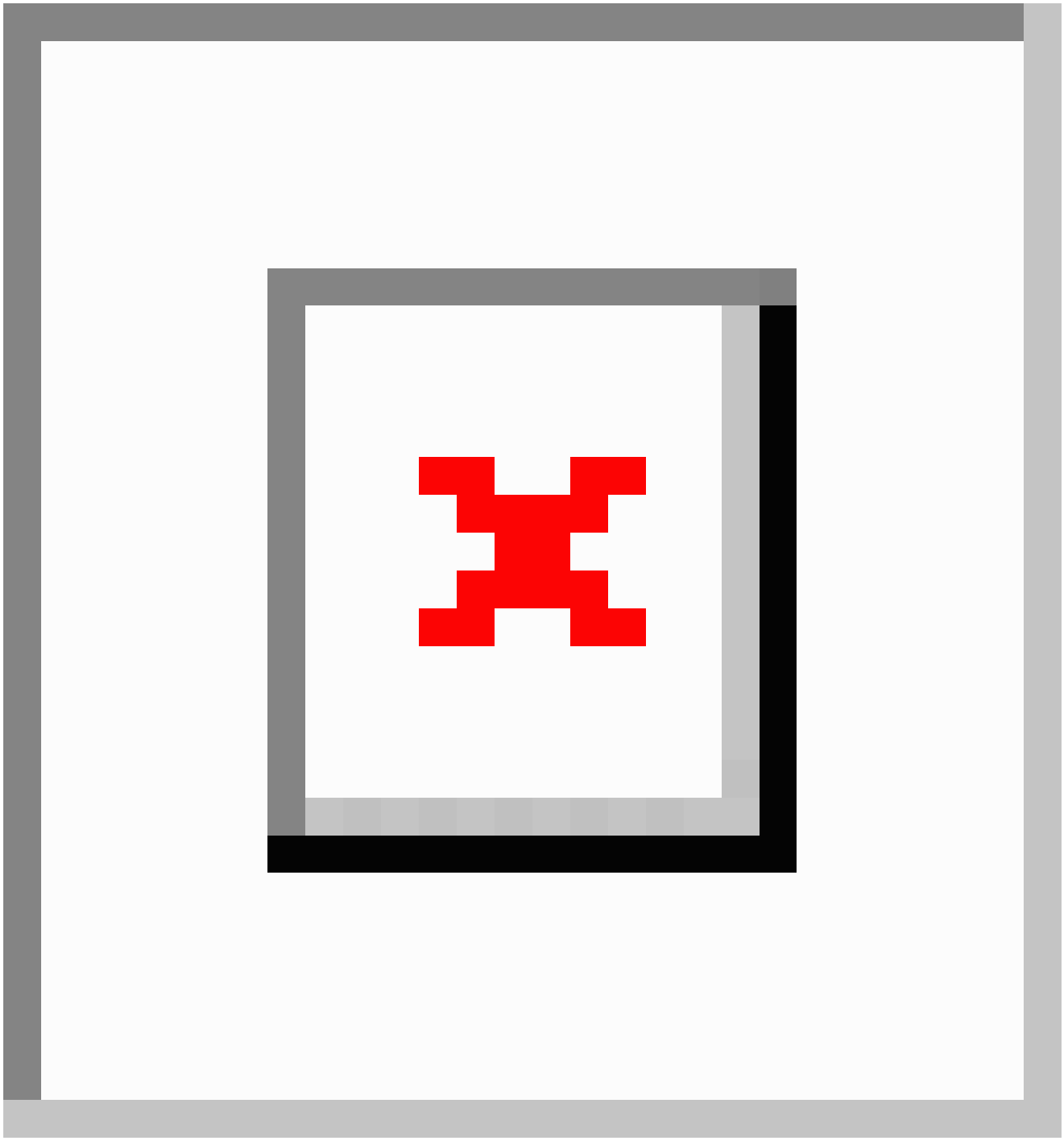
The prediction end point was defined by a threshold of SpO_2 between 89% and 92% for most of the studies (7/12, 58%) [15,17-20,23,24]. Thresholds of the partial pressure of oxygen, the arterial oxygen saturation, or the ratio of partial pressure of oxygen to the fraction of inspired oxygen were used in 3 other studies [16,21,22]. The remaining 2 studies assessed the presence and severity of hypoxia as defined by expert annotations and predicted functional markers of hypoxia, respectively [25,26]. Defined time frames for prediction included the length of a certain procedure [21,23-25], any time after

extubating [22], and a set time window of 5 to 30 minutes [15-18,20].

Performance

Most of the 12 studies reported sensitivity (n=9, 75%), specificity (n=8, 67%), or AUROC (n=9, 75%) as classification measures. Other performance indicators were PPV, NPV, area under the precision-recall curve, accuracy, and F_1-score. The most frequently reported performance measures of the best-performing model in each study are summarized in a heat map (Figure 3 [15-26]). The reported performance measures of 1 study were based on 10 individual patients since the focus of the study was to propose a performance metric and therefore have limited informative value [17]. One other study only reported the proportion of the mean averaged error to the range of values [26].

Figure 3. Heat map of performance measures, sorted by AUROC. The performance of the best-performing model in each study is presented. In the case of studies that examined multiple prediction outcomes, the outcome definition that is the most similar to the other studies was chosen for reporting. For studies stating 1 performance value per patient, the metrics represent the mean value. For 3 of the included studies, hypoxia prediction was not the main study aim [17,19,21]. The reported performance measures of 1 study were based on 10 individual patients and therefore have limited informative value. One study only reported the proportion of the mean averaged error to the range of values. AUROC: area under the receiver operating characteristics; NPV: negative predictive value; PPV: positive predictive value.



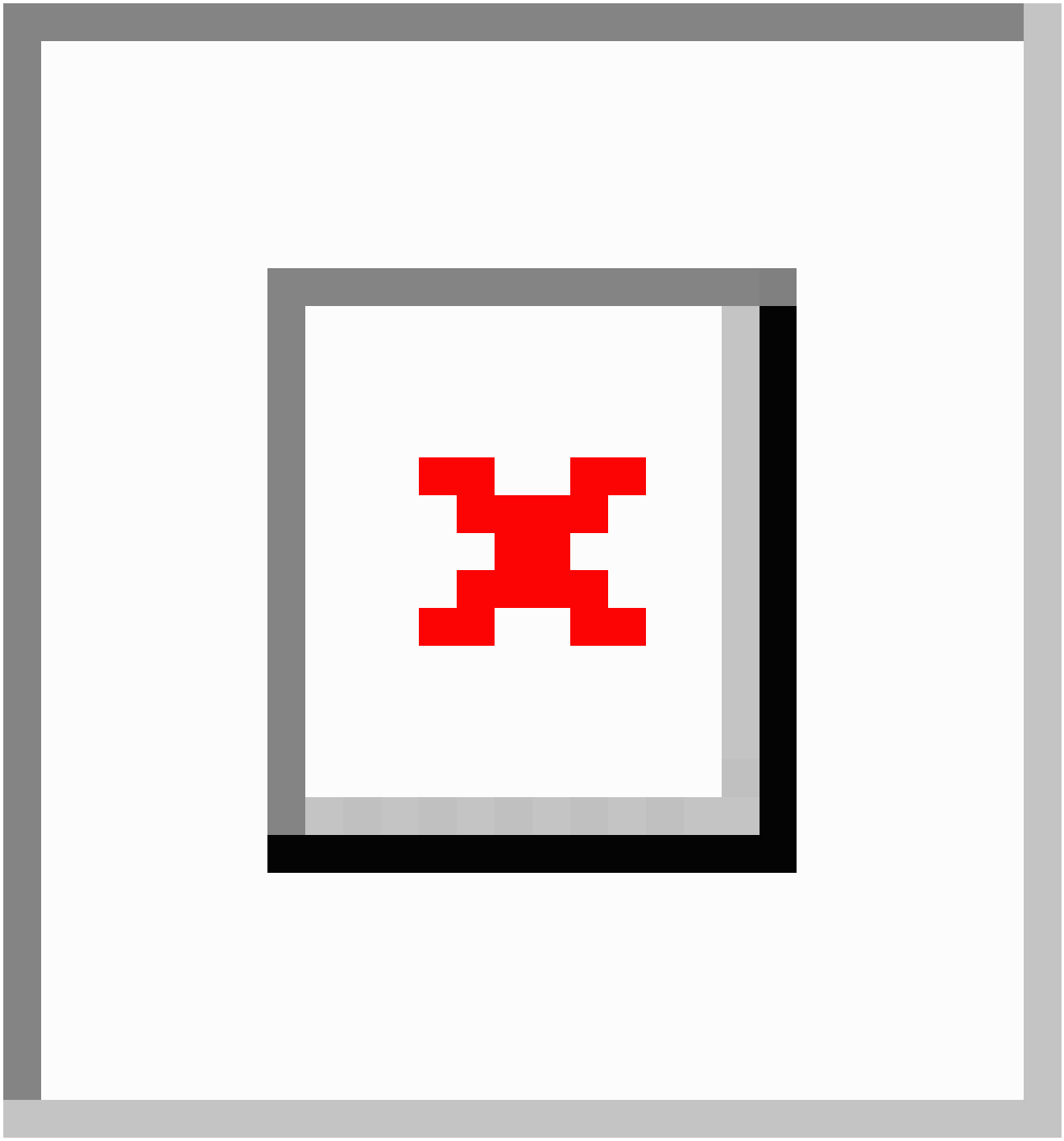
Of the 9 studies reporting AUROC, 8 (89%) showed a value higher than 0.75 [16-18,20-24]. This included 3 studies that showed a significant trade-off between sensitivity and specificity [17,21,24]. The overall performance was moderate or high with respect to classification metrics, both in studies performing the prediction task as the main study aim and in studies predicting hypoxia as a side or auxiliary goal. In studies drawing a comparison to anesthesiologist decisions, the prediction models alone or anesthesiologists using those models outperformed anesthesiologists without access to the model [18,20].

Deep learning and conventional ML are not directly comparable as they are not being applied on the same data set and the performance metrics are not consistently reported. However, in all studies comparing the 2 approaches, deep learning models showed similar or better performance than conventional ML models considering classification metrics [18,19,21,25]. Additionally, models only using prior SpO_2 data as a variable tended to outperform models using more clinical variables [15,17,18]. Two (67%) of the 3 studies only using prior SpO_2 data applied a long short-term memory (LSTM) algorithm, 1 of which was able to predict the detailed trend of the SpO_2 waveform [15,18]. Multitask learning for the prediction of related end points was implemented in 1 study, showing improved performance with an increasing number of tasks [19]. Approaches for providing explainability of their prediction outcome were presented in 2 studies, with 1 offering a real-time prediction tool displaying the contributing factors of an individual patient's hypoxemia risk within the next 5 minutes [16,20].

Risk-of-Bias Assessment

PROBAST was used to assess the risk of bias and applicability of each study. In the case of external validation, the assessment for that validation was performed separately. An overview of the overall and segment ratings of all 12 studies analyzed are shown in Figure 4. The overall risk of bias was rated as high or unclear for most of the studies (10/12, 83%) [16-21,23-26]. Unclear or high risk of bias ratings were mainly due to missing details of the procedure as well as unclear or unfitting timing of predictors or outcomes. External validation was only performed in 4 of the studies [15,16,19,21], whereas the other 8 studies relied on internal validation, primarily using random split samples and cross-validation [17,18,20,22-26].

Figure 4. Risk-of-bias assessment for all studies (n=12) based on 4 segments. The graph shows the number of studies with low, high, and unclear risk of bias by the author's assessment using PROBAST (Prediction Model Risk of Bias Assessment Tool).



Discussion

Principal Findings

In this systematic review, we identified and summarized 12 studies predicting hypoxic events or markers for hypoxia. The approaches proved to be highly diverse both in their assessment and definition of a hypoxic outcome as well as in the variables and model types used. Therefore, the comparability between studies was limited by the high variability of approaches, such as the variety of settings involving different influences on blood oxygen saturation (eg, sedation during surgery).

The data used to develop the models were primarily obtained from publicly available databases or directly from hospitals' AIMSs or electronic health record systems. Settings for the prediction included the OR, ICU, and general care units. The implemented ML models were based on both conventional ML and deep learning methods and assessed prediction end points defined as a threshold for blood oxygen measurements for most studies. Clinical variables used included patient characteristics, vital signs, and laboratory data. Blood oxygen data were the most applied model variables for hypoxia prediction.

The overall predictive performance of the presented models was moderate or high across the various settings. Deep learning approaches showed similar or better performance than

conventional ML approaches within the same studies. Models predicting hypoxia solely based on prior oximetry data tended to outperform models using more variables as inputs, with most of these studies using an LSTM algorithm.

The demonstrated trade-off between sensitivity and specificity of model performance highlights that it may be difficult to achieve both at the same time, especially when predicting medical events. This is a major caveat that holds true for a broad variety of diagnostic tests in medicine, such as D-dimers in investigating venous thromboembolism [29]. High specificity but low sensitivity, as demonstrated by 2 of the models, might, for example, result from missing relevant variables or an insufficient number of outcome events due to small sample sizes. An algorithm with high specificity may help to reduce unnecessary interventions, potentially leading to cost savings and minimizing patient inconvenience. However, in practice, an algorithm with that trade-off does not reliably detect patients with hypoxia who require immediate attention and may therefore be more appropriate as a decision support tool rather than a stand-alone diagnostic tool.

High sensitivity but low specificity on the other hand can, for example, be caused by the inclusion of variables that are highly associated with the presence of hypoxia but are not specific to hypoxia alone, or by the model being too sensitive and thus detecting subtle changes in nonhypoxic cases that are incorrectly classified as hypoxic. Practically, such a model could result in overalerting, disqualifying it for clinical application.

The informational value of many of the studies presented was limited due to a lack of external validation. In addition, more precise classification performance metrics were often not provided, thus not allowing for a meta-analysis. Unclear ratings were mostly due to missing information, particularly in the analysis segment. Comparability between studies was limited by the high variability of approaches, such as the variety of settings involving different influences on blood oxygen saturation (eg, sedation during surgery).

Applicability and Future Opportunities

The successful prediction of hypoxic events within a time frame of 5 or even 30 minutes into the future demonstrates the ability to provide sufficient lead time for crucial treatment interventions. Hence, these results suggest the potential of developing a helpful prediction tool, applicable in clinical practice, which complements the assessment of nurses and clinicians. Such a tool could be extended by a presentation and visualization of individual factors influencing the predicted outcome of hypoxia, as demonstrated by Lundberg et al [20]. The approach to make the model more understandable is useful both for more nuanced therapy strategies and for the general usability and acceptance of an ML tool for the prediction of hypoxia in the clinical setting.

While models with many features might have higher accuracy and might be able to capture more detailed and complex relationships between the features and the outcome of hypoxia, they also come with a higher complexity for use and are prone to overfitting [30]. Given the intended use of a predictive algorithm for making timely decisions that have immediate impact on the health status of patients, complex models with excessive features could impede their implementation in clinical practice. Additionally, utility might be reduced by patients missing 1 or more of these features. Therefore, the prediction results of LSTM models based only on previous SpO_2 values provide a foundation for further development and refinement of models using only a few, readily available, and noninvasive respiratory variables.

The results of Lam et al [19] suggest that multitask learning may contribute to higher predictive performance on related respiratory outcomes. Therefore, an approach for parallel prediction of several relevant intensive care parameters could provide a basis for further exploration. Opportunities for combined prediction include predictive models for the necessity of changes in ventilation, in airway pressure, or for increased risk of ventilation failure [31-33]. The prediction of hypoxia could also be embedded in a more general early warning score for related outcomes, for which ML mechanisms are already being applied [19,34-36]. In addition, the development of ML prediction models in a clinical context should include consideration of recent advances for the prediction of other unrelated health parameters and outcomes to avoid a complex system of different prediction systems, thus limiting the applicability and acceptability of these efforts. Forthcoming studies in this area should strive to accurately report performance details of their models, as well as to consistently define the end point of the prediction, to allow comparison with other approaches.

Limitations

This review focused on studies predicting hypoxic or hypoxemic events and therefore did not include studies predicting related outcomes (eg, blood oxygen saturation) without stating that aim of prediction. The comparability of predictive performance among the included studies was limited due to substantial differences in methodology, variables, and end point definition, precluding a meta-analysis from being conducted. An additional challenge arose from the fact that some studies, while including hypoxia predictions, did so as an auxiliary objective and not as their primary focus. Therefore, we focused on a qualitative summary and on demonstrating the variety of approaches taken. The generalizability of the results presented might be further restricted by the countries of origin being limited to the United States, Europe, and Asia.

Conclusion

Despite the large methodological variance of the studies presented, this review shows promising approaches for the prediction of hypoxia status, a factor that is highly informative for changes to a patient's state of health. Future studies must aim to improve the external validation of the predictive performance and, thus, verify the generalizability of the results to additional data sets. The applicability of validated predictive models for hypoxia risk should be proven by prospective studies in clinical practice.

Acknowledgments

This work was supported by the German Ministry of Education and Research (BMBF), Berlin (#01ZZ2005). The open access publication of this article was supported by the Open Access Fund of the Medical Faculty of the University of Augsburg.

Authors' Contributions

LCH, BPG, and LP initiated the project. LP and BPG conducted the search. LP, BPG, MK, SS, SZ, MS, SOR, CMW, SG, and JS performed the screening and review. LP and MS conducted the data extraction. LP carried out the synthesis and narrative summary with MS reviewing the process. LP, MK, MS, and LCH substantially contributed to the final manuscript. MK, BPG, JS, MS, and LCH provided constructive comments and discussion on the project. All authors carefully read and commented on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy, data sources, and clinical variables.
[DOCX File, 34 KB - medinform_v12i1e50642_app1.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File, 36 KB - medinform_v12i1e50642_app2.docx]

References

1. Bhutta BS, Alghoula F, Berim I. Hypoxia. In: StatPearls: StatPearls Publishing; 2023. [Medline: 29493941]
2. Pittman RN. Regulation of tissue oxygenation. Colloquium Series on Integrated Systems Physiology 2011;3(3):1-100. [doi: 10.4199/C00029ED1V01Y201103ISP017]
3. Sood S, Manaker S, Finlay G. Evaluation and management of the nonventilated, hospitalized adult patient with acute hypoxemia. UpToDate. 2022 Sep 8. URL: www.uptodate.com/contents/evaluation-and-management-of-the-nonventilated-hospitalized-adult-patient-with-acute-hypoxemia [accessed 2023-02-22]
4. Aronson LA. Hypoxemia. In: Atlee JL, editor. Complications in Anesthesia, 2nd edition: Saunders; 2007:637-640.
5. SRLF Trial Group. Hypoxemia in the ICU: prevalence, treatment, and outcome. Ann Intensive Care 2018 Aug 13;8(1):82. [doi: 10.1186/s13613-018-0424-4] [Medline: 30105416]
6. Akazawa M, Hashimoto K. Artificial intelligence in gynecologic cancers: current status and future challenges - a systematic review. Artif Intell Med 2021 Oct;120:102164. [doi: 10.1016/j.artmed.2021.102164] [Medline: 34629152]
7. Kuntz S, Krieghoff-Henning E, Kather JN, et al. Gastrointestinal cancer classification and prognostication from histology using deep learning: systematic review. Eur J Cancer 2021 Sep;155:200-215. [doi: 10.1016/j.ejca.2021.07.012] [Medline: 34391053]
8. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
9. Nadarajah R, Wu J, Frangi AF, Hogg D, Cowan C, Gale C. Predicting patient-level new-onset atrial fibrillation from population-based nationwide electronic health records: protocol of FIND-AF for developing a precision medicine prediction model using artificial intelligence. BMJ Open 2021 Nov 2;11(11):e052887. [doi: 10.1136/bmjopen-2021-052887] [Medline: 34728455]
10. Gunasekeran DV, Ting DSW, Tan GSW, Wong TY. Artificial intelligence for diabetic retinopathy screening, prediction and management. Curr Opin Ophthalmol 2020 Sep;31(5):357-365. [doi: 10.1097/ICU.0000000000000693] [Medline: 32740069]
11. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
12. Harzing AW. Publish or Perish. Harzing.com. 2016 Feb 6. URL: https://harzing.com/resources/publish-or-perish [accessed 2022-11-08]
13. Covidence - better systematic review management. Covidence. URL: www.covidence.org [accessed 2022-11-10]
14. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 1;170(1):W1-W33. [doi: 10.7326/M18-1377] [Medline: 30596876]
15. Annapragada AV, Greenstein JL, Bose SN, Winters BD, Sarma SV, Winslow RL. SWIFT: a deep learning approach to prediction of hypoxemic events in critically-ill patients using SpO2 waveform prediction. PLoS Comput Biol 2021 Dec 21;17(12):e1009712. [doi: 10.1371/journal.pcbi.1009712] [Medline: 34932550]

16. Chen H, Lundberg SM, Erion G, Kim JH, Lee SI. Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. NPJ Digit Med 2021 Dec 8;4(1):167. [doi: 10.1038/s41746-021-00536-y] [Medline: 34880410]
17. ElMoaqet H, Tilbury DM, Ramachandran SK. Evaluating predictions of critical oxygen desaturation events. Physiol Meas 2014 Apr;35(4):639-655. [doi: 10.1088/0967-3334/35/4/639] [Medline: 24621948]
18. Erion G, Chen H, Lundberg SM, Lee SI. Anesthesiologist-level forecasting of hypoxemia with only SpO2 data using deep learning. arXiv. Preprint posted online on Dec 2, 2017. [doi: 10.48550/arXiv.1712.00563]
19. Lam C, Thapa R, Maharjan J, et al. Multitask learning with recurrent neural networks for acute respiratory distress syndrome prediction using only electronic health record data: model development and validation study. JMIR Med Inform 2022 Jun 15;10(6):e36202. [doi: 10.2196/36202] [Medline: 35704370]
20. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct;2(10):749-760. [doi: 10.1038/s41551-018-0304-0] [Medline: 31001455]
21. Ren S, Zupetic JA, Tabary M, et al. Machine learning based algorithms to impute PaO2 from SpO2 values and development of an online calculator. Sci Rep 2022 May 17;12(1):8235. [doi: 10.1038/s41598-022-12419-7] [Medline: 35581469]
22. Xia M, Jin C, Cao S, et al. Development and validation of a machine-learning model for prediction of hypoxemia after extubation in intensive care units. Ann Transl Med 2022 May;10(10):577. [doi: 10.21037/atm-22-2118] [Medline: 35722375]
23. Geng W, Jia D, Wang Y, et al. A prediction model for hypoxemia during routine sedation for gastrointestinal endoscopy. Clinics (Sao Paulo) 2018 Nov 14;73:e513. [doi: 10.6061/clinics/2018/e513] [Medline: 30462756]
24. Geng W, Tang H, Sharma A, Zhao Y, Yan Y, Hong W. An artificial neural network model for prediction of hypoxemia during sedation for gastrointestinal endoscopy. J Int Med Res 2019 May;47(5):2097-2103. [doi: 10.1177/0300060519834459] [Medline: 30913936]
25. Sippl P, Ganslandt T, Prokosch HU, Muenster T, Toddenroth D. Machine learning models of post-intubation hypoxia during general anesthesia. Stud Health Technol Inform 2017;243:212-216. [doi: 10.3233/978-1-61499-808-2-212] [Medline: 28883203]
26. Statsenko Y, Habuza T, Talako T, et al. Deep learning-based automatic assessment of lung impairment in COVID-19 pneumonia: predicting markers of hypoxia with computer vision. Front Med (Lausanne) 2022 Jul 9;9:882190. [doi: 10.3389/fmed.2022.882190] [Medline: 35957860]
27. Simpao AF, Rehman MA. Anesthesia information management systems. Anesth Analg 2018 Jul;127(1):90-94. [doi: 10.1213/ANE.0000000000002545] [Medline: 29049075]
28. Shah NJ, Tremper KK, Kheterpal S. Anatomy of an anesthesia information management system. Anesthesiol Clin 2011 Sep;29(3):355-365. [doi: 10.1016/j.anclin.2011.05.013] [Medline: 21871398]
29. Weitz JI, Fredenburgh JC, Eikelboom JW. A test in context: D-dimer. J Am Coll Cardiol 2017 Nov 7;70(19):2411-2420. [doi: 10.1016/j.jacc.2017.09.024] [Medline: 29096812]
30. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. J Big Data 2020 Jul 23;7:52. [doi: 10.1186/s40537-020-00327-4]
31. Zhao QY, Wang H, Luo JC, et al. Development and validation of a machine-learning model for prediction of extubation failure in intensive care units. Front Med (Lausanne) 2021 May 17;8:676343. [doi: 10.3389/fmed.2021.676343] [Medline: 34079812]
32. Shashikumar SP, Wardi G, Paul P, et al. Development and prospective validation of a deep learning algorithm for predicting need for mechanical ventilation. Chest 2021 Jun;159(6):2264-2273. [doi: 10.1016/j.chest.2020.12.009] [Medline: 33345948]
33. Igarashi Y, Ogawa K, Nishimura K, Osawa S, Ohwada H, Yokobori S. Machine learning for predicting successful extubation in patients receiving mechanical ventilation. Front Med (Lausanne) 2022 Aug 11;9:961252. [doi: 10.3389/fmed.2022.961252] [Medline: 36035403]
34. Fang AHS, Lim WT, Balakrishnan T. Early warning score validation methodologies and performance metrics: a systematic review. BMC Med Inform Decis Mak 2020 Jun 18;20(1):111. [doi: 10.1186/s12911-020-01144-8] [Medline: 32552702]
35. Romero-Brufau S, Whitford D, Johnson MG, et al. Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS). J Am Med Inform Assoc 2021 Jun 12;28(6):1207-1215. [doi: 10.1093/jamia/ocaa347] [Medline: 33638343]
36. Winslow CJ, Edelson DP, Churpek MM, et al. The impact of a machine learning early warning score on hospital mortality: a multicenter clinical intervention trial. Crit Care Med 2022 Sep 1;50(9):1339-1347. [doi: 10.1097/CCM.0000000000005492] [Medline: 35452010]

Abbreviations

AI: artificial intelligence
AIMS: anesthesia information management system
AUROC: area under the receiver operating characteristics
ICU: intensive care unit
LSTM: long short-term memory
ML: machine learning

NPV: negative predictive value
OR: operating room
PaO_2: partial pressure of oxygen
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction Model Risk of Bias Assessment Tool
PROSPERO: International Prospective Register of Systematic Reviews
SpO_2: peripheral oxygen saturation

Edited by C Lovis; submitted 17.07.23; peer-reviewed by C Price, J Maharjan; revised version received 02.11.23; accepted 05.11.23; published 02.02.24.

Please cite as:
Pigat L, Geisler BP, Sheikhalishahi S, Sander J, Kaspar M, Schmutz M, Rohr SO, Wild CM, Goss S, Zaghdoudi S, Hinske LC
Predicting Hypoxia Using Machine Learning: Systematic Review
JMIR Med Inform 2024;12:e50642
URL: https://medinform.jmir.org/2024/1/e50642
doi:10.2196/50642

Review

Multicriteria Decision-Making in Diabetes Management and Decision Support: Systematic Review

Tahmineh Aldaghi[1*], MSc; Jan Muzik[2*], PhD

[1]Spin-off Companies and Research Results Commercialization Center, First Faculty of Medicine, Charles University, Prague, Czech Republic

[2]Department of Information and Communication Technologies in Medicine, Faculty of Biomedical Engineering, Czech Technical University, Prague, Czech Republic

[*]all authors contributed equally

Corresponding Author:
Jan Muzik, PhD
Department of Information and Communication Technologies in Medicine
Faculty of Biomedical Engineering
Czech Technical University
Studničkova 7
Prague, 128 00
Czech Republic
Phone: 420 777568945
Email: jan.muzik@cvut.cz

Abstract

Background: Diabetes mellitus prevalence is increasing among adults and children around the world. Diabetes care is complex; examining the diet, type of medication, diabetes recognition, and willingness to use self-management tools are just a few of the challenges faced by diabetes clinicians who should make decisions about them. Making the appropriate decisions will reduce the cost of treatment, decrease the mortality rate of diabetes, and improve the life quality of patients with diabetes. Effective decision-making is within the realm of multicriteria decision-making (MCDM) techniques.

Objective: The central objective of this study is to evaluate the effectiveness and applicability of MCDM methods and then introduce a novel categorization framework for their use in this field.

Methods: The literature search was focused on publications from 2003 to 2023. Finally, by applying the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method, 63 articles were selected and examined.

Results: The findings reveal that the use of MCDM methods in diabetes research can be categorized into 6 distinct groups: the selection of diabetes medications (19 publications), diabetes diagnosis (12 publications), meal recommendations (8 publications), diabetes management (14 publications), diabetes complication (7 publications), and estimation of diabetes prevalence (3 publications).

Conclusions: Our review showed a significant portion of the MCDM literature on diabetes. The research highlights the benefits of using MCDM techniques, which are practical and effective for a variety of diabetes challenges.

(JMIR Med Inform 2024;12:e47701) doi:10.2196/47701

KEYWORDS

analytical hierarchy process; diabetes management; diabetes recognition; glucose management; multi-criteria decision making; technique for order of preference by similarity to ideal solution; decision support; diabetes; diabetic; glucose; blood sugar; review methodology; systematic review; decision making; self-management; digital health tool

Introduction

Overview

Diabetes mellitus is a chronic disease that is characterized by impaired insulin production and action [1]. According to the etiopathology of diabetes, the 3 most common clinical categories are distinguished: type 1 diabetes, type 2 diabetes (T2D), and gestational diabetes mellitus [2,3]. In recent decades, diabetes prevalence has increased in both adults and children around the world. By 2035, there will be an estimated 592 million people worldwide with diabetes [4]. By 2040, this number is expected to rise to 642 million [5], and by 2045, there will be 783.2 million cases of diabetes worldwide [2]. According to the global

2021 findings of the International Diabetes Federation (IDF), 537 million adults are living with diabetes, and 3 in 4 of them reside in low- and middle-income countries. In 2021, a total of 6.7 million people died of diabetes, equating to 1 death every 5 seconds. The expenditure on diabetes-related health care is at least US $966 billion, and it has increased up to 316% over the last 15 years [2].

Diabetes is a chronic condition requiring continuous medical care and patient education to prevent severe complications and long-term risks. Managing diabetes involves addressing various aspects of the patient's health, including blood glucose monitoring, monitoring and managing carbohydrate intake, regular engagement in physical activity, and medication management. By understanding the disease's nuances and recognizing when it might become severe, people can take steps to protect their well-being. Thus, faster diagnosis of diabetes and its potential complications is crucial for both patients and health care providers [6]. General practitioners faced a significant problem when diagnosing diabetes, partly because patients displayed a wide range of signs and symptoms. This complex clinical environment confused general practitioners and changed the diagnostic procedure into a multiobjective health care decision-making challenge [7].

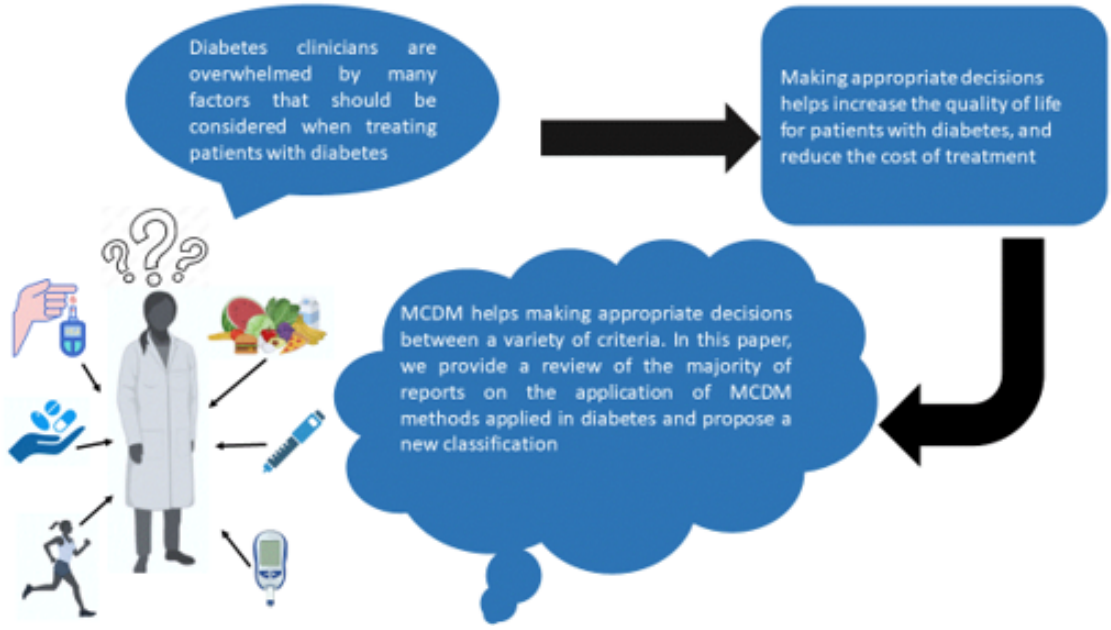
In addition to making informed decisions about the patient's health, endocrinologists and general practitioners should carefully assess various factors, including lifestyle choices, dietary habits, daily physical activity levels, insulin requirements, and the patient's willingness to embrace self-management technologies such as insulin pumps or pens, smart bracelets, continuous glucose monitoring, and mobile apps [8]. This comprehensive evaluation enables them to select the most appropriate treatment options. As an illustration, when it comes to managing hyperglycemia in patients with T2D, there is a diverse array of treatment options available. Currently, approximately 30 medications belonging to 9 distinct therapeutic categories have received approval for use, with ongoing research and development efforts yielding additional drugs and novel drug categories [9]. Due to the variety of options and guidelines from organizations such as the American Diabetes Association (ADA) [10], doctors often customize prescriptions using different doses and combinations for effective diabetes management [9]. The available medications vary in efficacy, safety, dosage, side effects, and cost. A lack of comparative information across these factors often leaves patients and physicians unable to make well-informed decisions [11]. The selection of diabetes medication presents itself as a multiobjective problem within the realm of health care decision-making [9].

Medical decision support could play a pivotal role in enhancing health care decision-making as it integrates pertinent, organized clinical knowledge and patient data into health-related decisions and processes [12]. Multiple stakeholders, including patients, health care providers, and those involved in patient care, can receive a mix of general clinical insights, patient-specific data, or both. Therefore, a quantitative approach that combines treatment benefits and drawbacks with individual preferences to effectively guide medical decisions could be multicriteria decision-making (MCDM) [13]. MCDM or multicriteria decision analysis (MCDA) is a valuable subdiscipline of operations research, particularly beneficial when dealing with multiple objectives, such as treatment-related outcomes, in benefit-risk analysis [14,15]. A typical MCDM problem consists of 4 key phases: option formulation, criteria selection, criteria weighting, and the decision-making process [16].

Objective

By considering the abovementioned factors, the primary aim of this research is to assess the use and practicality of MCDM methods in the context of diabetes. Our goal is to examine the various ways in which MCDM techniques have been used to study diabetes and present an innovative categorization of their applications in this field. Figure 1 demonstrates the graphical abstract of the paper.

Figure 1. Graphical abstract of the paper. MCDM: multicriteria decision-making.



Methods

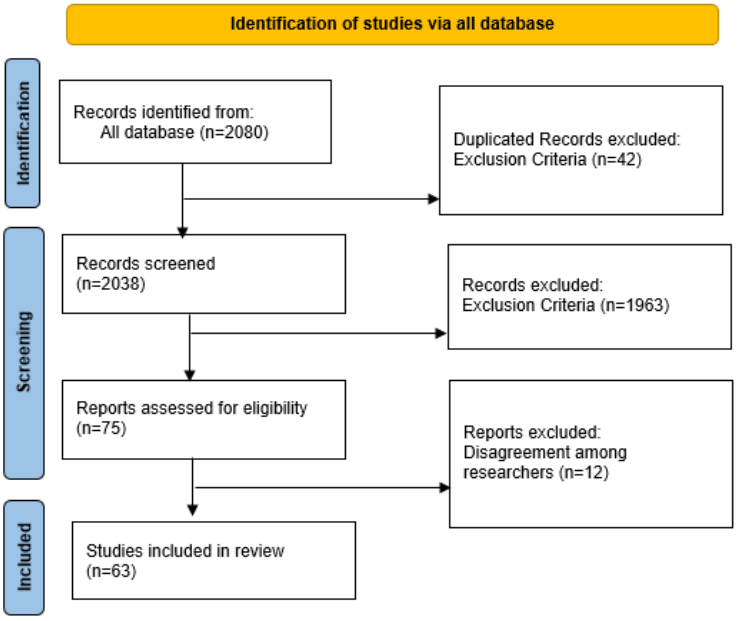
Search Strategies

A query was carried out on PubMed, Elsevier, Embase, MEDLINE, Scopus, MBC, Springer, IEEE, MDPI, Taylor and Francis Online, and Google Scholar based on published articles. The keywords for our paper were extracted from Medical Subject Headings (MeSH). The keywords “diabetes” and “glucose” were combined with MCDM techniques terms such as TOPSIS, AHP, and multi-criteria-decision-making using the Boolean operator AND/OR. The specific query searched was: ((diabetes OR glucose) AND (AHP OR TOPSIS OR MCDM OR multi-criteria-decision-making)).

Inclusion and Exclusion Criteria

We initially eliminated any duplicate articles from various sources after receiving the results of an initial collection of relevant articles and then manually inspected the remaining articles to assess them under the inclusion criteria. The inclusion criteria were any English papers published between 2003 and 2023. Research, review, conference, and case report articles with an abstract or full text were taken into account. Non-English articles and other research forms, such as letters to editors and brief messages, were excluded. Out of almost 2210 articles, only 63 were found and chosen based on keywords and all of our criteria. The article selection process was based on PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses; Figure 2) [17].

Figure 2. PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) flowchart.



Results

Overview

Based on Figure 2, after removing duplicates and examining according to the inclusion and exclusion criteria, 63 publications were included in the final evaluation. Based on our investigation to reveal the frequency of publications in databases, it became clear that most of the publications were indexed in Google Scholar, with 60 publications; PubMed, with 17 publications; and Springer and IEEE, with 8 and 7 publications, respectively.

We initially provided a concise overview of MCDM and its techniques, followed by the presentation of our research findings gathered from reviewing publications.

MCDM Techniques Overview

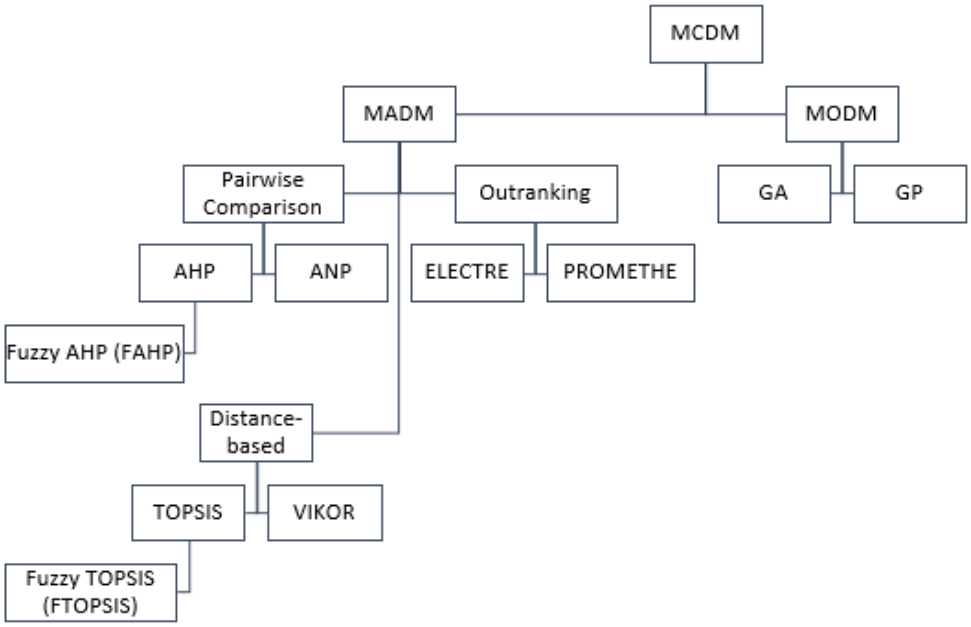
Since so many choices in our modern lives depend on a multitude of factors, the decision can be made by giving various criteria varying weights, which is done by expert groups. Determining the structure and explicitly evaluating several criteria is crucial. Therefore, constructing and resolving multicriteria planning and decision-making challenges is referred to as MCDM. As a result, MCDM is composed of a set of numerous criteria, a set of alternatives, and some sort of comparison between them [18-20].

No alternative optimizes all criteria uniformly in multicriteria optimization assignments. Any solution to the multicriteria task that enhances a specific criterion can be examined, but the task must ultimately have a preferred option. The decision maker must provide more details to select the best decision. Throughout its brief history of about 50 years, MCDM has been an interesting study topic [20]. There are 2 categories of MCDM approaches: multiattribute decision-making (MADM) and multiobjective decision-making (MODM) [19,20].

In order to find the optimal answer, decision makers in MADM choose to categorize, rank, or prioritize a limited number of choices. Pairwise comparison, outranking, and distance-based approaches are the 3 basic methods used in MADM. Pairwise comparison involves evaluating and contrasting the weights of several criteria using a base scale. Analytic hierarchy process (AHP) and analytical network process (ANP) are frequently used in pairwise comparison [21]. Outranking approaches offer a variety of options and determine whether one option has any sort of dominance over the others [22]; instances of outranking techniques include Elimination Et Choix Traduisant la Realité

(ELECTRE) and preference ranking organization method for enrichment of evaluations (PROMETHEE) [21]. The solution with the shortest distance to the ideal point is considered the best according to distance-based techniques, which measure the distance a solution is from the ideal point. The technique for order of preference by similarity to ideal solution (TOPSIS) and ViseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) are 2 popular distance-based methodologies [21]. Unlike MADM, MODM handles situations where there are many decision makers and an infinite number of possibilities. All of these MCDM methods are presented in Figure 3. The most efficient MCDM techniques are introduced in the following sections.

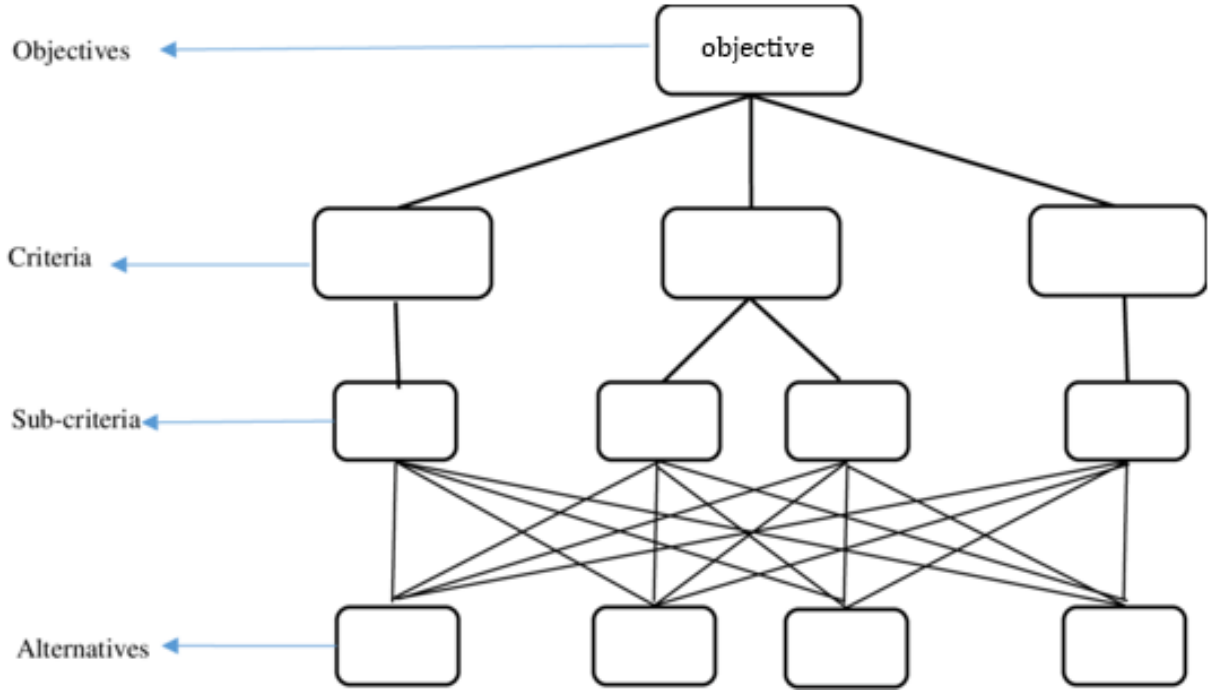
Figure 3. Hierarchical structures of MCDM methods. AHP: analytic hierarchy process; ANP: analytical network process; ELECTRE: Elimination Et Choix Traduisant la Realité; GA: genetic algorithm; GP: goal programming; MADM: multiattribute decision-making; MCDM: multicriteria decision-making; MODM: multiobjective decision-making; PROMETHEE: preference ranking organization method for enrichment of evaluations; TOPSIS: technique for order of preference by similarity to ideal solution; VIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje.



AHP Method

Saaty [23] was the first to introduce the AHP. As shown in Figure 4, AHP includes the decision's objective at the top, the criteria and subcriteria in the middle, and the collection of alternatives at the bottom [7]. The key benefits of AHP are its scalability and ease of usage. AHP can be applied using Excel (Microsoft) or web-based tools such as Transparent Choice, SpiceLogic, Decerns MCDA, MATLAB (MathWorks), R (R Core Team), and Super Decisions.

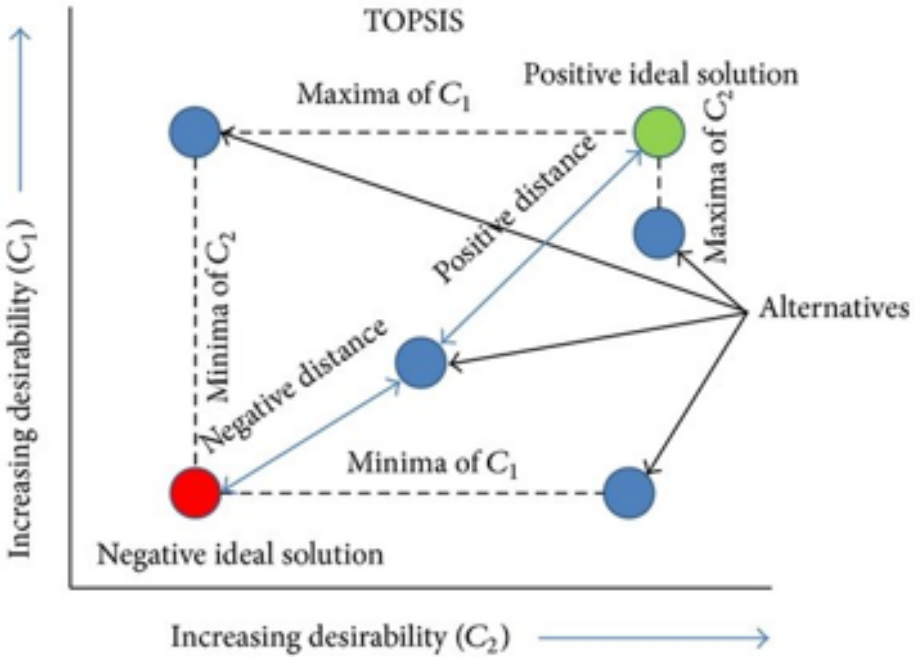
Figure 4. Hierarchical structure of analytic hierarchy process.



TOPSIS Method

As shown in Figure 5, TOPSIS is a distance-based technique that Hwang and Yoon [24] proposed in 1981. The TOPSIS technique makes it easy to define the positive and negative ideal solutions by presuming that each criterion tends to monotonically increase or reduce use. A Euclidean distance approach is suggested to assess how closely the alternatives resemble the ideal solution. The preferred order of the alternatives will be determined by a series of comparisons of their relative distances. The general principle behind this approach is that the optimal option should be closest to the ideal solution and the farthest distance from the negative ideal solution. In the ideal solution, the ideal solution has the best attribute values, maximizes the benefit criteria, and minimizes the cost criteria. In the negative ideal solution, the negative solution has the worst attribute values, maximizes the cost criteria, and minimizes the benefit criteria [19,21].

Figure 5. TOPSIS method. TOPSIS: technique for order of preference by similarity to ideal solution.



ANP Method

Due to the inability of AHP to produce an adequate rating with a limited number of possibilities, the majority of organizations do not use it often. Therefore, Saaty [25] suggested ANP as a continuation of AHP. Decision makers are capable of making

decisions in difficult situations, according to ANP's capability [21].

Weighting Methods

One of the crucial phases of MCDM problems is determining the weights of the criterion [26]. Several weighing techniques can be divided into the following groups: (1) subjective weighting method: AHP, Weighted Sum Model (WSM) [27], and Weighted Product Model (WPM) [27]; (2) objective weighting method: Entropy method [28] and Criteria Importance Through Intercriteria Correlation (CRITIC) [28]; and (3) integrated method: step-wise weigh assessment ratio analysis (SWARA) [29] and Weighted Aggregated Sum Product Assessment (WASPAS) [28].

Following a thorough analysis of all of the MCDM publications in the field of diabetes research during a 2-decade period, it was evident that, starting in 2016, the number of publications in this area has been steadily rising, reaching 10 in 2022.

Then, a new classification of the applications of MCDM approaches in diabetes was proposed: (1) selection of diabetes medication, (2) diagnosis of diabetes, (3) meal recommendation for diabetes, (4) diabetes management, (5) diabetes complication, and (6) estimation of diabetes prevalence.

Selection of Diabetes Medication

Table 1 shows that approximately 30% (n=19/63) of the publications focused on using MCDM techniques to determine the optimal diabetes medication among various options. Notably, AHP and fuzzy AHP, with 6 and 4 mentions, respectively, were the most frequently used methods.

Table 1. Diabetes medication publications.

Reference	Methods	Objective	Results
Maruthur et al [14]	AHP[a]	Select oral T2D[b] medications	Sitagliptin, sulfonylureas, and pioglitazone
Eghbali-Zarch et al [29]	SWARA[c] method, ratio analysis, and the FMULTIMOORA[d] method	Choose the pharmacological treatment for T2D	Metformin should be used as the first-line medication, followed by sulfonylurea, glucagon-like peptide-1 receptor agonist, dipeptidyl peptidase-4 inhibitor, and insulin
Eghbali-Zarch et al [28]	WASPAS[e], entropy, and CRITIC[f]	Determine the final ranking of the medications	Proposed a model to help endocrinologist to choose the best medicine
Zhang et al [30]	TOPSIS[g]	Ranking of diabetes medicines	CDSS[h] can assist young doctors and nonspecialty physicians with medication prescriptions
Maruthur et al [31]	AHP	Select oral T2D medications	AHP will aid, support, and enhance the ability of decision makers to make evidence-based informed decisions consistent with their values and preferences
Nag and Helal [32]	Fuzzy AHP and AHP	Classification of diabetic medications	Fuzzy AHP model can better handle the ambiguity of decision makers
Chen et al [33]	Entropy	Choose pharmaceuticals	AGI[i], DPP4[j], MET[k], Glinide, SU[l], and TZD[m]
Wang et al [34]	AHP and ANP[n]	Combine different clinical, economic, and medical decision-making elements	Modifying one's lifestyle, taking metformin, and receiving insulin injections
Bao et al [35]	MCDA[o]	Assess medicine for diabetes	Five DPP4 inhibitors was valuable
Onar and Ibil [36]	Fuzzy AHP	Considered the best oral antidiabetic	Proposed a decision support system
Zhang et al [37]	MCDA	Examine the Mudan Granules	The new medication was acceptable
Cai et al [38]	AHP	Evaluate strains of the efficacy of the LAB[p] with possible antidiabetic capabilities	Potential antidiabetic effect
Sekar et al [39]	Fuzzy PROMETHEE[q]	Choose the best course of therapy	Giving the high peace of treatment to the most affected people
Mühlbacher et al [40]	AHP and BWS[r]	Evaluate patients' preferences for various T2D treatment parameters	Proposed a model
Mahat and Ahmad [41]	Fuzzy AHP	Identify and choose the most efficient thermal massage treatment session	Number of therapy sessions (per day) was the most important factor
Pan et al [42]	Fuzzy AHP	Determine the weights of the various physiological factors	The mathematical model of exercise rehabilitation program for patients with diabetes was established
Rani et al [43]	COPRAS[s]	Select T2D medication treatment	Developed a new formula-based PFSs[t] and evaluated its feasibility by applying the model on selecting the T2D pharmacological therapy
Balubaid and Basheikh [44]	AHP	Developed a mathematical decision-making model that prioritizes the available diabetes medication based on criteria	Metformin, pioglitazone, sitagliptin, and glimepiride were ranked first, second, third, and fourth, respectively
Mühlbacher et al [45]	AHP and BWS	Examine the key patient-related decision criteria involved in the medicinal treatment of T2D	For oral antidiabetes-treated patient groups and insulin-treated patient groups, HbA1c[u] level, delay of insulin therapy, and occurrence of hypoglycemia were ranked first, second, and third, respectively

[a]AHP: analytic hierarchy process.

[b]T2D: type 2 diabetes.

[c]SWARA: step-wise weigh assessment ratio analysis.

[d]FMULTIMOORA: full multiplicative form.

[e]WASPAS: Weighted Aggregated Sum Product Assessment.

[f]CRITIC: Criteria Importance Through Intercriteria Correlation.

[g]TOPSIS: technique for order of preference by similarity to ideal solution.

[h]CDSS: clinical decision support system.
[i]AGI: α-glucosidase.
[j]DDP4: dipeptidyl peptidase-4.
[k]MET: meglitinide.
[l]SU: sulfonylureas.
[m]TZD: thiazolidinedione.
[n]ANP: analytical network process.
[o]MCDA: multicriteria decision analysis.
[p]LAB: lactic acid bacteria.
[q]PROMETHEE: preference ranking organization method for enrichment of evaluations.
[r]BWS: best–worst-scaling.
[s]COPRAS: Complex Proportional Assessment.
[t]PFS: Pythagorean Fuzzy Set.
[u]HbA1c: hemoglobin A1c.

Diagnosis of Diabetes

Table 2 displays that roughly 19% (12/63) of the publications centered on the application of MCDM techniques for aiding general practitioners and endocrinologists in diagnosing diabetes. Among these, AHP and TOPSIS were the most commonly cited methods, with 4 and 3 mentions, respectively.

Table 2. Diabetes diagnosis publications.

Reference	Methods	Objective	Risk factors	Results
Zulqarnain et al [6]	TOPSIS[a]	Investigate the prevalence of diabetes among women and men	Age, weight, height, BMI, systolic and diastolic BP[b], urine creatinine, albuminuria, and ACR[c]	Female patients were more likely to develop diabetes
Abdulkareem et al [7]	Fuzzy AHP[d]	Predict diabetes risks	Weakness, obesity, delayed healing, alopecia, muscle stiffness, polydipsia, polyuria, visual blurring, sudden weight loss, and itching	FAHP[e] model is an excellent tool for diagnosing medical disorders based on many criteria
Abbasi et al [46]	AHP	Identify the most significant risk factors for GDM[f]	A history of GDM or impaired glucose tolerance in previous pregnancies and a history of macrosomia in the infant	N/A[g]
Yas et al [47]	Fuzzy TOPSIS	Identify the symptoms of diabetes	Age, pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, and diabetes pedigree function	Proposed a framework to recognize the symptoms of disease
Amin-Naseri and Neshat [48]	AHP	Determine the likelihood of developing T2D[h]	FBS[i] index, PRF[j], BMI, diet, age, BP, gender, family history, and smoking status	DIBAR[k], a knowledge-based expert system
El-Sappagh et al [49]	Fuzzy AHP	Diagnosis of diabetes	N/A	Created a new, systematically interpretable FRBS[l] framework
Baha et al [50]	AHP	Diagnosis of diabetes	Heredity, sex, ethnicity, age, impaired glucose tolerance, gestational diabetes, and so forth	Recognized top 3 most important risk factors: heredity, obesity, and physical inactivity
Sharma and Sharma [51]	EDAS[m]	Forecast diabetes	N/A	Combined MCDM[n] with machine-learning techniques to find the best forecasting model
Malapane et al [52]	WPM[o]	Forecast diabetes	N/A	Combined WPM method with machine learning to select the best model
Felix et al [53]	TOPSIS	Identification of the most important T2D risk factors in the Pima Indian database	Blood glucose, BP, blood cholesterol, obesity, blindness, physical inactivity	Blindness, obesity, and inactivity were the risk factors with greatest impact
Sankar and Jeyaraj [54]	AHP	Forecast diabetes in women	N/A	Propose a model for predicting diabetes among women
Bondor and Mureşan [55]	TOPSIS	Solve the problem of multicollinearity between criteria in diabetes diagnosis	N/A	Proposed a new algorithm which removed the multicollinearity among criteria

[a]TOPSIS: technique for order of preference by similarity to ideal solution.

[b]BP: blood pressure.

[c]ACR: albumin creatinine ratio.

[d]AHP: analytic hierarchy process.

[e]FAHP: fuzzy analytic hierarchy process.

[f]GDM: gestational diabetes mellitus.

[g]N/A: not applicable.

[h]T2D: type 2 diabetes.

[i]FBS: fasting blood sugar.

[j]PRF: physical risk factors.

[k]DIBAR: Created Diabetes Risk Assessment.

[l]FRBS: fuzzy rule-based systems.

[m]EDAS: evaluation based on distance for average solution.

[n]MCDM: multicriteria decision-making.

[o]WPM: Weighted Product Model.

Meal Recommendation for Diabetes

According to Table 3, a total of 8 (13%) out of 63 publications focused on using MCDM techniques to assist people with diabetes in making the healthiest food choices from their food options, considering factors such as fat content, carbohydrate content, and calorie count. Among these, AHP was mentioned most frequently, with 6 instances.

Table 3. Meal recommendation publications.

Reference	Methods	Objective	Criteria	Results
Gaikwad et al [56]	AHP[a]	Recommend a particular ice cream for patients with diabetes	Sugar, cholesterol, dietary fiber, and proteins	Ben & Jerry's Butter Pecan was enriched with all 4 criteria
Sharawat and Dubey [57]	AHP	Find out the best diet for a patient with diabetes among 3 alternatives: solid food, liquid food, and fluid food	Calories, body fat, healthy carbs, and dietary needs	Solid food was selected as the best
Santoso et al [58]	Fuzzy AHP	Designed a new yogurt product for patients with diabetes	N/A[b]	N/A
Zadeh et al [59]	AHP	Proposed a personalized meal-planning strategy	N/A	Proposed an affordable and culturally appropriate meals that would provide all the nutrition needed for a diabetic while still being mindful of calories and carbs
Gulint and Kadam [60]	AHP and TOPSIS[c]	Recommended shakes and ice cream for patients with diabetes	Sugar, cholesterol, carbs, fat, protein, and dietary fiber	Selected a type of ice cream that satisfies all criteria
Gaikwad et al [61]	ANP[d]	Recommendation of a particular ice cream	Sugar, calories, cholesterol, and proteins	Selected a type of ice cream that satisfies all criteria
Gaikwad et al [62]	AHP	Recommendation of a particular ice cream	N/A	Proposed a model combination of AHP-GA[e] and AHP-CI[f] to recommend an ice cream to patients with diabetes
Gaikwad et al [63]	AHP	Recommendation of a particular ice cream	Sugar, protein, cholesterol, and dietary fiber	Patient having a high sugar level of 262 mg/dl can consume an ice cream lower sugar like Breyers butter almond, also patient with low sugar level of 77 mg/dl can consume high sugar ice cream like Breyers

[a]AHP: analytic hierarchy process.

[b]N/A: not applicable.

[c]TOPSIS: technique for order of preference by similarity to ideal solution.

[d]ANP: analytical network process.

[e]AHP-CI: analytic hierarchy process–cohort intelligence.

[f]AHP-GA: analytic hierarchy process–genetic algorithm.

Diabetes Management

Based on Table 4, additional applications of MCDM techniques, particularly AHP methods, in diabetes management (14/63, 22%) encompass tasks such as identifying ideal locations for diabetes clinics, allocating resources for diabetes care, assessing the current diabetes applications, and constructing models to prioritize criteria that bolster the safety of the insulin supply chain.

Table 4. Diabetes management publications.

Reference	Method	Results
Gupta et al [64]	TOPSIS[a], VIKOR[b], PROMETHEE II[c]	Assess current mHealth[d] applications for T2D[e], including Glucose Buddy, mySugr, Diabetes: M, Blood Glucose Tracker, and OneTouch Reveal
Wang et al [65]	ANP[f] and CRITIC[g]	Assess the influence of social support on T2DM[h] self-management
Mishra et al [66]	AHP[i]	Created and used the SCP[j] assessment methodology for Indian diabetes clinic
Mishra [67]	AHP	Developed a customized service quality assessment model for diabetes care
Mishra [68]	Fuzzy TOPSIS	Proposed 3 alternatives for the placement of a diabetes clinic using the SLP[k] method
Byun et al [69]	AHP	Improving the treatment compliance of patients with diabetes
Mehrotra and Kim [70]	New multicriterion, robust weighted-sum methodology	Calculate the amount of funding allocated to diabetes preventive initiatives across the United States to reduce the weighted sum of diabetes prevalence and outcomes caused by improper health expenditure
Haji et al [71]	AHP and TOPSIS	Create a model that can prioritize and pick the optimal criterion for optimizing insulin safety
Suka et al [72]	AHP	Described a clinical decision support system that enhance dynamic decision-making
Fico et al [73]	AHP	Selected the best tool for screening and managing T2D
Long and Centor [74]	AHP	Assess the relative significance of 4 frequently used diabetes quality indicators: measuring HbA1c[l], measuring LDL[m], performing a dilated eye examination, and performing a foot examination
Gajdoš et al [75]	TOPSIS	Proposed a concept of chronic care management, which could increase effectiveness and reduce the cost of health care provided to patients with T2D
Gupta et al [76]	CODAS-FAHP[n] and MOORA-FAHP[o]	Assess the usability of mHealth applications to monitor T2D by developing 2 hybrid decision-making methods
Chang et al [77]	Delphi-AHP	Recommended a Delphi-AHP framework to establish agreement in creating a decision-making algorithm for evaluating the balance of benefits and risks associated with the use of complementary and alternative medicine for diabetes

[a]TOPSIS: technique for order of preference by similarity to ideal solution.

[b]VIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje.

[c]PROMETHEE II: preference ranking organization method for enrichment of evaluation II.

[d]mHealth: mobile health.

[e]T2D: type 2 diabetes.

[f]ANP: analytical network process.

[g]CRITIC: Criteria Importance Through Intercriteria Correlation

[h]T2DM: type 2 diabetes mellitus.

[i]AHP: analytic hierarchy process.

[j]SCP: Supply Chain Partnership.

[k]SLP: Systematic Layout Planning.

[l]HbA1c: hemoglobin A1c.

[m]LDL: low-density lipoprotein.

[n]CODAS-FAHP: combine distance-based assessment-fuzzy AHP.

[o]MOORA-FAHP: multiobjective optimization on the basis of ratio analysis-fuzzy AHP.

Diabetes Complication

T2D is a significant global public health issue, characterized by 2 categories of harm: macrovascular (involving large arteries) and microvascular (involving small blood vessels). Macrovascular disease such as strokes and microvascular diseases such as retinopathy, nephropathy, and neuropathy [7]. MCDM techniques, especially TOPSIS, as shown in Table 5, are used to assist endocrinologists and general practitioners in analyzing the severity of these complications, forecasting their likelihood of occurrence, and pinpointing the risk factors for them (n=7).

Table 5. Diabetes complication diagnosis publications.

Reference	Methods	Objective	Criteria	Complications	Results
Ebrahimi and Ahmadi [78]	Fuzzy TOPSIS[a]	Analyzed the severity caused by diabetes	High cholesterol, high BP[b], obesity, physical inactivity, smoking, family history, age, and sex	Neuropathy, diabetic retinopathy, cardiovascular disease, kidney disease, foot ulcer, and amputation	Cardiovascular disease was the most important complication in the problem
Ahmadi and Ebrahimi [79]	MCDM[c]	Assessed the severity of difficulties caused by diabetes	Ischemic heart disease, heart failure, heart stroke, ketoacidosis, diabetic ulcer, neuropathy, and lower extremely amputation	Cardiovascular disease, diabetic ketoacidosis, lower extremity complications, and lower extremity amputation	Proposed a new hybrid algorithm that calculate the severity of damage caused by diabetes
Bondor et al [80]	TOPSIS	Identification of the risk factors in kidney disease	Urinary albumin per creatinine ratio and glomerular filtration	Diabetic kidney	Rank the risk factors of microalbuminuria and eGFR[d] to evaluate the risk factor for CKD[e]
Ahmed et al [81]	TOPSIS and entropy	Detection of DR[f] through machine learning and TOPSIS models	Criteria of TOPSIS model: AUC[g], accuracy, precision, F1-score, recall, TPR[h], FNR[i], FPR[j], TNR[k], and time	DR	According to TOPSIS, Adaboost model ranks at the best model to detect DR
Bondor et al [82]	VIKOR[l]	Rank risk factors of diabetic kidney disease	Serum adiponectin, triglycerides, SBP, duration of diabetes and age, Malondialdehyde, and HDL[m]-cholesterol	Diabetic kidney	Identification of diabetic kidney disease risk factors
Alassery et al [83]	Fuzzy AHP[n] and Fuzzy TOPSIS	Determine the impact of mental health in patients with diabetes	BMI, SBP, DBP[o], age, height, exercise	Mental health	The model showed the applicability and impact of mental health in patients with diabetes
Wang et al [84]	AHP	Relieve the pain in patients with diabetes	N/A[p]	Diabetic neuropathy and foot ulcers	Selection of shoe lasts for footwear design to help relieve the pain associated with diabetic neuropathy and foot ulcers

[a]TOPSIS: technique for order of preference by similarity to ideal solution.

[b]BP: blood pressure.

[c]MCDM: multicriteria decision-making.

[d]GFR: estimated glomerular filtration rate.

[e]CKD: chronic kidney disease.

[f]DR: diabetic retinopathy.

[g]AUC: area under the curve.

[h]TPR: true positive rate.

[i]FNR: false negative rate.

[j]FPR: false positive rate.

[k]TNR: true negative rate.

[l]VIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje.

[m]HDL: high-density lipoprotein.

[n]AHP: analytic hierarchy process.

[o]DBP: diastolic blood pressure.

[p]N/A: not applicable.

Discussion

Principal Findings

Given the multitude of choices involved in selecting diabetes medication, meal planning, nutrient intake, diabetes management apps, and speedy diagnosis, endocrinologists, general practitioners, and individuals with diabetes, along with their caregivers, need guidance to make informed decisions. MCDM is a quantitative approach that effectively integrates treatment benefits and drawbacks, as well as individual preferences, to facilitate sound medical decision-making in these complex situations. Consequently, we embarked on an evaluation of the effectiveness of MCDM methods in the context of diabetes.

Based on a notable upward trend in publications within the realm of using MCDM methods in diabetes research over the last 2 decades, this underscores the growing interest among researchers in applying MCDM methods to address diabetes-related challenges. Furthermore, the majority of these publications (n=19) focus on diabetes treatment selection [14,28-45]. Diabetes management (n=14), diagnosis of diabetes (n=12), meal recommendation (n=8), diabetes complications (n=7), and global estimation (n=3) are in the later ranks. This outcome highlights the efficacy of using MCDM methods in the process of choosing diabetes medications.

All MCDM methods in diabetes are classified into 13 groups. AHP is ranked first, having been used in 25 articles. AHP is designed to help individuals and groups make complex decisions by breaking them into a hierarchical structure, comparing and weighting criteria and alternatives, and deriving a rational choice based on these comparisons [7,85,24]. AHP can be applied to diabetes issues and decision-making in several ways including treatment selection [14,31,32,34,36,38-42,44,45], diabetes diagnosis [46,48-50,54], dietary planning [56-60,62,63], diabetes management [66,67,69,71-74,77], complication diagnosis [84], and estimating diabetes prevalence [4,5]. TOPSIS and fuzzy AHP with 9 and 8 publications are in the next ranks, respectively.

As observed, 6 distinct weighting algorithms were recognized, with the Entropy approach ranking highest. The final component in our proposed classification pertains to estimating diabetes prevalence. In a 2013 study, researchers used logistic regression and AHP techniques to produce smoothed age-specific occurrence estimates for adults aged 20 to 79 years. These estimates were then used to calculate population projections for the years 2013 and 2035, foreseeing an increase in the number of individuals with diabetes to 592 million by 2035 [4]. In another investigation conducted by the IDF in 2015, AHP and logistic regression methods were used to estimate that there were 415 million people (ranging from 340 million to 536 million) with diabetes. Projections indicate that this figure is expected to reach 642 million (ranging from 521 million to 829 million) by 2040 [5].

Conclusions

One of the most serious health problems of the 21st century, whose prevalence is rapidly increasing, is diabetes mellitus. Almost all areas of diabetes research have seen significant progress to date, particularly in the areas of medication selection, meal selection, diabetes management applications, use of continuous glucose monitoring, and closed-loop system. The advancement of technology has expanded the scope of decision-making responsibilities for general practitioners in the initial stages of patient care. Determining the most optimal choice among numerous options falls within the domain of MCDM.

In this research, for the first time, we reviewed the majority of MCDM papers for diabetes and considered 2 important issues in the field of diabetes: examining the usability of MCDM techniques in diabetes and proposing a new classification of applications of MCDM methods in diabetes. Our study highlights that the use of MCDM techniques extends beyond the realm of diabetes medication selection. These methods hold promise for diverse applications, spanning meal planning, diabetes diagnosis, and addressing diabetes-related challenges. This includes tasks such as selecting optimal diabetes management applications from a wide range of options, identifying ideal locations for diabetes clinics, and efficiently allocating resources for diabetes care. Moreover, the analysis reveals that AHP is the preferred and widely embraced strategy and approach, primarily owing to its straightforward structure and user-friendliness. We firmly believe that the adoption of MCDM approaches offers advantages to a broad spectrum of stakeholders, including patients with diabetes, endocrinologists, general practitioners, caregivers, and health care policy makers. These techniques have the potential to serve as valuable tools for general practitioners, assisting in quicker diabetes diagnosis and more accurate medication selection, ultimately reducing patient costs and lifestyle concerns.

Acknowledgments

This research was supported by the project TN02000067—Future Electronics for Industry 4.0 and Medical 4.0 is cofinanced from the state budget by the Technology Agency of the Czech Republic under the National Centers of Competence: support programme for applied research, experimental development, and innovation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 35 KB - medinform_v12i1e47701_app1.docx]

References

1. Forouhi NG, Wareham NJ. Epidemiology of diabetes. Medicine 2010;38(11):602-606. [doi: 10.1016/j.mpmed.2010.08.007]
2. IDF diabetes atlas 2021—10th edition. International Diabetes Federation. URL: https://diabetesatlas.org/atlas/tenth-edition/ [accessed 2023-12-29]

3. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
4. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. Diabetes Res Clin Pract 2014;103(2):137-149 [FREE Full text] [doi: 10.1016/j.diabres.2013.11.002] [Medline: 24630390]
5. Ogurtsova K, da Rocha Fernandes JD, Huang Y, Linnenkamp U, Guariguata L, Cho NH, et al. IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Res Clin Pract 2017;128:40-50. [doi: 10.1016/j.diabres.2017.03.024] [Medline: 28437734]
6. Zulqarnain M, Dayan F, Saeed M. TOPSIS analysis for the prediction of diabetes based on general characteristics of humans. Int J Pharm Sci Res 2018;9(7):2932-2939 [FREE Full text] [doi: 10.13040/IJPSR.0975-8232.9(7).2932-2939]
7. Abdulkareem SA, Radhi HY, Fadil YA, Mahdi H. Soft computing techniques for early diabetes prediction. Indones J Electr Eng Comput Sci 2022;25(2):1167-1176 [FREE Full text] [doi: 10.11591/ijeecs.v25.i2.pp1167-1176]
8. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. J Med Internet Res 2018;20(5):e10775 [FREE Full text] [doi: 10.2196/10775] [Medline: 29848472]
9. Grant RW, Wexler DJ, Watson AJ, Lester WT, Cagliero E, Campbell EG, et al. How doctors choose medications to treat type 2 diabetes: a national survey of specialists and academic generalists. Diabetes Care 2007;30(6):1448-1453 [FREE Full text] [doi: 10.2337/dc06-2499] [Medline: 17337497]
10. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care 2014;37(Suppl 1):S81-S90 [FREE Full text] [doi: 10.2337/dc14-S081] [Medline: 24357215]
11. Montori VM. Selecting the right drug treatment for adults with type 2 diabetes. BMJ 2016;352:i1663. [doi: 10.1136/bmj.i1663] [Medline: 27029501]
12. Diabetes medication choice decision conversation aid. Welcome to the Diabetes Medication Choice Decision Conversation Aid. URL: https://diabetesdecisionaid.mayoclinic.org/index [accessed 2023-09-07]
13. Dolan JG. Multi-criteria clinical decision support: a primer on the use of multiple criteria decision making methods to promote evidence-based, patient-centered healthcare. Patient 2010;3(4):229-248 [FREE Full text] [doi: 10.2165/11539470-000000000-00000] [Medline: 21394218]
14. Maruthur NM, Joy SM, Dolan JG, Shihab HM, Singh S. Use of the analytic hierarchy process for medication decision-making in type 2 diabetes. PLoS One 2015;10(5):e0126625 [FREE Full text] [doi: 10.1371/journal.pone.0126625] [Medline: 26000636]
15. Peteiro-Barral D, Remeseiro B, Méndez R, Penedo MG. Evaluation of an automatic dry eye test using MCDM methods and rank correlation. Med Biol Eng Comput 2017;55(4):527-536. [doi: 10.1007/s11517-016-1534-5] [Medline: 27311605]
16. Adhikary P, Kundu S. MCDA or MCDM based selection of transmission line conductor: small hydropower project planning and development. Int J Eng Res Appl 2014;4(2):357-361 [FREE Full text]
17. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151(4):264-269, W64 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
18. Borissova D. An overview of multi-criteria decision making models and software systems. In: Atanassov KT, editor. Research in Computer Science in the Bulgarian Academy of Sciences. Cham, Switzerland: Springer International Publishing; 2021:305-323.
19. Aruldoss M, Lakshmi TM, Venkatesan VP. A survey on multi criteria decision making methods and its applications. Am J Inf Syst 2013;1(1):31-43 [FREE Full text] [doi: 10.12691/ajis-1-1-5]
20. Singh A, Malik SK. Major MCDM techniques and their application-a review. IOSR J Eng 2014;4(5):15-25 [FREE Full text] [doi: 10.9790/3021-04521525]
21. Azhar NA, Radzi NAM, Ahmad WSHMW. Multi-criteria decision making: a systematic review. Recent Adv Electr Electron Eng 2021;14(8):779-801 [FREE Full text] [doi: 10.2174/2352096514666211029112443]
22. Kangas J, Kangas A, Leskinen P, Pykäläinen J. MCDM methods in strategic planning of forestry on state‐owned lands in Finland: applications and experiences. Multi Criteria Decision Anal 2002;10(5):257-271. [doi: 10.1002/mcda.306]
23. Saaty TL. A scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology 1977 Jun;15(3):234-281. [doi: 10.1016/0022-2496(77)90033-5]
24. Hwang CL, Yoon K. Methods for multiple attribute decision making. In: Multiple Attribute Decision Making: Methods and Applications: A State-of-the-art Survey. Berlin Heidelberg: Springer; 1981:58-191.
25. Saaty TL. Decision Making with Dependence and Feedback: The Analytic Network Process. Pittsburgh: RWS publications; 1996.
26. Pamučar D, Stević Ž, Sremac S. A new model for determining weight coefficients of criteria in MCDM models: Full Consistency Method (FUCOM). Symmetry 2018;10(9):393 [FREE Full text] [doi: 10.3390/sym10090393]
27. Triantaphyllou E. Multi-criteria decision making methods. In: Multi-Criteria Decision Making Methods: A Comparative Study. Boston, MA: Springer US; 2000:5-21.

28. Eghbali-Zarch M, Tavakkoli-Moghaddam R, Esfahanian F, Masoud S. Prioritizing the glucose-lowering medicines for type 2 diabetes by an extended fuzzy decision-making approach with target-based attributes. Med Biol Eng Comput 2022;60(8):2423-2444. [doi: 10.1007/s11517-022-02602-3] [Medline: 35776373]
29. Eghbali-Zarch M, Tavakkoli-Moghaddam R, Esfahanian F, Sepehri MM, Azaron A. Pharmacological therapy selection of type 2 diabetes based on the SWARA and modified MULTIMOORA methods under a fuzzy environment. Artif Intell Med 2018;87:20-33. [doi: 10.1016/j.artmed.2018.03.003] [Medline: 29606521]
30. Zhang Y, McCoy RG, Mason JE, Smith SA, Shah ND, Denton BT. Second-line agents for glycemic control for type 2 diabetes: are newer agents better? Diabetes Care 2014;37(5):1338-1345 [FREE Full text] [doi: 10.2337/dc13-1901] [Medline: 24574345]
31. Maruthur NM, Joy S, Dolan J, Segal JB, Shihab HM, Singh S. Systematic assessment of benefits and risks: study protocol for a multi-criteria decision analysis using the analytic hierarchy process for comparative effectiveness research. F1000Res 2013;2:160 [FREE Full text] [doi: 10.12688/f1000research.2-160.v1] [Medline: 24555077]
32. Nag K, Helal M. Multicriteria inventory classification of diabetes drugs using a comparison of AHP and fuzzy AHP models. : IEEE; 2018 Presented at: 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM); December 16-19, 2018; Bangkok, Thailand p. 1456-1460. [doi: 10.1109/ieem.2018.8607678]
33. Chen RC, Chiu JY, Batj CT. The recommendation of medicines based on multiple criteria decision making and domain ontology—an example of anti-diabetic medicines. : IEEE; 2011 Presented at: 2011 International Conference on Machine Learning and Cybernetics; July 10-13, 2011; Guilin, China p. 27-32. [doi: 10.1109/icmlc.2011.6016682]
34. Wang M, Liu YW, Li X. Type-2 diabetes management using analytic hierarchy process and analytic network process. : IEEE; 2014 Presented at: Proceedings of the 11th IEEE International Conference on Networking, Sensing and Control; April 07-09, 2014; Miami, FL, USA p. 655-660. [doi: 10.1109/ICNSC.2014.6819703]
35. Bao Y, Gao B, Meng M, Ge B, Yang Y, Ding C, et al. Impact on decision making framework for medicine purchasing in Chinese public hospital decision-making: determining the value of five Dipeptidyl Peptidase 4 (DPP-4) inhibitors. BMC Health Serv Res 2021;21(1):807 [FREE Full text] [doi: 10.1186/s12913-021-06827-0] [Medline: 34384428]
36. Onar SC, Ibil EH. A decision support system proposition for type-2 diabetes mellitus treatment using spherical fuzzy AHP method. In: Tolga AC, Oztaysi B, Kahraman C, Sari IU, Cebi S, Onar SC, editors. Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, Held August 24-26, 2021. Volume 2. Cham, Switzerland: Springer International Publishing; 2021:749-756.
37. Zhang LD, Cui X, Liu FM, Xie YM, Zhang Q. Clinical comprehensive evaluation of Mudan Granules in treatment of diabetic peripheral neuropathy with qi-deficiency and collateral stagnation syndrome. Zhongguo Zhong Yao Za Zhi 2021;46(23):6078-6086. [doi: 10.19540/j.cnki.cjcmm.20210930.501] [Medline: 34951235]
38. Cai T, Wu H, Qin J, Qiao J, Yang Y, Wu Y, et al. In vitro evaluation by PCA and AHP of potential antidiabetic properties of lactic acid bacteria isolated from traditional fermented food. LWT 2019;115:108455. [doi: 10.1016/j.lwt.2019.108455]
39. Sekar KR, Yogapriya S, Anand NS, Venkataraman V. Ranking diabetic mellitus using improved PROMETHEE hesitant fuzzy for healthcare systems. In: Chen JIZ, Hemanth J, Bestak R, editors. Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020. Singapore: Springer Nature; 2021:709-724.
40. Mühlbacher AC, Bethge S, Kaczynski A, Juhnke C. Patients preferences regarding the treatment of type II diabetes mellitus: comparison of best-worst scaling and analytic hierarchy process. Value Health 2013;16(7):A446 [FREE Full text] [doi: 10.1016/j.jval.2013.08.707]
41. Mahat N, Ahmad S. Selection of the best thermal massage treatment for diabetes by using fuzzy analytical hierarchy process. J Comput Res Innov 2018;2(1):23-28 [FREE Full text] [doi: 10.24191/jcrinn.v2i1.25]
42. Pan D, Wang K, Zhou Z, Liu X, Shen J. FAHP-based mathematical model for exercise rehabilitation management of diabetes mellitus. ArXiv. Preprint posted online on January 7 2022 [FREE Full text] [doi: 10.48550/arXiv.2201.07884]
43. Rani P, Mishra AR, Mardani A. An extended Pythagorean fuzzy complex proportional assessment approach with new entropy and score function: application in pharmacological therapy selection for type 2 diabetes. Appl Soft Comput 2020;94:106441. [doi: 10.1016/j.asoc.2020.106441]
44. Balubaid MA, Basheikh MA. Using the analytic hierarchy process to prioritize alternative medicine: selecting the most suitable medicine for patients with diabetes. Int J Basic Appl Sci 2016;5(1):67 [FREE Full text] [doi: 10.14419/ijbas.v5i1.5607]
45. Mühlbacher AC, Bethge S, Kaczynski A, Juhnke C. Objective criteria in the medicinal therapy for type II diabetes: an analysis of the patients' perspective with analytic hierarchy process and best-worst scaling. Gesundheitswesen 2016;78(5):326-336. [doi: 10.1055/s-0034-1390474] [Medline: 25853782]
46. Abbasi M, Khorasani ZM, Etminani K, Rahmanvand R. Determination of the most important risk factors of gestational diabetes in Iran by group analytical hierarchy process. Int J Reprod Biomed 2017;15(2):109-114 [FREE Full text] [Medline: 28462403]
47. Yas QM, Adday BN, Abed AS. Evaluation multi diabetes mellitus symptoms by integrated fuzzy-based MCDM approach. Turk J Comput Math Educ 2021;12(13):4069-4082 [FREE Full text]

48. Amin-Naseri MR, Neshat N. An expert system based on analytical hierarchy process for Diabetes Risk Assessment (DIABRA). In: Wang G, Chai Y, Tan Y, Shi Y, editors. Advances in Swarm Intelligence, Part II: Second International Conference, ICSI 2011, Chongqing, China, June 12-15, 2011, Proceedings, Part II. Berlin Heidelberg: Springer; 2011:252-259.
49. El-Sappagh S, Alonso JM, Ali F, Ali A, Jang J, Kwak K. An ontology-based interpretable fuzzy decision support system for diabetes diagnosis. IEEE Access 2018;6:37371-37394 [FREE Full text] [doi: 10.1109/access.2018.2852004]
50. Baha BY, Wajiga GM, Blamah NV, Adewumi AO. Analytical hierarchy process model for severity of risk factors associated with type 2 diabetes. Sci Res Essays 2013;8(39):1906-1910 [FREE Full text]
51. Sharma S, Sharma B. EDAS based selection of machine learning algorithm for diabetes detection. : IEEE; 2020 Presented at: 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART); December 04-05, 2020; Moradabad, India p. 240-244.
52. Malapane J, Doorsamy W, Paul BS. Prediction analysis using weighted product method to compare machine learning algorithms for diabetes disease. Int J Res Eng 2022 Sep 04;5(9):49-53.
53. Felix A, Kumar RS, Parthiban A. Soft computing decision making system to analyze the risk factors of T2DM. AIP Conf Proc 2019;2112:020086-1-020086-12 [FREE Full text] [doi: 10.1063/1.5112271]
54. Sankar A, Jeyaraj GT. Extreme learning machine and K-means clustering for the improvement of link prediction in social networks using analytic hierarchy process. Int J Enterp Netw Manag 2019;10(3/4):371-388. [doi: 10.1504/ijenm.2019.10024740]
55. Bondor CI, Mureşan A. Correlated criteria in decision models: recurrent application of TOPSIS method. Appl Med Inform 2012;30(1):55-63 [FREE Full text]
56. Gaikwad SM, Mulay P, Joshi RR. Analytical hierarchy process to recommend an ice cream to a diabetic patient based on sugar content in it. Procedia Comput Sci 2015;50:64-72 [FREE Full text] [doi: 10.1016/j.procs.2015.04.062]
57. Sharawat K, Dubey SK. Diet recommendation for diabetic patients using MCDM approach. In: Gehlot A, Singh R, Choudhury S, editors. Intelligent Communication, Control and Devices: Proceedings of ICICCD 2017. Singapore: Springer Nature; 2018:239-246.
58. Santoso I, Sa'adah M, Wijana S. QFD and fuzzy AHP for formulating product concept of probiotic beverages for diabetic. TELKOMNIKA 2017;15(1):391-398 [FREE Full text] [doi: 10.12928/telkomnika.v15i1.3555]
59. Zadeh MSAT, Li J, Alian S. Personalized meal planning for diabetic patients using a multi-criteria decision- making approach. : IEEE; 2019 Presented at: 2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom); October 14-16, 2019; Bogota, Colombia p. 1-6. [doi: 10.1109/healthcom46333.2019.9009593]
60. Gulint G, Kadam K. Recommending food replacement shakes along with ice cream for diabetic patients using AHP and TOPSIS to control blood glucose level. Int J Eng Trends Technol 2016;34(5):243-251 [FREE Full text] [doi: 10.14445/22315381/ijett-v34p250]
61. Gaikwad SM, Joshi RR, Mulay P. Analytical Network Process (ANP) to recommend an ice cream to a diabetic patient. Int J Comput Appl 2015;121(12):49-52 [FREE Full text] [doi: 10.5120/21596-4692]
62. Gaikwad SM, Joshi RR, Kulkarni AJ. Cohort intelligence and genetic algorithm along with AHP to recommend an ice cream to a diabetic patient. In: Swarm, Evolutionary, and Memetic Computing: 6th International Conference, SEMCCO 2015, Hyderabad, India, December 18-19, 2015, Revised Selected Papers. Cham: Springer International Publishing; 2016:40-49.
63. Gaikwad SM, Joshi R, Gaikwad SM. Modified analytical hierarchy process to recommend an ice cream to a diabetic patient. 2016 Presented at: ICTCS '16: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies; March 4-5, 2016; Udaipur, India p. 1-5. [doi: 10.1145/2905055.2905198]
64. Gupta K, Roy S, Poonia RC, Nayak SR, Kumar R, Alzahrani KJ, et al. Evaluating the usability of mHealth applications on type 2 diabetes mellitus using various MCDM methods. Healthcare (Basel) 2021;10(1):4 [FREE Full text] [doi: 10.3390/healthcare10010004] [Medline: 35052167]
65. Wang X, He L, Zhu K, Zhang S, Xin L, Xu W, et al. An integrated model to evaluate the impact of social support on improving self-management of type 2 diabetes mellitus. BMC Med Inform Decis Mak 2019;19(1):197 [FREE Full text] [doi: 10.1186/s12911-019-0914-9] [Medline: 31640691]
66. Mishra V, Samuel C, Sharma SK. Supply chain partnership assessment of a diabetes clinic. Int J Health Care Qual Assur 2018;31(6):646-658. [doi: 10.1108/IJHCQA-06-2017-0113] [Medline: 29954271]
67. Mishra V. Customized quality assessment framework for diabetes care. Int J Qual Res 2020;14(1):129-146 [FREE Full text] [doi: 10.24874/ijqr14.01-09]
68. Mishra V. Planning and selection of facility layout in healthcare services. Hosp Top 2022:1-9. [doi: 10.1080/00185868.2022.2088433] [Medline: 35758293]
69. Byun DH, Chang RS, Park MB, Son HR, Kim CB. Prioritizing community-based intervention programs for improving treatment compliance of patients with chronic diseases: applying an analytic hierarchy process. Int J Environ Res Public Health 2021;18(2):455 [FREE Full text] [doi: 10.3390/ijerph18020455] [Medline: 33430108]
70. Mehrotra S, Kim K. Outcome based state budget allocation for diabetes prevention programs using multi-criteria optimization with robust weights. Health Care Manag Sci 2011;14(4):324-337. [doi: 10.1007/s10729-011-9166-7] [Medline: 21674143]

71. Haji M, Kerbache L, Al-Ansari T. Evaluating the performance of a safe insulin supply chain using the AHP-TOPSIS approach. Processes 2022;10(11):2203 [FREE Full text] [doi: 10.3390/pr10112203]
72. Suka M, Ichimura T, Yoshida K. Clinical decision support system applied the analytic hierarchy process. In: Palade V, Howlett RJ, Jain L, editors. Knowledge-Based Intelligent Information and Engineering Systems, LNCS 2774. Berlin Heidelberg: Springer; 2003:417-423.
73. Fico G, Cancela J, Arredondo MT, Dagliati A, Sacchi L, Segagni D, et al. User requirements for incorporating diabetes modeling techniques in disease management tools. In: Lackovic I, Vasic D, editors. 6th European Conference of the International Federation for Medical and Biological Engineering. IFMBE Proceedings, vol 45. Cham: Springer; 2015:992-995.
74. Long MD, Centor R. 236 utilizing pairwise comparisons to determine relative importance of diabetes guidelines: a comparison of physician and patient perspectives. J Investig Med 2005;53(1):S294 [FREE Full text] [doi: 10.2310/6650.2005.00006.235]
75. Gajdoš O, Juřičková I, Otawova R. Health technology assessment models utilized in the chronic care management. In: Ortuño F, Rojas I, editors. Bioinformatics and Biomedical Engineering. IWBBIO 2015. Lecture Notes in Computer Science, vol 9043. Cham: Springer; 2015:54-65.
76. Gupta K, Roy S, Poonia RC, Kumar R, Nayak SR, Altameem A, et al. Multi-criteria usability evaluation of mHealth applications on type 2 diabetes mellitus using two hybrid MCDM models: CODAS-FAHP and MOORA-FAHP. Appl Sci 2022;12(9):4156 [FREE Full text] [doi: 10.3390/app12094156]
77. Chang HY, Lo CL, Chang HL. Development of the benefit-risk assessment of complementary and alternative medicine use in people with diabetes: a Delphi-analytic hierarchy process approach. Comput Inform Nurs 2021;39(7):384-391 [FREE Full text] [doi: 10.1097/CIN.0000000000000749] [Medline: 33871384]
78. Ebrahimi M, Ahmadi K. Diabetes-related complications severity analysis based on hybrid fuzzy multi-criteria decision making approaches. Iran J Med Inform 2017;6(1):11 [FREE Full text] [doi: 10.24200/ijmi.v6i1.129]
79. Ahmadi K, Ebrahimi M. A novel algorithm based on information diffusion and fuzzy MADM methods for analysis of damages caused by diabetes crisis. Appl Soft Comput 2019;76:205-220. [doi: 10.1016/j.asoc.2018.12.004]
80. Bondor CI, Kacso IM, Lenghel AR, Mureşan A. Hierarchy of risk factors for chronic kidney disease in patients with type 2 diabetes mellitus. : IEEE; 2012 Presented at: 2012 IEEE 8th International Conference on Intelligent Computer Communication and Processing; August 30-September 01, 2012; Cluj-Napoca, Romania p. 103-106. [doi: 10.1109/iccp.2012.6356170]
81. Ahmed S, Roy S, Alam GR. Benchmarking and selecting optimal diabetic retinopathy detecting machine learning model using entropy and TOPSIS method. : IEEE; 2021 Presented at: 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME); October 07-08, 2021; Mauritius, Mauritius p. 1-6. [doi: 10.1109/iceccme52200.2021.9591153]
82. Bondor CI, Kacso IM, Lenghel A, Istrate D, Muresan A. VIKOR method for diabetic nephropathy risk factors analysis. Appl Med Inform 2013;32(1):43-52 [FREE Full text]
83. Alassery F, Alzahrani A, Khan AI, Khan A, Nadeem M, Ansari MTJ. Quantitative evaluation of mental-health in type-2 diabetes patients through computational model. Intell Autom Soft Comput 2022;32(3):1701-1715 [FREE Full text] [doi: 10.32604/iasc.2022.023314]
84. Wang CC, Yang CH, Wang CS, Xu D, Huang BS. Artificial neural networks in the selection of shoe lasts for people with mild diabetes. Med Eng Phys 2019;64:37-45. [doi: 10.1016/j.medengphy.2018.12.014] [Medline: 30655221]
85. Jain R, Kathuria A, Mukhopadhyay D, Gupta M. Fuzzy MCDM: application in disease risk and prediction. In: Devi KG, Rath M, Linh NTD, editors. Artificial Intelligence Trends for Data Analytics Using Machine Learning and Deep Learning Approaches. Boca Raton, FL: CRC Press; 2020:55-70.

Abbreviations

ADA: American Diabetes Association
AHP: analytic hierarchy process
ANP: analytical network process
CRITIC: Criteria Importance Through Intercriteria Correlation
ELECTRE: Elimination Et Choix Traduisant la Realité
IDF: International Diabetes Federation
MADM: multiattribute decision-making
MCDA: multicriteria decision-analysis
MCDM: multicriteria decision-making
MeSH: Medical Subject Headings
MODM: multiobjective decision-making
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analyses
PROMETHEE: preference ranking organization method for enrichment of evaluations
SWARA: step-wise weigh assessment ratio analysis
TOPSIS: technique for order of preference by similarity to ideal solution

T2D: type 2 diabetes
VIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje
WASPAS: Weighted Aggregated Sum Product Assessment
WPM: Weighted Product Model
WSM: Weighted Sum Model

Edited by A Castonguay; submitted 29.03.23; peer-reviewed by E Nazarie, J Sussman, A Kandwal, A Ranusch; comments to author 31.08.23; revised version received 24.10.23; accepted 11.12.23; published 01.02.24.

Please cite as:
Aldaghi T, Muzik J
Multicriteria Decision-Making in Diabetes Management and Decision Support: Systematic Review
JMIR Med Inform 2024;12:e47701
URL: https://medinform.jmir.org/2024/1/e47701
doi:10.2196/47701
PMID:38300703

Review

Frameworks, Dimensions, Definitions of Aspects, and Assessment Methods for the Appraisal of Quality of Health Data for Secondary Use: Comprehensive Overview of Reviews

Jens Declerck[1,2], MSc; Dipak Kalra[1,2], Prof Dr; Robert Vander Stichele[3], Prof Dr; Pascal Coorevits[1], Prof Dr

[1]Department of Public Health and Primary Care, Unit of Medical Informatics and Statistics, Ghent University, Ghent, Belgium

[2]The European Institute for Innovation through Health Data, Ghent, Belgium

[3]Faculty of Medicine and Health Sciences, Heymans Institute of Pharmacology, Ghent, Belgium

Corresponding Author:
Jens Declerck, MSc
Department of Public Health and Primary Care
Unit of Medical Informatics and Statistics
Ghent University
Campus UZ-Ghent, Entrance 42, 6th Floor
Corneel Heymanslaan 10
Ghent, 9000
Belgium
Phone: 32 93323628
Email: jens.declerck@ugent.be

Abstract

Background: Health care has not reached the full potential of the secondary use of health data because of—among other issues—concerns about the quality of the data being used. The shift toward digital health has led to an increase in the volume of health data. However, this increase in quantity has not been matched by a proportional improvement in the quality of health data.

Objective: This review aims to offer a comprehensive overview of the existing frameworks for data quality dimensions and assessment methods for the secondary use of health data. In addition, it aims to consolidate the results into a unified framework.

Methods: A review of reviews was conducted including reviews describing frameworks of data quality dimensions and their assessment methods, specifically from a secondary use perspective. Reviews were excluded if they were not related to the health care ecosystem, lacked relevant information related to our research objective, and were published in languages other than English.

Results: A total of 22 reviews were included, comprising 22 frameworks, with 23 different terms for dimensions, and 62 definitions of dimensions. All dimensions were mapped toward the data quality framework of the European Institute for Innovation through Health Data. In total, 8 reviews mentioned 38 different assessment methods, pertaining to 31 definitions of the dimensions.

Conclusions: The findings in this review revealed a lack of consensus in the literature regarding the terminology, definitions, and assessment methods for data quality dimensions. This creates ambiguity and difficulties in developing specific assessment methods. This study goes a step further by assigning all observed definitions to a consolidated framework of 9 data quality dimensions.

(JMIR Med Inform 2024;12:e51560) doi:10.2196/51560

KEYWORDS

data quality; data quality dimensions; data quality assessment; secondary use; data quality framework; fit for purpose

Introduction

To face the multiple challenges within our health care system, the secondary use of health data holds multiple advantages: it could increase patient safety, provide insights into person-centered care, and foster innovation and clinical research. To maximize these benefits, the health care ecosystem is investing rapidly in primary sources, such as electronic health records (EHRs) and personalized health monitoring, as well as in secondary sources, such as health registries, health information systems, and digital health technologies, to effectively manage illnesses and health risks and improve health care outcomes [1]. These investments have led to large volumes

of complex real-world data. However, health care is not obtaining the full potential of the secondary use of health data [2,3] because of—among other issues—concerns about the quality of the data being used [4,5]. Errors in the collection of health data are common. Studies have reported that at least half of EHR notes may contain an error leading to low-quality data [6-11]. The transition to digital health has produced more health data but not to the same extent as an increase in the quality of health data [12]. This will impede the potentially positive impact of digitalization on patient safety [13], patient care [14], decision-making [15], and clinical research [16].

The literature is replete with various definitions of data quality. One of the most used definitions for data quality comes from Juran et al [17], who defined data quality as "data that are fit for use in their intended operational, decision-making, planning, and strategic roles." According to the International Organization for Standardization (ISO) definition, quality is "the capacity of an ensemble of intrinsic characteristics to satisfy requirements" (ISO 9000-2015). DAMA International (The Global Data Management Community: a leading international association involving both business and technical data management professionals) adapts this definition to a data context: "data quality is the degree to which the data dimensions meet requirements." These definitions emphasize the subjectivity and context dependency of data quality [18]. Owing to this "fit for purpose" principle, the quality of data may be adequate when used for one specific task but not for another.

For example, when health data collected for primary use setting, such as blood pressure, are reused for different purposes, the adequacy of their quality can vary. For managing hypertension, the data's accuracy and completeness may be considered adequate. However, if the same data are reused for research, for example, in a clinical trial evaluating the effectiveness of an antihypertensive, more precise and standardized measurements methods are needed. From the perspective of secondary use, data are of sufficient quality when they serve the needs of the specific goals of the reuser [4].

To ensure that the data are of high quality, they must meet some fundamental measurable characteristics (eg, data must be complete, correct, and up to date). These characteristics are called data quality dimensions, and several authors have attempted to formulate a complex multidimensional framework of data quality. Kahn et al [19] developed a data quality framework containing conformance, completeness, and plausibility as the main data quality dimensions. This framework was the result of 2 stakeholder meetings in which data quality terms and definitions were grouped into an overall conceptual framework. The i~HD (European Institute for Innovation through Health Data) prioritized 9 data quality dimensions as most important to assess the quality of health data [20]. These dimensions were selected during a series of workshops with clinical care, clinical research, and ICT leads from 70 European hospitals. In addition, it is well known that there are several published reviews in which the results of individual quality assessment studies were collated into a new single framework of data quality dimensions. However, the results of these reviews have not yet been evaluated. Therefore, answering the "fit for purpose" question and establishing effective methods to assess data quality remain a challenge [21].

The primary objective of this review is to provide a thorough overview of data quality frameworks and their associated assessment methods, with a specific focus on the secondary use of health data, as presented in published reviews. As a secondary aim, we seek to align and consolidate the findings into a unified framework that captures the most crucial aspects of quality with a definition along with their corresponding assessment methods and requirements for testing.

Methods

Overview

We conducted a review of reviews to gain insights into data quality related to the secondary use of health data. In this review of reviews, we applied the Equator recommendations from the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines proposed by Page et al [22]. As our work is primarily a review of reviews, we included only the items from these guidelines that were applicable. Abstracts were sourced by searching the PubMed, Embase, Web of Science, and SAGE databases. The search was conducted in April 2023, and only reviews published between 1995 and April 2023 were included. We used specific search terms that were aligned with the aim of our study. To ensure comprehensiveness, the search terms were expanded by searching for synonyms and relevant key terms. The following concepts were used: "data quality" or "data accuracy," combined with "dimensions," "quality improvement," "data collection," "health information interoperability," "health information systems," "public health information," "quality assurance," and "delivery of health care." Textbox 1 illustrates an example of the search strategy used in PubMed. To ensure the completeness of the review, the literature search spanned multiple databases. All keywords and search queries were adapted and modified to suit the requirements of these various databases (Multimedia Appendix 1).

Textbox 1. Search query used.

("data quality" OR "Data Accuracy"[Mesh]) AND (dimensions OR "Quality Improvement"[Mesh] OR "Data Collection/standards"[Mesh] OR "Health Information Interoperability/standards"[Mesh] OR "Health Information Systems/standards"[Mesh] OR "Public Health Informatics/standards" OR "Quality Assurance, Health Care/standards"[Mesh] OR "Delivery of Health Care/standards"[Mesh]) Filters: Review, Systematic Review

Inclusion and Exclusion Criteria

We included review articles that described and discussed frameworks of data quality dimensions and their assessment methods, especially from a secondary use perspective. Reviews were excluded if they were (1) not specifically related to the health care ecosystem, (2) lacked relevant information related to our research objective (no definition of dimensions), or (3) published in languages other than English.

Selection of Articles

One reviewer (JD) screened the titles and abstracts of 982 articles from the literature searches and excluded 940 reviews. Two reviewers (RVS and JD) independently performed full-text screening of the remaining 42 reviews. Disagreements between the 2 reviewers were resolved by consulting a third reviewer (DK). After full-text screening, 20 articles were excluded because they did not meet the inclusion criteria. A total of 22 articles were included in this review.

Data Extraction

All included articles were imported into EndNote 20 (Clarivate). Data abstraction was conducted independently by 2 reviewers (RVS and JD). Disagreements between the 2 reviewers were resolved by consulting a third reviewer (DK). The information extracted from the reviews included various details, including the authors, publication year, research objectives, specific data source used, scope of secondary use, terminology used for the data quality dimensions, their corresponding definitions, and the measurement methods used.

Data Synthesis

To bring clarity to the diverse dimensions and definitions scattered throughout the literature, we labeled the observed definitions of dimensions from the reviews as “aspects.” We then used the framework of the i~HD. This framework underwent extensive validation through a large-scale exercise and was published [20]. It will now serve as a reference framework for mapping the diverse literature in the field. This overarching framework comprised 9 loosely delineated data quality dimensions (Textbox 2, [20]). Each observed definition of a data quality dimension was mapped onto a dimension of this reference framework. This mapping process was collaborative and required consensus among the reviewers. This consolidation is intended to offer a more coherent and unified perspective on data quality for secondary use.

Textbox 2. Consolidated data quality framework of the European Institute for Innovation through Health Data [20].

Data quality dimension and definition

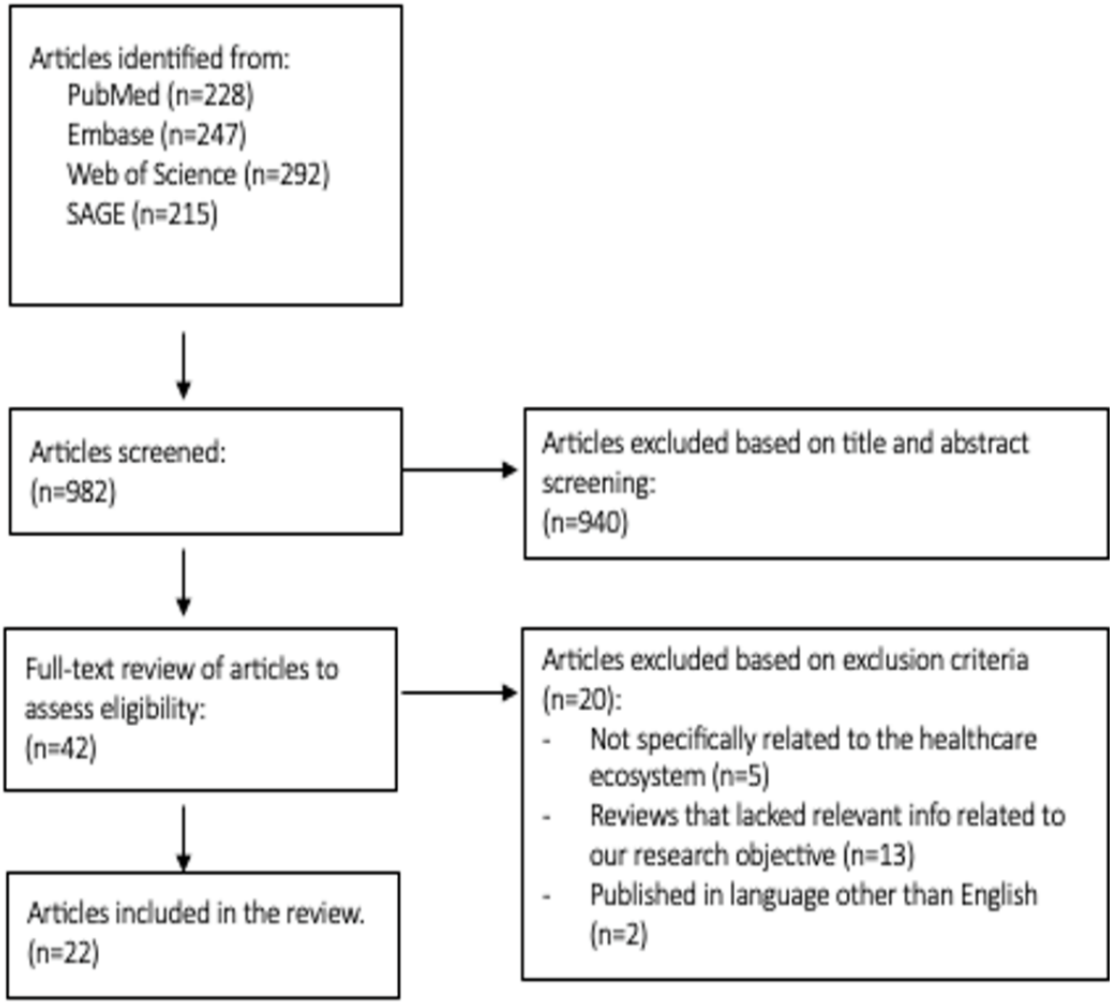
- Completeness: the extent to which data are present
- Consistency: the extent to which data satisfy constraints
- Correctness: the extent to which data are true and unbiased
- Timeliness: the extent to which data are promptly processed and up to date
- Stability: the extent to which data are comparable among sources and over time
- Contextualization: the extent to which data are annotated with acquisition context
- Representativeness: the extent to which data are representative of intended use
- Trustworthiness: the extent to which data can be trusted based on the owner’s reputation
- Uniqueness: the extent to which data are not duplicated

Results

Search Process

Figure 1 summarizes the literature review process and the articles included and excluded at every stage of the review using the PRISMA guidelines. It is important to note that this was not a systematic review of clinical trials; rather, it was an overview of existing reviews. As such, it synthesizes and analyzes the findings from multiple reviews on the topic of interest. A total of 22 articles were included in this review. The 22 reviews included systematic reviews (4/22, 18%) [23-26], scoping reviews (2/22, 9%) [27,28], and narrative reviews (16/22, 73%) [4,29-43]. All the reviews were published between 1995 and 2023. Of the 20 excluded reviews, 5 (25%) were excluded because they were not specific to the health care ecosystem [18,44-47], 13 (65%) lacked relevant information related to our research objective [6-18], and 2 (10%) were published in a language other than English [48,49].

Figure 1. The process of selecting articles.



Data Sources

Of the 22 reviews, 10 (45%) discussed data quality pertaining to a registry [25-27,34-36,40-43] and 4 (18%) to a network of EHRs [4,24,29,33]. Of the 22 reviews, 4 (18%) discussed the quality of public health informatics systems [37,38], real-world data repositories [31], and clinical research informatics tools [30]. Of the 22 reviews, 4 (18%) did not specify their data source [23,28,32,39].

Observed Frameworks for Data Quality Dimensions

In the initial phase of our study, we conducted a comprehensive review of 22 selected reviews, each presenting a distinct framework for understanding data quality dimensions. Across these reviews, the number of dimensions varied widely, ranging from 1 to 14 (median 4, IQR 2-5). The terminology used was diverse, yielding 23 different terms for dimensions and 62 unique definitions. A detailed overview, including data sources, data quality dimensions, and definitions, is provided in Multimedia Appendix 2 [4,23-43]. Figure S1 in Multimedia Appendix 3 presents the frequency of all dimensions in each review along with the variety of definitions associated with each dimension.

Data Synthesis: Constructing a Consolidated Data Quality Framework For Secondary Use

Overview

Table 1 presents all dimensions mentioned in the included reviews, with their definitions, mapped toward each of the 9 data quality dimensions in the framework of i~HD.

Table 1. Mapping of data quality aspects toward i~HD (European Institute for Innovation through Health Data) data quality framework.

i~HD data quality dimensions and aspects as mentioned in the reviews	Definition
Completeness	
Completeness [30,32,33,39]	The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.
Completeness [24,29,39]	This focuses on features that describe the frequencies of data attributes present in a data set without reference to data values.
Completeness [27,35,42]	The extent to which all necessary data that could have been registered have been registered.
Completeness [34,41]	The extent to which all the incident cases occurring in the population are included in the registry database.
Completeness [43]	The completeness of data values can be divided between mandatory and optional data fields.
Completeness [23]	The absence of data at a single moment over time or when measured at multiple moments over time.
Completeness [4]	Is a truth of a patient present in the EHR[a]?
Completeness [26]	All necessary data are provided.
Completeness [25]	Defined as the presence of recorded data points for each variable.
Plausibility [31]	Focuses on features that describe the frequencies of data attributes present in a data set without reference to data values.
Capture [27,35]	The extent to which all necessary cases that could have been registered have been registered.
Consistency	
Accuracy [43]	The accuracy of data values can be divided into syntactic and semantic values.
Consistency [43]	Data inconsistencies occur when values in ≥2 data fields are in conflict.
Consistency [39]	Representation of data values is the same in all cases.
Consistency [26]	Data are logical across data points.
Consistency [32]	The degree to which data have attributes that are free from contradiction and are coherent with other data in a specific content of use.
Consistency [23]	Absence of differences between data items representing the same objects based on specific information requirements.
Consistency [30]	Refers to the extent to which data are applicable and helpful to the task at hand.
Correctness [26]	Data are within the specified value domains.
Comparability [34,40]	The extent to which coding and classification procedures at a registry, together with the definitions of recoding and reporting specific data terms, adhere to the agreed international guidelines.
Validity [30]	Refers to information that does not conform to a specific format or does not follow business rules.
Concordance [32]	The data are concordant when there was agreement or comparability between data elements.
Conformance [29,31]	Focuses on data quality features that describe the compliance of the representation of data against internal or external formatting, relational, or computational definitions.
Conformance [24]	Whether the values that are present meet syntactic or structural constraints.
Correctness	
Accuracy [27,35,42]	The extent to which registered data are in conformity to the truth.
Accuracy [32,33]	The extent to which data are correct and reliable.
Accuracy [23]	The degree to which data reveal the truth about the event being described.
Accuracy [26]	Data conform to a verifiable source.
Accuracy [30]	Refers to the degree to which information accurately reflects an event or object described.
Correctness [4,24]	Is an element that is present in the EHR true?
Correctness [39]	The free-of-error dimension.
Plausibility [4]	Does an element in the EHR makes sense in the light of other knowledge about what that element is measuring?

i~HD data quality dimensions and aspects as mentioned in the reviews	Definition
Plausibility [24]	This focuses on actual values as a representation of a real-world object or conceptual construct by examining the distribution and density of values or by comparing multiple values that have an expected relationship with each other.
Plausibility [29]	Focuses on features that describe the believability or truthfulness of data values.
Validity [34,40]	Defined as the proportion of cases in a data set with a given characteristic which truly have the attribute.
Uniqueness	
Redundancy [32]	Data contain no redundant values.
Stability	
Consistency [33]	Representations of data values remain the same in multiple data items in multiple locations.
Consistency [24]	Refers to the consistency of data at the specified level of detail for the study's purpose, both within individual databases and across multiple data sets.
Currency [43]	Data currency is important for those data fields that involve information that may change over time.
Comparability [24]	This is the similarity in data quality and availability for specific data elements used in measure across different entities, such as health plans, physicians, or data sources.
Concordance [4,24]	Is there agreement between elements in the EHR or between the EHR and another data source?
Information loss and degradation [24]	The loss and degradation of information content over time.
Timeliness	
Timeliness [30,33,39]	The extent to which information is up to date for the task at hand.
Timeliness [27,34,40]	Related to the rapidity at which a registry can collect, process, and report sufficiently reliable and complete data.
Timeliness [26]	Data are available when needed.
Currency [4]	Is an element in the EHR a relevant representation of the patient's state at a given point in time?
Currency [32]	The degree to which data have attributes that are of the right age in a specific context of use.
Currency [24]	Data were considered current if they were recorded in the EHR within a reasonable period following a measurement or if they were representative of the patient's state at a desired time of interest.
Currency [23]	The degree to which data represent reality from the required point in time.
Accessibility [33]	The extent to which data are available or easily and quickly retrievable.
Contextualization	
Understandability [24]	The ease with which a user can understand the data.
Understandability [30]	Refers to the degree to which the data can be comprehended.
Contextual validity [23]	Assessment of data quality is dependent on the task at hand.
Flexibility [24]	The extent to which data are expandable, adaptable, and easily applied to many tasks.
Trustworthiness	
Security [24,39]	Personal data are not corrupted, and access is suitably controlled to ensure privacy and confidentiality.
Representation	
Relevance [24,39]	The extent to which information is applicable and helpful for the task at hand.
Precision [26]	Data value is specific.

[a]EHR: electronic health record.

Completeness

The first data quality dimension relates to the completeness of data. Among the 22 reviews included, 20 (91%) highlighted the significance of completeness [4,23-27,29-35,39,41-43]. Of these 20 reviews, 17 (85%) used the term completeness to refer to this dimension [4,23-27,29-35,39,41-43], whereas the remaining 3 (15%) used the terms plausibility [31] and capture [27,35].

On the basis of the definitions of completeness, we can conclude that this dimension contains 2 main aspects. First, completeness related to the data level. The most used definition related to this aspect is the extent to which information is not missing [30,32,33,39]. Other reviews focused more on features that describe the frequencies of data attributes present in a data set without reference to data values [24,29,39]. Shivasabesan et al [25], for example, defined completeness as the presence of

recorded data points for each variable. A second aspect for completeness relates more to a case level, in which all the incident cases occurring in the population are included [27,34,35,41].

Consistency

The second data quality dimension concerns the consistency of the data. Among the 22 selected reviews, 11 (50%) highlighted the importance of consistency [23,24,26,29-32,34,39,40,43]. Although various frameworks acknowledge this as a crucial aspect of data quality, achieving a consensus on terminology and definition has proven challenging. Notably, some reviews used different terminologies to describe identical concepts associated with consistency [26,30,32,43]. Of the 11 reviews, 6 (55%) used the term consistency to describe this dimension [23,26,30,32,39,43], whereas 3 (27%) used conformance [24,29,31] and 2 (18%) referred to comparability [34,40]. Of the 11 reviews, 3 (27%) used distinct terms: accuracy [43], validity [30], and concordance [32]. Most definitions focus on data quality features that describe the compliance of the representation of data with internal or external formatting, relational, or computational definitions [29,31]. Of the 11 reviews, 2 (18%) provided a specific definition of consistency concerning registry data, concentrating on the extent to which coding and classification procedures, along with the definitions or recording and reporting of specific data terms, adhere to the agreed international guidelines [34,40]. Furthermore, Bian et al [24] concentrated on whether the values present meet syntactic or structural constraints in their definition, whereas Liaw et al [39] defined consistency as the extent to which the representation of data values is consistent across all cases.

Correctness

The third data quality dimension relates to the correctness of the data. Of the 22 reviews, 14 (64%) highlighted the importance of correctness [4,23,24,26,27,29,30,32-35,39,40,42]. Of the 14 reviews, 2 (14%) used 2 different dimensions to describe the same concept of correctness [4,24]. Accuracy was the most frequently used term within these frameworks [23,26,27,32,33,35,42]. In addition, other terms used included correctness [4,24,39], plausibility [4,24,29], and validity [34,40]. In general, this dimension assesses the degree to which the recorded data align with the truth [27,35,42], ensuring correctness and reliability [32,33]. Of the 14 reviews, 2 (14%) provided a specific definition of correctness concerning EHR data, emphasizing that the element collected is true [4,24]. Furthermore, of the 14 reviews, 2 (14%) defined correctness more at a data set level, defining it as the proportion of cases in a data set with a given characteristic that genuinely possess the attribute [34,40]. These reviews specifically referred to this measure as validity. Nevertheless, the use of the term validity was not consistent across the literature; it was also used to define consistency. For instance, AbuHalimeh [30] used validity to describe the degree to which information adheres to a predefined format or complies with the established business rules.

Timeliness

The fourth data quality dimension concerns the timeliness of the data. Among the 22 selected reviews, 11 (50%) underscored the importance of this data quality dimension [4,23,24,26,27,30,32-34,39,40]. Of the 11 reviews, 7 (64%) explicitly used the term timeliness [26,27,30,33,34,39,40], whereas 4 (36%) referred to it as currency [4,23,24,32]. Mashoufi et al [33] used the terms accessibility and timeliness to explain the same concept. Broadly, timeliness describes how promptly information is processed or how up to date the information is. Most reviews emphasized timeliness as the extent to which information is up to date for the task at hand [30,33,39]. For instance, Weiskopf and Weng [4] provided a specific definition for EHR data, stating that an element should be a relevant representation of the patient's state at a given point in time. Other reviews defined timeliness as the speed at which data can be collected, processed, and reported [27,34,40]. Similarly, Porgo et al [26] defined timeliness as the extent to which data are available when needed.

Stability

The fifth data quality dimension concerns the stability of the data. Among the 22 included reviews, 4 (18%) acknowledged the significance of stability [4,24,33,43]. The most frequently used terms for this dimension are consistency [24,33] and concordance [24]. In addition, other terms used include currency [43], comparability [24], and information loss and degradation [24]. Bian et al [24] explored this aspect of data quality by using multiple terminologies to capture its multifaceted nature: stability, consistency, concordance, and information loss and degradation. This dimension, in general, encompasses 2 distinct aspects. First, it underscores the importance of data values that remain consistent across multiple sources and locations [4,24,33]. Alternatively, as described by Bian et al [24], it refers to the similarity in data quality for specific data elements used in measurements across different entities, such as health plans, physicians, or other data sources. Second, it addresses temporal changes in data that are collected over time. For instance, Lindquist [43] highlighted the importance of stability in data fields that involve information that may change over time. The term consistency is used across different data quality dimensions, but it holds different meanings depending on the context. When discussing the dimension of stability, consistency refers to the comparability of data across different sources. This ensures that information remains uniform when aggregated or compared. Compared with the consistency dimension, the term relates to the internal coherence of data within a single data set, which relates to the absence of contradiction and compliance with certain constraints. The results indicate the same ambiguity in terms of currency. When associated with stability, currency refers to the longitudinal aspect of variables. In contrast, within the dimension of timeliness, currency is concerned with the aspect if data are up to date.

Contextualization

The sixth data quality dimension revolves around the context of the data. Of the 22 reviews analyzed, 3 (14%) specifically addressed this aspect within their framework [23,24,30]. The most used term was understandability [24,30]. In contrast, Syed et al [23] used the term contextual validity, and Bian et al [24] referred to flexibility and understandability for defining the same concept. Broadly speaking, contextualization pertains to

whether the data are annotated with their acquisition context, which is a crucial factor for the correct interpretation of results. As defined by Bian et al [24], this dimension relates to the ease with which a user can understand data. In addition, AbuHalimeh [30] refers to the degree to which data can be comprehended.

Representation

The seventh dimension of data quality focuses on the representation of the data. Of the 22 reviews examined, 3 (14%) specifically highlighted the importance of this dimension [24,26,39]. Of the 3 reviews, 2 (67%) used the term relevance [24,39], whereas Porgo et al [26] used the term precision. Broadly speaking, representativeness assesses whether the information is applicable and helpful for the task at hand [24,39]. In more specific terms, as defined by Porgo et al [26], representativeness relates to the extent to which data values are specific to the task at hand.

Trustworthiness

The eighth dimension of data quality relates to the trustworthiness of the data. Of the 22 reviews, only 2 (9%) considered this dimension in their review [24,39]. In both cases, trustworthiness was defined as the extent to which data are free from corruption, and access was appropriately controlled to ensure privacy and confidentiality.

Uniqueness

The final dimension of data quality relates to the uniqueness of the data. Of the 22 reviews, only 1 (5%) referred to this aspect [32]. Uniqueness is evaluated based on whether there are no duplications or redundant data present in a data set.

Observed Data Quality Assessment Methods

Overview

Of the 22 selected reviews, only 8 (36%) mentioned data quality assessment methods [4,24,32,34,35,39-41]. Assessment methods were defined for 15 (65%) of the 23 data quality dimensions. The number of assessment methods per dimension ranged from 1 to 15 (median 3, IQR 1-5). There was no consensus on which method to use for assessing data quality dimensions. Figure S2 in Multimedia Appendix 3 presents the frequency of the dimensions assessed in each review, along with the number of different data quality assessment methods.

In the following section, we harmonize these assessment methods with our consolidated framework. This provides a comprehensive overview linking the assessment methods to the primary data quality dimensions from the previous section. Table 2 provides an overview of all data quality assessment techniques and their definitions. Textbox 3 presents an overview of all assessment methods mentioned in the literature and mapped toward the i~HD data quality framework.

Table 2. Overview of all data quality assessment methods with definitions.

Assessment M[a]	Assessment technique in reviews	Explanation
M1	Linkages—other data sets	• Percentage of eligible population included in the data set.
M2	Comparison of distributions	• Difference in means and other statistics.
M3	Case duplication	• Number and percentage or cases with >1 record.
M4	Completeness of variables	• Percentage of cases with complete observations of each variable.
M5	Completeness of cases	• Percentage of cases with complete observations for all variables.
M6	Distribution comparison	• Distributions or summary statistics of aggregated data from the data set are compared with the expected distributions for the clinical concepts of interest.
M7	Gold standard	• A data set drawn from another source or multiple sources is used as a gold standard.
M8	Historic data methods	• Stability of incidence rates over time. • Comparison of incidence rates in different populations. • Shape of age-specific curves. • Incidence rates of childhood curves.
M9	M:I[b]	• Comparing the number of deaths, sourced independently from the registry, with the number of new cases recorded for a specific period.
M10	Number of sources and notifications per case	• Using many sources reduces the possibility of diagnoses going unreported, thus increasing the completeness of cases.
M11	Capture-recapture method	• A statistical method using multiple independent samples to estimate the size of an entire population.
M12	Death certificate method	• This method requires that death certificate cases can be explicitly identified by the data set and makes use of the M:I ratio to estimate the proportion of the initially unregistered cases.
M13	Histological verification of diagnosis	• The percentage of cases morphologically verified is a measure of the completeness of the diagnostic information.
M14	Independent case ascertainment	• Rescreening the sources used to detect any case missing during the registration process.
M15	Data element agreement	• Two or more elements within a data set are compared to check if they report the same or compatible information.
M16	Data source agreement	• Data from the data set are cross-referenced with another source to check for agreement.
M17	Conformance check	• Check the uniqueness of objects that should not be duplicated; the data set agreement with prespecified or additional structural constraints, and the agreement of object concepts and formats granularity between ≥2 data sources.
M18	Element presence	• A determination is made as to whether or not desired or expected data elements are present.
M19	Not specified	• Number of consistent values and number of total values.
M20	International standards for classification and coding	• For example, neoplasms, the International Classification of Diseases for Oncology provides coding of topography, morphology, behavior, and grade.
M21	Incidence rate	• Not specified
M22	Multiple primaries	• The extent that a distinction must be made between those that are new cases and those that represent an extension or recurrence of an existing one.
M23	Incidental diagnosis	• Screening aims to detect cases that are asymptomatic. • Autopsy diagnosis without any suspicion of diagnosed case before death.

Assessment M[a]	Assessment technique in reviews	Explanation
M24	Not specified	• 1=ratio of violations of specific consistency type to the total number of consistency checks.
M25	Validity check	• Data in the data set are assessed using various techniques that determine of the values "make sense."
M26	Reabstracting and recoding	• Reabstracting describes the process of independently reabstracting records from a given source, coding the data, and comparing the abstracted and coded data with the information recorded in the database. For each reabstracted data item, the auditor's codes are compared with the original codes to identify discrepancies. • Recoding involves independently reassigning codes to abstracted text information and evaluating the level of agreement with records already in the database.
M27	Missing information	• The proportion of registered cases with unknown values for various data items.
M28	Internal consistency	• The proportion of registered cases with unknown values for various data items.
M29	Domain check	• Proportion of observations outside plausible range (%).
M30	Interrater variability	• Proportion of observations in agreement (%). • Kappa statistics.
M31	Log review	• Information on the actual data entry practices (eg, dates, times, and edits) is examined.
M32	Syntactic accuracy	• Not specified.
M33	Log review	• Information on the actual data entry practices (eg, dates, times, and edits) is examined. • Time at which data are stored in the system. • Time of last update. • User survey.
M34	Not specified	• Ratio: number of reports sent on time divided by total reports.
M35	Not specified	• Ratio: number of data values divided by the overall number of values.
M36	Time to availability	• The interval between date of diagnosis (or date of incidence) and the date the case was available in the registry or data set.
M37	Security analyses	• Analyses of access reports.
M38	Not specified	• Descriptive qualitative measures with group interviews and interpreted with grounded theory.

[a]M: method.

[b]M:I: mortality:incidence ratio.

Textbox 3. Mapping of assessment methods (Ms) toward data quality framework of the European Institute for Innovation through Health Data.

Completeness

- Capture [35]
  - M1: linkages—other data sets
  - M2: comparison of distributions
  - M3: case duplication
- Completeness [35]
  - M4: completeness of variables
  - M5: completeness of cases
- Completeness [32]
  - M4: completeness of variables
  - M6:distribution comparison
  - M7: gold standard
  - M5: completeness of cases
- Completeness [34]
  - M8: historic data methods
  - M9: mortality:incidence ratio (M:I)
  - M10: number of sources and notifications per case
  - M11: capture-recapture method
  - M12: death certificate method
- Completeness [41]
  - M8: historic data methods
  - M9: M:I
  - M10: number of sources and notifications per case
  - M11: capture-recapture method
  - M12: death certificate method
  - M13: histological verification of diagnosis
  - M14: independent case ascertainment
- Completeness [4]
  - M4: completeness of variables
  - M6: distribution comparison
  - M7: gold standard
  - M15: data element agreement
  - M16: data source agreement
- Completeness [24]
  - M4: completeness of variables
  - M6: distribution comparison
  - M7: gold standard
  - M17: conformance check

Consistency

- Conformance [24]

- M18: element presence
- M17: conformance check

- Concordance [32]
  - M15: data element agreement
  - M19: not specified
- Consistency [32]
  - M16: data source agreement
- Comparability [40]
  - M20: international standards for classification and coding
  - M21: incidence rate
  - M22: multiple primaries
  - M23: incidental diagnosis
  - M24: not specified
- Comparability [34]
  - M20: international standards for classification and coding
- Consistency [39]
  - M24: not specified

Correctness

- Correctness [4]
  - M7: gold standard
  - M15: data element agreement
- Plausibility [4]
  - M6: distribution comparison
  - M25: validity check
  - M31: log review
  - M16: data source agreement
- Validity [40]
  - M26: reabstracting and recoding
  - M13: histological verification of diagnosis
  - M27: missing information
  - M28: internal consistency
  - M12: death certificate method
- Validity [34]
  - M13: histological verification of diagnosis
  - M12: death certificate method
- Accuracy [35]
  - M7: gold standard
  - M28: internal consistency
  - M29: domain check

- M30: interrater variability
- Correctness [24]
  - M25: validity check
- Accuracy [32]
  - M7: gold standard
  - M32: syntactic accuracy

Stability

- Concordance [4]
  - M15: data element agreement
  - M16: data source agreement
  - M6: distribution comparison
- Comparability [24]
  - M18: element presence
- Consistency [24]
  - M17: conformance check
- Consistency [32]
  - M15: data element agreement
  - M16: data source agreement

Timeliness

- Currency [32]
  - M33: log review
- Currency [4]
  - M33: log review
- Timeliness [39]
  - M34: not specified
  - M35: not specified
- Currency [24]
  - M18: element presence
- Timeliness [40]
  - M36: time to availability

Trustworthiness

- Security [24,39]
  - M37: security analyses

Representation

- Relevance [39]
  - M38: not specified

Completeness

Among the 20 reviews that defined data quality dimensions related to completeness, 6 (30%) incorporated data quality assessment methods into their framework [4,24,32,34,35,41]. These 6 reviews collectively introduced 17 different data quality assessment methods. Some reviews (4/6, 67%) mentioned multiple methods to evaluate completeness, which highlights the absence of a consensus within the literature regarding the most suitable approach. The most frequently used method in the literature for assessing completeness was the examination of variable completeness [4,24,32,35]. This method involved calculating the percentage of cases that had complete observations for each variable within the data set. In 3 reviews [4,24,32], researchers opted to compare the distributions or summary statistics of aggregated data from the data set with the expected distributions for the clinical concepts of interest. Another approach found in 3 reviews involved the use of a gold standard to evaluate completeness [4,24,32]. This method relied on external knowledge and entailed comparing the data set under examination with data drawn from other sources or multiple sources.

Consistency

Among the 15 reviews highlighting the significance of consistency, 6 (40%) defined data quality assessment methods [4,24,32,34,39,40]. In these 6 reviews, a total of 10 distinct data quality assessment methods were defined. The most used method involved calculating the ratio of violations of specific consistency types to the total number of consistency checks [32,39]. There were 2 categories established for this assessment. First, internal consistency, which focuses on the most commonly used data type, format, or label within the data set. Second, external consistency, which centered on whether data types, formats, or labels could be mapped to a relevant reference terminology or data dictionary. Another common assessment method was the implementation of international standards for classification and coding standards [34,40]. This addressed specific oncology and suggested coding for topography, morphology, behavior, and grade. Liaw et al [39] defined an assessment method in which ≥2 elements within a data set are compared to check if they report compatible information.

Correctness

Among the 16 reviews underscoring the importance of correctness, 6 (38%) detailed data quality assessment methods [4,24,32,34,35,40]. Collectively, these 6 reviews proposed 15 different techniques. Prominent among these were histological verification [34,40], where the percentage of morphologically verified values served as an indicator of diagnosis correctness. Another frequently used technique was the use of validity checks [4], involving various methods to assess whether the data set values “make sense.” Three additional reviews opted for a comparative approach, benchmarking data against a gold standard and calculating the sensitivity, specificity, and accuracy scores [4,32,35]. Interestingly, there is an overlap between consistency and completeness as data quality dimensions in the assessment of correctness. For instance, Weiskopf and Weng [4] defined data element agreement as an assessment for this dimension, whereas Bray and Parkin [40] evaluated the proportion of registered cases with unknown values for specific items as a correctness assessment method.

Stability

Among the 7 reviews emphasizing the importance of stability of the data, only 3 (43%) discussed assessment techniques that address this dimension [4,24,39]. These 3 reviews collectively outlined 5 different techniques. Notably, there was no predominant technique. Specifically, Weiskopf and Weng [4] used several techniques to assess data stability, including an overlap with other dimensions, by using data element agreement. Another technique introduced in the same review was data source agreement, involving the comparison of data from different data sets from distinct sources.

Timeliness

Of the 12 reviews focusing on the timeliness of data, 5 (42%) delved into assessment techniques for this data quality dimension [4,24,32,39,40]. Across these reviews, 5 distinct assessment techniques were discussed. The most commonly used technique was the use of a log review [4,39]. This method involved collecting information that provides details on data entry, the time of data storage, the last update of the data, or when the data were accessed. In addition, Bray and Parkin [40] assessed timeliness by calculating the interval between the date of diagnosis (or date of incidence) and the date the case was available in the registry or data set.

Trustworthiness

In the 2 reviews that considered trustworthiness as a data quality dimension, both used the same assessment technique [24,39]. This method involves the analysis of access reports as a security analysis, providing insight into the trustworthiness of the data.

Representation

In 1 review that addressed the representation dimension as a data quality aspect, only 1 assessment method was mentioned. Liaw et al [39] introduced descriptive qualitative measures through group interviews to determine whether the data accurately represented the intended use.

Uniqueness and Contextualization

No assessment methods were mentioned for these data quality dimensions.

Discussion

Principal Findings

This first review of reviews regarding the quality of health data for secondary use offers an overview of the frameworks of data quality dimensions and their assessment methods, as presented in published reviews. There is no consensus in the literature on the specific terminology and definitions of terms. Similarly, the methodologies used to assess these terms vary widely and are often not described in sufficient detail. Comparability, plausibility, validity, and concordance are the 4 aspects classified under different consolidated dimensions, depending on their definitions. This variability underscores the prevailing discrepancies and the urgent need for harmonized definitions. Almost none of the reviews explicitly refer to requirements of

quality for the context of the data collection. Building on the insights gathered from these reviews, our consolidated framework organizes the numerous observed definitions into 9 main data quality dimensions, aiming to bring coherence to the fragmented landscape.

Health data in primary sources refer to data produced in the process of providing real-time and direct care to an individual [50], with the purpose of improving the care process. A secondary source captures data collected by someone other than the primary user and can be used for other purposes (eg, research, quality measurement, and public health) [50]. The included reviews discussed data quality for secondary use. However, the quality of health data in secondary systems is a function of the primary sources from which they originate, the quality of the process to transfer and transform the primary data to the secondary source, and the quality of the secondary source itself. The transfer and transformation of primary data to secondary sources implies the standardization, aggregation, and streamlining of health data. This can be considered as an export-transform-load (ETL) process with its own data quality implications. When discussing data quality dimensions and assessment methods, research should consider these different stages within the data life cycle, a distinction seldom made in the literature. For example, Prang et al [27] defined completeness within the context of a registry, which can be regarded as a secondary source. In this context, completeness was defined as the degree to which all potentially registrable data had been registered. The definition for completeness by Bian et al [24] pertains to an EHR, which is considered a primary source. Here, the emphasis was on describing the frequencies of data attributes. Both papers emphasized the importance of completeness, but they approached this dimension from different perspectives within the data life cycle.

This fragmented landscape regarding terminology and definition of data quality dimensions, the lack of distinction between quality in primary and secondary data and in the ETL process, and the lack of consideration for the context allows room for interpretation, leading to difficulties in developing assessment methods. In our included articles, only 8 (36%) out of 22 reviews mentioned and defined assessment methods [4,24,32,34,35,39-41]. However, the results showed that the described assessment methods are limited by a lack of well-defined and standardized metrics that can quantitatively or qualitatively measure the quality of data across various dimensions and often suffer from inadequate translation of these dimensions into explicit requirements for primary and secondary data and the ETL process, considering the purpose of the data collection of the secondary source. Both the DAMA and ISO emphasize in their definition of data quality that requirements serve as the translation of dimensions. Data quality dimensions refer to a broad context or characteristics of data that are used to assess the quality of data. Data quality requirements are derived from data quality dimensions and specify the specific criteria or standards that data must meet to be considered high-quality data. These requirements define the specific thresholds that need to be achieved for each dimension. However, our results show that the focus of the literature lies in defining dimensions and frameworks, rather than adequately developing these essential data quality requirements.

To avoid further problems and ambiguities, it is important to understand the purpose, context, and limitations of the data and data sources to establish a comprehensive view on the quality of the data. Rather than pursuing an elusive quest in the literature for a rigid framework defined by a fixed number of dimensions and precise definitions, future research should shift its focus toward defining and developing specific data quality requirements tailored to each use case. This approach should consider various stages within the data life cycle. For example, when defining a specific completeness requirement for a secondary use case, it will impact the way data are generated at the primary source and how they are transformed and transferred between the primary and secondary sources. Creating explicit requirements that align with the purpose of each use case along with well-defined criteria and thresholds can foster the development of precise assessment methods for each dimension. Moreover, formulating these use case requirements will facilitate addressing the fundamental question of whether health data are fit for purpose, thus determining if they are of a sufficient quality.

Limitations

The strength of a review of reviews methodology is to provide a comprehensive overview of the current state of knowledge. However, it is important to acknowledge that this approach may have limitations, particularly in identifying new studies that have not yet undergone review or inclusion in the existing body of literature. Terms such as “information quality,” “error check,” “data check,” “data validation,” and “data cleaning” are commonly associated with the concept of data quality, particularly in older research papers. However, we did not include these terms in our search query because subsequent checking using these terms did not reveal any additional reviews that met our inclusion criteria. Furthermore, this overview focused on published reviews. Important information can also be found in grey literature [51,52] and in studies that collect stakeholders’ opinions on the quality of health data [20]. Finally, none of the included reviews discussed patient-generated data or data generated by wearables. Given the increasing adoption and use of these sources in health care, it is becoming important to consider their impact on data quality. Developing assessment methods that are applicable to these emerging data sources is an important area for further research.

Although having a consolidated reference framework of data quality dimensions and aspects is valuable, it is also of great importance to define specific data quality requirements for each relevant aspect within a single quality dimension. These requirements should specify the desired quality level to be achieved in a given percentage of the primary sources, based on the purpose of the data collection or a particular real-world data study. Once these requirements are clearly articulated, appropriate measurement methods can be determined, thereby ensuring the comprehensive analysis of secondary data collection for its suitability for a specific purpose.

Conclusions

The absence of a consensus in the literature regarding the precise terminology and definitions of data quality dimensions has resulted in ambiguity and challenges in creating specific assessment methods. This review of reviews offers an overview of data quality dimensions, along with the definitions and assessment methods used in these reviews. This study goes a step further by assigning all observed definitions to a consolidated framework of 9 data quality dimensions. Further research is needed to complete the collection of aspects within each quality dimension, with the elaboration of a full set of assessment methods, and the establishment of specific requirements to evaluate the suitability for the purpose of secondary data collection systems.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search items by database.
[DOCX File , 21 KB - medinform_v12i1e51560_app1.docx]

Multimedia Appendix 2
Data sources, data quality aspects, and definitions reported in the 22 publications included in the review.
[DOCX File , 46 KB - medinform_v12i1e51560_app2.docx]

Multimedia Appendix 3
The frequency of all dimensions with definitions in each review and assessment methods per dimension.
[DOCX File , 169 KB - medinform_v12i1e51560_app3.docx]

Multimedia Appendix 4
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 65 KB - medinform_v12i1e51560_app4.pdf]

References

1. Duncan R, Eden R, Woods L, Wong I, Sullivan C. Synthesizing dimensions of digital maturity in hospitals: systematic review. J Med Internet Res 2022 Mar 30;24(3):e32994 [FREE Full text] [doi: 10.2196/32994] [Medline: 35353050]
2. Eden R, Burton-Jones A, Scott I, Staib A, Sullivan C. Effects of eHealth on hospital practice: synthesis of the current literature. Aust Health Rev 2018 Sep;42(5):568-578. [doi: 10.1071/AH17255] [Medline: 29986809]
3. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A survey of the literature on unintended consequences associated with health information technology: 2014–2015. Yearb Med Inform 2018 Mar 06;25(01):13-29. [doi: 10.15265/iy-2016-036]
4. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
5. Feder SL. Data quality in electronic health records research: quality domains and assessment methods. West J Nurs Res 2018 May;40(5):753-766. [doi: 10.1177/0193945916689084] [Medline: 28322657]
6. Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. JAMA Netw Open 2020 Jun 01;3(6):e205867 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.5867] [Medline: 32515797]
7. Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes. An evaluation of input errors. Methods Inf Med 2003;42(1):61-67. [Medline: 12695797]
8. Suresh G. Don't believe everything you read in the patient's chart. Pediatrics 2003 May;111(5 Pt 1):1108-1109. [doi: 10.1542/peds.111.5.1108] [Medline: 12728099]
9. Kaboli PJ, McClimon BJ, Hoth AB, Barnett MJ. Assessing the accuracy of computerized medication histories. Am J Manag Care 2004 Nov;10(11 Pt 2):872-877 [FREE Full text] [Medline: 15609741]
10. Staroselsky M, Volk LA, Tsurikova R, Newmark LP, Lippincott M, Litvak I, et al. An effort to improve electronic health record medication list accuracy between visits: patients' and physicians' response. Int J Med Inform 2008 Mar;77(3):153-160. [doi: 10.1016/j.ijmedinf.2007.03.001] [Medline: 17434337]
11. Yadav S, Kazanji N, Paudel S, Falatko J, Shoichet S, Maddens M, et al. Comparison of accuracy of physical examination findings in initial progress notes between paper charts and a newly implemented electronic health record. J Am Med Inform Assoc 2017 Jan;24(1):140-144 [FREE Full text] [doi: 10.1093/jamia/ocw067] [Medline: 27357831]
12. Darko-Yawson S, Ellingsen G. Assessing and improving EHRs data quality through a socio-technical approach. Procedia Comput Sci 2016;98:243-250. [doi: 10.1016/j.procs.2016.09.039]

13. Wang Z, Penning M, Zozus M. Analysis of anesthesia screens for rule-based data quality assessment opportunities. Stud Health Technol Inform 2019;257:473-478 [FREE Full text] [Medline: 30741242]
14. Puttkammer N, Baseman JG, Devine EB, Valles JS, Hyppolite N, Garilus F, et al. An assessment of data quality in a multi-site electronic medical record system in Haiti. Int J Med Inform 2016 Feb;86:104-116. [doi: 10.1016/j.ijmedinf.2015.11.003] [Medline: 26620698]
15. Wiebe N, Xu Y, Shaheen AA, Eastwood C, Boussat B, Quan H. Indicators of missing Electronic Medical Record (EMR) discharge summaries: a retrospective study on Canadian data. Int J Popul Data Sci 2020 Dec 11;5(1):1352 [FREE Full text] [doi: 10.23889/ijpds.v5i3.1352] [Medline: 34007880]
16. von Lucadou M, Ganslandt T, Prokosch HU, Toddenroth D. Feasibility analysis of conducting observational studies with the electronic health record. BMC Med Inform Decis Mak 2019 Oct 28;19(1):202 [FREE Full text] [doi: 10.1186/s12911-019-0939-0] [Medline: 31660955]
17. Juran JM, Gryna FM, Bingham RS. Quality Control Handbook. New York, NY: McGraw-Hill; 1974.
18. Ehrlinger L, Wöß W. A survey of data quality measurement and monitoring tools. Front Big Data 2022;5:850611 [FREE Full text] [doi: 10.3389/fdata.2022.850611] [Medline: 35434611]
19. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
20. Aerts H, Kalra D, Saez C, Ramírez-Anguita JM, Mayer MA, Garcia-Gomez JM, et al. Is the quality of hospital EHR data sufficient to evidence its ICHOM outcomes performance in heart failure? A pilot evaluation. medRxiv. Preprint posted online February 5, 2021 2021. [doi: 10.1101/2021.02.04.21250990]
21. Ge M, Helfert M. A review of information quality research - develop a research agenda. In: Proceedings of the 2007 MIT International Conference on Information Quality. 2007 Presented at: MIT ICIQ '07; November 9-11, 2007; Cambridge, MA URL: http://mitiq.mit.edu/iciq/pdf/a%20review%20of%20information%20quality%20research.pdf
22. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
23. Syed R, Eden R, Makasi T, Chukwudi I, Mamudu A, Kamalpour M, et al. Digital health data quality issues: systematic review. J Med Internet Res 2023 Mar 31;25:e42615 [FREE Full text] [doi: 10.2196/42615] [Medline: 37000497]
24. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc 2020 Dec 09;27(12):1999-2010 [FREE Full text] [doi: 10.1093/jamia/ocaa245] [Medline: 33166397]
25. Shivasabesan G, Mitra B, O'Reilly GM. Missing data in trauma registries: a systematic review. Injury 2018 Sep;49(9):1641-1647. [doi: 10.1016/j.injury.2018.03.035] [Medline: 29678306]
26. Porgo TV, Moore L, Tardif PA. Evidence of data quality in trauma registries: a systematic review. J Trauma Acute Care Surg 2016 Apr;80(4):648-658. [doi: 10.1097/TA.0000000000000970] [Medline: 26881490]
27. Prang KH, Karanatsios B, Verbunt E, Wong HL, Yeung J, Kelaher M, et al. Clinical registries data quality attributes to support registry-based randomised controlled trials: a scoping review. Contemp Clin Trials 2022 Aug;119:106843. [doi: 10.1016/j.cct.2022.106843] [Medline: 35792338]
28. Nesca M, Katz A, Leung C, Lix L. A scoping review of preprocessing methods for unstructured text data to assess data quality. Int J Popul Data Sci 2022 Oct 05;7(1):1-15 [FREE Full text] [doi: 10.23889/ijpds.v7i1.1757]
29. Ozonze O, Scott PJ, Hopgood AA. Automating electronic health record data quality assessment. J Med Syst 2023 Feb 13;47(1):23 [FREE Full text] [doi: 10.1007/s10916-022-01892-2] [Medline: 36781551]
30. AbuHalimeh A. Improving data quality in clinical research informatics tools. Front Big Data 2022;5:871897 [FREE Full text] [doi: 10.3389/fdata.2022.871897] [Medline: 35574572]
31. Liaw S, Guo JG, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. J Am Med Inform Assoc 2021 Jul 14;28(7):1591-1599 [FREE Full text] [doi: 10.1093/jamia/ocaa340] [Medline: 33496785]
32. Rajan NS, Gouripeddi R, Mo P, Madsen RK, Facelli JC. Towards a content agnostic computable knowledge repository for data quality assessment. Comput Methods Programs Biomed 2019 Aug;177:193-201. [doi: 10.1016/j.cmpb.2019.05.017] [Medline: 31319948]
33. Mashoufi M, Ayatollahi H, Khorasani-Zavareh D. A review of data quality assessment in emergency medical services. Open Med Inform J 2018 May 31;12(1):19-32 [FREE Full text] [doi: 10.2174/1874431101812010019] [Medline: 29997708]
34. Fung JW, Lim SBL, Zheng H, Ho WY, Lee BG, Chow KY, et al. Data quality at the Singapore cancer registry: an overview of comparability, completeness, validity and timeliness. Cancer Epidemiol 2016 Aug;43:76-86. [doi: 10.1016/j.canep.2016.06.006] [Medline: 27399312]
35. O'Reilly GM, Gabbe B, Moore L, Cameron PA. Classifying, measuring and improving the quality of data in trauma registries: a review of the literature. Injury 2016 Mar;47(3):559-567. [doi: 10.1016/j.injury.2016.01.007] [Medline: 26830127]
36. Stausberg J, Nasseh D, Nonnemacher M. Measuring data quality: a review of the literature between 2005 and 2013. Stud Health Technol Inform 2015;210:712-716. [Medline: 25991245]

37. Chen H, Yu P, Hailey D, Wang N. Methods for assessing the quality of data in public health information systems: a critical review. Stud Health Technol Inform 2014;204:13-18. [Medline: 25087521]
38. Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. Int J Environ Res Public Health 2014 May 14;11(5):5170-5207 [FREE Full text] [doi: 10.3390/ijerph110505170] [Medline: 24830450]
39. Liaw ST, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, et al. Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature. Int J Med Inform 2013 Jan;82(1):10-24. [doi: 10.1016/j.ijmedinf.2012.10.001] [Medline: 23122633]
40. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer 2009 Mar;45(5):747-755. [doi: 10.1016/j.ejca.2008.11.032] [Medline: 19117750]
41. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. Eur J Cancer 2009 Mar;45(5):756-764. [doi: 10.1016/j.ejca.2008.11.033] [Medline: 19128954]
42. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002;9(6):600-611 [FREE Full text] [doi: 10.1197/jamia.m1087] [Medline: 12386111]
43. Lindquist M. Data quality management in pharmacovigilance. Drug Saf 2004;27(12):857-870. [doi: 10.2165/00002018-200427120-00003] [Medline: 15366974]
44. Haug A. Understanding the differences across data quality classifications: a literature review and guidelines for future research. Ind Manag Data Syst 2021 Aug 24;121(12):2651-2671. [doi: 10.1108/imds-12-2020-0756]
45. Triki Z, Bshary R. A proposal to enhance data quality and FAIRness. Ethol 2022 Aug 02;128(9):647-651. [doi: 10.1111/eth.13320]
46. Šlibar B, Oreški D, Begičević Ređep NB. Importance of the open data assessment: an insight into the (meta) data quality dimensions. SAGE Open 2021 Jun 15;11(2):215824402110231. [doi: 10.1177/21582440211023178]
47. Verma R. Data quality and clinical audit. Intensive Care Med 2012 Aug;13(8):397-399. [doi: 10.1016/j.mpaic.2012.05.009]
48. Lima CR, Schramm JM, Coeli CM, da Silva ME. [Review of data quality dimensions and applied methods in the evaluation of health information systems]. Cad Saude Publica 2009 Oct;25(10):2095-2109 [FREE Full text] [doi: 10.1590/s0102-311x2009001000002] [Medline: 19851611]
49. Correia LO, Padilha BM, Vasconcelos SM. [Methods for assessing the completeness of data in health information systems in Brazil: a systematic review]. Cien Saude Colet 2014 Nov;19(11):4467-4478 [FREE Full text] [doi: 10.1590/1413-812320141911.02822013] [Medline: 25351313]
50. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
51. European health data space data quality framework. European Union's 3rd Health Programme. 2022. URL: https://tehdas.eu/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf [accessed 2024-01-29]
52. Data quality framework for EU medicines regulation. European Medicines Agency. 2023. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf [accessed 2024-01-29]

Abbreviations

EHR: electronic health record
ETL: export-transform-load
i~HD: European Institute for Innovation through Health Data
ISO: International Organization for Standardization
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by C Lovis; submitted 03.08.23; peer-reviewed by D Courvoisier, Z Wang; comments to author 16.09.23; revised version received 07.11.23; accepted 09.01.24; published 06.03.24.

Please cite as:
Declerck J, Kalra D, Vander Stichele R, Coorevits P
Frameworks, Dimensions, Definitions of Aspects, and Assessment Methods for the Appraisal of Quality of Health Data for Secondary Use: Comprehensive Overview of Reviews
JMIR Med Inform 2024;12:e51560
URL: https://medinform.jmir.org/2024/1/e51560
doi:10.2196/51560
PMID:38446534

Review

The Key Digital Tool Features of Complex Telehealth Interventions Used for Type 2 Diabetes Self-Management and Monitoring With Health Professional Involvement: Scoping Review

Choumous Mannoubi[1,2], RDT, MSc; Dahlia Kairy[1,2], PT, PhD; Karla Vanessa Menezes[1,2], PT, PhD; Sophie Desroches[3,4,5], RD, PhD; Geraldine Layani[6,7], MSc, MD; Brigitte Vachon[1,8], OTR, PhD

[1]School of Rehabilitation, Université de Montréal, Montreal, QC, Canada

[2]Centre interdisciplinaire en readaptation du Montreal Métropolitain, Institut Universitaire sur la readaptation en déficience physique de Montreal, Montréal, QC, Canada

[3]Institute of Nutrition and Functional Foods, Université Laval, Quebec, QC, Canada

[4]Centre nutrition, sante´ et socie´te´ NUTRISS, Université Laval, Québec, QC, Canada

[5]School of Nutrition, Université Laval, Québec, QC, Canada

[6]Centre de recherche du centre hospitalier de l'universite de Montreal, Montréal, QC, Canada

[7]Département de médecine de famille et de médecine d'urgence, Universté de Montréal, Montreal, QC, Canada

[8]Centre de recherche de l'Institut universitaire en santé mentale de Montréal, Centre integre de sante et de services sociaux de l'Est-de-l'ile-de-Montreal, Montréal, QC, Canada

Corresponding Author:
Choumous Mannoubi, RDT, MSc
School of Rehabilitation
Université de Montréal
7077, avenue du Parc
Montreal, QC, H3N 1X7
Canada
Phone: 1 5143436111
Email: cmannoubi@gmail.com

Abstract

Background: Therapeutic education and patient self-management are crucial in diabetes prevention and treatment. Improving diabetes self-management requires multidisciplinary team intervention, nutrition education that facilitates self-management, informed decision-making, and the organization and delivery of appropriate health care services. The emergence of telehealth services has provided the public with various tools for educating themselves and for evaluating, monitoring, and improving their health and nutrition-related behaviors. Combining health technologies with clinical expertise, social support, and health professional involvement could help persons living with diabetes improve their disease self-management skills and prevent its long-term consequences.

Objective: This scoping review's primary objective was to identify the key digital tool features of complex telehealth interventions used for type 2 diabetes or prediabetes self-management and monitoring with health professional involvement that help improve health outcomes. A secondary objective was to identify how these key features are developed and combined.

Methods: A 5-step scoping review methodology was used to map relevant literature published between January 1, 2010 and March 31, 2022. Electronic searches were performed in the MEDLINE, CINAHL, and Embase databases. The searches were limited to scientific publications in English and French that either described the conceptual development of a complex telehealth intervention that combined self-management and monitoring with health professional involvement or evaluated its effects on the therapeutic management of patients with type 2 diabetes or prediabetes. Three reviewers independently identified the articles and extracted the data.

Results: The results of 42 studies on complex telehealth interventions combining diabetes self-management and monitoring with the involvement of at least 1 health professional were synthesized. The health professionals participating in these studies were physicians, dietitians, nurses, and psychologists. The digital tools involved were smartphone apps or web-based interfaces that could be used with medical devices. We classified the features of these technologies into eight categories, depending on the intervention objective: (1) monitoring of glycemia levels, (2) physical activity monitoring, (3) medication monitoring, (4) diet

monitoring, (5) therapeutic education, (6) health professional support, (7) other health data monitoring, and (8) health care management. The patient-logged data revealed behavior patterns that should be modified to improve health outcomes. These technologies, used with health professional involvement, patient self-management, and therapeutic education, translate into better control of glycemia levels and the adoption of healthier lifestyles. Likewise, they seem to improve monitoring by health professionals and foster multidisciplinary collaboration through data sharing and the development of more concise automatically generated reports.

Conclusions: This scoping review synthesizes multiple studies that describe the development and evaluation of complex telehealth interventions used in combination with health professional support. It suggests that combining different digital tools that incorporate diabetes self-management and monitoring features with a health professional's advice and interaction results in more effective interventions and outcomes.

(JMIR Med Inform 2024;12:e46699) doi:10.2196/46699

KEYWORDS

telehealth; telemedicine; telenutrition; telemonitoring; electronic coaching; e-coaching; scoping review; type 2 diabetes; prediabetes; diabetes management; diabetes self-management; mobile phone

Introduction

Diabetes and Nutrition

The prevalence of diabetes in Canada is constantly rising, and related health expenditures are among the highest in the world. In 2018, approximately 8% of the Canadian population was living with this disease, and it is predicted that in 2025, a total of 5 million people will be affected (ie, 12.1% of the population) [1,2]. According to estimates, type 2 diabetes accounts for 90% of all diabetes diagnoses in the general population, type 1 diabetes accounts for 9%, and other kinds of diabetes account for 1% [3]. The prevalence of diabetes has been closely linked to dietary and lifestyle factors prevalent within the country, such as high rates of obesity and sedentary behavior coupled with a diet often rich in processed foods. However, best practice guidelines suggest that the onset of type 2 diabetes can be delayed or prevented using early lifestyle change interventions. As prediabetes is characterized by elevated blood glucose levels that do not yet meet the diagnostic criteria for diabetes, the therapeutic management of diabetes and prediabetes is similar [4,5]. In both cases, a comprehensive approach is required to better control glycemia levels [6,7]. Many factors are involved in preventing the disease and achieving better disease control, such as changing lifestyles through education, supporting self-management, and preventing the development and progression of complications [8]. The Diabetes Canada clinical practice guidelines recommend that individuals with diabetes receive personalized nutrition counseling by a registered dietitian to optimize glycemic control and weight management [3]. Strategies include caloric reduction for individuals who are overweight; the incorporation of low glycemic index carbohydrates; and the adoption of a Mediterranean, Nordic, Dietary Approaches to Stop Hypertension (DASH), or vegetarian diet because they are rich in protective foods [3]. These interventions are supported by evidence demonstrating improvements in glycated hemoglobin (HbA_{1c}) levels, metabolic outcomes, and reductions in hospitalization rates. As stated in the Diabetes Canada clinical practice guidelines, the care offered should be organized around the needs of people with diabetes (and of their families and close friends) because patients must be active participants for optimal engagement in self-managing their condition [4,8]. This active patient participation must be facilitated by a multidisciplinary team (nurses, dietitians, and physicians) that offers education and self-management support. Changing dietary behaviors poses a considerable challenge for people living with diabetes, yet it is a vital means of preventing the associated complications [4]. Monitoring with a dietitian's involvement has proven effective in supporting such behavior changes [4]. Again according to the Diabetes Canada clinical practice guidelines, all people living with diabetes should receive the services of a dietitian [4]. It has been shown that diet monitoring with a dietitian's involvement can alone reduce HbA_{1c} levels by 1% to 2% [4]. In addition, recent evidence underscores the advantages of using telehealth to foster adherence to medical recommendations and self-management [4,5,9]. Scientific literature has shown the benefits of telehealth in Canada for diabetes management [3,10]. These technological innovations facilitate patient monitoring and promote the use of different interventions that can support lifestyle changes through, for example, remote support, the telemonitoring of glycemia levels, reminders about taking medication, and the use of a food diary. These innovations also allow this information to be shared with the health care team. In 2018, the Diabetes Canada clinical practice guidelines advocated for the use of telehealth in disease management programs to improve self-management in underserved communities and to facilitate consultation with specialized teams, highlighting its effectiveness and the importance of integrating it into shared care models [3].

Telehealth and Diabetes Self-Management

Telehealth refers to "the use of communications and information technology to deliver health and health care services and information over large and small distances" [11]. In the same field of application, telemedicine refers to the exchange of medical information using information and communication technologies to improve a patient's health condition and is delivered by at least 1 health professional [12]. Telemedicine services are provided using various means, including the telephone, internet, email, mobile apps, SMS text messaging, photographs, and videos. New technologies are revolutionizing the health care field by creating new prospects for various care delivery modalities [13]. They are thus paving the way for

innovations and represent a real benefit in the face of new health care challenges, such as the aging population, rising health care costs, and the unprecedented challenges posed by pandemics such as the COVID-19 pandemic [6]. Particularly in Canada, the public health care system faces challenges often associated with overcrowded clinics, long wait times, and limited resources [7]. Through remote consultations and continuous monitoring, telehealth has the potential to relieve pressure on health care facilities, improving resource allocation and optimizing patient flow management in the public health care system. As such, telehealth would be a pertinent response to public health organizational challenges in the Canadian context, where the universal health care system aims to provide equitable and accessible care to all residents.

The day-to-day management of type 2 diabetes can be a complex challenge. Patients must monitor their blood glucose levels regularly, take medication on a precise schedule, adopt a balanced diet, and maintain adequate physical activity [7]. However, these requirements can be difficult to meet owing to time constraints, a lack of knowledge, or limited resources. In addition, fluctuations in blood glucose levels can occur unpredictably, increasing the risk of short- and long-term complications [7]. In particular, nutrition plays a fundamental role in diabetes management. Dietary monitoring, nutrition education, and the personalization of dietary recommendations are key aspects in optimizing health outcomes for patients with diabetes. Using digital technologies, it is possible to offer ongoing personalized nutrition support, enabling patients to make informed dietary decisions and maintain adequate glycemic control.

Recent evidence points to the enormous potential of using health technologies to facilitate access to care, patient adherence to their treatment plan, and self-management [14]. Many experts point out that diabetes is a chronic disease best adapted to self-management through telehealth [14-19]. Technological innovations have been developed to support lifestyle changes and facilitate patient monitoring. Telehealth offers a range of potential benefits for people with type 2 diabetes. Continuous monitoring of blood glucose levels using connected sensors enables patients to receive real-time information on their blood glucose levels and be alerted to abnormal variations [2,3]. This enables them to take immediate action to correct blood glucose levels and avoid complications. In addition, telehealth facilitates access to specialized care by enabling patients to consult health professionals remotely. This reduces geographic barriers and enables patients to receive personalized advice, education, and support tailored to their specific needs [9]. Regular monitoring and feedback as well as the use of digital tools encourage patients to better understand their condition, make informed decisions, and improve their quality of life [8]. According to recent systematic reviews and meta-analyses, these telehealth interventions involving everyday web-based and mobile technologies help reduce HbA_{1c} levels, allow for better daily glycemic control, promote an increase in physical activity, and improve dietary habits [20,21]. Connected blood glucose meters enable more convenient and accurate monitoring of blood glucose levels, whereas web-based platforms offer a web-based space for education, support, and communication with health professionals [14,15]. Teleconsultation enables patients to consult their physicians and specialists remotely, reducing travel and time constraints [15,16].

Combining self-management technologies with clinical expertise, social support, and health professional involvement can allow the development of telehealth solutions better adapted to the therapeutic management of patients with a chronic disease. Telehealth interventions using this combination are therefore expanding [22], but they present both advantages and limitations [12]. Telehealth enables improved care coordination, personalized interventions, and tailored patient education. However, it can lead to an increased workload for health care providers and raise data privacy concerns. The tension between interventions focused on service delivery and those involving health care providers highlights the importance of striking a balance between patient autonomy and medical expertise. An integrated collaborative approach involving both patients and health care providers may offer the best digital health outcomes. However, further studies are needed to fill the gaps in the literature, focusing on comparative studies with usual care, the evaluation of adherence, and long-term accessibility to optimize the use of telehealth in the self-management of type 2 diabetes.

To the best of our knowledge, no literature review has been conducted to identify the key digital tool features of such interventions. Nonetheless, improving knowledge on this subject could advance the development of more effective telehealth interventions for people with diabetes.

The primary objective of this scoping review was to identify the key digital tool features of complex telehealth interventions used for diabetes self-management and monitoring with health professional involvement that help improve health outcomes. The secondary objective was to identify how these key features should be developed and combined to optimize their contribution to improving health outcomes. Although our review draws from global scientific literature, the intent is to inform the future development of telehealth technologies, with a particular emphasis on the Canadian health care context. This focus stems from the recognition that although universal principles may guide the development of digital health tools, the specific features and their implementation must be tailored to meet the unique needs, regulations, and health care infrastructure of Canada. Our review aims to explicitly identify the characteristics of digital tools that have been shown to be effective in improving patient engagement, improving self-management, and leading to better health outcomes in diabetes care. By systematically cataloging these characteristics, we can provide a model for the design, development, and implementation of future telehealth interventions, provided we keep in mind specific requirements of the Canadian health care context, such as compliance with telehealth policies, local health care, patient privacy laws, and existing health IT infrastructure. In this study, *improving health outcomes* encompasses both the positive effects of the intervention on behavior changes (eg, eating healthier foods or performing physical activity) and the positive impacts on the health condition (eg, improved blood glucose levels or blood pressure).

Methods

Overview

Scoping reviews exhaustively synthesize the evidence to map a vast, complex, or emerging field of study and identify gaps in the literature, ultimately highlighting priorities for future studies in the field [23]. We chose this method because telehealth has emerged in different formats and offers solutions to various pathologies. We structured our scoping review according to the five steps developed by Arksey and O'Malley [24] and the revisions made by Levac et al [25]: (1) identifying the research question; (2) identifying relevant studies; (3) selecting the studies; (4) charting the data; and (5) collating, summarizing, and reporting the results. The procedure, which is described in the following subsections, was conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist (Multimedia Appendix 1) to ensure rigorous and transparent reporting of the methodology and findings [26]. Several additional recommendations made by Levac et al [25] were also followed: clearly articulate the research question for the scoping review, have 2 researchers independently review the full articles to determine their inclusion, have the research team collectively develop the data-charting form, and continually extract data.

Identifying the Research Questions

This review seeks to answer the following research questions:

1. What are the key digital tool features of complex telehealth interventions used for diabetes self-management and monitoring with health professional involvement that help improve health outcomes?
2. How should these key features be developed and combined to help improve health outcomes?

These questions stem from the lack of consensus in scientific literature on the conceptual development, implementation, and evaluation of telehealth solutions. The research questions and objectives were developed based on the research team's expertise and a preliminary analysis of the literature on the subject. In accordance with scoping review methodology, this review included studies that used different approaches and research designs.

In this review, we applied the World Health Organization definition of telemedicine: "The delivery of health care services, where distance is a critical factor, by all health professionals using information and communication technologies for the exchange of valid information for diagnosis, treatment and prevention of disease and injuries, research and evaluation, and for the continuing education of health care providers, all in the interests of advancing the health of individuals and their communities." Furthermore, in the context of telehealth technology, the term *features* refers to the various components or tools that enable the various activities associated with remote health care delivery.

Identifying and Selecting the Studies

The search strategy was developed in collaboration with a Université de Montréal librarian specializing in health. The keywords based on *telehealth*, *nutrition*, and *diabetes* were identified by examining relevant articles, their references, and the associated keywords (Multimedia Appendix 2). A systematic search was performed in the MEDLINE, CINAHL, and Embase databases, covering the period from January1, 2010, to March 31, 2022. Our search efforts were focused on these databases because they are repositories where studies related to health and nutrition can be found. Only articles published since January 1, 2010, were selected to account for the widespread adoption of smartphones. By extending our review to cover more than a decade, we were able to capture the significant developments in mobile apps and smartphone use, which are pivotal in digital health. We also perused the bibliographies of the included articles to identify any additional studies. Only articles published in peer-reviewed scientific journals were examined. As proposed by the framework developed by Arksey and O'Malley [24], a quality assessment was not performed because it is not deemed essential for exploratory studies. The methodological rigor of the published articles was not an inclusion or exclusion criterion; instead, the articles were examined to substantiate the results and the discussion.

Given the rapid development of new technologies, only articles on complex telehealth interventions for managing diabetes published in the 12 years covering the period from January 1, 2010, to March 31, 2022, were retained. We used an iterative process to develop the inclusion and exclusion criteria during our searches to ensure a selection of studies more closely aligned with the research question. The searches were limited to scientific publications in English and French that either described the conceptual development of a complex telehealth intervention combining self-management and monitoring with health professional involvement or evaluated its effects on the therapeutic management of patients with type 2 diabetes or prediabetes. For inclusion in this review, the complex interventions had to be digital, have a patient interface, and concern type 2 diabetes or prediabetes self-management or monitoring. We excluded studies (1) not using a nutritional approach to investigate telehealth interventions, (2) involving a single component, (3) not integrating at least 1 health professional, (4) concerning type 1 diabetes or gestational diabetes, (5) involving populations aged <18 years, and (6) lacking empirical data (eg, literature reviews). All search results were imported into the Covidence reference management software (Veritas Health Innovation Ltd), and duplicates were removed [27].

The review team comprised CM, DG, KVM, and BV. These 4 researchers determined the inclusion of relevant studies based on the title and abstract; CM and BV determined the selection based on the full-text articles. Differences were discussed in detail until a consensus was reached. The full texts of the relevant articles were retrieved for more in-depth analysis (CM).

Charting the Data

The research team developed a data extraction table. It included the following information: study characteristics (eg, title,

participants, the results of interest, and effectiveness), intervention characteristics (eg, a brief description of the intervention, the components of self-management, and the components of monitoring with health professional involvement), and the benefits and limitations of both the intervention and the study according to the authors or reviewers.

Collecting, Summarizing, and Reporting the Results

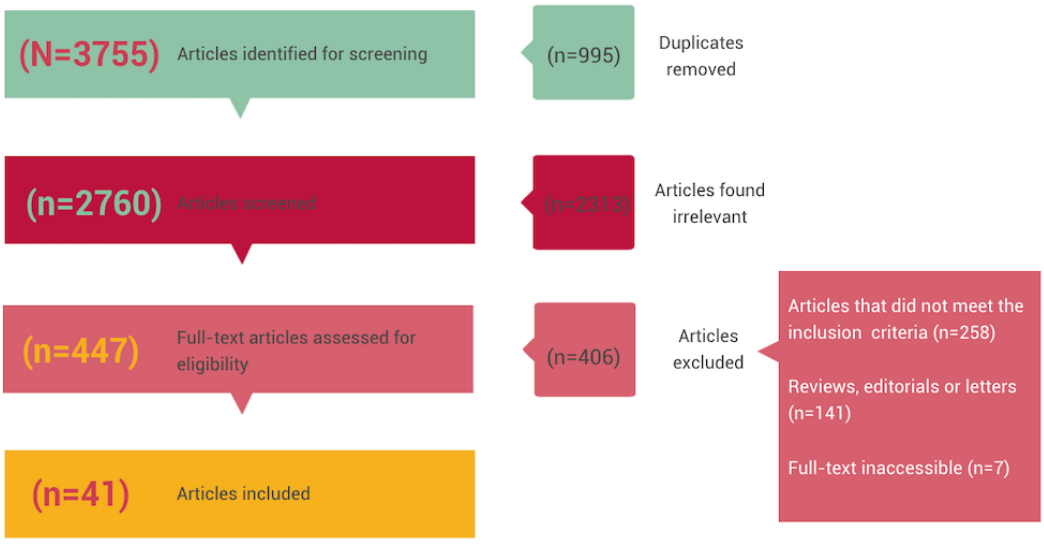
Again according to the framework developed by Arksey and O'Malley [24] and the revisions by Levac et al [25], descriptive web-based abstracts and thematic analyses performed with NVivo software (release 1.7; Lumivero) were used for data analysis, yielding an approach resembling that of a narrative review. In conducting our thematic analysis, we adopted a qualitative approach to discern the impact of telehealth interventions with health professionals on the health outcomes of patients with diabetes. Through meticulous data immersion and iterative coding, we identified recurring patterns that we then shaped into themes. An initial list of these codes, forming a codebook, was iteratively refined during the data analysis process [28]. Once the codes were established, it enabled a comprehensive review of their interrelationships, aiding in the identification of the key digital tool features of complex telehealth interventions used for diabetes self-management and monitoring with health professional involvement that help improve health outcomes. These themes were refined against the data set to ensure coherence and direct relation to our research objectives. By integrating concrete examples from the data, we were able to provide a rich, detailed description of the telehealth features, thereby adding depth to our findings and ensuring that they were both representative of real-world practices and aligned with our research questions.

Results

Overview

The database searches identified 3755 articles, from which 995 (26.5%) duplicates were removed. The 2760 remaining articles underwent an initial screening based on the abstract and title, after which 2313 (83.8%) were excluded. The full-text screening involved assessing 447 articles, of which 406 (90.8%) were deemed ineligible because the studies did not meet the inclusion criteria (n=258, 63.7%); were literature reviews, editorials, or letters (n=141, 34.8%); or the full texts were inaccessible (n=7, 1.7%; Figure 1). Thus, of the 3755 articles identified from the database searches, 42 (1.12%) were ultimately included in this scoping review (Multimedia Appendix 3 [29-70]). The qualitative analysis of the 42 articles using NVivo (release 1.7) yielded the coding of 1520 references, divided among 113 codes.

Figure 1. Flow diagram of study selection.



Characteristics of the Studies

The 42 studies were published between January 1, 2010, and March 31, 2022, with as many as 28 (67%) published within the past 6 years [29-56]. We found that, in 2021, nearly twice as many articles were published on the topic as in each of the previous 4 years (Figure 2).

Information on complex telehealth interventions used for diabetes self-management and monitoring with health professional involvement was obtained for 18 countries. Of the 42 studies, 11 (26%) were conducted in the United States [30,31,39,48-51,57-59]; 5 (12%) in South Korea [37,41,44,60,61]; 4 (10%) in Singapore [29,43,46,55]; 4 (10%) in Norway [32,62,63]; 3 (7%) in the United Kingdom [33,35,38]; 3 (7%) in Germany [40,45,56]; 2 (5%) in China [47,64]; and 1 (2%) each in Australia [54], South Africa [65], Spain [66], Iran [52], Italy [67], Japan [42], Lebanon [34], Slovenia [36], Switzerland, and Taiwan [68] (Figure 3).

Figure 2. Years in which the studies were published. Each circle represents 1 study.

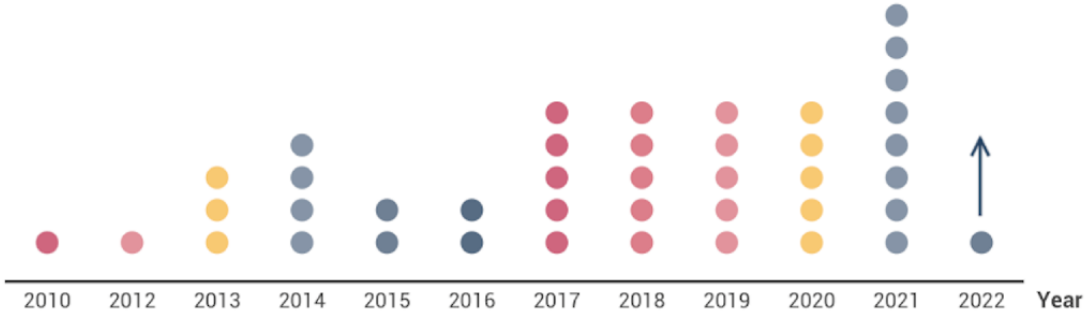
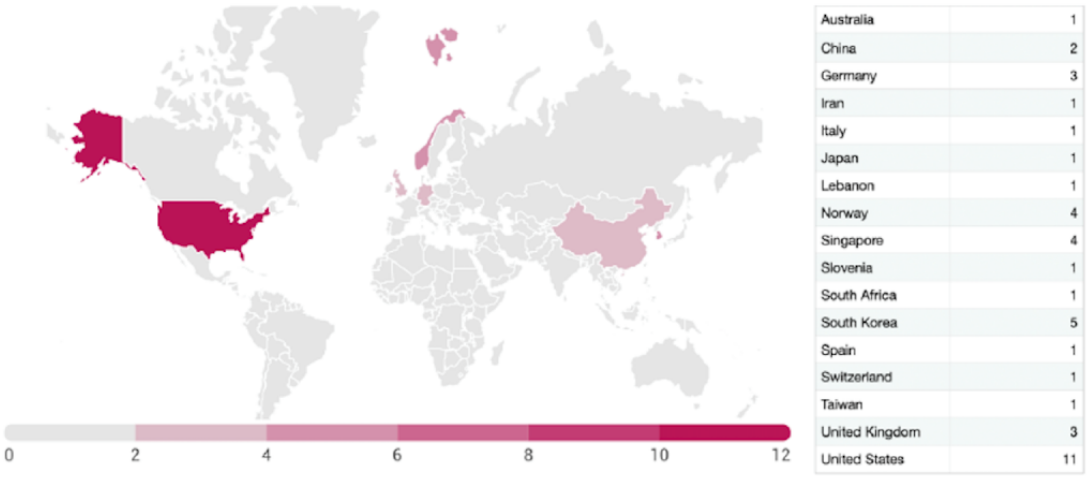


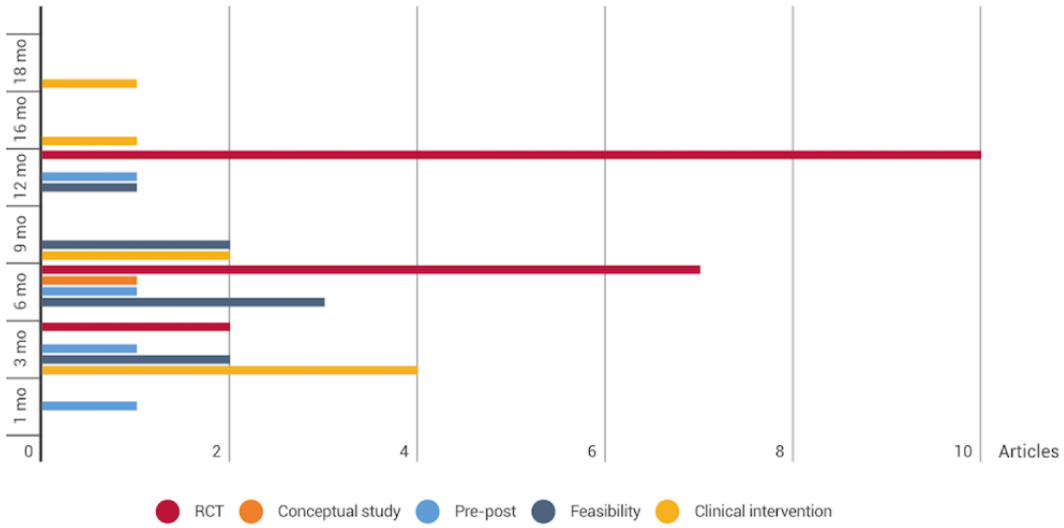
Figure 3. Countries in which the studies were published. The dots represent articles and the x-axis denotes the years.



General Characteristics of the Intervention

One-third (14/42, 33%) of the studies were randomized controlled trials [35,36,40-42,46,47,55,57-59,62,64,67], with most of them (n=12, 86%) ranging from 6 months to 1 year in duration. Of the 42 studies, 9 (21%) were feasibility studies, with the interventions ranging from 3 months to 1 year in duration [33,38,39,43,44,51,53,56,71]; 8 (19%) were interventional studies, with the interventions ranging from 3 to 18 months in duration [30,31,34,48,50,54,63,68]; 5 (12%) were conceptual studies lasting 6 months [45,49,52,65,66]; and 4 (10%) were pre-post studies, in which the interventions ranged from 1 month to 1 year in duration [29,37,60,61] (Figure 4).

Figure 4. Durations of the interventions. RCT: randomized controlled trial.



Health Professional Involvement

Of the 42 studies, 21 (50%) included physicians [29-32,36,40-42,45,47-51,53,58,60,64,66,68,69], 16 (38%) involved dietitians [29,31,33,35,39,43,46,54,56,59,60,65,68-70,72], 12 (29%) involved nurses [31,32,36,41,55,58-62,68,69], 4 (10%) involved psychologists [31,33,67,69], 4 (10%) involved physical educators [29,33,35,60], and 3 (7%) involved case

managers [36,60,68]. Finally, of the 42 studies, 13 (31%) involved a multidisciplinary team [29,31-33,35,36,41,58-60,68,69,72], and 22 (52%) involved only 1 clinician [30,39,40,42,43,45-51,53-56,61,62,64-66,70] (Multimedia Appendix 4 [29-70]).

Characteristics of Digital Self-Management

The interventions under study involved the use of a mobile app [29-34,36-39,41-46,49,51-53,55,59-65,67,69,70] or a web portal [32,35,36,40-42,44,45,47,48,52,57,58,60,63,66-68,70], usually coupled with a blood glucose meter to optimize diabetes self-management [37,38,40,41,43,44,46,48-51,53,55,57,59-64,68,69]. Other Bluetooth-connected devices were used in some of the interventions (10/42, 24%), such as a Bluetooth-connected weight scale [31,40,42,43,46,48], a pedometer [40,43,45], an accelerometer [33,42], a Bluetooth-connected smartwatch [49], and a tensiometer [42].

The types of data collected concerned the monitoring of glycemia levels through, for example, the visualization of a blood sugar curve over time [29,32,37,38,41,43-49,51-53,55,59-66,68,69]; physical activity monitoring using, for example, a pedometer [33,35,40,43,45,46,56,61,67]; diet monitoring using, for example, a food diary [29-33,35,37,39,41,43-45,47,49,52,53,55,56,59,61-63,67-69]; medication monitoring through, for example, adherence monitoring or the possibility of issuing remote prescriptions [30,50]; and other health data monitoring (weight, BMI, and laboratory tests) [29,32,33,40,43,45-47,53,56,60,65,66]. Other features made it possible to ensure continuity of care by, for example, generating reports [34,38,42,45,47,52,60,66,67,70]; supporting therapeutic patient education; and ensuring support from a health professional to help patients learn and develop skills to independently manage their chronic disease and improve their quality of life [15,16].

On the basis of our analysis of the literature, we classified the key digital tool features that can have a positive impact on intervention outcomes into eight categories: (1) monitoring of glycemia levels, (2) diet monitoring, (3) physical activity monitoring, (4) medication monitoring, (5) therapeutic education, (6) health professional support, (7) other health data monitoring, and (8) health care management (Multimedia Appendix 5 [29-70]).

Key Digital Tool Features With Positive Impacts on the Health Condition

Monitoring of Glycemia Levels

Of the 42 studies, 22 (52%) incorporated a blood glucose meter to precisely monitor blood glucose levels during interventions; the blood glucose meter allowed the visual tracking of blood sugar curves by the patient and health professionals [37,38,40,41,43,44,46,48-51,53,55,57,59-64,68,69]. In addition, 3 (7%) of the 42 studies included blood glucose meters permitting real-time continuous blood glucose monitoring [29,30,50].

Of the 42 studies, 4 (10%) included an alert system [36,52,68,73]: "The online diabetes self-management system sent an SMS text message to care providers when the data exceeded the alerting range" [68]; "The application automatically sent users reminders by simple e-mail and SMS: 'Please enter your blood sugar/or other parameters into the eDiabetes application'" [36]. Of the 42 studies, 9 (21%) included a bolus dosing system [32,38,45,55,57-59,66,74]: "An optional bolus dosing feature was available as an algorithm on the e-diary that allowed the patient to generate a premeal bolus insulin dose" [57]. Of the 42 studies, 2 (5%) allowed the remote prescription of real-time continuous blood glucose monitoring devices [30,50].

The 42 studies used different indicators to collect glycemic control data, such as (1) HbA_{1c} levels in 27 (64%) studies, monitored through blood tests [29-34,36,40,41,43,44,46-48,51,54,55,57-60,62-64,67-69]; (2) blood glucose levels in 24 (57%) studies, monitored using data recorded by a blood glucose meter or a blood test [32-34,36,40,41,43,44,46-48,50,53-55,57-60,63,64,66,68,69]; and (3) hypoglycemia events in 4 (10%) studies [55,57,60,69], based on self-reports or alert systems after the recording of blood glucose levels with a blood glucose meter. All interventional studies included in the review reported a reduction of between 0.433 mmol/L and 1.554 mmol/L in fasting blood glucose levels. The studies reported a statistically significant decrease in HbA_{1c} levels ranging from 0.5% to 1.65% [34,40,41,57,68], as well as a drop of up to 1.554 mmol/L in blood glucose levels [29-31,36,41-44,46,48,56,57,59,61,62,68,69].

Diet Monitoring

Of the 42 studies, 13 (31%) included a meal planning system, with features such as generating shopping lists and recipes and calculating caloric intake [29,31,35,38,43,44,46,48,49,52,53,65,68]; and 27 (64%) included a food diary system that could be shared with the health professional for comment [29-33,35,37-39,41,43-47,49,52,53,55,56,59,61-63,67-69]. Patients logged their data using a list of foods or by taking photographs. A caloric intake–counting feature was available in 11 (26%) of the 42 studies [29,35,38,43,44,48,49,52,53,65,68]. Of the 42 studies, 5 (12%) included a carbohydrate-counting system [32,46,49,53,66]: "The app provided an automated individualized calorie limit which was computed based on body weight, gender, age and activity level. The total daily carbohydrate intake was restricted to 40% of total daily calories" [46]; "From the nutrition screen, the test persons manually entered carbohydrate values for their meals or scanned products to import the carbohydrate data into the app" [53]. Of the 42 studies, 18 (43%) included pedagogical material, particularly nutrition education and knowledge evaluation [31,33-37,46-48,51,52,55,59,60,63,64,66,69]. To collect data on diet, the studies used the data logged on mobile or internet platforms or obtained from food diaries, 24-hour reminders, or calorie counting [32,46,49,53,66]. The health professionals evaluated diet quality using the shared data or validated questionnaires (eg, the Healthy Eating Index). The studies reported a better understanding of nutritional issues, greater confidence in maintaining a healthy diet, and an improvement in dietary behavior [30,31,40,41,44,61,68,70].

Physical Activity Monitoring

Of the 42 studies, 6 (14%) monitored physical activity using a Bluetooth-connected device (Bluetooth-connected watch [49], pedometer [40,43,45], or accelerometer [33,42]), 8 (19%) used step counting via a Bluetooth-connected pedometer or a smartphone-integrated feature [33,35,40,45,46,56,61,67], and 16 (38%) included a graphic monitoring tool for monitoring physical activity [29,32,35,37,41,43-45,49,52,55,62-65,69]. These graphs were generated automatically using pedometer data or after patients' manual logging of their activities based on a list of predefined physical activities. A caloric expenditure–counting feature was often available: "Type, time, and intensity of any completed physical activity, which could be translated into calories burned. (BCT: prompt self-monitoring of behavior; provide feedback on performance)" [35]. The studies used data logged on mobile or internet platforms and obtained from pedometers, accelerometers, or self-reported physical activity diaries to collect physical activity data. These data made it possible to adjust the automated recommendation messages and the messages from the health professionals with whom the data were shared. The studies reported a trend toward increased weekly physical activity owing to the technology-motivated engagement (eg, Chen et al [68] report a significant increase in physical activity; $P<.001$) [30-38,41-47,50,54,55,57,60,62-64,66-69].

Medication Monitoring

Of the 42 studies, 16 (38%) included a medication adherence–tracking device [30,32,37,38,41,45,49,52,53,55,59,61,63-65,68], half of which (n=8, 50%) had a reminder feature [32,37,45,52,55,61,63,68]. Of the 42 studies, 6 (14%) included an insulin dose–adjustment device used by the health professional or patient (eg, using a bolus dose algorithm) [29,40,48,57,66,69]. Regarding the medication data collected, of the 42 studies, 6 (14%) reported medication adjustments [29,40,48,57,66,69], 7 (17%) analyzed the monitoring of prescribed insulin doses [30,32,55,57,58,66,68], and 5 (12%) administered questionnaires on medication adherence [31,34,50,57,67]. Finally, 4 (10%) of the 42 studies reported decreased oral antidiabetic doses after the interventions [31,40,48,68].

Therapeutic Education

Patients were provided various pedagogical tools to support their therapeutic education in 20 (48%) of the 42 studies [31,33-37,43,46-48,51,52,55,59,60,63,64,66,69,70]. Among these 20 studies, web-based course modules were used in 4 (20%) [43,48,63,66]. Other tools were used to advance nutritional literacy [31,35,46,59]; or the tools talked about or referred to relevant articles on topics such as using a blood glucose meter, diabetes complications, physical activity, and tobacco use [33,35,36,43,46,48,52,55,59,66,69]. Finally, 2 (10%) of the 20 studies proposed meditation or mindfulness exercises [51,55]. Personalized recommendation tools were used in 11 (26%) of the 42 studies [29,37,45-48,51,52,60,63,66]. These recommendations were either delivered by a health professional after an analysis of the patient's logged data, generated automatically by an artificial intelligence algorithm, or planned according to a therapeutic education protocol. The pedagogical materials were often supported by electronic notebook tools where patients could jot down topics to discuss with their health professionals [52,64,67].

Health Professional Involvement

Among the 42 studies, communication between the health professional and patient was ensured through a chat feature in 13 (31%) studies [31,35,43,46,47,49,51,52,59,60,63,66,68], by email in 7 (17%) studies [31,33,36,43,66,67,71], by SMS text messaging in 14 (33%) studies [29,36,37,41,42,44,48,54,58,62,63,67-69], by telephone calls in 13 (31%) studies [31,33,40,48,55-58,60,62,67-69], and by videoconferencing in 4 (10%) studies [55,60,67,68].

Of the 42 studies, 33 (79%) included a tool for displaying patient data [30,32-37,39,41-49,51-59,63-66,68-70], one-third of them (n=11, 33%) in real time, in the form of a graphic report. Of the 42 studies, 3 (7%) included a decision support tool [34,45,64], whereas 12 (29%) included a tool for setting and monitoring therapeutic goals that could be shared by the care provider and patient [29,32,33,37,45,46,53,59,61,63,68,69].

Other Health Data Monitoring

The monitoring of other health data concerned weight loss. Of the 42 studies, 16 (38%) monitored weight using a graphic representation over time [29,31-33,40,42,43,45-48,53,56,60,65,66]. Of these 16 studies, 6 (38%) collected automated data using a Bluetooth-connected weight scale [31,40,42,43,46,48]. In addition, 7 (17%) of the 42 studies enabled the sharing of blood test results [38,40,60,61,64,65,68]. Kobayashi et al [42] used a Bluetooth-connected tensiometer to transmit blood pressure readings to a cloud-based server, making it possible to summarize and present the data to patients and their primary care physicians to promote self-management, monitoring, and follow-up. The studies reported a statistically significant reduction in weight ranging from 3 to 6.2 kg [29,40,43,46,56,60] and in BMI ranging from 1.6 kg/m^2 to 4 kg/m^2 [29,34,42,48,56,60].

Health Care Management

Of the 42 studies, 20 (48%) included personal spaces in their technologies [31-35,38,40,42,45,47,49,51-53,61,63,66-68,70]. In these spaces, it was possible to view a dashboard summarizing the logged health data, monitor exchanges with health professionals, and generate reports that could be shared by the patient and downloaded by the health professionals for inclusion in the medical file [34,38,42,45,47,52,60,66,67,70]. Social support was promoted through links to social networks in 6 (14%) of the 42 studies [31,35,37,41,44,48]. Of the 42 studies, 6 (14%) included a web-based appointment scheduling tool, facilitating monitoring and follow-up by the health professionals [32,33,35,53,64,67]. Finally, Holmen et al [69] made technical support available 7 days a week to users of their technology.

Combination of Interventions

Studies showing significant positive results were those combining the involvement of a health professional with the monitoring of glycemia levels, diet, physical activity, and medication [41,57,61]. Of the 42 studies, 1 (2%) combined support from a health professional with the monitoring of

glycemia levels, diet, and physical activity; therapeutic education; and a follow-up of body weight [29]. Some of the studies (7/42, 17%) only added to the involvement of a health professional the monitoring of glycemia levels and physical activity (n=1, 14%) [40], the monitoring of glycemia levels alone (n=2, 29%) [51,58], diet and medication monitoring with therapeutic education (n=1, 14%) [31] or without therapeutic education (n=1, 14%) [35], diet monitoring and therapeutic education (n=1, 14%) [70], and physical activity and body weight monitoring (n=1, 14%) [42]. Of the 42 studies, 2 (5%) with positive significant results evaluated the combination of a health professional and the monitoring of glycemia levels, diet, and medication (n=1, 50%) [30] and therapeutic education and body weight follow-up (n=1, 50%) [34]. Most often (23/42, 55%), the combined strategies involved a health professional and the monitoring of glycemia levels and diet (Multimedia Appendix 6 [29-31,34,35,40-42,51,57,58,61,70]).

Discussion

Principal Findings

This study mapped telehealth interventions tailored to the needs of patients with type 2 diabetes supported by a health professional. This review—despite the range of scientific literature available; the complex nature of these interventions; and the heterogeneity of study designs, populations, organizational care contexts, measures, and result indicators used—revealed a trend suggesting the effectiveness of telehealth interventions with health professional involvement in improving health outcomes. The use of everyday technologies in these interventions could facilitate their accessibility and usability, which would facilitate their implementation in the longer term. On the basis of our exploration of the literature, we were able to classify the key features of digital tools that may have a positive effect on intervention outcomes into eight categories: (1) monitoring of glycemia levels, (2) diet monitoring, (3) physical activity monitoring, (4) medication monitoring, (5) therapeutic education, (6) health professional support, (7) other health data monitoring, and (8) health care management (Figure 5).

The duration of the interventions varied significantly among the studies, with interventions lasting 1 month to 18 months. A recent meta-analysis on the effectiveness of telemedicine application for chronic diseases found that for people living with type 2 diabetes, HbA_{1c} levels began to decrease after up to 12 months of telemedicine intervention compared with interventions lasting 6 months [75]. These results were also supported in a study by Timpel et al [76], where HbA_{1c} levels began to decrease in participants after 12 months of long-term telemedicine intervention. Given that the HbA_{1c} level is a recognized indicator of glycemic control over a retrospective period, reflecting average blood glucose levels over approximately 3 months, it is regarded as a standard for assessing the effectiveness of long-term diabetes interventions [77]. This measure offers a more stable view of a patient's glycemia levels than instantaneous measurements, which can be influenced by many immediate factors [77]. Longer interventions could allow for more accurate adjustments in treatments and disease management behaviors as well as provide enough time for these changes to result in improvements in glycemic control.

The health professionals involved in these studies were primarily physicians, dietitians, and nurses. Nearly half (19/42, 45%) of the studies involved a multidisciplinary care team [29,31,34,37,38,41,44,48,50,52-54,57-59,63,67,68,71] (Multimedia Appendix 4). The studies showed that health technologies could help optimize the therapeutic education and monitoring of people living with type 2 diabetes through collecting and sharing information between consultations. Care provider personnel would thus be better able to focus on other aspects of their practice during consultations. Some of the interventions (4/42, 10%) used a videoconferencing platform for consultations with the health professional to make the exchanges more natural and pleasant [55,60,67,68]. A recent narrative review that included 12 randomized controlled trials assessing the effectiveness of telemedicine versus conventional counseling, demonstrated that the counseling and monitoring of patients living with diabetes via telemedicine was more effective than conventional counseling [78]. Similarly, health technologies could help improve the efficiency of practical tasks performed by health professionals, for example, by producing more concise automatically generated reports that can be shared among the care team, thus fostering interdisciplinary monitoring and follow-up. They also offer the possibility of monitoring patients in real time and sharing targeted information with them, thereby facilitating timely adjustments. Telehealth tools enable the continuous monitoring of blood glucose levels, physical activity, diet, medication intake, and other health indicators. This enables patients and health care providers to quickly detect fluctuations in blood sugar levels and take appropriate action to maintain optimal control of blood sugar levels [1]. The features of telehealth tools can provide personalized recommendations and advice based on each patient's specific data [2]; for example, patients can receive medication reminders, nutritional advice tailored to their dietary preferences, and suggestions for physical activities based on their condition and health goals [2]. Telehealth tools offer educational resources and information on type 2 diabetes [3]. Patients can access educational materials, explanatory videos, meal plans, and tips to improve their understanding of the disease and its management [3]. This promotes patient empowerment by enabling them to actively participate in the management of their health [3-5]. Telehealth tools can include features such as appointment reminders, food diaries, and physical activity logs. These features help patients track their progress, stay engaged with their treatment, and maintain their motivation [3,5].

Our findings are in line with the chronic care model [79]. Telehealth interventions, as observed in our study, frequently incorporate goal-setting tools that empower patients to set and track health-related objectives, aligning with the model's emphasis on self-management support. In addition, our results underscore the vital role of health professional support within telehealth interventions, enabling remote monitoring and timely guidance, consistent with the model's focus on patient-centered care. Social support emerged in our findings, with patients benefiting from the encouragement of their social networks—a

concept aligned with the chronic care model's recognition of involving the patient's social support system. Finally, our research highlights the inclusion of educational materials in telehealth interventions, providing patients with essential knowledge about their condition, in line with the model's emphasis on patient education. Together, these elements within telehealth strategies contribute to patient empowerment, improved self-management, and enhanced outcomes for the management of chronic conditions such as diabetes, emphasizing the importance of a comprehensive approach to health care delivery, even in remote or web-based settings.

However, there are also potential limitations and challenges associated with the use of telehealth tools for the management of type 2 diabetes. The use of telehealth tools may be limited by internet access, technological skills, and the availability of the necessary devices [2,3]. Populations that have been historically marginalized or disadvantaged may face digital disparities, limiting their ability to benefit fully from these tools. It is thus essential to recognize that some patients may require additional human support. Interaction with health care providers may be necessary to obtain answers to questions, resolve problems, and receive emotional support. Furthermore, the use of telehealth tools involves the collection, storage, and sharing of sensitive health data. It is crucial to implement robust security measures to protect data confidentiality and prevent privacy breaches [2,8]. Telehealth tools use monitoring devices to collect data, such as blood glucose meters or continuous blood glucose monitoring sensors. However, these devices can have technical limitations and measurement errors, which can affect the accuracy of the data collected and potentially influence treatment decisions [5,8]. Given that diabetes management is characterized by a long process of therapeutic education, monitoring, and follow-up, technological support would be a helpful asset in primary health care because it would help maintain motivation [29,37,40,46,54,61,70] through the use of numerous tools (goal-setting tools and shared decision-making support tools, recipes, informational content, etc), by facilitating interactions with a health professional, and by promoting access to care (eg, with the possibility of using multilingual resources).

Figure 5. Classification of digital features for diabetes self-management and monitoring.

Recommendations for Future Designs

Telehealth offers many opportunities for diabetes self-management and monitoring, enabling patients to benefit from remote care, continuous monitoring, and personalized support. The use of continuous blood glucose monitoring devices, mobile apps, web-based platforms, and other technologies facilitates the collection and tracking of diabetes-related data [9]. The introduction of web-based educational resources, web-based learning modules, and self-help tools to help patients better understand their disease as well as manage their diet, physical activity, medication, and monitoring of blood glucose levels promotes patient self-management and empowerment [10,11]. In addition, web-based support via secure messaging to answer patients' questions and respond to their concerns supports therapeutic education and keeps them engaged. Indeed, technology developers will need to set up clear and effective communication channels between patients and health professionals. This may include web-based consultations, secure message exchanges, and regular reports on patient progress [11]. Finally, it will be important to consider the integration of these telehealth interventions into existing health care systems, ensuring coordination and continuity of care. It will be necessary to ensure that data collected by remote monitoring devices are accessible to health professionals and integrated into patients' medical records [12].

Limitations of Included Studies

The studies identified in this review involved voluntary patient participation. In particular, the studies favored individuals with good technology literacy. The selection bias inherent in voluntary patient participation and the preference for technology-literate individuals suggest that the findings might not be generalizable to the broader population of people with diabetes. The indicators used to assess the effectiveness of the interventions were primarily dietary intake; clinical indicators such as glycemia levels, HbA_{1c} levels, blood pressure, and cholesterol levels; physical activity; medication adherence; motivation; and the use of telehealth technology. Although positive changes in these indicators were noted in most clinical results, this may translate into something other than rigorous clinical parameters. Different strategies were used to collect data, notably involving innovative digital tools (although these

tools did not undergo a validation study). In addition, lifestyle changes (dietary planning and physical activity) were measured using the patient self-administered digital questionnaires, leaving the door open to all biases inherent in self-reporting. A meta-analysis of these data would help inform a position in this regard.

The heterogeneity of the included studies posed a real challenge in interpreting the results. Aside from the various methods used, which yielded different levels of evidence, the interventions were based mainly on effecting behavior changes through therapeutic education supported by digital tools and a health professional; yet, none of the studies assessed the impact on the results of the context within which these technologies were used, such as concurrent public health policies (eg, diabetes or obesity prevention campaigns, the promotion of a balanced diet, physical activity, or tobacco use).

Moreover, the literature states that 90% of people with diabetes have at least 1 other chronic disease. Nonetheless, few interventions have provided the integrated management of diabetes and other pathologies. Specifically, renal and cardiac risks have not always been assessed. The multipathological context should be systematically considered when designing studies because multiple medication use (eg, sulfonylureas and insulin) can cause iatrogenic hypoglycemia and influence the clinical parameters [80-82]. Similarly, the different stages of diabetes severity should be documented to foster a more accurate interpretation of the results.

The varying durations of the interventions, ranging from 1 month to 18 months, and the differing technologies used emphasize that outcomes such as improvements in HbA_{1c} levels are not uniform across all studies. The positive association observed with longer interventions and the reduction in HbA_{1c} levels may not hold true in every context or for every patient demographic. The role of health professionals in these interventions is undoubtedly significant, but the translation of these findings into practice must consider the individual needs and circumstances of diverse patient populations, including access issues and technological literacy. The integration of everyday technologies seems promising for broader implementation; however, this assumption requires careful consideration of the digital disparities that may exist, particularly among groups that have been historically marginalized or disadvantaged.

Strengths and Limitations of This Review

To further leverage the qualitative nature of the content analyzed in the studies, we performed a descriptive content analysis of the data using NVivo (release 1.7). This allowed us to supplement our research with a narrative account of the selected studies. The abundance of literature on the subject attests to a worldwide questioning of digital health policies. The COVID-19 pandemic led to a doubling of the number of annual publications on the topic of telehealth interventions used for type 2 diabetes or prediabetes self-management and monitoring with health professional involvement. Given the rapid development of technologies and research, which has only escalated in recent years, a systematic review would help provide invaluable data on the effectiveness of these interventions. This scoping review included studies published in peer-reviewed journals and is thus subject to publication bias owing to the well-documented notion that researchers and journals tend to publish positive results. In addition, we limited ourselves to selecting studies published in French or English from 2010 given the rapid pace of technological development and the consequent rapid increase in the literature. Future researchers should consider more inclusive approaches, such as conducting systematic reviews that encompass gray literature and unpublished studies. This ensures a more comprehensive and unbiased overview of existing literature on the topic.

The results of this review did not allow us to identify how the 8 key digital tool features should be developed and combined to help improve health outcomes. However, the strategy most often combined with telehealth interventions facilitating interaction with health professionals was the monitoring of glycemia levels, diet, and physical activity. A few of the studies (7/42, 17%) also included medication monitoring and therapeutic education. Future studies should perform in-depth analyses of the usability and acceptability of these technologies to highlight the design issues and shed light on health policies.

The diversity of the interventions analyzed underscores the necessity to acknowledge the unique challenges and issues inherent to each specific population. Such issues can encompass socioeconomic factors, cultural differences, accessibility to health services, and varying levels of health literacy, all of which can significantly influence the effectiveness of interventions; for instance, interventions that succeed in urban environments with high connectivity and technologically savvy populations may not yield identical results in rural or low-income areas where internet access is scarce and digital literacy is an issue. Moreover, the cultural context may impact patient engagement and the suitability of educational materials. Each population may hold distinct health beliefs, practices, and priorities, which must be considered during the design and implementation of health interventions. Recognizing these disparities is critical to understanding why results from 1 group cannot be generalized to another. Public health strategies must develop resource allocation policies and create interventions focused on the users' needs. Hence, although telehealth presents a promising avenue for improving diabetes management, its application must be nuanced and considerate of the public health challenges unique to each specific population to be truly effective and equitable.

Future Research Prospects

With regard to gaps in the literature, some questions require further research. This scoping review revealed a need for long-term implementation studies, possibly because telehealth programs require a less-structured time commitment and could be used over extended periods. Long-term evaluation studies are also needed to facilitate the implementation of telehealth interventions. Further studies on adherence and engagement could explore the factors that influence patients' adherence to telehealth interventions and their engagement in diabetes self-management. These studies will also help to identify effective strategies for encouraging patients' active participation and maintaining their motivation over the long term. Evaluation

frameworks should incorporate reports on participant engagement and satisfaction, acceptability, security, and costs into future telehealth interventions because these will facilitate their translation into clinical practice. In addition, the measurement of the effects of interventions should include measures other than clinical data, such as patient-reported experience measures and patient-reported outcome measures to ensure that these interventions are meeting the needs of patients. In addition, multimorbidity was mentioned by only a few of the included studies (7/42, 17%) and warrants further research to assess the impact of these interventions on health [34,49,54,56,65,67,70]. Additional studies could define standardized assessment criteria for telehealth interventions that support the therapeutic management of patients with diabetes and multiple comorbidities. The impact of equity of access to care on the use of telehealth interventions for populations considered vulnerable, including populations with low-income status, rural or remote populations, and culturally diverse groups, will need to be studied. A better understanding of these impacts will help identify potential barriers and strategies to reduce disparities and improve equitable access to telehealth [12]. Finally, it will be vital to evaluate the effectiveness of integrating telehealth interventions into existing health care systems, including collaboration among health professionals, data sharing, and care coordination. This will help distinguish best practices for the successful integration of telehealth into clinical care and existing health care systems [12]. Of the 42 studies, 3 (7%) assessed the impact on the cost of care [48,58,64]. The macroeconomic implications of these telehealth interventions for health care systems warrant future studies to shed clearer light on health policies. Finally, the COVID-19 pandemic has revealed the various structural and organizational shortcomings of health care around the globe. It has also accelerated the dissemination and adoption of digital tools and advanced the digital ambitions of governments worldwide. The abundance of publications means that future studies can perform a meta-analysis of randomized controlled trials. Our analysis underscores the critical role of multidisciplinary health care teams and promotes the integration of ubiquitous technologies into daily health management practices to achieve superior patient outcomes. Furthermore, this review stresses the necessity of considering the long-term viability of telehealth solutions, patient adherence, and the seamless incorporation of these solutions into current health care frameworks in subsequent research.

Finally, although we included studies conducted in different parts of the world in this scoping review, we did not find relevant studies conducted in Canada, indicating an opportunity for research tailored to the Canadian context. For the implementation of future telehealth interventions to improve diabetes management in Canada, it is recommended to consider the specificities of the Canadian health care system, such as the heterogeneity of its organization across different provinces, the diversity of its population, and its varied health resources. It would be wise to design personalized interventions that address the unique needs of patients with diabetes within the Canadian population, particularly in Indigenous communities that are disproportionately affected by diabetes, including linguistic and cultural considerations. Strategies for equitable access to telehealth technologies for populations that have been historically marginalized or those living in remote areas should also be considered. Training health professionals in telehealth tools and best practices for web-based care is equally essential. Moreover, interdisciplinary and intersectoral collaboration would be beneficial to effectively integrate telehealth into primary care, allowing for coordinated and consistent follow-up. Finally, by anticipating challenges related to privacy and data security, interventions should incorporate robust security measures to protect sensitive patient information while focusing on a personalized approach and the development of patient-centered interventions and technologies.

Conclusions

This review systematically maps out the effectiveness of telehealth interventions for managing type 2 diabetes, with a focus on the enhanced outcomes gained through the involvement of health professionals. It presents a detailed categorization of the pivotal characteristics of digital tools into 8 distinct areas that significantly influence the success of these interventions. The evidence-based data suggest that participation in sustained telehealth interventions with health professional involvement helps improve health outcomes and type 2 diabetes–related behavior, reducing the risks of complications. However, despite our identification of the key digital tool features of these interventions, it remains to be seen how to combine and translate them into long-term usable components in specific care contexts. Nonetheless, the results are promising for future health care because they point to consolidating care through a single platform, which could improve patients' quality of life while encouraging active self-management. They also shed light on developing evidence-based telehealth programs that can be adapted to specific care contexts and offer decision makers more effective options for funding diabetes management programs. Ultimately, this review aims to enrich the understanding of telehealth's role in diabetes care and to outline specific domains for future research that will inform policy making and the advancement of telehealth practices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 517 KB - medinform_v12i1e46699_app1.pdf]

Multimedia Appendix 2
Search strategy.
[DOCX File , 14 KB - medinform_v12i1e46699_app2.docx]

Multimedia Appendix 3
Data extraction table.
[XLSX File (Microsoft Excel File), 18 KB - medinform_v12i1e46699_app3.xlsx]

Multimedia Appendix 4
Health professional involvement.
[XLSX File (Microsoft Excel File), 13 KB - medinform_v12i1e46699_app4.xlsx]

Multimedia Appendix 5
Digital features of the interventions.
[XLSX File (Microsoft Excel File), 13 KB - medinform_v12i1e46699_app5.xlsx]

Multimedia Appendix 6
Studies showing significant positive health outcomes.
[XLSX File (Microsoft Excel File), 12 KB - medinform_v12i1e46699_app6.xlsx]

References

1. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 2004 May;27(5):1047-1053. [doi: 10.2337/diacare.27.5.1047] [Medline: 15111519]
2. IDF diabetes atlas 7th edition. IDF Diabetes Atlas. 2015. URL: https://diabetesatlas.org/atlas/seventh-edition/ [accessed 2024-01-19]
3. Diabetes Canada. Diabetes Canada 2018 clinical practice guidelines for the prevention and management of diabetes in Canada. Can J Diabetes 2018;42(1):A1-A18, S1-S326 [FREE Full text]
4. Ivers NM, Jiang M, Alloo J, Singer A, Ngui D, Casey CG, et al. Diabetes Canada 2018 clinical practice guidelines: key messages for family physicians caring for patients living with type 2 diabetes. Can Fam Physician 2019 Jan;65(1):14-24 [FREE Full text] [Medline: 30674509]
5. American Diabetes Association Professional Practice Committee. 6. Glycemic targets: standards of medical care in diabetes-2022. Diabetes Care 2022 Jan 01;45(Suppl 1):S83-S96. [doi: 10.2337/dc22-S006] [Medline: 34964868]
6. Bhattacherjee A, Hikmet N, Menachemi N, Kayhan VO, Brooks RG. The differential performance effects of healthcare information technology adoption. Inf Syst Manag 2006 Dec 22;24(1):5-14. [doi: 10.1080/10580530601036778]
7. Kabir MJ, Heidari A, Honarvar MR, Khatirnamani Z, Rafiei N. Challenges in the implementation of an electronic referral system: a qualitative study in the Iranian context. Int J Health Plann Manage 2023 Jan 21;38(1):69-84. [doi: 10.1002/hpm.3563] [Medline: 35988065]
8. Nathan DM, Cleary PA, Backlund JY, Genuth SM, Lachin JM, Orchard TJ, et al. Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. N Engl J Med 2005 Dec 22;353(25):2643-2653 [FREE Full text] [doi: 10.1056/NEJMoa052187] [Medline: 16371630]
9. Toma T, Athanasiou T, Harling L, Darzi A, Ashrafian H. Online social networking services in the management of patients with diabetes mellitus: systematic review and meta-analysis of randomised controlled trials. Diabetes Res Clin Pract 2014 Nov;106(2):200-211 [FREE Full text] [doi: 10.1016/j.diabres.2014.06.008] [Medline: 25043399]
10. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J, et al. Telehealth for global emergencies: implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2020 Jun;26(5):309-313 [FREE Full text] [doi: 10.1177/1357633X20916567] [Medline: 32196391]
11. Picot J, Cradduck T. The telehealth industry in Canada: industry profile and capability analysis. The Keston Group and Infotelmed Communications Inc. 2000 Mar 30. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c7f26186673239a22d731d13857780ba0a5c7918 [accessed 2024-01-19]
12. Kairy D. Télérééducation, téléréadaptation et e-santé : définition, évolution et diversité des points de vue sur ces concepts. Edimark.fr. 2017 Jan 20. URL: https://www.edimark.fr/revues/actualites-en-mpr/n-1-janvier-2017/telereeducation-telereadaptation-et-e-sante-definition-evolution-et-diversite-des-points-de-vue-sur-ces-concepts [accessed 2024-01-19]
13. Yaya S, Raffelini C. [Technological transformations and evolution of the medical practice: current status, issues and perspectives for the development of telemedicine]. Rev Med Brux 2009;30(2):83-91. [Medline: 19517904]
14. Ramadas A, Quek KF, Chan CK, Oldenburg B. Web-based interventions for the management of type 2 diabetes mellitus: a systematic review of recent evidence. Int J Med Inform 2011 Jun;80(6):389-405. [doi: 10.1016/j.ijmedinf.2011.02.002] [Medline: 21481632]

15. Crosson JC, Ohman-Strickland PA, Cohen DJ, Clark EC, Crabtree BF. Typical electronic health record use in primary care practices and the quality of diabetes care. Ann Fam Med 2012 May 14;10(3):221-227 [FREE Full text] [doi: 10.1370/afm.1370] [Medline: 22585886]
16. Tenforde M, Nowacki A, Jain A, Hickner J. The association between personal health record use and diabetes quality measures. J Gen Intern Med 2012 Apr 18;27(4):420-424 [FREE Full text] [doi: 10.1007/s11606-011-1889-0] [Medline: 22005937]
17. Marcolino MS, Maia JX, Alkmim MB, Boersma E, Ribeiro AL. Telemedicine application in the care of diabetes patients: systematic review and meta-analysis. PLoS One 2013 Nov 8;8(11):e79246 [FREE Full text] [doi: 10.1371/journal.pone.0079246] [Medline: 24250826]
18. Malanda UL, Welschen LM, Riphagen II, Dekker JM, Nijpels G, Bot SD. Self-monitoring of blood glucose in patients with type 2 diabetes mellitus who are not using insulin. Cochrane Database Syst Rev 2012 Jan 18;1:CD005060. [doi: 10.1002/14651858.CD005060.pub3] [Medline: 22258959]
19. Pal K, Eastwood SV, Michie S, Farmer AJ, Barnard ML, Peacock R, et al. Computer-based diabetes self-management interventions for adults with type 2 diabetes mellitus. Cochrane Database Syst Rev 2013 Mar 28;2013(3):CD008776 [FREE Full text] [doi: 10.1002/14651858.CD008776.pub2] [Medline: 23543567]
20. Howland C, Wakefield B. Assessing telehealth interventions for physical activity and sedentary behavior self-management in adults with type 2 diabetes mellitus: an integrative review. Res Nurs Health 2021 Feb 22;44(1):92-110 [FREE Full text] [doi: 10.1002/nur.22077] [Medline: 33091168]
21. Anderson A, O'Connell SS, Thomas C, Chimmanamada R. Telehealth interventions to improve diabetes management among Black and Hispanic patients: a systematic review and meta-analysis. J Racial Ethn Health Disparities 2022 Dec 09;9(6):2375-2386 [FREE Full text] [doi: 10.1007/s40615-021-01174-6] [Medline: 35000144]
22. American Diabetes Association Professional Practice Committee. 7. Diabetes technology: standards of medical care in diabetes-2022. Diabetes Care 2022 Jan 01;45(Suppl 1):S97-112. [doi: 10.2337/dc22-S007] [Medline: 34964871]
23. May CR, Finch TL, Cornford J, Exley C, Gately C, Kirk S, et al. Integrating telecare for chronic disease management in the community: what needs to be done? BMC Health Serv Res 2011 May 27;11(1):131 [FREE Full text] [doi: 10.1186/1472-6963-11-131] [Medline: 21619596]
24. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
25. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
27. Babineau J. Product review: covidence (systematic review software). J Can Health Libr Assoc 2014 Aug 01;35(2):68-71. [doi: 10.5596/c14-016]
28. Mak S, Thomas A. Steps for conducting a scoping review. J Grad Med Educ 2022 Oct;14(5):565-567 [FREE Full text] [doi: 10.4300/JGME-D-22-00621.1] [Medline: 36274762]
29. Ang IY, Tan KX, Tan C, Tan CH, Kwek JW, Tay J, et al. A personalized mobile health program for type 2 diabetes during the COVID-19 pandemic: single-group pre-post study. JMIR Diabetes 2021 Jul 09;6(3):e25820 [FREE Full text] [doi: 10.2196/25820] [Medline: 34111018]
30. Bergenstal RM, Layne JE, Zisser H, Gabbay RA, Barleen NA, Lee AA, et al. Remote application and use of real-time continuous glucose monitoring by adults with type 2 diabetes in a virtual diabetes clinic. Diabetes Technol Ther 2021 Feb 01;23(2):128-132 [FREE Full text] [doi: 10.1089/dia.2020.0396] [Medline: 33026839]
31. Berman MA, Guthrie NL, Edwards KL, Appelbaum KJ, Njike VY, Eisenberg DM, et al. Change in glycemic control with use of a digital therapeutic in adults with type 2 diabetes: cohort study. JMIR Diabetes 2018 Feb 14;3(1):e4 [FREE Full text] [doi: 10.2196/diabetes.9591] [Medline: 30291074]
32. Bradway M, Giordanengo A, Joakimsen R, Hansen AH, Grøttland A, Hartvigsen G, et al. Measuring the effects of sharing mobile health data during diabetes consultations: protocol for a mixed method study. JMIR Res Protoc 2020 Feb 10;9(2):e16657 [FREE Full text] [doi: 10.2196/16657] [Medline: 32039818]
33. Cassidy S, Okwose N, Scragg J, Houghton D, Ashley K, Trenell MI, et al. Assessing the feasibility and acceptability of changing health for the management of prediabetes: protocol for a pilot study of a digital behavioural intervention. Pilot Feasibility Stud 2019 Nov 26;5(1):139 [FREE Full text] [doi: 10.1186/s40814-019-0519-1] [Medline: 31788325]
34. Doocy S, Paik KE, Lyles E, Hei Tam H, Fahed Z, Winkler E, et al. Guidelines and mHealth to improve quality of hypertension and type 2 diabetes care for vulnerable populations in Lebanon: longitudinal cohort study. JMIR Mhealth Uhealth 2017 Oct 18;5(10):e158 [FREE Full text] [doi: 10.2196/mhealth.7745] [Medline: 29046266]
35. Haste A, Adamson AJ, McColl E, Araujo-Soares V, Bell R. Web-based weight loss intervention for men with type 2 diabetes: pilot randomized controlled trial. JMIR Diabetes 2017 Jul 07;2(2):e14 [FREE Full text] [doi: 10.2196/diabetes.7430] [Medline: 30291100]

36. Iljaž R, Brodnik A, Zrimec T, Cukjati I. E-healthcare for diabetes mellitus type 2 patients - a randomised controlled trial in Slovenia. Zdr Varst 2017 Sep;56(3):150-157 [FREE Full text] [doi: 10.1515/sjph-2017-0020] [Medline: 28713443]
37. Jeon E, Park HA. Experiences of patients with a diabetes self-care app developed based on the information-motivation-behavioral skills model: before-and-after study. JMIR Diabetes 2019 Apr 18;4(2):e11590 [FREE Full text] [doi: 10.2196/11590] [Medline: 30998218]
38. Johnston P. Monitoring of blood glucose levels, ketones and insulin bolus advice using 4SURE products and app-based technology. Br J Nurs 2022 Jan 13;31(1):34-39. [doi: 10.12968/bjon.2022.31.1.34] [Medline: 35019739]
39. Jung H, Demiris G, Tarczy-Hornoch P, Zachry M. A novel food record app for dietary assessments among older adults with type 2 diabetes: development and usability study. JMIR Form Res 2021 Feb 17;5(2):e14760 [FREE Full text] [doi: 10.2196/14760] [Medline: 33493129]
40. Kempf K, Altpeter B, Berger J, Reuß O, Fuchs M, Schneider M, et al. Efficacy of the telemedical lifestyle intervention program TeLiPro in advanced stages of type 2 diabetes: a randomized controlled trial. Diabetes Care 2017 Jul;40(7):863-871. [doi: 10.2337/dc17-0303] [Medline: 28500214]
41. Lee DY, Yoo SH, Min KP, Park CY. Effect of voluntary participation on mobile health care in diabetes management: randomized controlled open-label trial. JMIR Mhealth Uhealth 2020 Sep 18;8(9):e19153 [FREE Full text] [doi: 10.2196/19153] [Medline: 32945775]
42. Kobayashi T, Tsushita K, Nomura E, Muramoto A, Kato A, Eguchi Y, et al. Automated feedback messages with Shichifukujin characters using IoT system-improved glycemic control in people with diabetes: a prospective, multicenter randomized controlled trial. J Diabetes Sci Technol 2019 Jul 20;13(4):796-798 [FREE Full text] [doi: 10.1177/1932296819851785] [Medline: 31104490]
43. Koot D, Goh PS, Lim RS, Tian Y, Yau TY, Tan NC, et al. A mobile lifestyle management program (GlycoLeap) for people with type 2 diabetes: single-arm feasibility study. JMIR Mhealth Uhealth 2019 May 24;7(5):e12965 [FREE Full text] [doi: 10.2196/12965] [Medline: 31127720]
44. Ku EJ, Park JI, Jeon HJ, Oh T, Choi HJ. Clinical efficacy and plausibility of a smartphone-based integrated online real-time diabetes care system via glucose and diet data management: a pilot study. Intern Med J 2020 Dec 22;50(12):1524-1532. [doi: 10.1111/imj.14738] [Medline: 31904890]
45. Lamprinos I, Demski H, Mantwill S, Kabak Y, Hildebrand C, Ploessnig M. Modular ICT-based patient empowerment framework for self-management of diabetes: design perspectives and validation results. Int J Med Inform 2016 Jul;91:31-43. [doi: 10.1016/j.ijmedinf.2016.04.006] [Medline: 27185507]
46. Lim SL, Ong KW, Johal J, Han CY, Yap QV, Chan YH, et al. Effect of a smartphone app on weight change and metabolic outcomes in Asian adults with type 2 diabetes: a randomized clinical trial. JAMA Netw Open 2021 Jun 01;4(6):e2112417 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.12417] [Medline: 34081137]
47. Liu Y, Yu Z, Sun H. Treatment effect of type 2 diabetes patients in outpatient department based on blockchain electronic mobile medical app. J Healthc Eng 2021 Mar 1;2021:6693810 [FREE Full text] [doi: 10.1155/2021/6693810] [Medline: 33728034]
48. McKenzie AL, Hallberg SJ, Creighton BC, Volk BM, Link TM, Abner MK, et al. A novel intervention including individualized nutritional recommendations reduces hemoglobin A1c level, medication use, and weight in type 2 diabetes. JMIR Diabetes 2017 Mar 07;2(1):e5 [FREE Full text] [doi: 10.2196/diabetes.6981] [Medline: 30291062]
49. Modave F, Bian J, Rosenberg E, Mendoza T, Liang Z, Bhosale R, et al. DiaFit: the development of a smart app for patients with type 2 diabetes and obesity. JMIR Diabetes 2016 Dec 13;1(2):e5 [FREE Full text] [doi: 10.2196/diabetes.6662] [Medline: 29388609]
50. Polonsky WH, Layne JE, Parkin CG, Kusiak CM, Barleen NA, Miller DP, et al. Impact of participation in a virtual diabetes clinic on diabetes-related distress in individuals with type 2 diabetes. Clin Diabetes 2020 Oct;38(4):357-362 [FREE Full text] [doi: 10.2337/cd19-0105] [Medline: 33132505]
51. Quinn CC, Butler EC, Swasey KK, Shardell MD, Terrin MD, Barr EA, et al. Mobile diabetes intervention study of patient engagement and impact on blood glucose: mixed methods analysis. JMIR Mhealth Uhealth 2018 Feb 02;6(2):e31 [FREE Full text] [doi: 10.2196/mhealth.9265] [Medline: 29396389]
52. Salari R, R Niakan Kalhori S, GhaziSaeedi M, Jeddi M, Nazari M, Fatehi F. Mobile-based and cloud-based system for self-management of people with type 2 diabetes: development and usability evaluation. J Med Internet Res 2021 Jun 02;23(6):e18167 [FREE Full text] [doi: 10.2196/18167] [Medline: 34076579]
53. Schmocker KS, Zwahlen FS, Denecke K. Mobile app for simplifying life with diabetes: technical description and usability study of GlucoMan. JMIR Diabetes 2018 Feb 26;3(1):e6 [FREE Full text] [doi: 10.2196/diabetes.8160] [Medline: 30291070]
54. Schusterbauer V, Feitek D, Kastner P, Toplak H. Two-stage evaluation of a telehealth nutrition management service in support of diabesity therapy. Stud Health Technol Inform 2018;248:314-321. [Medline: 29726453]
55. Wang W, Seah B, Jiang Y, Lopez V, Tan C, Lim ST, et al. A randomized controlled trial on a nurse-led smartphone-based self-management programme for people with poorly controlled type 2 diabetes: a study protocol. J Adv Nurs 2018 Jan;74(1):190-200. [doi: 10.1111/jan.13394] [Medline: 28727183]
56. Zaharia OP, Kupriyanova Y, Karusheva Y, Markgraf DF, Kantartzis K, Birkenfeld AL, et al. Improving insulin sensitivity, liver steatosis and fibrosis in type 2 diabetes by a food-based digital education-assisted lifestyle intervention program: a

feasibility study. Eur J Nutr 2021 Oct;60(7):3811-3818 [FREE Full text] [doi: 10.1007/s00394-021-02521-3] [Medline: 33839905]
57. Bastyr EJ3, Zhang S, Mou J, Hackett AP, Raymond SA, Chang AM. Performance of an electronic diary system for intensive insulin management in global diabetes clinical trials. Diabetes Technol Ther 2015 Aug;17(8):571-579 [FREE Full text] [doi: 10.1089/dia.2014.0407] [Medline: 25826466]
58. Levy NK, Moynihan V, Nilo A, Singer K, Etiebet MA, Bernik L, et al. The mobile insulin titration intervention (MITI) study: innovative chronic disease management of diabetes. J Gen Internal Med 2015;30:S547-S548.
59. Tang PC, Overhage JM, Chan AS, Brown NL, Aghighi B, Entwistle MP, et al. Online disease management of diabetes: engaging and motivating patients online with enhanced resources-diabetes (EMPOWER-D), a randomized controlled trial. J Am Med Inform Assoc 2013 May 01;20(3):526-534 [FREE Full text] [doi: 10.1136/amiajnl-2012-001263] [Medline: 23171659]
60. Chung YS, Kim Y, Lee CH. Effectiveness of the smart care service for diabetes management. Healthc Inform Res 2014 Oct;20(4):288-294 [FREE Full text] [doi: 10.4258/hir.2014.20.4.288] [Medline: 25405065]
61. Kim Y, Lee H, Seo JM. Integrated diabetes self-management program using smartphone application: a randomized controlled trial. West J Nurs Res 2022 Apr 03;44(4):383-394. [doi: 10.1177/0193945921994912] [Medline: 33655794]
62. Torbjørnsen A, Jenum AK, Småstuen MC, Arsand E, Holmen H, Wahl AK, et al. A low-intensity mobile health intervention with and without health counseling for persons with type 2 diabetes, part 1: baseline and short-term results from a randomized controlled trial in the Norwegian part of renewing health. JMIR Mhealth Uhealth 2014 Dec 11;2(4):e52 [FREE Full text] [doi: 10.2196/mhealth.3535] [Medline: 25499592]
63. Nes AA, van Dulmen S, Eide E, Finset A, Kristjánsdóttir OB, Steen IS, et al. The development and feasibility of a web-based intervention with diaries and situational feedback via smartphone to support self-management in patients with diabetes type 2. Diabetes Res Clin Pract 2012 Sep;97(3):385-393. [doi: 10.1016/j.diabres.2012.04.019] [Medline: 22578890]
64. Jia W, Zhang P, Duolikun N, Zhu D, Li H, Bao Y, et al. Study protocol for the road to hierarchical diabetes management at primary care (ROADMAP) study in China: a cluster randomised controlled trial. BMJ Open 2020 Jan 06;10(1):e032734 [FREE Full text] [doi: 10.1136/bmjopen-2019-032734] [Medline: 31911516]
65. Esau N, Koen N, Herselman MG. Adaptation of the RenalSmart® web-based application for the dietary management of patients with diabetic nephropathy. South Afr J Clin Nutr 2016 May 31;26(3):132-140. [doi: 10.1080/16070658.2013.11734457]
66. Hidalgo JI, Maqueda E, Risco-Martín JL, Cuesta-Infante A, Colmenar JM, Nobel J. glUCModel: a monitoring and modeling system for chronic diseases applied to diabetes. J Biomed Inform 2014 Apr;48:183-192 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.015] [Medline: 24407050]
67. Castelnuovo G, Manzoni GM, Cuzziol P, Cesa GL, Tuzzi C, Villa V, et al. TECNOB: study design of a randomized controlled trial of a multidisciplinary telecare intervention for obese patients with type-2 diabetes. BMC Public Health 2010 Apr 23;10(1):204 [FREE Full text] [doi: 10.1186/1471-2458-10-204] [Medline: 20416042]
68. Chen L, Chuang LM, Chang CH, Wang CS, Wang IC, Chung Y, et al. Evaluating self-management behaviors of diabetic patients in a telehealthcare program: longitudinal study over 18 months. J Med Internet Res 2013 Dec 09;15(12):e266 [FREE Full text] [doi: 10.2196/jmir.2699] [Medline: 24323283]
69. Holmen H, Torbjørnsen A, Wahl AK, Jenum AK, Småstuen MC, Arsand E, et al. A mobile health intervention for self-management and lifestyle change for persons with type 2 diabetes, part 2: one-year results from the Norwegian randomized controlled trial renewing health. JMIR Mhealth Uhealth 2014 Dec 11;2(4):e57 [FREE Full text] [doi: 10.2196/mhealth.3882] [Medline: 25499872]
70. Chang AR, Bailey-Davis L, Yule C, Kwiecen S, Graboski E, Juraschek S, et al. Abstract P289: effects of dietary app-supported tele-counseling on sodium intake, diet quality, and blood pressure in patients with diabetes and kidney disease. Circulation 2019 Mar 05;139(Suppl_1):AP289. [doi: 10.1161/circ.139.suppl_1.p289]
71. Bradway M, Pfuhl G, Joakimsen R, Ribu L, Grøttland A, Årsand E. Analysing mHealth usage logs in RCTs: explaining participants' interactions with type 2 diabetes self-management tools. PLoS One 2018 Aug 30;13(8):e0203202 [FREE Full text] [doi: 10.1371/journal.pone.0203202] [Medline: 30161248]
72. Castelnuovo G, Manzoni GM, Cuzziol P, Cesa GL, Corti S, Tuzzi C, et al. TECNOB study: ad interim results of a randomized controlled trial of a multidisciplinary telecare intervention for obese patients with type-2 diabetes. Clin Pract Epidemiol Ment Health 2011 Mar 04;7(1):44-50 [FREE Full text] [doi: 10.2174/1745017901107010044] [Medline: 21559233]
73. Bird D, Oldenburg B, Cassimatis M, Russell A, Ash S, Courtney MD, et al. Randomised controlled trial of an automated, interactive telephone intervention to improve type 2 diabetes self-management (Telephone-Linked Care Diabetes Project): study protocol. BMC Public Health 2010 Oct 12;10(1):599 [FREE Full text] [doi: 10.1186/1471-2458-10-599] [Medline: 20937148]
74. Kesavadev J, Saboo B, Shankar A, Krishnan G, Jothydev S. Telemedicine for diabetes care: an Indian perspective - feasibility and efficacy. Indian J Endocrinol Metab 2015;19(6):764-769 [FREE Full text] [doi: 10.4103/2230-8210.167560] [Medline: 26693425]

75. Ma Y, Zhao C, Zhao Y, Lu J, Jiang H, Cao Y, et al. Telemedicine application in patients with chronic disease: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2022 Apr 19;22(1):105 [FREE Full text] [doi: 10.1186/s12911-022-01845-2] [Medline: 35440082]
76. Timpel P, Oswald S, Schwarz PE, Harst L. Mapping the evidence on the effectiveness of telemedicine interventions in diabetes, dyslipidemia, and hypertension: an umbrella review of systematic reviews and meta-analyses. J Med Internet Res 2020 Mar 18;22(3):e16791 [FREE Full text] [doi: 10.2196/16791] [Medline: 32186516]
77. Sherwani SI, Khan HA, Ekhzaimy A, Masood A, Sakharkar MK. Significance of HbA1c test in diagnosis and prognosis of diabetic patients. Biomark Insights 2016 Jul 03;11. [doi: 10.4137/bmi.s38440]
78. Kusuma CF, Aristawidya L, Susanti CP, Kautsar AP. A review of the effectiveness of telemedicine in glycemic control in diabetes mellitus patients. Medicine (Baltimore) 2022 Dec 02;101(48):e32028 [FREE Full text] [doi: 10.1097/MD.0000000000032028] [Medline: 36482628]
79. Bodenheimer T, Wagner EH, Grumbach K. Improving primary care for patients with chronic illness. JAMA 2002 Oct 09;288(14):1775-1779. [doi: 10.1001/jama.288.14.1775] [Medline: 12365965]
80. Nourine I. Influence des comorbidités sur la prise en charge du diabète de type 2 de la personne âgée. Universite de Lorraine. 2016 Mar 8. URL: https://hal.univ-lorraine.fr/hal-01932239/document [accessed 2024-01-19]
81. UK Hypoglycaemia Study Group. Risk of hypoglycaemia in types 1 and 2 diabetes: effects of treatment modalities and their duration. Diabetologia 2007 Jun 6;50(6):1140-1147. [doi: 10.1007/s00125-007-0599-y] [Medline: 17415551]
82. Budnitz DS, Lovegrove MC, Shehab N, Richards CL. Emergency hospitalizations for adverse drug events in older Americans. N Engl J Med 2011 Nov 24;365(21):2002-2012. [doi: 10.1056/nejmsa1103053]

Abbreviations

DASH: Dietary Approaches to Stop Hypertension
HbA_{1c}: glycated hemoglobin
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by A Castonguay; submitted 10.03.23; peer-reviewed by MH Kurniawan, N Mungoli, J Ross; comments to author 03.05.23; revised version received 21.09.23; accepted 07.12.23; published 13.03.24.

Please cite as:
Mannoubi C, Kairy D, Menezes KV, Desroches S, Layani G, Vachon B
The Key Digital Tool Features of Complex Telehealth Interventions Used for Type 2 Diabetes Self-Management and Monitoring With Health Professional Involvement: Scoping Review
JMIR Med Inform 2024;12:e46699
URL: https://medinform.jmir.org/2024/1/e46699
doi:10.2196/46699
PMID:38477979

Review

Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review

Aditya Singhal[1], MSc; Nikita Neveditsin[2], MSc; Hasnaat Tanveer[3], BSc; Vijay Mago[4], PhD

[1]Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada

[2]Department of Mathematics and Computing Science, Saint Mary's University, Halifax, NS, Canada

[3]Faculty of Mathematics, University of Waterloo, Waterloo, ON, Canada

[4]School of Health Policy and Management, York University, Toronto, ON, Canada

Corresponding Author:
Nikita Neveditsin, MSc
Department of Mathematics and Computing Science
Saint Mary's University
923 Robie Street
Halifax, NS, B3H 3C3
Canada
Phone: 1 902 420 5893
Email: Nikita.Neveditsin@smu.ca

Abstract

Background: The use of social media for disseminating health care information has become increasingly prevalent, making the expanding role of artificial intelligence (AI) and machine learning in this process both significant and inevitable. This development raises numerous ethical concerns. This study explored the ethical use of AI and machine learning in the context of health care information on social media platforms (SMPs). It critically examined these technologies from the perspectives of fairness, accountability, transparency, and ethics (FATE), emphasizing computational and methodological approaches that ensure their responsible application.

Objective: This study aims to identify, compare, and synthesize existing solutions that address the components of FATE in AI applications in health care on SMPs. Through an in-depth exploration of computational methods, approaches, and evaluation metrics used in various initiatives, we sought to elucidate the current state of the art and identify existing gaps. Furthermore, we assessed the strength of the evidence supporting each identified solution and discussed the implications of our findings for future research and practice. In doing so, we made a unique contribution to the field by highlighting areas that require further exploration and innovation.

Methods: Our research methodology involved a comprehensive literature search across PubMed, Web of Science, and Google Scholar. We used strategic searches through specific filters to identify relevant research papers published since 2012 focusing on the intersection and union of different literature sets. The inclusion criteria were centered on studies that primarily addressed FATE in health care discussions on SMPs; those presenting empirical results; and those covering definitions, computational methods, approaches, and evaluation metrics.

Results: Our findings present a nuanced breakdown of the FATE principles, aligning them where applicable with the American Medical Informatics Association ethical guidelines. By dividing these principles into dedicated sections, we detailed specific computational methods and conceptual approaches tailored to enforcing FATE in AI-driven health care on SMPs. This segmentation facilitated a deeper understanding of the intricate relationship among the FATE principles and highlighted the practical challenges encountered in their application. It underscored the pioneering contributions of our study to the discourse on ethical AI in health care on SMPs, emphasizing the complex interplay and the limitations faced in implementing these principles effectively.

Conclusions: Despite the existence of diverse approaches and metrics to address FATE issues in AI for health care on SMPs, challenges persist. The application of these approaches often intersects with additional ethical considerations, occasionally leading to conflicts. Our review highlights the lack of a unified, comprehensive solution for fully and effectively integrating FATE principles in this domain. This gap necessitates careful consideration of the ethical trade-offs involved in deploying existing methods and underscores the need for ongoing research.

(JMIR Med Inform 2024;12:e50048) doi:10.2196/50048

KEYWORDS

fairness, accountability, transparency, and ethics; artificial intelligence; social media; health care

Introduction

Background

Machine learning (ML) algorithms have become pervasive in today's world, influencing a wide range of fields, from governance and financial decision-making to medical diagnosis and security assessment. These technologies depend on artificial intelligence (AI) and ML to provide results, offering clear advantages in terms of speed and cost-effectiveness for businesses over time [1]. However, as AI research progresses rapidly, the importance of ensuring that its development and deployment adhere to ethical principles has become paramount.

User data on social media platforms (SMPs) can reveal patterns, trends, and behaviors. Platforms such as Twitter (X Corp) are predominantly used by younger individuals and those residing in urban areas [2]. These platforms often impose age restrictions, leading to a potential bias in algorithms trained on their data toward younger, urban demographics. Social media presents a rich source of data invaluable for health research [3], yet using these data without proper consent poses ethical concerns. Furthermore, social media content is influenced by various social factors and should not always be interpreted at face value. For example, certain topics may engage users from specific regions or demographic groups more than others [4], rendering the data less universally applicable. An additional challenge is the trustworthiness of these data. The issue of bias is further exacerbated when AI or ML software is proprietary with a closed source code, making it challenging to analyze and understand the reasons behind biased decisions [3].

The spread of both misinformation and disinformation is a significant concern on social media [5,6], a problem that became particularly acute during the COVID-19 pandemic. False claims about vaccine safety contributed to public mistrust and hesitancy, undermining efforts to control the virus. In tackling this issue, AI tools have been deployed to sift through information and spotlight reliable content for users [7]. These AI systems are trained using health data from trustworthy sources, ensuring the dissemination of scientifically sound information. On the bright side, social media provides a venue for disseminating new health information, offering valuable insights for the health sector [8]. However, the inherent challenges of social media, such as verifying information authenticity and the risk of spreading misinformation, require careful management to guarantee that the health information shared is accurate and reliable.

Fairness, accountability, transparency, and ethics (FATE) research focuses on evaluating the fairness and transparency of AI and ML models, developing accountability metrics, and designing ethical frameworks [9]. Incorporating a human in the loop is one approach to upholding ethical principles in algorithmic processes. For example, in the case of the Correctional Offender Management Profiling for Alternative Sanctions system used within the US judicial system to predict the likelihood of a prisoner reoffending after release, it is recommended that a judge first review the AI's decision to ensure its accuracy. In summary, recognizing the inherent biases in AI and ML, the implementation of systematic models is crucial for maintaining accountability. Efforts in computer science are directed toward enhancing the transparency of AI and ML, which helps uncover the decision-making processes, identify biases, and hold systems accountable for failures [10,11].

Motivation

The American Medical Informatics Association (AMIA) has delineated a comprehensive set of ethical principles for the governance of AI [12] building on the foundations laid out in the Belmont Report [13]: autonomy, beneficence, nonmaleficence, and justice. These principles are critical for the responsible application of AI in monitoring health care–related data on SMPs [7]. The AMIA expanded these principles to include 6 technical aspects—explainability, interpretability, fairness, dependability, auditability, and knowledge management—as well as 3 organizational principles: benevolence, transparency, and accountability. Furthermore, it incorporated special considerations for vulnerable populations, AI research, and user education [12]. Our review emphasized the concept of FATE, which is prevalent in the AI and ML community [14], and discussed its alignment with the principles outlined by the AMIA.

The discourse on AI ethics is notably influenced by geographic and socioeconomic contexts [15]. There has been extensive debate regarding the best practices for evaluating work produced by explanatory AI and conducting gap analyses on model interpretability in AI [16,17]. Recent advancements in ML interpretability have also been subject to review [18]. Table 1 provides a summary of existing studies that discuss FATE in various contexts. These studies reveal a substantial research gap in understanding how the principles of FATE are integrated within the realm of AI in health care on SMPs. Notably, none of the studies have thoroughly investigated the computational methods commonly used to assess the components of FATE and their intricate interrelationships in this domain.

To bridge the identified research gap, this study focused on three pivotal research questions (RQs):

1. What existing solutions address FATE in the context of health care on SMPs? (RQ 1)
2. How do these solutions identified in response to RQ 1 compare with each other in terms of computational methods, approaches, and evaluation metrics? (RQ 2)
3. What is the strength of the evidence supporting these various solutions? (RQ 3)

Our aim was to enrich the domain of FATE by exploring the array of techniques, methods, and solutions that facilitate social media interventions in health care settings while pinpointing gaps in the current body of literature. This study encompassed the definitions, computational methods, approaches, and evaluation metrics pertinent to FATE in AI along with an

examination of FATE in data sets. The novelty of our research lies in delivering a comprehensive analysis of metrics, computational solutions, and the application of FATE principles specifically within the realm of SMPs. This includes a focus on uncovering further research directions and challenges at the confluence of health care, computer science, and social science.

Table 1. An overview of existing studies focusing on fairness, accountability, transparency, and ethics.

Study	Fairness			Accountability			Transparency			Ethics		
	A[a]	B[b]	C[c]	A	B	C	A	B	C	A	B	C
Mehrabi et al [1], 2021	✓	✓	✓									
Golder et al [19], 2017											✓	✓
Bear Don't Walk et al [20], 2022	✓	✓	✓									
Attard-Frost et al [21], 2022	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Wieringa [9], 2020				✓	✓	✓						
Adadi and Berrada [22], 2018								✓	✓			
Diogo et al [18], 2019	✓					✓	✓	✓	✓			✓
Chakraborty et al [17], 2017	✓			✓			✓		✓			
Hagerty and Rubinov [15], 2019										✓		✓
Vian and Kohler [23], 2016				✓			✓					

[a]Definitions.

[b]Computational methods and approaches.

[c]Evaluation metrics.

Methods

Research Methodology

Our research methodology was grounded in the approach presented by Kofod-Petersen [24] and adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [25]. We used 2 search databases, PubMed and Web of Science, to ensure the reproducibility of the search results in the identification of records. PubMed was chosen for its comprehensive coverage of biomedical literature, providing direct access to the most recent research in health care and its intersections with AI, rendering it indispensable for studies focused on the FATE principles in the domain. Web of Science was selected for its interdisciplinary scope, diversity of publication sources, and rigorous citation analysis, offering a broad and authoritative overview of global research trends and impacts across computer science, social sciences, and health care. In addition, we used Google Scholar, which is recognized as the most comprehensive repository of scholarly articles [26], known for its inclusivity and extensive coverage across multiple disciplines. However, due to the lack of reproducibility of the search results on Google Scholar, we classified it as *other source* for record identification, as shown in Figure 1. Our search across these databases was conducted without any language restrictions, ensuring a comprehensive and inclusive review of the relevant literature.

We conducted a strategic search using Table 2 as a filter to identify research papers pertinent to our review. The table was designed to allow for customization of groups for retrieving varied sets of literature, aiming to find the intersection among these sets. For group 1, we selected "fairness," "accountability," "transparency," and "ethics." These keywords, being integral components of the FATE framework, were an obvious choice for our search queries. In group 2, we identified "natural language processing" and "artificial intelligence" as our keywords. The selection of "natural language processing" was justified by the predominance of textual data on SMPs, necessitating algorithms adept at processing natural language. The inclusion of "artificial intelligence" reflected its broad applicability beyond traditional ML applications. Given that AI encompasses a wide range of advanced technologies, including sophisticated natural language processing (NLP) techniques, its inclusion ensured the comprehensive coverage of relevant studies. Finally, the terms "social media" and "healthcare" were directly pertinent to our review, making their inclusion essential. Consequently, our aim was to encompass a wide spectrum of studies relevant to the topic of our review.

On the basis of Table 1, our initial strategy involved using the intersection of groups as follows: ([group 1, search term 1 ∩ group 2, search term 1] AND [group 1, search term 1 ∩ group 2, search term 2]) ∩ ([group 1, search term 1 ∩ group 3, search term 1] AND [group 1, search term 1 ∩ group 3, search term 2]), which, for simplicity, we condensed to (group 1, search term 1 ∩ group 2, search term 1 ∩ group 2, search term 2 ∩ group 3, search term 1 ∩ group 3, search term 2), as outlined in the search query presented in Textbox 1.

For our queries, we implemented year-based filtering in PubMed and conducted a parallel topic search in Web of Science, limiting the results to articles published since 2012. However, this approach yielded only 2 publications from each database, a tally considered inadequate for our purposes. Consequently, we opted to broaden our search by applying the union of 2 intersections. The initial formula ([group 1, search term 1 ∩ group 2, search

term 1] AND [group 1, search term 1 ∩ group 2, search term 2]) ∪ ([group 1, search term 1 ∩ group 3, search term 1] AND [group 1, search term 1 ∩ group 3, search term 2]) was streamlined to group 1, search term 1 ∩ ([group 2, search term 1 ∩ group 2, search term 2] ∪ [group 3, search term 1 ∩ group 3, search term 2]), as detailed in the search query in Textbox 2, while maintaining the same year range.

Our search queries resulted in 442 records from PubMed and 327 records from Web of Science, as shown in Figure 1. Subsequently, we eliminated duplicates across the 3 sources, consolidating the findings into 672 records for initial screening. During the screening phase, we applied specific inclusion criteria based on an analysis of titles and abstracts to refine the selection: (1) the study primarily addressed FATE principles in the context of health care on SMPs (inclusion criterion 1); (2) the study reported empirical findings (inclusion criterion 2); (3) the study elaborated on definitions, computational methods, approaches, and evaluation metrics (inclusion criterion 3).

This process narrowed down the field to 172 records eligible for full-text assessment. At this stage, we applied our quality criteria to further assess eligibility: (1) we confirmed through full-text screening that the study adhered to inclusion criteria 1, 2, and 3 (quality criterion 1); (2) the study articulated a clear research objective (quality criterion 2).

Ultimately, this led to the selection of 135 articles for inclusion in our review. The complete list of these articles is available in Multimedia Appendix 1 [1-3,5-11,15-23,26-141].

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram for record selection.

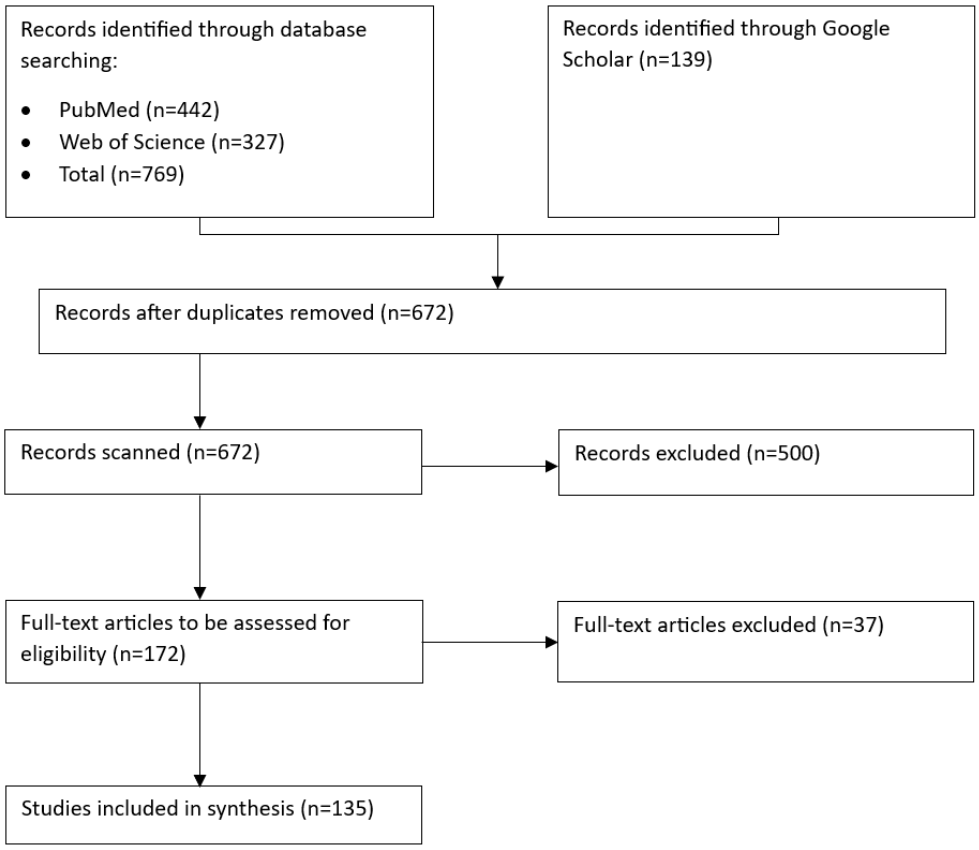


Table 2. Search strategy for finding research articles.

	G1[a]	G2[b]	G3[c]
T1[d]	Quality[e]	Natural language processing	Social media
T2[f]	N/A[g]	Artificial intelligence	Health care

[a]G1: group 1.

[b]G2: group 2.

[c]G3: group 3.

[d]T1: search term 1.

[e]{Fairness, Accountability, Transparency, Ethics}

[f]T2: search term 2.

[g]N/A: not applicable.

Textbox 1. The initial query to the databases.

- ("Fairness" OR "Accountability" OR "Transparency" OR "Ethics") AND ("NLP" or "Natural Language Processing") AND ("AI" OR "Artificial Intelligence") AND ("Healthcare" AND "Social Media")

Textbox 2. Modified query to the databases.

- ("Fairness" OR "Accountability" OR "Transparency" OR "Ethics") AND ((("NLP" or "Natural Language Processing") AND ("AI" OR "Artificial Intelligence")) OR ("Healthcare" AND "Social Media"))

Data Items and Data-Charting

In our review, we incorporated the following data items: (1) approaches and definitions related to each component of FATE; (2) mathematical formulations and algorithms designed to address FATE; (3) methodologies for the integration of FATE principles into AI and ML systems, particularly within health care settings on SMPs; (4) characteristics of the AI or ML systems under study, encompassing their type, application areas within health care, and the specific roles that SMPs play in these systems; (5) outcomes from the formal evaluation or assessment of FATE aspects within the studies, such as their impact on decision-making processes; (6) challenges and barriers reported in the implementation of FATE principles in AI or ML systems; (7) use of frameworks or tools developed to support or evaluate FATE in AI and ML systems; and (8) engagement of stakeholders throughout the AI and ML system's life cycle, including their perspectives on FATE.

The data-charting process involved 3 researchers, each independently extracting pertinent data from the selected sources with a particular focus on the aforementioned data items. For methodical organization and analysis, the extracted information was documented in Microsoft Excel spreadsheets (Microsoft Corp). These spreadsheets were organized alphabetically by the last name of the first author of each article and included references to the corresponding data items as presented in the studies. To consolidate the compiled data, one researcher was tasked with merging the information from these spreadsheets. This step aimed to synthesize the data and ensure a coherent presentation of our findings. The merging process entailed a thorough review and amalgamation of the data charted by each researcher, emphasizing the consolidation of similar approaches and methodologies as identified in the studies.

Results

Definitions, Computational Methods, and Approaches to Fairness

Overview

The understanding of fairness among the public is diverse [26]. The AMIA classifies fairness as a technical principle, emphasizing its importance in creating AI systems that are free from bias and discrimination [12]. This study reviewed various approaches to achieving fairness, with a particular focus on perspectives that facilitate the quantification of fairness in the context of AI for health care on SMPs. The mathematical formulations used to measure fairness are presented in Multimedia Appendix 2 [27-32,142,143]. The following subsections offer a comprehensive examination of approaches to ensure fairness.

Calibrated Fairness

Calibrated fairness seeks to balance providing equal opportunities for all individuals with accommodating their distinct differences and needs [33]. For instance, in the context of social media, a calibrated fair algorithm aims to ensure equal access to opportunities, such as visibility for all users, while also considering specific factors, such as language or location, to offer a personalized experience. In health care, such an algorithm would ensure that all patients have access to the same standard of care yet take into account variables such as age and health status to tailor the best possible treatment plan. The objective is to find a balance between treating everyone equally and acknowledging individual differences to achieve the most equitable outcomes. Fairness metrics, including the false positive rate difference [29] and the equal opportunity difference [34], are used to evaluate the degree of calibrated fairness. Common computational methods used to achieve calibrated fairness include the following: (1) preprocessing—modifying the original data set to diminish or eliminate the impact of sensitive attributes (eg, gender and ethnic background) on the outcome of an ML model [35]; (2) in-processing—integrating fairness constraints into the model's training process to ensure calibration with respect to sensitive attributes [35]; (3) postprocessing—adjusting the model's output after training to calibrate it in relation to sensitive attributes [35]; (3) adversarial training—training the model on adversarial examples, which are designed to test the model's fairness in predictions [36].

Each of the approaches to achieving calibrated fairness in AI systems has a specific application context that is influenced by various factors. Preprocessing aims to directly mitigate biases in the data before the model's training phase but may present challenges in preserving the integrity of the original data, potentially resulting in the loss of important information. In contrast, in-processing involves the integration of fairness constraints during the model's learning process, which, while aiming to ensure fairness, might compromise model performance due to the added constraints. Postprocessing, which adjusts the model's outputs after training, may appear as a straightforward solution but often falls short in addressing the root causes of bias, thus providing a superficial fix. Adversarial training stands out as a promising approach by challenging the model's fairness through specially designed examples; however, its effective implementation can be complex and resource intensive. Each method has inherent trade-offs between fairness, accuracy, and complexity. The choice among them depends on the specific

circumstances of the application, including the nature of the data, the criticality of the decision-making context, and the specific fairness objectives.

Statistical Fairness

Statistical fairness considers various factors, including demographic information, that may be pertinent to the concept of fairness within a specific context. Among the widely recognized statistical definitions of fairness are demographic parity, equal opportunity, and equal treatment [37]. The measure of "demographic parity" is used to reduce data bias by incorporating penalty functions into matrix-factorization objectives [38], whereas the "equal opportunity" metric is crucial for ensuring that decisions are devoid of bias [39]. In the realm of social media, individual notions of fairness might encompass issues such as unbiased content moderation, equitable representation of diverse perspectives and voices, and transparency in the algorithms used for content curation and ranking. Common approaches for measuring statistical fairness include the following: (1) equalized odds—this approach evaluates fairness by examining the differences in true positive and false positive rates across various groups [40]; (2) theorem of equal treatment—this approach assesses fairness by comparing how similar individuals from different groups are treated [41].

Moreover, several toolkits have been developed for measuring statistical fairness in ML and AI models. For instance, Aequitas, as introduced by Saleiro et al [42], generates reports aiding in equitable decision-making by policy makers and ML researchers. The AI Fairness 360 toolkit [43] provides metrics and algorithms designed to reduce statistical biases that lead to the unfair treatment of various groups by ML models [44]. Another toolkit, Fairlearn [45], offers algorithms aimed at addressing disparities in the treatment of different demographic groups by an ML model.

Intersectional Fairness

This approach integrates multiple intersecting identity facets, such as race, gender, and socioeconomic status, into decision-making processes concerning individuals [46]. Its objective is to guarantee equitable treatment for all stakeholders, recognizing that the confluence of these identities may exacerbate marginalization and discrimination. Within the realm of social media, an algorithm designed with intersectional fairness in mind ensures that content is neither recommended nor censored in a manner that is prejudiced against a user's race, gender, or socioeconomic status. Similarly, in health care, an algorithm that incorporates intersectional fairness aims to prevent the disproportionate allocation of medical treatments and resources. Intersectional fairness can be operationalized using the worst-case disparity method, which involves evaluating each subgroup individually and comparing the best and worst outcomes to ascertain the precision of the fairness score. Subsequently, the ratio of the maximum to minimum scores is calculated, with a ratio nearing 1 indicating a more equitable outcome [46]. Other prevalent methods and strategies for achieving intersectional fairness include the following: (1) constraint-based methods—these are designed to honor specific fairness constraints, such as providing equal treatment to different groups identified by multiple attributes, through mathematical optimization [47]; (2) causal inference methods—these aim to ensure that the algorithm's outputs are unbiased by examining the causal relationships between inputs and outputs [48]; (3) decision trees and rule-based systems—these are used to guarantee that the algorithm's decisions are informed by relevant factors and free from bias [49].

Constraint-based methods are adept at enforcing predefined fairness goals; however, the complexity of defining and optimizing these goals poses a significant challenge. In contrast to constraint-based methods, causal inference methods do not necessitate predefined fairness constraints but require a thorough comprehension of the data at hand. Erroneous assumptions regarding causality can result in flawed assessments of fairness. Decision trees and rule-based systems, owing to their interpretability, facilitate the understanding of algorithmic decisions. However, their simplicity may be a limitation as they may not adequately address the complexities inherent in various data sets. To mitigate some of the discussed shortcomings, supervised ranking, unsupervised regression, and reinforcement in fairness evaluation can be approached through pairwise evaluation [50]. This technique involves assessing an AI model's performance by comparing its outputs against a preselected set of input data pairs.

Definitions, Computational Methods, and Approaches to Accountability

Overview

The AMIA considers accountability a fundamental organizational principle, stressing that organizations should bear the responsibility for continuously monitoring AI systems. This includes identifying, reporting, and managing potential risks. Furthermore, organizations are expected to implement strategies for risk mitigation and establish a system for the submission and resolution of complaints related to AI operations [12]. In the following subsections, we explore prevalent views on accountability within the ML and AI community. In addition, we provide summaries of the measurements for different accountability components as identified in the reviewed literature, which can be found in Multimedia Appendix 3 [51-54,144].

Legal Accountability

Legal accountability encompasses the obligations of entities involved in designing, developing, deploying, and using AI systems for health care purposes on social media [55]. This responsibility includes ensuring that AI systems are developed and used in compliance with relevant laws and regulations in addition to addressing any adverse effects or impacts that might arise from their use. Legal accountability also covers issues such as data protection and privacy along with the duty to prevent the use of AI systems for discriminatory or unethical purposes. Commonly used conceptual methods for achieving legal accountability include the following: (1) transparency—this method involves making AI systems transparent, ensuring that their decision-making processes are explainable and comprehensible [56] (there are existing

frameworks designed to enhance transparency in the accountability of textual models [57]); (2) documentation—this involves maintaining detailed records of the systems' design, development, and testing processes, as well as documenting the data used for training them [58] (an initiative toward accountability is the implementation of model cards, which are intended to outline an ML model's limitations and disclose any biases that it may be susceptible to [59]); (3) adjudication—this refers to the creation of procedures for addressing disputes and grievances associated with the use of ML and AI systems [60].

Overall, the pursuit of legal accountability should be carefully balanced with the autonomy of stakeholders and must not hinder innovation.

Ethical Accountability

Ethical accountability ensures that AI systems make decisions that are transparent, justifiable, and aligned with societal values [61]. This encompasses addressing data privacy, securing informed consent, and preventing the perpetuation of existing biases and discrimination. Ethical concerns specific to the use of AI in health care include safeguarding patient privacy, handling sensitive health data responsibly, and avoiding the reinforcement of existing health disparities [62]. Common strategies for achieving ethical accountability include the following: (1) ethical impact assessment—this approach entails assessing the ethical risks and benefits of the system and weighing the trade-offs between them [63]; (2) value alignment—this strategy involves embedding ethical principles and values into the design and development of the system, ensuring that its operations are in harmony with these values [64]; (3) transparency and explanation—this is accomplished by offering clear, understandable explanations of the system's functionality and making its data and algorithms openly available [65]; (4) stakeholder engagement—this involves the active participation of a diverse group of stakeholders, including users, developers, and experts, in all phases of the AI or ML system's life cycle [66].

When crafting ethical AI for disseminating health care–related information on social media, the application of these methodologies varies according to specific tasks. Ethical impact assessments, for instance, are valuable for evaluating the potential advantages, such as enhanced patient engagement via personalized dissemination of health care information, against risks, including privacy breaches and the spread of misinformation. The value alignment method plays a crucial role in pinpointing essential ethical values such as patient privacy, information accuracy, nondiscrimination, and accessibility. This method also supports the performance of regular audits to verify that AI systems continuously reflect these ethical standards. Finally, approaches to stakeholder engagement establish a platform for transparent and continuous communication between stakeholders and developers, thereby promoting a cooperative atmosphere in development.

Technical Accountability

Technical accountability ensures that developers and designers of AI and ML systems are held responsible for maintaining standards of security, privacy, and functionality [67]. This responsibility encompasses the implementation of adequate mechanisms to monitor and manage AI algorithms and address arising technical issues. Within the realms of social media and health care, technical accountability further entails the use of AI technologies to foster ethical decision-making, safeguard user privacy, and ensure that decisions are made fairly and transparently [68]. Common strategies for achieving technical accountability include the following: (1) logging—the practice of recording all inputs, outputs, and decisions to trace the system's performance and pinpoint potential problems [69]; (2) auditing—conducting evaluations to check the system's performance, detect biases, and ensure compliance with ethical and legal standards [70].

Both logging and auditing play critical roles in the development of ethical AI for health care information on social media, each with its unique benefits and challenges. Logging, which captures the inputs, outputs, and decisions of an AI system, is vital for tracking system performance. Nonetheless, the retention of detailed logs, especially those involving sensitive health care information, may introduce privacy concerns and necessitate careful consideration of data protection strategies. Auditing, essential for upholding ethical and legal norms, demands expertise and considerable time to effectively scrutinize complex AI systems. In addition, frameworks designed to enhance AI system accountability are in use. An example is Pandora [71], representing a significant move toward achieving a holistic approach to accountable AI systems.

Societal Accountability

Societal accountability entails the obligation of stakeholders to ensure that their AI systems align with societal values and interests [72]. This encompasses addressing privacy, transparency, and fairness issues, along with considering the wider social, cultural, and economic impacts that AI systems may have. Achieving societal accountability may require stakeholders to participate in public consultations, develop ethical and transparent regulations and standards for AI use, and enhance public understanding of AI system functionalities and applications. Essentially, it advocates for the development and use of AI systems under the principles of responsible innovation, with society's interests considered at every life cycle stage.

Methods for ensuring societal accountability include the following: (1) regulation and standardization—creating regulations and standards for AI system design and use can help hold these systems accountable to society, safeguarding the rights and interests of all stakeholders [73]; (2) public-private partnerships—fostering collaboration among government agencies, private-sector companies, and other entities to promote the societal accountability of AI and ML systems [74].

To ensure accountability, integrating transparency and fairness into algorithms, designing systems with privacy considerations, and conducting regular audits and evaluations to review AI system performance is critical. Researchers have suggested approaches for holding companies accountable for their AI-related actions [9]. They emphasize the importance of pinpointing specific decision makers within a company responsible for any errors, a crucial step for ensuring equitable

accountability. The entity or individuals determining accountability should possess comprehensive knowledge of legal, political, administrative, professional, and social viewpoints regarding the error to guarantee fair and unbiased judgments. Moreover, the consequences imposed on decision makers should be appropriately matched to their areas of responsibility, considering each individual's level of responsibility within the company's hierarchy when deciding on these consequences.

Definitions, Computational Methods, and Approaches to Transparency

Overview

According to the AMIA, transparency is an organizational principle that asserts that an AI system must operate impartially, not favoring its host organization. This principle ensures fairness, treating all stakeholders equally without privileging any party. Moreover, transparency requires stakeholders to be clearly informed that they are interacting with an AI system and not a human [12]. Adadi and Berrada [22] presented a nuanced view on transparency, defining it as the degree to which the workings of an AI system are comprehensible to humans. This definition encompasses providing explanations for the system's decision-making processes, clarifying the data used for system training, and certifying the system's neutrality and nondiscriminatory nature. The balancing act between transparency and privacy presents challenges. For instance, in the analysis of mental health data on SMPs, the difficulty does not lie in pinpointing user-specific attributes (as data are often aggregated) but in the application of these data [75]. Here, transparency intersects with the ethical principle of autonomy, which demands that systems protect individual independence, treat users respectfully, and secure informed consent [12]. Guaranteeing autonomy is particularly crucial in the deployment of AI-powered depression detection systems on social networks [76]. The following subsections will delve into the nuances of transparency in AI, emphasizing the importance of openness in data and algorithmic procedures. This focus is particularly critical in the context of data derived from SMPs. We also introduce some metrics for assessing transparency in Multimedia Appendix 4 [77-81].

Algorithmic Transparency

Algorithmic transparency is the clarity with which one can comprehend the manner in which an AI algorithm or model produces its outputs or decisions [82]. Within the context of AI for health care on SMPs, transparency entails the ability to lucidly grasp the processes and methodologies used in the creation, dissemination, and evaluation of social media interventions for health care objectives [83]. This encompasses an understanding of the data sources that inform these interventions, the algorithms or models that analyze the data and generate the interventions, and the criteria for assessing intervention effectiveness. Algorithmic transparency is crucial for identifying and addressing potential biases or errors in interventions and fostering trust among stakeholders, including patients, health care providers, and regulatory bodies. Several computational techniques can enhance algorithmic transparency: (1) feature importance analysis—this technique identifies the most impactful features or variables in the model's output, shedding light on the decision-making process [84]; (2) model interpretability—this involves designing models whose outputs are easily understood and interpreted by humans [85] (for instance, decision trees and logistic regression models are more interpretable compared to more complex models [86]; detailed discussions of model interpretability will follow in a dedicated subsection); (3) explanation generation—this technique produces explanations for a model's outputs, offering insights into its decision-making process through visualizations or natural language descriptions [87].

Feature importance analysis enhances the comprehension of a model's decision-making process, yet it may not fully elucidate the complex interactions among features or their combined effect on the model's decisions, especially in the case of sophisticated deep neural networks. Models that are inherently interpretable, such as decision trees and logistic regression, promote user trust and facilitate the validation of model behaviors. However, these models might not offer the same level of power and precision as more complex models such as deep neural networks, which restricts their effectiveness in analyzing health care–related social media interactions. On the other hand, explanation generation seeks to clarify the model's reasoning for stakeholders. Nonetheless, guaranteeing that these explanations are both accurate and reflective of the model's inner workings poses a considerable challenge.

Data Transparency

Data transparency pertains to the comprehensibility of how data are collected, stored, and used in the development of an AI system [88]. Within the realm of AI for health care on SMPs, data transparency delineates the degree to which health care organizations and providers maintain openness and clarity regarding the collection, storage, and use of patient data [89]. This aspect is critical to the design and implementation of social media campaigns, encompassing the provision of explicit information to patients about the nature of the data being collected, their intended uses, the entities granted access, and the measures in place for their protection. By adopting a transparent approach to data collection and use, health care organizations can foster trust among patients and encourage more robust engagement in social media–driven health interventions. Such transparency can significantly enhance patient health outcomes as individuals are more inclined to engage in interventions in which they feel informed, comfortable, and confident. Examples of computational methods to enhance data transparency include the following: (1) data visualization—this method entails the creation of graphical representations of data to simplify user understanding and interpretation [90]; (2) data profiling—this process analyzes data to ascertain their structure, quality, and content, aiding in the identification of issues such as missing values and inconsistencies [91]; (3) data lineage analysis and provenance tracking—this approach tracks the movement of data through various systems and processes to verify their accuracy and reliability [81,92].

A critical consideration in implementing any of the data transparency methods is ensuring that the autonomy and privacy of all stakeholders are upheld.

Process Transparency

Process transparency denotes the capability to comprehend the procedures involved in the development and deployment of an AI system, including the testing and validation methodologies used [93]. Within the sphere of social media and health care, this notion extends to the clarity of decision-making processes that govern the prioritization, display, and dissemination of health-related information on SMPs. This encompasses transparency regarding the algorithms and computational methods used to curate and showcase health-related content as well as the policies and guidelines governing the moderation of user-generated content pertaining to health. Enhancing process transparency allows users to place greater trust in the information and interventions presented to them and affords researchers increased confidence in the data they examine. Several computational techniques can facilitate enhanced process transparency in AI systems: (1) auditability and monitoring—this involves integrating auditing and monitoring functions within the AI system, including tracking the system's performance, detecting biases or other ethical concerns, and pinpointing instances of underperformance [94]; (2) open-source development—this entails the open and transparent creation of AI systems, where the code, data, and models are made accessible to the public. Such transparency fosters enhanced scrutiny and accountability of the system by external parties, including regulators and the general public [95].

Adopting these methods while recognizing their limitations and taking into account additional ethical considerations can foster greater transparency in AI applications for health care interventions on SMPs.

Explainability and Interpretability

According to the AMIA, the concepts of explainability and interpretability in AI are closely intertwined in the context of transparency. Explainability necessitates that AI developers articulate the functions of AI systems using language appropriate to the context, ensuring that users have a clear understanding of the system's intended use, scope, and limitations. Conversely, interpretability concentrates on the system's capability to elucidate its decision-making processes [12]. It is common for researchers to use the terms explainability and interpretability interchangeably [18,96].

In the realm of social media interventions for health care, explainability and interpretability pertain to comprehending how an AI system processes social media data, identifies pertinent information, and bases its recommendations or decisions on those data [97]. Research conducted by Amann et al [98] delves into the explainability aspects of AI in health care from 4 perspectives: technological, medical, legal, and that of the patient. The authors highlighted the critical role of explainability in the medical domain, arguing that its absence could compromise fundamental ethical values in medicine and public health. The pursuit of explainability and interpretability in AI systems remains a vibrant area of research. For AI systems that apply social media interventions in health care, various methods, including feature selection techniques and visualizations, can facilitate a deeper understanding among health care professionals of the AI system's underlying mechanisms and the factors influencing its decision-making process. As Barredo Arrieta et al [99] noted, techniques for interpretability in AI involve the design of models with clear and comprehensible features, which can aid in identifying the factors that impact the AI's decisions, thus simplifying the understanding and explanation of the outcomes. The existing computational approaches to achieving explainability and interpretability include the following: (1) partial dependence plots (PDPs) [98,100]—PDPs elucidate the relationship between specific input variables and the predicted outcome, offering insights into the rationale behind an AI model's decisions; (2) local interpretable model-agnostic explanations (LIME)—LIME elucidates the outputs of ML models by creating a simpler, interpretable model that approximates the behavior of the original model [101]; (3) Shapley additive explanations (SHAP)—unlike LIME, SHAP explains the outputs of ML models by calculating the contribution of each input feature to the final output [102]; (4) counterfactual explanations—this approach identifies the minimal changes required in the input features to alter the model's output, providing insights into alternative decision pathways [103]; (5) using mathematical structures for analyzing ML model parameters—techniques such as concept activation vectors, t-distributed stochastic neighbor embedding, and singular vector canonical correlation analysis are used for this purpose [104]; (6) attention visualization [105]—techniques for visualizing attention in transformer-based language models used across various NLP tasks on SMPs help reveal the models' inner workings and potential biases; (7) explanation generation—this involves creating natural language or visual explanations for an AI system's decisions (using techniques such as saliency maps, LIME [101], and SHAP [102] in conjunction with NLP methods enhances the generation of comprehensible explanations); (8) applying inherently interpretable models—models such as fuzzy decision trees, which graphically depict the decision-making process akin to standard decision trees, clarify how decisions are made and identify the most influential factors [106]; (9) model distillation—this technique trains a simpler model to approximate the decision boundaries of a more complex model, thereby facilitating the creation of an interpretable model while maintaining the original's performance [107].

While all the aforementioned methods significantly contribute to the explainability and interpretability of AI and ML systems in this domain, it is crucial to recognize their inherent limitations in practical applications. Specifically, PDPs may face challenges with complex unstructured data such as natural language. SHAP can become computationally intensive when dealing with a large number of input features, which is typical in complex models. LIME might yield inconsistent outcomes, and the interpretations from attention visualization techniques necessitate detailed analysis by experts. Explanation generation, which is often dependent on the aforementioned methods, can inherit their flaws, potentially resulting in misleading explanations. Finally, models that are inherently interpretable or refined through distillation techniques might oversimplify,

failing to fully encapsulate the complexities of health care interventions on SMPs.

Definitions, Computational Methods, and Approaches to Ethics

Overview

Ethics encompasses a wide range of considerations, many of which align with the AI principles recognized by the AMIA. In the realm of AI, ethics generally pertains to the study and practice of crafting and applying AI technologies in ways that are fair, transparent, and advantageous to all stakeholders [108]. The objective of ethical AI is to ensure that AI systems and their decisions are in harmony with human values, uphold fundamental human rights, and do not cause harm or discrimination to individuals or groups. This encompasses issues related to privacy, data protection, bias, accountability, and explainability [109].

Within the sphere of social media, the digital surveillance of public health data from SMPs should adhere to several key principles: (1) beneficence, ensuring that surveillance contributes to better public health outcomes; (2) nonmaleficence, ensuring that the use of data does not undermine public trust; (3) autonomy, either through the informed consent of users or by anonymizing personal details; (4) equity, ensuring equal access for individuals to public health interventions; and (5) efficiency, advocating for legal frameworks that guarantee continuous access to web platforms and the algorithms that guide decision-making [110]. AI-mediated health care interventions must consider affordability and equity across the wider population. In addition, health-related data gathered from social platforms need to be scrutinized for various biases such as population and behavioral biases using appropriate metrics [111]. The following subsections offer insights into different ethical viewpoints and the methods used to evaluate how well AI systems align with these ethical standards. We also present summaries of quantifications of key ethical elements in Multimedia Appendix 5 [112-115].

Philosophical Ethics

Our review concentrated primarily on the practical application of ethical principles in AI rather than exploring the purely philosophical dimensions of ethics. Consequently, this subsection focuses on a set of general ethical principles directly pertinent to AI. Kazim and Koshiyama [116] examined various philosophical aspects of ethics and supported a human-centric approach to AI. This perspective underscores the significance of designing and using AI systems in ways that uphold human autonomy, dignity, and privacy [116]. Within the realm of health care interventions on social media, the philosophical ethics of AI can be specifically perceived as the application of ethical principles and values to the development and use of AI-powered tools and technologies [117]. This entails scrutinizing the potential benefits and risks associated with using AI to gather, analyze, and interpret health-related data from SMPs. It also involves ensuring that the deployment of such technologies adheres to the ethical principles recognized by the AMIA, including autonomy, beneficence, and nonmaleficence [12]. The ultimate goal is to foster the development and use of AI technologies that enhance health outcomes while minimizing the potential risks and harms that could emerge from their application. Examples of computational methods and models for addressing philosophical ethics include the following: (1) Methods and models focused on the simulation and modeling of ethical dilemmas, such as those using model-based control and Pavlovian mechanisms, are instrumental. These approaches offer valuable insights into the likely outcomes of diverse ethical decisions [118]. (2) Game theory experiments serve as a pivotal means to model and analyze decision-making processes in social contexts, encompassing ethical dilemmas. Notable examples of these experiments include the ultimatum game, the trust game, and the prisoner's dilemma [119]. (3) The field of data analytics provides methods and models that leverage statistical methods and ML algorithms to scrutinize data. This analysis aims to unearth patterns or insights pertinent to ethical questions or dilemmas [120].

Overall, while methods and models for simulating and modeling ethical dilemmas are capable of effectively representing various scenarios and predicting outcomes, there is a risk that they might oversimplify the complexities inherent in real-world ethics and fail to fully encapsulate the nuances of human ethical reasoning. Although game theory experiments provide insightful perspectives on human behavior in ethical dilemmas, they possess an abstract nature that may limit their practical applicability in realistic situations. Moreover, the efficacy of data analytics methods is heavily dependent on the quality and quantity of the available data. Thus, the application of these methodologies in AI for health care–related interventions on social media should be approached with caution. It is essential to ensure that such applications are in alignment with broader ethical principles.

Professional Ethics

In the context of health care interventions via social media, professional ethics refers to a set of guidelines and principles that guide the behavior of health care professionals engaging with social media as part of their practice [121]. These guidelines may cover aspects such as patient privacy; confidentiality; informed consent; and the appropriate use of SMPs for disseminating health information, which includes avoiding conflicts of interest or biased behavior [122]. Algorithms that are designed to detect and flag fraudulent behavior among stakeholders can play a crucial role in identifying potential breaches of professional ethics [123]. Various modeling approaches, such as the living laboratory model, can support the development of health care professional ethics on SMPs [124]. Some researchers call for the development and implementation of local policies at health care organizations to govern the social media activities of health care professionals, highlighting the significant risks associated with the dissemination of information in health care–related social media endeavors [125].

While enforcing professional ethics is vital, it poses challenges, particularly when the methods used may infringe on the autonomy of stakeholders. The strategies mentioned, although essential for upholding ethics, could inadvertently overstep

boundaries, thus eliciting concerns regarding the autonomy and privacy of the individuals involved.

Legal Ethics

Legal ethics refers to the ethical considerations related to complying with the laws, regulations, and policies surrounding health care data privacy and security. This encompasses safeguarding the confidentiality of patient data, adhering to informed consent and data-sharing agreements, and complying with relevant legal and ethical standards [126,127]. Furthermore, it necessitates ensuring that AI models used in social media interventions for health care are developed and used in conformity with applicable regulations and standards. The existing regulatory and ethical oversight frameworks include the following: (1) the Health Insurance Portability and Accountability Act (HIPAA)—this framework is dedicated to implementing privacy regulations for health care data [145]; (2) the General Data Protection Regulation (GDPR)—it mandates compliance with data protection laws and adherence to other relevant legal and regulatory frameworks governing the use of AI in health care and social media interventions [128]; (3) ethical review boards—advocating for Ethics by Design, this approach involves integrating the services of an ethical review board into the development process of any product within an organization [129].

Both HIPAA and the GDPR are pivotal in the realm of data protection; however, they face intrinsic limitations, with HIPAA being constrained by jurisdictional reach and the GDPR being constrained by the specific subjects it safeguards. The Ethics by Design concept encourages the responsible and ethical development of AI. Nonetheless, this approach could potentially decelerate the innovation process due to the additional layer of review and oversight required during the deployment phase.

Other Ethical Considerations

Guttman [130] highlighted a range of ethical concerns tied to health promotion and communication interventions, including issues related to autonomy, equity, the digital divide, consent, and the risk of unintended adverse effects such as stigmatization of certain groups through the use of derogatory terms to describe their medical conditions. The author stressed the importance of identifying and addressing these issues in the context of health care–related communication interventions [130]. This involves safeguarding the privacy and confidentiality of patient data, respecting patient autonomy and consent, and ensuring that the use of SMPs does not harm the patient [131]. Gagnon and Sabus [132] recognized the concerns that health care professionals may have regarding the use of SMPs due to potential factual inaccuracies. Nevertheless, they argued that using social media in health care does not inherently breach ethical principles as long as evidence-based practices are followed, digital professionalism is upheld through controlled information sharing, and the potential benefits of disseminated information outweigh the risks [132].

Bhatia-Lin et al [133] suggested a rubric approach for the ethical use of SMPs in research that is applicable to health care–associated research involving social media surveillance. Wright [63] introduced a framework for assessing the ethical implications of a wide range of technologies whose comprehensiveness renders it a suitable baseline for evaluating the ethical implications of using AI in social media and health care contexts. Various tools, methods, and approaches can aid in ensuring the ethical use of AI within the health care domain on SMPs: (1) data visualization tools—these tools are designed to present complex ethical data in a clear and accessible manner, thus aiding health care professionals and other stakeholders in understanding and making informed decisions [134]; (2) sentiment analysis of social media posts related to health care interventions—this technique identifies ethical issues and concerns, such as biases or stigmatization of certain patient groups, by analyzing the sentiment of social media content [135]; (3) crowdsourcing platforms for ethical feedback—these platforms are developed to gather insights from a wide range of individuals on the ethical implications of AI systems and their recommendations, ensuring the inclusion of diverse perspectives and values (this approach highlights potential ethical concerns that development teams may otherwise overlook [136]); (4) fairness-aware ML algorithms—these algorithms are designed to address and mitigate unfairness in both the training data and the algorithmic decision-making process with the goal of promoting equity [137]; (5) privacy-preserving data analysis—this method emphasizes the protection of sensitive data from unauthorized access while enabling meaningful analysis, thus balancing privacy with utility [138,139]; (6) human-in-the-loop approaches by incorporating human oversight and decision-making into AI systems, these approaches aim to ensure that technology aligns with social values and ethical principles, thereby promoting responsible use [140]; (7) value-sensitive design—this approach focuses on identifying and integrating social values and ethical principles into the design and development of AI systems, thereby promoting their alignment with societal ethics [141].

In summary, each method has distinct applications and limitations. For instance, sentiment analysis of health care–related social media posts is effective in identifying ethical issues such as biases or stigmatization, yet it is susceptible to misinterpretation due to the inherent ambiguity of natural language. On the other hand, human-in-the-loop approaches may introduce subjectivity and diminish the efficiency of automated systems. Consequently, stakeholders involved in applying AI in social media within the health care domain should be cognizant of these methods' inherent limitations before implementation.

Discussion

Principal Findings and Future Research Directions

Overview

Health care providers leverage social media to advertise their services, engage with individuals, and cultivate community bonds [146]. SMPs enable medical professionals to interact with patients and gather feedback, thereby enhancing patient care. Moreover, social media acts as a medium for health promotion via peer support and disease awareness initiatives and enabling web-based consultations between physicians and patients [147]. To combat misinformation, implementing

rigorous fact-checking measures is imperative for the dissemination of accurate health information. It is also vital to oversee the use of these platforms by health professionals to ensure the protection of patient confidentiality.

The key findings of this study are outlined in the following sections.

RQ 1: What Existing Solutions Address FATE in the Context of Health Care on SMPs?

There are 4 identified solutions to FATE in health care discussions on SMPs. First, fairness in this domain is tackled through calibrated, statistical, and intersectional approaches. Calibrated fairness seeks to balance equal opportunities with individual differences, such as language or location. Statistical fairness uses demographic data to prevent biases. Intersectional fairness examines various aspects of an individual's identity. Second, accountability in health care on SMPs is ensured by adhering to legal standards, incorporating ethical principles into system design, and maintaining technical functionality and privacy, as well as through societal regulation and standardization. These measures include protecting data privacy, preventing discriminatory or unethical use of AI systems, conducting ethical impact assessments, enhancing transparency, involving stakeholders, carrying out audits and evaluations, and holding decision makers responsible. Third, transparency in AI within health care on social media emphasizes the importance of understanding AI systems, including their algorithms, data sources, and decision-making processes. Transparency is vital for comprehending how interventions are crafted, disseminated, and assessed, playing a significant role in identifying and rectifying biases or errors, fostering trust among stakeholders, and improving participation in social media–based health interventions. Fourth, ethics in health care on SMPs focuses on the development of AI technologies that are fair, transparent, and beneficial. This encompasses considerations of privacy, data protection, bias, accountability, and explainability. Upholding professional and social ethics, such as ensuring patient privacy and autonomy, is crucial. The primary aim is to guarantee the ethical use of AI in health care on SMPs while reducing potential risks and adverse effects.

RQ 2: How Do the Different Solutions Identified in Response to RQ 1 Compare to Each Other in Terms of Computational Methods, Approaches, and Evaluation Metrics?

The various solutions identified in response to RQ 1 can be compared based on computational methods, approaches, and evaluation metrics. These solutions encompass strategies for achieving calibrated, statistical, and intersectional fairness through a variety of computational methods, including data preprocessing, postprocessing, adversarial training, and decision tree use. Key evaluation metrics for assessing these solutions are equal opportunity and equalized odds. Accountability can be examined from multiple perspectives: legal accountability, achieved through regulatory measures and public-private partnerships; technical accountability, emphasizing logging and auditing; and ethical accountability, focusing on the identification of ethical risks through methods such as ethical impact assessments, value alignment, and stakeholder engagement. Transparency is attainable through several strategies: algorithmic transparency, data transparency, process transparency, and the interpretability and explainability of models. Enhancements in algorithmic transparency can be achieved through feature importance analysis, interpretability techniques for models, and the generation of explanations. Data transparency improvements are facilitated by data visualization, profiling, lineage analysis, and provenance tracking. Process transparency can be bolstered by auditability, monitoring, and adoption of open-source development practices. Although interpretability and explainability remain burgeoning research areas, there is a diverse range of methods for attaining these goals, each suitable for specific contexts. The promotion of ethics in health care on SMPs involves the use of simulation, modeling, data analytics, sentiment analysis, crowdsourcing, and automated systems considering both professional and social ethics.

RQ 3: What Is the Strength of the Evidence Supporting the Different Solutions?

The strength of the evidence supporting the solutions is variable and influenced by research quality, methodology, and the statistical significance of the findings. Concepts such as calibrated, statistical, and intersectional fairness are grounded in substantial research. Computational methods, including data preprocessing, adversarial training, and the use of decision trees, are widely adopted, although the extent of evidence supporting their efficacy varies. Evaluation metrics such as equal opportunity and equalized odds rely on well-established statistical measures, but their applicability and effectiveness can differ across studies. Within the ethics domain of health care on SMPs, the principles of privacy protection and bias mitigation are robustly supported by research; however, the evidence for the effectiveness of specific solutions may vary. Techniques such as simulation, modeling, data analytics, and crowdsourcing are commonly used, with their success dependent on the specific application context. Due to the rapidly evolving nature of this field, consulting current and reputable sources is essential for accessing the latest research findings.

The findings from this study contribute to the evolving landscape of AI applications within health care on SMPs by enhancing the understanding of the ethical considerations essential for deploying AI in health care. They delineate practical pathways for leveraging social media to improve patient care and engagement. This study offers insights into achieving fairness in this domain through calibrated, statistical, and intersectional approaches, presenting methodologies that balance personalized care with broader demographic considerations and effectively address biases. It identifies accountability measures such as transparency, documentation, adjudication, stakeholder engagement, logging, and auditing as essential for the design and regulation of AI, ensuring its responsible use in health care contexts. Achieving public transparency presents technical and practical challenges; however, entities involved in AI applications within health care should provide comprehensive reports on decision-making factors, data origins and use, and solid scientific evidence supporting their decisions to stakeholders upon request. Finally, ethical considerations,

encompassing philosophical, professional, and legal dimensions, should drive the implementation of the 3 core components of FATE: fairness, accountability, and transparency.

Our study identified several research gaps in AI systems within health care on SMPs. First, primary challenge in the integration of AI and health care on SMPs is the collection and use of data that accurately represent diverse populations without inherent biases. Trustworthy data sets are crucial for training large language models for clinical applications, yet these data sets often lack diversity in key demographics such as age, ethnicity, or medical history. This shortfall can result in AI predictions that disproportionately benefit certain groups. Moreover, the process of obtaining informed consent on SMPs is complicated by the limited understanding users have of how their data might be used for health care research. A common scenario involves the use of patient-generated data from web-based health forums or social media support groups where consent is ambiguously defined, thereby raising ethical and privacy concerns. Second, the operationalization of the broad set of ethical principles defined by the AMIA into a cohesive FATE framework presents significant challenges. The pursuit of a unified approach that addresses the components of FATE simultaneously is hampered by potential conflicts among these principles. For example, increasing transparency by making AI decision-making processes more comprehensible can inadvertently risk patient privacy and system security by exposing sensitive data or proprietary algorithms. Third, the application of FATE principles in real-world health care interventions on SMPs is critically underdocumented. There is a notable absence of comprehensive case studies that detail the implementation, challenges, and outcomes of ethical frameworks in practice. Such documentation is essential for grasping how theoretical ethical considerations are translated into practical impacts and for pinpointing areas that need adjustment when applying these principles. The effectiveness and ethical considerations of AI-driven public health campaigns on platforms such as Twitter and Facebook, for instance, are largely unexplored in a manner that would provide actionable insights into their real-world impact and ethical ramifications. Fourth, the current landscape of evaluating FATE in AI systems, particularly at the intersection of health care and social media, is characterized by a lack of methods that can be universally applied across different models and data types. The specific challenges of the health care domain on SMPs, which include the necessity to analyze diverse data formats in real time, call for the development of model-agnostic tools for ethical assessment. Most existing methods are designed for particular models or data types and do not comprehensively address the wide range of health care applications on social media. Furthermore, there is an absence of a clear strategy for assessing the impact of various AI-assisted interactions between health care and social media domains.

Given the identified gaps, our study proposes 5 research directions. First, research should focus on the development of comprehensive models that integrate the FATE framework with the broader ethical principles outlined by the AMIA. This involves pioneering methodologies that ensure a balanced consideration of all ethical dimensions, aiming to uphold each without compromising the significance or effectiveness of the others. For medical professionals and researchers, this direction represents a shift toward creating AI systems in health care that are both technologically advanced and ethically robust, ensuring equitable and responsible AI use in patient care and data management. Second, investigations are needed into merging computational methods with ethical evaluations to devise sophisticated mathematical formulations capable of quantitatively assessing ethical components in AI applications within health care on SMPs. By developing robust metrics and evaluation frameworks, researchers can bridge the theoretical ethical considerations with practical computational methods. This effort aims to facilitate the integration of ethical principles into the design and evaluation of AI technologies, ensuring that they meet the highest standards of medical ethics and patient care. Third, exploration is required into ethical trade-offs by focusing on understanding and mitigating inherent conflicts between different ethical components within the FATE framework. By systematically examining these trade-offs, research could aim to find innovative solutions that minimize conflicts, such as between transparency and privacy or between fairness and accountability. For the medical and research community, acknowledging and navigating these trade-offs is crucial for the development and implementation of AI systems that are both ethically responsible and effective in achieving health care goals. Fourth, investigation is necessary into the application of FATE principles in real health care interventions on SMPs. This direction seeks to understand the ethical impact of these technologies on users and society. Focusing on the ethical implications of AI-driven health care solutions, from patient engagement strategies to public health campaigns on social media, this research direction aims to ensure that they positively contribute to user well-being and societal health standards. Fifth, a strategic approach should be identified to evaluate the impact of AI-assisted interactions within health care and social media from a FATE perspective. This includes analyzing these interactions to develop universal, model-agnostic metrics that assess the ethical dimensions of AI applications across various platforms. Once established, such metrics could be integrated into social networks, guiding the regulation of AI use in health care on SMPs. For medical professionals and researchers, these metrics would provide a framework for consistently evaluating and ensuring the ethical integrity of AI technologies, promoting safer and more beneficial health care interactions on social media.

Limitations

The primary limitation of our study stems from the scarcity of comprehensive research that thoroughly explores all dimensions of FATE in the context of AI applications in health care on SMPs. This scarcity reflects not only existing research gaps but also the early stage of scholarly inquiry in this interdisciplinary area. Consequently, our review may not fully encapsulate the complex and multidimensional nature of how FATE intersect and manifest in the deployment of AI within health care settings on social media. This limitation is significant because it suggests that our understanding of FATE issues in this context may rely on an incomplete picture, thus impacting the generalizability of our findings across all potential AI applications in health care on social media.

In addition, identifying the precise population of studies relevant to FATE in AI and health care on SMPs is made more challenging by the heterogeneity and dynamism of SMPs as well as the diversity of AI applications within health care. SMPs are rapidly evolving, introducing new functionalities and altering user interactions, which in turn influences how AI technologies can be applied and examined within these contexts. The challenge of compiling a representative collection of studies that fully encompasses this range contributes to potential gaps in our review, limiting the degree to which our findings can be seen as representative of the field as a whole.

Moreover, the fast-paced advancement of technology, along with the continual evolution of both SMPs and AI, imposes a temporal limitation on our study. Research that was up-to-date at the time of our review may soon become outdated as new technologies emerge and existing ones advance. This swift pace of change implies that the ethical challenges identified today may evolve, new challenges may surface, and previously proposed solutions may become obsolete or less applicable. Therefore, the applicability of our findings is inherently limited by this temporal aspect, underscoring the necessity for ongoing research to continuously refresh our understanding of FATE within AI in health care on SMPs.

Conclusions

Our review sheds light on the current state of FATE in health care AI as applied to SMPs. It offers a critical analysis of the methodologies, computational techniques, and evaluative strategies used in various studies. By examining the successes and identifying the shortcomings of current practices, this review stimulates further innovation in the field. It challenges existing paradigms on how AI technologies can be both technologically advanced and ethically robust, ensuring fairness, accountability, and transparency in their application.

The practical implications of this work are substantial. First, it guides future research by identifying recent trends and research gaps, suggesting that researchers focus on creating more robust, fair, and ethical AI systems. This includes using diverse data sets that more accurately represent the global population and using evaluation metrics that comprehensively assess the systems' impacts on all stakeholders. Second, this review underscores the importance of integrating FATE principles throughout the AI system development life cycle, from conceptualization to deployment. For practitioners in health care and technology, this signifies a move toward more inclusive, transparent, and ethically guided development processes. Such a transition not only addresses biases and accountability issues but also boosts patient trust and engagement with AI-driven health care solutions on social media.

Third, the insights from this review are invaluable for policy makers and regulatory bodies, aiding in the creation of nuanced regulations and guidelines that ensure that AI technologies positively contribute to health care outcomes without compromising ethical standards or patient rights. Furthermore, by simplifying complex concepts, this review acts as an educational tool for a broad audience, including health care providers, AI developers, patients, and the general public. Raising awareness about the importance of FATE in health care AI fosters more informed participation in discussions and decision-making regarding AI use in health care, particularly on SMPs.

Ultimately, this study aids in the pursuit of ethical development and deployment of AI systems in health care. By providing an in-depth analysis of the current achievements and future directions for FATE in health care AI on social media, it advocates for the adoption of best practices that balance ethical considerations with technological innovations. The implications of this study extend beyond academia, affecting how AI technologies are conceptualized, developed, and implemented in health care on social media, thereby shaping a future where AI-driven health care solutions are not only effective and innovative but also ethically responsible, equitable, and transparent. This ensures that these technologies serve the best interests of society.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Studies selected for the review.
[DOCX File , 29 KB - medinform_v12i1e50048_app1.docx]

Multimedia Appendix 2
Fairness evaluation metrics with mathematical formulation.
[DOCX File , 27 KB - medinform_v12i1e50048_app2.docx]

Multimedia Appendix 3
Accountability evaluation metrics with mathematical formulation.

[DOCX File , 20 KB - medinform_v12i1e50048_app3.docx]

Multimedia Appendix 4
Transparency evaluation metrics with mathematical formulation.
[DOCX File , 20 KB - medinform_v12i1e50048_app4.docx]

Multimedia Appendix 5
Ethics evaluation metrics with mathematical formulation.
[DOCX File , 18 KB - medinform_v12i1e50048_app5.docx]

References

1. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021 Jul 13;54(6):1-35. [doi: 10.1145/3457607]
2. Mashhadi A, Winder SG, Lia EH, Wood SA. No walk in the park: the viability and fairness of social media analysis for parks and recreational policy making. Proc Int AAAI Conf Web Soc Media 2021 May 22;15(1):409-420. [doi: 10.1609/icwsm.v15i1.18071]
3. Leonelli S, Lovell R, Wheeler BW, Fleming L, Williams H. From FAIR data to fair data use: methodological data fairness in health-related social media research. Big Data Soc 2021 May 03;8(1). [doi: 10.1177/20539517211010310]
4. Singhal A, Baxi MK, Mago V. Synergy between public and private health care organizations during COVID-19 on Twitter: sentiment and engagement analysis using forecasting models. JMIR Med Inform 2022 Aug 18;10(8):e37829 [FREE Full text] [doi: 10.2196/37829] [Medline: 35849795]
5. Kington RS, Arnesen S, Chou WY, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021:10.31478/202107a [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]
6. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med 2018 May 28;7(6):121 [FREE Full text] [doi: 10.3390/jcm7060121] [Medline: 29843360]
7. Flores L, Young SD. Ethical considerations in the application of artificial intelligence to monitor social media for COVID-19 data. Minds Mach (Dordr) 2022;32(4):759-768 [FREE Full text] [doi: 10.1007/s11023-022-09610-0] [Medline: 36042870]
8. Pirraglia PA, Kravitz RL. Social media: new opportunities, new ethical concerns. J Gen Intern Med 2013 Feb 8;28(2):165-166 [FREE Full text] [doi: 10.1007/s11606-012-2288-x] [Medline: 23225258]
9. Wieringa M. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020 Presented at: FAT* '20; January 27-30, 2020; Barcelona, Spain URL: https://dl.acm.org/doi/abs/10.1145/3351095.3372833 [doi: 10.1145/3351095.3372833]
10. Hutchinson B, Smart A, Hanna A, Denton E, Greer C, Kjartansson O, et al. Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual event, Canada. [doi: 10.1145/3442188.3445918]
11. Johnson SL. AI, machine learning, and ethics in health care. J Leg Med 2019;39(4):427-441. [doi: 10.1080/01947648.2019.1690604] [Medline: 31940250]
12. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, et al. Defining AMIA's artificial intelligence principles. J Am Med Inform Assoc 2022 Mar 15;29(4):585-591 [FREE Full text] [doi: 10.1093/jamia/ocac006] [Medline: 35190824]
13. The Belmont report. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html [accessed 2023-12-05]
14. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud 2021 Feb;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
15. Hagerty A, Rubinov I. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv Preprint posted online July 18, 2019 [FREE Full text] [doi: 10.48550/arXiv.1907.07892]
16. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018 Presented at: DSAA 2018; October 1-3, 2018; Turin, Italy. [doi: 10.1109/dsaa.2018.00018]
17. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: a survey of results. In: Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. 2017 Presented at: SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017; August 4-8, 2017; San Francisco, CA. [doi: 10.1109/uic-atc.2017.8397411]

18. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. Electronics 2019 Jul 26;8(8):832. [doi: 10.3390/electronics8080832]
19. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res 2017 Jun 06;19(6):e195 [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]
20. Bear Don't Walk OJ4, Reyes Nieva H, Lee SS, Elhadad N. A scoping review of ethics considerations in clinical natural language processing. JAMIA Open 2022 Jul;5(2):ooac039 [FREE Full text] [doi: 10.1093/jamiaopen/ooac039] [Medline: 35663112]
21. Attard-Frost B, De los Ríos A, Walters DR. The ethics of AI business practices: a review of 47 AI ethics guidelines. AI Ethics 2022 Apr 13;3(2):389-406. [doi: 10.1007/s43681-022-00156-6]
22. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138-52160. [doi: 10.1109/access.2018.2870052]
23. Vian T, Kohler JC. Medicines Transparency Alliance (MeTA): pathways to transparency, accountability, and access: cross-case analysis and review of phase II. World Health Organization. 2016 May 25. URL: https://tinyurl.com/3vhjyysd [accessed 2023-12-05]
24. Kofod-Petersen A. How to do a structured literature review in computer science. Norwegian University of Science and Technology. 2018. URL: https://research.idi.ntnu.no/aimasters/files/SLR_HowTo2018.pdf [accessed 2024-03-13]
25. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
26. Saha D, Schumann C, Mcelfresh DC, Dickerson JP, Mazurek ML, Tschantz MC. Measuring non-expert comprehension of machine learning fairness metrics. In: Proceedings of the 37th International Conference on Machine Learning. 2020 Presented at: PMLR 2020; July 13-18, 2020; Virtual event URL: https://proceedings.mlr.press/v119/saha20c.html [doi: 10.1145/3375627.3375819]
27. Mehrabi N, Gupta U, Morstatter F, Steeg GV, Galstyan A. Attributing fair decisions with attention interventions. In: Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022). 2022 Presented at: TrustNLP 2022; July 14, 2022; Seattle, WA. [doi: 10.18653/v1/2022.trustnlp-1.2]
28. Hertweck C, Heitz C, Loi M. On the moral justification of statistical parity. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual event. [doi: 10.1145/3442188.3445936]
29. Yao H, Chen Y, Ye Q, Jin X, Ren X. Refining language models with compositional explanations. arXiv Preprint posted online March 18, 2021 [FREE Full text]
30. Markoulidakis I, Kopsiaftis G, Rallis I, Georgoulas I. Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. In: Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference. 2021 Presented at: PETRA '21; June 29-July 2, 2021; Corfu, Greece. [doi: 10.1145/3453892.3461323]
31. Vergeer P, van Schaik Y, Sjerps M. Measuring calibration of likelihood-ratio systems: a comparison of four metrics, including a new metric devPAV. Forensic Sci Int 2021 Apr;321:110722. [doi: 10.1016/j.forsciint.2021.110722] [Medline: 33684845]
32. Lagioia F, Rovatti R, Sartor G. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. AI Soc 2022 Apr 28;38(2):459-478. [doi: 10.1007/s00146-022-01441-y]
33. Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y. How do fairness definitions fare?: examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019 Presented at: AIES '19; January 27-28, 2019; Honolulu, HI. [doi: 10.1145/3306618.3314248]
34. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open 2021 Apr 01;4(4):e213909 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.3909] [Medline: 33856478]
35. Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, et al. Algorithmic fairness in computational medicine. EBioMedicine 2022 Oct;84:104250 [FREE Full text] [doi: 10.1016/j.ebiom.2022.104250] [Medline: 36084616]
36. Tao G, Sun W, Han T, Fang C, Zhang X. RULER: discriminative and iterative adversarial training for deep neural network fairness. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2022 Presented at: ESEC/FSE '22; November 14-18, 2022; Singapore. [doi: 10.1145/3540250.3549169]
37. Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. Commun ACM 2020 Apr 20;63(5):82-89. [doi: 10.1145/3376898]
38. Yao S, Huang B. Beyond parity: fairness objectives for collaborative filtering. arXiv Preprint posted online March 24, 2017 [FREE Full text]
39. Zhang Y, Zhou L. Fairness assessment for artificial intelligence in financial industry. arXiv Preprint posted online December 16, 2019 [FREE Full text] [doi: 10.5260/chara.21.2.8]

40. Ghassami A, Khodadadian S, Kiyavash N. Fairness in supervised learning: an information theoretic approach. arXiv Preprint posted online January 13, 2018 [FREE Full text] [doi: 10.1109/isit.2018.8437807]
41. Malawski M. A note on equal treatment and symmetry of values. In: Nguyen NT, Kowalczyk R, Mercik J, Motylska-Kuźma A, editors. Transactions on Computational Collective Intelligence XXXV. Berlin, Heidelberg: Springer; 2020.
42. Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, et al. Aequitas: a bias and fairness audit toolkit. arXiv Preprint posted online November 14, 2018 [FREE Full text] [doi: 10.48550/arXiv.1811.05577]
43. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 2019 Jul 1;63(4/5):4:1-:15. [doi: 10.1147/jrd.2019.2942287]
44. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Sep;6(9):856-864 [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
45. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, et al. Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft. 2020 Sep 22. URL: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf [accessed 2023-12-05]
46. Ghosh A, Genuit L, Reagan M. Characterizing intersectional group fairness with worst-case comparisons. arXiv Preprint posted online January 05, 2021 [FREE Full text]
47. Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP. Fairness constraints: a flexible approach for fair classification. J Mach Learn Res 2019;20(75):1-42.
48. Chakraborti T, Patra A, Noble JA. Contrastive fairness in machine learning. IEEE Lett Comput Soc 2020 Jul 7;3(2):38-41. [doi: 10.1109/locs.2020.3007845]
49. Rosenfeld A, Richardson A. Explainability in human–agent systems. Auton Agent Multi-Agent Syst 2019 May 13;33:673-705. [doi: 10.1007/s10458-019-09408-y]
50. Narasimhan H, Cotter A, Gupta M, Wang S. Pairwise fairness for ranking and regression. Proc AAAI Conf Artif Intell 2020 Apr 03;34(04):5248-5255. [doi: 10.1609/aaai.v34i04.5970]
51. Kaur D, Uslu S, Durresi A, Badve S, Dundar M. Trustworthy explainability acceptance: a new metric to measure the trustworthiness of interpretable ai medical diagnostic systems. In: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems. 2021 Presented at: CISIS-2021; July 1-3, 2021; Asan, Korea. [doi: 10.1007/978-3-030-79725-6_4]
52. Bucher M, Herbin S, Jurie F. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: Proceedings of the Computer Vision – ECCV 2016. 2016 Presented at: ECCV 2016; October 11-14, 2016; Amsterdam, The Netherlands. [doi: 10.1007/978-3-319-46454-1_44]
53. Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T. Improved precision and recall metric for assessing generative models. arXiv Preprint posted online April 15, 2019 [FREE Full text]
54. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv Preprint posted online August 13, 2020 [FREE Full text] [doi: 10.48550/arXiv.2008.05756]
55. Zaki MM, Jena AB, Chandra A. Supporting value-based health care - aligning financial and legal accountability. N Engl J Med 2021 Sep 09;385(11):965-967. [doi: 10.1056/NEJMp2105625] [Medline: 34478249]
56. Blacklaws C. Algorithms: transparency and accountability. Philos Trans A Math Phys Eng Sci 2018 Sep 13;376(2128):20170351. [doi: 10.1098/rsta.2017.0351] [Medline: 30082299]
57. Kim B, Park J, Suh J. Transparency and accountability in AI decision support: explaining and visualizing convolutional neural networks for text information. Decis Support Syst 2020 Jul;134:113302. [doi: 10.1016/j.dss.2020.113302]
58. Dubberley S, Murray D, Koenig A. Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability. Oxford, UK: Oxford University Press; 2020.
59. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019 Presented at: FAT* '19; January 29-31, 2019; Atlanta, GA. [doi: 10.1145/3287560.3287596]
60. King J. The instrumental value of legal accountability. In: Accountability in the Contemporary Constitution. Oxford, UK: Oxford University Press; 2013.
61. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019 Nov 04;1:501-507. [doi: 10.1038/s42256-019-0114-4]
62. Kass NE, Faden RR. Ethics and learning health care: the essential roles of engagement, transparency, and accountability. Learn Health Syst 2018 Sep 18;2(4):e10066 [FREE Full text] [doi: 10.1002/lrh2.10066] [Medline: 31245590]
63. Wright D. A framework for the ethical impact assessment of information technology. Ethics Inf Technol 2010 Jul 8;13:199-226. [doi: 10.1007/s10676-010-9242-6]
64. Arnold T, Kasenberg D, Scheutz M. Value alignment or misalignment – what will keep systems accountable? Association for the Advancement of Artificial Intelligence. 2017. URL: https://hrilab.tufts.edu/publications/arnoldetal17aiethics.pdf [accessed 2023-12-05]

65. Iyer R, Li Y, Li H, Lewis M, Sundar R, Sycara K. Transparency and explanation in deep reinforcement learning neural networks. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Presented at: AIES '18; February 2-3, 2018; New Orleans, LA. [doi: 10.1145/3278721.3278776]
66. Fukuda‐Parr S, Gibbons E. Emerging consensus on 'ethical AI': human rights critique of stakeholder guidelines. Glob Policy 2021 Jun 19;12(S6):32-44. [doi: 10.1111/1758-5899.12965]
67. Wachter S, Mittelstadt B, Floridi L. Transparent, explainable, and accountable AI for robotics. Sci Robot 2017 May 31;2(6):eaan6080. [doi: 10.1126/scirobotics.aan6080] [Medline: 33157874]
68. Ozga J. The politics of accountability. J Educ Change 2020;21:19-35. [doi: 10.1007/s10833-019-09354-2]
69. Ko RK, Kirchberg M, Lee BS. From system-centric to data-centric logging - accountability, trust and security in cloud computing. In: Proceedings of the Defense Science Research Conference and Expo. 2011 Presented at: DSR 2011; August 3-5, 2011; Singapore. [doi: 10.1109/dsr.2011.6026885]
70. Raji I, Smart A, White R, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020 Presented at: FAT* '20; January 27-30, 2020; Barcelona, Spain. [doi: 10.1145/3351095.3372873]
71. Nushi B, Kamar E, Horvitz E. Towards accountable AI: hybrid human-machine analyses for characterizing system failure. Proc AAAI Conf Hum Comput Crowdsourc 2018 Jun 15;6(1):126-135. [doi: 10.1609/hcomp.v6i1.13337]
72. Vesnic-Alujevic L, Nascimento S, Pólvora A. Societal and ethical impacts of artificial intelligence: critical notes on European policy frameworks. Telecommun Policy 2020 Jul;44(6):101961. [doi: 10.1016/j.telpol.2020.101961]
73. Kerikmäe T, Pärn-Lee E. Legal dilemmas of Estonian artificial intelligence strategy: in between of e-society and global race. AI Soc 2020 Jul 01;36:561-572. [doi: 10.1007/s00146-020-01009-8]
74. Reich MR. The core roles of transparency and accountability in the governance of global health public-private partnerships. Health Syst Reform 2018;4(3):239-248. [doi: 10.1080/23288604.2018.1465880] [Medline: 30207904]
75. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016 Jun;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
76. Laacke S, Mueller R, Schomerus G, Salloch S. Artificial intelligence, social media and depression. A new concept of health-related digital autonomy. Am J Bioeth 2021 Jul;21(7):4-20. [doi: 10.1080/15265161.2020.1863515] [Medline: 33393864]
77. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013 Oct;46(5):830-836 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.010] [Medline: 23820016]
78. Crawley AW, Divi N, Smolinski MS. Using timeliness metrics to track progress and identify gaps in disease surveillance. Health Secur 2021;19(3):309-317. [doi: 10.1089/hs.2020.0139] [Medline: 33891487]
79. Zhai C, Cohen WW, Lafferty J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. ACM SIGIR Forum 2015 Jun 23;49(1):2-9. [doi: 10.1145/2795403.2795405]
80. Weiss DJ, Nelson A, Gibson HS, Temperley W, Peedell S, Lieber A, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature 2018 Jan 18;553(7688):333-336 [FREE Full text] [doi: 10.1038/nature25181] [Medline: 29320477]
81. Burgess K, Hart D, Elsayed A, Cerny T, Bures M, Tisnovsky P. Visualizing architectural evolution via provenance tracking: a systematic review. In: Proceedings of the Conference on Research in Adaptive and Convergent Systems. 2022 Presented at: RACS '22; October 3-6, 2022; Virtual event. [doi: 10.1145/3538641.3561493]
82. Diakopoulos N, Koliska M. Algorithmic transparency in the news media. Digit J 2016 Jul 27;5(7):809-828. [doi: 10.1080/21670811.2016.1208053]
83. Stellefson M, Paige SR, Chaney BH, Chaney JD. Evolving role of social media in health promotion: updated responsibilities for health education specialists. Int J Environ Res Public Health 2020 Feb 12;17(4):1153 [FREE Full text] [doi: 10.3390/ijerph17041153] [Medline: 32059561]
84. Valko M, Hauskrecht M. Feature importance analysis for patient management decisions. Stud Health Technol Inform 2010;160(Pt 2):861-865 [FREE Full text] [Medline: 20841808]
85. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 2018 Jun 01;16(3):31-57. [doi: 10.1145/3236386.3241340]
86. Slack D, Friedler SA, Scheidegger C, Dutta Roy C. Assessing the local interpretability of machine learning models. arXiv Preprint posted online February 9, 2019 [FREE Full text]
87. Stepin I, Alonso JM, Catala A, Pereira-Farina M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 2021 Jan 13;9:11974-12001. [doi: 10.1109/access.2021.3051315]
88. Bertino E, Merrill S, Nesen A, Utz C. Redefining data transparency: a multidimensional approach. Computer 2019 Jan;52(1):16-26. [doi: 10.1109/MC.2018.2890190]
89. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
90. Li Q. Overview of data visualization. In: Embodying Data. Singapore: Springer; Jun 20, 2020.

91. Azeroual O, Saake G, Schallehn E. Analyzing data quality issues in research information systems via data profiling. Int J Inf Manage 2018 Aug;41:50-56. [doi: 10.1016/j.ijinfomgt.2018.02.007]
92. Tang M, Shao S, Yang W, Liang Y, Yu Y, Saha B, et al. SAC: a system for big data lineage tracking. In: Proceedings of the IEEE 35th International Conference on Data Engineering (ICDE). 2019 Presented at: ICDE 2019; April 8-11, 2019; Macao, China. [doi: 10.1109/icde.2019.00215]
93. Leslie D. Understanding artificial intelligence ethics and safety. arXiv Preprint posted online June 11, 2019 [FREE Full text]
94. Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst 2020 Oct 16;10(4):1-31. [doi: 10.1145/3419764]
95. Brundage M, Avin S, Wang J, Belfield H, Krueger D, Hadfield G, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv Preprint posted online April 15, 2020 [FREE Full text] [doi: 10.48550/ARXIV.2004.07213]
96. Janssen M, Hartog M, Matheus R, Yi Ding A, Kuk G. Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. Soc Sci Comput Rev 2020 Dec 28;40(2):478-493. [doi: 10.1177/0894439320980118]
97. Paredes JN, Teze JC, Martinez MV, Simari GI. The HEIC application framework for implementing XAI-based socio-technical systems. Online Soc Netw Media 2022 Nov;32:100239. [doi: 10.1016/j.osnem.2022.100239]
98. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
99. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
100. Xiong Z, Cui Y, Liu Z, Zhao Y, Hu M, Hu J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Comput Materials Sci 2020 Jan;171:109203. [doi: 10.1016/j.commatsci.2019.109203]
101. Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach Learn Knowl Extr 2021 Jun 30;3(3):525-541. [doi: 10.3390/make3030027]
102. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv Preprint posted online May 2, 2017 [FREE Full text]
103. Sokol K, Flach P. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In: Proceedings of the AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019. 2019 Presented at: SafeAI 2019; January 27, 2019; Honolulu, HI.
104. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021 Nov;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
105. Vig J. A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019 Presented at: ACL 2019; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-3007]
106. Fan CY, Chang PC, Lin JJ, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput 2011 Jan;11(1):632-644. [doi: 10.1016/j.asoc.2009.12.023]
107. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A 2019 Oct 29;116(44):22071-22080 [FREE Full text] [doi: 10.1073/pnas.1900654116] [Medline: 31619572]
108. Leikas J, Koivisto R, Gotcheva N. Ethical framework for designing autonomous intelligent systems. J Open Innov Technol Mark Complex 2019 Mar;5(1):18. [doi: 10.3390/joitmc5010018]
109. Latonero M. Governing artificial intelligence: upholding human rights and dignity. Data & Society. 2018. URL: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf [accessed 2023-12-05]
110. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 02;41:101-118 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
111. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data 2019;2:13 [FREE Full text] [doi: 10.3389/fdata.2019.00013] [Medline: 33693336]
112. Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Presented at: AIES '18; February 2-3, 2018; New Orleans, LA. [doi: 10.1145/3278721.3278729]
113. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD '15; August 10-13, 2015; Sydney, Australia. [doi: 10.1145/2783258.2783311]

114. Mendes R, Cunha M, Vilela JP, Beresford AR. Enhancing user privacy in mobile devices through prediction of privacy preferences. In: Proceedings of the 27th European Symposium on Research in Computer Security. 2022 Presented at: ESORICS 2022; September 26-30, 2022; Copenhagen, Denmark. [doi: 10.1007/978-3-031-17140-6_8]
115. Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: Proceedings of the IEEE Symposium on Security and Privacy (SP). 2016 Presented at: SP 2016; May 22-26, 2016; San Jose, CA. [doi: 10.1109/sp.2016.42]
116. Kazim E, Koshiyama AS. A high-level overview of AI ethics. Patterns (N Y) 2021 Sep 10;2(9):100314 [FREE Full text] [doi: 10.1016/j.patter.2021.100314] [Medline: 34553166]
117. Nebeker C, Parrish EM, Graham S. The AI-powered digital health sector: ethical and regulatory considerations when developing digital mental health tools for the older adult demographic. In: Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues. Cham, Switzerland: Springer; 2022:159-176.
118. Crockett MJ. Models of morality. Trends Cogn Sci 2013 Aug;17(8):363-366 [FREE Full text] [doi: 10.1016/j.tics.2013.06.005] [Medline: 23845564]
119. Colman AM. Game Theory and its Applications: In the Social and Biological Sciences. London, UK: Psychology Press; 1995.
120. Someh IA, Davern M, Breidbach C, Shanks G. Ethical issues in big data analytics: a stakeholder perspective. Commun Assoc Inf Syst 2019 May;44(34):718-747 [FREE Full text] [doi: 10.17705/1CAIS.04434]
121. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520 [FREE Full text] [Medline: 25083128]
122. Ponce SB, M Barry M, S Dizon D, S Katz M, Murphy M, Teplinsky E, et al. Netiquette for social media engagement for oncology professionals. Future Oncol 2022 Mar;18(9):1133-1141 [FREE Full text] [doi: 10.2217/fon-2021-1366] [Medline: 35109663]
123. Drabiak K, Wolfson J. What should health care organizations do to reduce billing fraud and abuse? AMA J Ethics 2020 Mar 01;22(3):E221-E231 [FREE Full text] [doi: 10.1001/amajethics.2020.221] [Medline: 32220269]
124. Neville P, Waylen A. Social media and dentistry: some reflections on e-professionalism. Br Dent J 2015 Apr 24;218(8):475-478. [doi: 10.1038/sj.bdj.2015.294] [Medline: 25908363]
125. Ennis-O-Connor M, Mannion R. Social media networks and leadership ethics in healthcare. Healthc Manage Forum 2020 May;33(3):145-148. [doi: 10.1177/0840470419893773] [Medline: 31884833]
126. Garg T, Shrigiriwar A. Managing expectations: how to navigate legal and ethical boundaries in the era of social media. Clin Imaging 2021 Apr;72:175-177. [doi: 10.1016/j.clinimag.2020.11.005] [Medline: 33296827]
127. Kalkman S, Mostert M, Gerlinger C, van Delden JJ, van Thiel GJ. Responsible data sharing in international health research: a systematic review of principles and norms. BMC Med Ethics 2019 Mar 28;20(1):21 [FREE Full text] [doi: 10.1186/s12910-019-0359-9] [Medline: 30922290]
128. Sharma S. Data Privacy and GDPR Handbook. Hoboken, NJ: John Wiley & Sons; 2019.
129. Leidner JL, Plachouras V. Ethical by design: ethics best practices for natural language processing. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. 2017 Presented at: EthNLP@EACL; April 4, 2017; Valencia, Spain. [doi: 10.18653/v1/w17-1604]
130. Guttman N. Ethical issues in health promotion and communication interventions. Oxford Research Encyclopedias Communication. 2017 Feb 27. URL: https://www.academia.edu/100398082/Ethical_Issues_in_Health_Promotion_and_Communication_Interventions [accessed 2023-12-05]
131. Denecke K, Bamidis P, Bond C, Gabarron E, Househ M, Lau AY, et al. Ethical issues of social media usage in healthcare. Yearb Med Inform 2015 Aug 13;10(1):137-147 [FREE Full text] [doi: 10.15265/IY-2015-001] [Medline: 26293861]
132. Gagnon K, Sabus C. Professionalism in a digital age: opportunities and considerations for using social media in health care. Phys Ther 2015 Mar;95(3):406-414. [doi: 10.2522/ptj.20130227] [Medline: 24903111]
133. Bhatia-Lin A, Boon-Dooley A, Roberts MK, Pronai C, Fisher D, Parker L, et al. Ethical and regulatory considerations for using social media platforms to locate and track research participants. Am J Bioeth 2019 Jun;19(6):47-61 [FREE Full text] [doi: 10.1080/15265161.2019.1602176] [Medline: 31135323]
134. Davis K, Patterson D. Ethics of Big Data. Sebastopol, CA: O'Reilly Media; Sep 2012.
135. Livingston JD, Milne T, Fang ML, Amari E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. Addiction 2012 Jan;107(1):39-50 [FREE Full text] [doi: 10.1111/j.1360-0443.2011.03601.x] [Medline: 21815959]
136. Jakesch M, Buçinca Z, Amershi S, Olteanu A. How different groups prioritize ethical values for responsible AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022 Presented at: FAccT '22; June 21-24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3533097]
137. Pastaltzidis I, Dimitriou N, Quezada-Tavarez K, Aidinlis S, Marquenie T, Gurzawska A, et al. Data augmentation for fairness-aware machine learning: preventing algorithmic bias in law enforcement systems. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022 Presented at: FAccT '22; June 21-24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3534644]

138. Keshk M, Moustafa N, Sitnikova E, Turnbull B. Privacy-preserving big data analytics for cyber-physical systems. Wireless Netw 2018 Dec 20;28(3):1241-1249. [doi: 10.1007/s11276-018-01912-5]
139. Kayaalp M. Patient privacy in the era of big data. Balkan Med J 2018 Jan 20;35(1):8-17 [FREE Full text] [doi: 10.4274/balkanmedj.2017.0966] [Medline: 28903886]
140. Enarsson T, Enqvist L, Naarttijärvi M. Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. Inf Commun Technol Law 2021 Jul 27;31(1):123-153. [doi: 10.1080/13600834.2021.1958860]
141. Umbrello S, van de Poel I. Mapping value sensitive design onto AI for social good principles. AI Ethics 2021 Feb 01;1(3):283-296 [FREE Full text] [doi: 10.1007/s43681-021-00038-3] [Medline: 34790942]
142. Hossin M, Sulaiman MN, Mustapha A, Mustapha N, Rahmat RW. A hybrid evaluation metric for optimizing classifier. In: Proceedings of the 3rd Conference on Data Mining and Optimization (DMO). 2011 Presented at: DMO 2011; June 28-29, 2011; Putrajaya, Malaysia. [doi: 10.1109/dmo.2011.5976522]
143. Nguyen AT, Raff E, Nicholas C, Holt J. Leveraging uncertainty for improved static malware detection under extreme false positive constraints. arXiv Preprint posted online August 9, 2021 [FREE Full text] [doi: 10.48550/arXiv.2108.04081]
144. Jadon S. A survey of loss functions for semantic segmentation. In: Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. 2020 Presented at: CIBCB 2020; October 27-29, 2020; Via del Mar, Chile. [doi: 10.1109/cibcb48159.2020.9277638]
145. Hansen E. HIPAA (Health Insurance Portability and Accountability Act) rules: federal and state enforcement. Med Interface 1997 Aug;10(8):96-8, 101. [Medline: 10169779]
146. Grajales FJ3, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014 Feb 11;16(2):e13 [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
147. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]

Abbreviations

AI: artificial intelligence
AMIA: American Medical Informatics Association
FATE: fairness, accountability, transparency, and ethics
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LIME: local interpretable model-agnostic explanations
ML: machine learning
NLP: natural language processing
PDP: partial dependence plot
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RQ: research question
SHAP: Shapley additive explanations
SMP: social media platform

Edited by A Castonguay; submitted 18.06.23; peer-reviewed by G Randhawa, D Valdes, M Arab-Zozani; comments to author 28.10.23; revised version received 21.12.23; accepted 15.02.24; published 03.04.24.

Please cite as:
Singhal A, Neveditsin N, Tanveer H, Mago V
Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review
JMIR Med Inform 2024;12:e50048
URL: https://medinform.jmir.org/2024/1/e50048
doi:10.2196/50048
PMID:38568737

Review

Evaluating the Prevalence of Burnout Among Health Care Professionals Related to Electronic Health Record Use: Systematic Review and Meta-Analysis

Yuxuan Wu[1], MMed; Mingyue Wu[2], MSc; Changyu Wang[3], BSc; Jie Lin[4*], MSN; Jialin Liu[1,2*], MD; Siru Liu[5], PhD

[1]Department of Medical Informatics, West China Hospital, Sichuan University, Chengdu, China

[2]Information Center, West China Hospital, Sichuan University, Chengdu, China

[3]West China College of Stomatology, Sichuan University, Chengdu, China

[4]Department of Oral Implantology, West China Hospital of Stomatology, Sichuan University, Chengdu, China

[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Information Center
West China Hospital
Sichuan University
37 Guoxue Road
Chengdu, 610041
China
Phone: 86 28 85422306
Fax: 86 28 85582944
Email: Dljl8@163.com

Abstract

Background: Burnout among health care professionals is a significant concern, with detrimental effects on health care service quality and patient outcomes. The use of the electronic health record (EHR) system has been identified as a significant contributor to burnout among health care professionals.

Objective: This systematic review and meta-analysis aims to assess the prevalence of burnout among health care professionals associated with the use of the EHR system, thereby providing evidence to improve health information systems and develop strategies to measure and mitigate burnout.

Methods: We conducted a comprehensive search of the PubMed, Embase, and Web of Science databases for English-language peer-reviewed articles published between January 1, 2009, and December 31, 2022. Two independent reviewers applied inclusion and exclusion criteria, and study quality was assessed using the Joanna Briggs Institute checklist and the Newcastle-Ottawa Scale. Meta-analyses were performed using R (version 4.1.3; R Foundation for Statistical Computing), with EndNote X7 (Clarivate) for reference management.

Results: The review included 32 cross-sectional studies and 5 case-control studies with a total of 66,556 participants, mainly physicians and registered nurses. The pooled prevalence of burnout among health care professionals in cross-sectional studies was 40.4% (95% CI 37.5%-43.2%). Case-control studies indicated a higher likelihood of burnout among health care professionals who spent more time on EHR-related tasks outside work (odds ratio 2.43, 95% CI 2.31-2.57).

Conclusions: The findings highlight the association between the increased use of the EHR system and burnout among health care professionals. Potential solutions include optimizing EHR systems, implementing automated dictation or note-taking, employing scribes to reduce documentation burden, and leveraging artificial intelligence to enhance EHR system efficiency and reduce the risk of burnout.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42021281173; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021281173

(JMIR Med Inform 2024;12:e54811) doi:10.2196/54811

KEYWORDS

clinical decision support system; electronic health record; electronic medical record; health information technology; alert fatigue; burnout; health care professionals; health care service; EHR; systematic review; meta-analysis; health information system; clinician burnout; health informatics

Introduction

The integration of electronic health record (EHR) into health care systems marks the beginning of a new era in medical information management, with significant potential benefits for patient care, clinical decision-making, and administrative efficiency [1,2]. EHR systems are central to the modern health care infrastructure [3]. Along with these benefits, however, the widespread adoption of EHR systems has raised concerns about the well-being of health care professionals [4,5]. Unintended consequences, such as burnout among health care professionals, technology-related errors, and increased safety risks, have been associated with EHR use [4,6,7]. In addition, a notable part of the problems with EHR systems in the United States is the need to provide additional documentation for insurance companies [8].

Within the realm of EHR use, burnout among health care professionals, characterized by emotional exhaustion, depersonalization, and a diminished sense of personal accomplishment, has emerged as a critical concern [9,10]. Burnout among health care professionals has become a pressing public health concern [11-13]. Some studies have reported an average burnout prevalence of 44% [2], with rates exceeding 80% in some specific settings and departments [4,5] such as primary care and emergency departments. This pervasive problem affects not only health care professionals but also patients, with negative consequences such as reduced quality of care and increased medical errors and psychological problems [14-17]. The estimated annual cost of burnout among health care professionals due to medical negligence and staff turnover exceeds US $4 billion [18].

The phenomenon of burnout among health care professionals goes beyond individual distress and has significant implications for patient safety, quality of care, and overall health system performance [14,15,19]. Understanding the prevalence and underlying factors of EHR-related burnout among health care professionals is critical to developing effective interventions and policy adaptations. These interventions are essential to mitigate this burden and ensure the long-term sustainability of EHR implementation in health care [19,20]. The increase in EHR-related burnout among health care professionals reflects a multifaceted interplay of factors, including increased documentation requirements, cumbersome user interfaces, and the rapid pace of technological development [9,16,18].

This systematic review and meta-analysis aims to provide a comprehensive assessment of the existing literature on EHR-related provider burnout. It seeks to capture the full extent of burnout, identify its causes, and provide evidence-based support and recommendations to alleviate this pervasive problem. In addition, we hypothesize that specific features of EHR systems, such as user interface design or increased documentation requirements, may contribute to provider burnout. We hope that this work will serve as a guide for health care organizations, policy makers, and EHR developers in developing interventions and technological improvements that prioritize the well-being of health care professionals. In doing so, we can promote a sustainable and resilient health care system while harnessing the potential benefits of EHR systems to improve patient care.

Methods

Study Guidelines

We focused on studies that directly measured burnout, as it is often considered in existing research to be a distinct emotional state, separate from depression or anxiety. This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines [21] and was registered with PROSPERO (CRD42021281173). Details of the guidelines and registration can be found in Multimedia Appendices 1 and 2, respectively.

Definitions

Our definitions of burnout were based on the Maslach Burnout Inventory-Human Services Survey instrument (MBI-HSS) [13,22], which characterizes burnout with high emotional exhaustion as a score ≥27, high depersonalization as >10, and low personal accomplishment as <33. Across the included studies, burnout was defined inconsistently, with definitions ranging from any one of the 3 items to all 3 items. In cases where the same study examined multiple definitions of burnout, we used the most common definition (high emotional exhaustion, high depersonalization, and low personal accomplishment) for the meta-analysis. For alternative definitions, such as those from the Stanford Physician Wellness Survey [23] or mini-Z [24], only outcomes explicitly described as burnout were documented. We categorized studies according to the measurement tool and definition of burnout.

Search Strategy

We systematically searched PubMed, Embase, and Web of Science to identify relevant peer-reviewed English language studies published between January 1, 2009, and December 31, 2022. We used several search terms to capture EHR systems, including “electronic health record” and its abbreviation “EHR,” as well as “electronic medical record (EMR)” and “computerized physician order entry (CPOE).” To capture the phenomenon of burnout, we used terms such as “burnout,” “alert fatigue,” and “exhaustion.” In defining our study participants, we considered a spectrum of health care professionals, including “doctor,” “clinician,” “physician,” “surgeon,” “medical staff,” and “health care provider.” On June 30, 2023, the researchers conducted a literature search in databases such as PubMed, Embase, and Web of Science, following the previously established search strategy. No papers were found that met the inclusion criteria for this review.

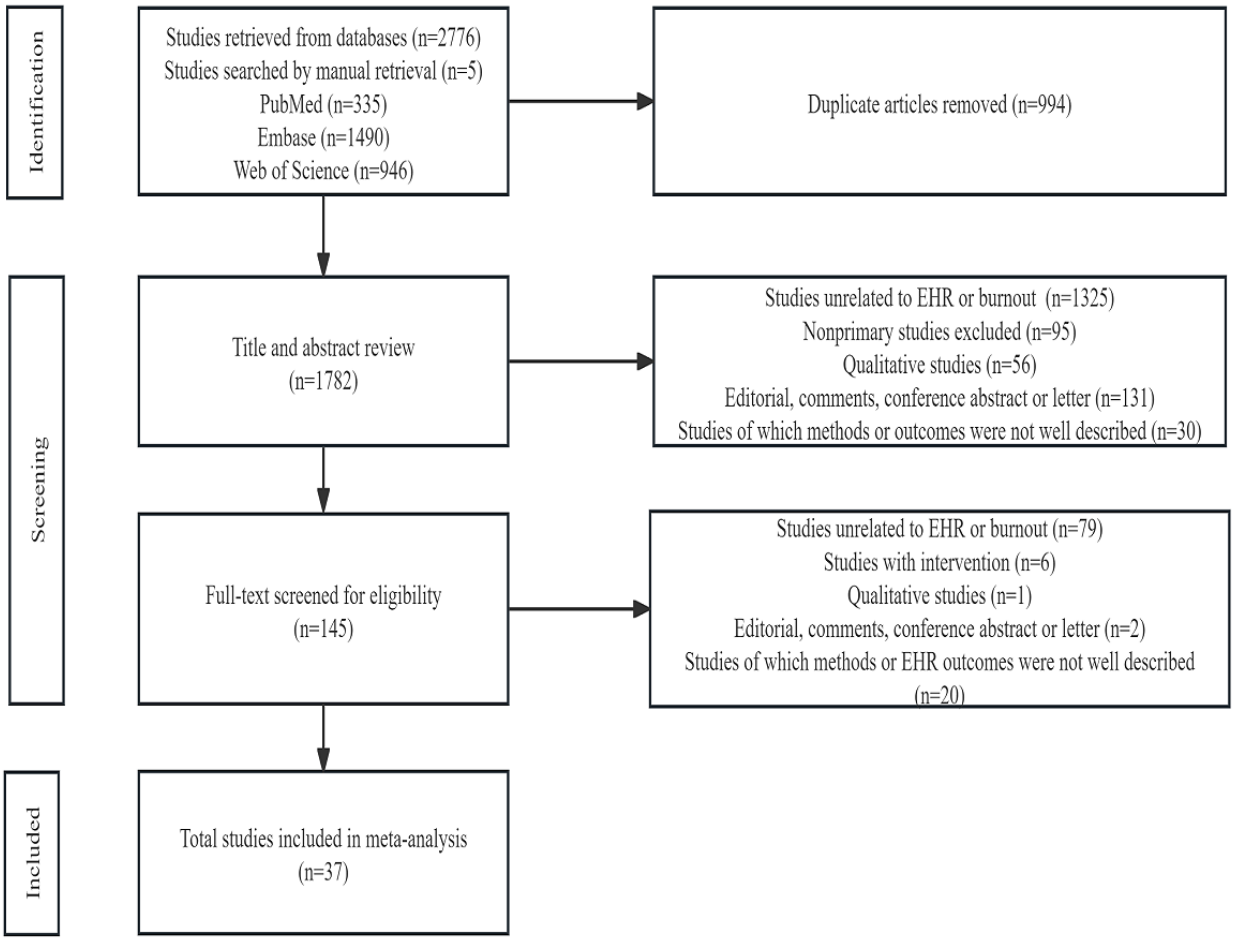
The terms were combined using Boolean logic and then filtered by publication date and language (English). A full description of the search strategy can be found in Multimedia Appendix 3. In addition, we carefully examined the references of each article and manually added 5 relevant references to our review list. Duplicate studies were systematically excluded from consideration.

Inclusion and Exclusion Criteria

Figure 1 shows the search and selection process. We applied strict inclusion and exclusion criteria to identify original and observational studies relevant to our research objectives. We included studies that examined general EHR use or specific supporting systems such as computerized physician order entry. We focused on studies that directly assessed burnout among health care professionals and individual psychological responses to EHR systems. Our review included the following types of research: cohort studies, case-control studies, and cross-sectional studies. EHR-related burnout was assessed using validated tools such as the MBI-HSS, the mini-Z, or other similar measures. The following publication types were excluded: abstracts, editorials, letters, reviews, commentaries, guidelines, and studies by non–health care professionals. In addition, studies were excluded if the necessary data could not be obtained from the corresponding author. We also excluded studies that repeated data already published in the literature.

Two reviewers independently screened all titles and abstracts to assess for relevance. Full texts of articles identified for further review were then assessed against the inclusion criteria. In cases of disagreement about the study eligibility of studies, a third reviewer was consulted for resolution.

Figure 1. Flowchart of study selection. EHR: electronic health record.



Data Extraction and Synthesis

For the included studies, we extracted relevant information including study design, geographical region, study duration, medical specialties involved, sample size, and relevant outcomes. The main outcomes included the prevalence of burnout in cross-sectional studies, the odds ratio (OR) along with its 95% CI in case-control studies, and the factors influencing burnout. We also documented the specific tools or measures used to assess these outcomes.

Risk of Bias Assessment

Two reviewers assessed the integrity, confirmability, and quality of the cross-sectional studies using the Joanna Briggs Institute (JBI) checklist [25] and the Newcastle-Ottawa Scale (NOS) [26] for the included case-control studies. Details of these assessments can be found in Multimedia Appendices 4-6, respectively.

Statistical Analysis

Meta-analysis was performed using R software (version 4.1.3; R Foundation for Statistical Computing). Heterogeneity was

calculated using the Cochran *Q* test, and statistical significance was set at $P<.05$. If there was no statistical heterogeneity ($I^2<50\%$), the fixed-effects model was used to pool results; otherwise ($I^2>50\%$) the random-effects model was used [27,28]. We grouped the main outcomes according to the predictor and moderator factors described by the participants and derived from the outcome reports. Continuous variables were summarized using the mean and standardized mean difference, whereas rates were extracted for categorical variables. For cross-sectional studies, the effect size measure was the prevalence of burnout and its corresponding 95% CI. For case-control studies, the effect of EHR was assessed using the pooled OR and its 95% CI. Publication bias was analyzed using the Egger test [29] and the trim-fill funnel plot. A sensitivity analysis was performed for each omitted method to determine the robustness and reliability of the results.

Results

Characteristics of the Included Studies

After reviewing a total of 2776 studies, 37 were selected for inclusion in our analysis (Figure 1) according to the predefined criteria and after elimination of duplicates. The baseline characteristics of the selected studies are summarized in Tables 1 and 2. For further details see Multimedia Appendix 7 [6,30-60].

The studies included in our review covered the period from 2009 to 2022 and included regions in both Canada and the United States. They involved a total of 66,556 health care professionals. The sample sizes of these studies varied widely from 84 to 25,018 participants, and the response rates ranged from 8.9% to 73.0%.

The primary measure used to assess burnout in the majority of studies was the MBI-HSS, which was used in 17 of 37 studies (46%). In addition, the mini-Z scale was used in 10 studies (27%). Notably, 2 studies using the MBI-HSS used cutoff definitions for burnout subcomponents that followed the standardized recommendations of the MBI-HSS.

Table 1. Characteristics of the cross-sectional studies.

Author	Data collection	Region	Participants	Sample (total)	Burnout cases	Burnout prevalence (%)
Tawfik et al [30]	2011	United States	Physicians and other clinician staff	6560	3586	54.66
Shanafelt et al [31]	2014	United States	Physicians	1934	517	26.73
Tawfik et al [32]	2015	United States	Physicians and other clinician staff	1165	624	53.56
Olson et al [33]	2016	United States	Physicians	282	127	45.04
Tai-Seale et al [34]	2016	United States	Physicians	107	41	38.32
Apaydin et al [35]	2016	United States	Physicians and other clinician staff	110	44	40
Livaudais et al [36]	2016	United States	Physicians and other clinician staff	557	267	47.94
Tran et al [37]	2017	United States	Physicians and other clinician staff	1792	465	25.95
Marckini et al [38]	2017	Canada and United States	Physicians	919	331	36.02
Gardner et al [39]	2017	United States	Physicians	208	51	24.52
Hilliard et al [40]	2017	United States	Physicians and other clinician staff	422	116	27.49
Higgins et al [41]	2017	United States	Residents	116	62	53.45
Czernik et al [42]	2017	United States	Residents	163	81	49.69
Hauer et al [43]	2018	United States	Physicians	122	44	36.07
Gajra et al [44]	2018	United States	Physicians	2468	539	21.84
Adler-Milstein et al [45]	2018	United States	Physicians	100	52	52
Somerson et al [46]	2018	United States	Residents	128	65	50.78
Melnick et al [47]	2018	United States	Physicians	203	78	38.42
Coleman et al [48]	2018	United States	Physicians	870	397	45.63
Abraham et al [49]	2018	United States	Nurses	368	134	36.41
Kondrich et al [50]	2018	Canada and United States	Physicians	872	360	41.28
Kroth et al [51]	2019	United States	Physicians and other clinician staff	856	276	32.24
Tajirian et al[6]	2019	Canada	Physicians and trainee	222	84	37.84
Mandeville et al [52]	2019	United States	Physicians and other clinician staff	396	100	25.25
Tiwari et al [53]	2019	United States	Physicians and other medical staff	15,505	5065	32.67
Sinha et al [54]	2019	United States	Physicians	103	41	39.81
Anderson et al [55]	2019	United States	Physicians and trainee	756	373	49.34
Nair et al [56]	2019	United States	Physicians	281	127	45.20
Jha et al [57]	2020	United States	Physicians and other medical staff	230	86	37.39
Esmaeilzadeh and Mirzaei [58]	2020	Iran	Physicians and other medical staff	416	206	49.52
Holzer et al [59]	2020	United States	Physicians and trainee	84	30	35.71
Wilkie et al [60]	2021	Canada	Physicians	457	106	23.19

Table 2. Characteristics of the case-control studies.

Author	Data collection	Participants	Region	Exposure	Sample (total)	Burnout cases	OR[a] (95% CI)
Eschenroeder et al [61]	2020	Physicians	United States	After-hours EHR[b] charting time per week >6 hours	25,018	7616	2.43 (2.30-2.57)
Sharp et al [62]	2019	Physician trainees	United States	Working hours per week >70 hours	502	159	2.80 (1.78-4.40)
Peccoralo et al [63]	2019	Clinical faculty	United States	Time spent on EHR outside work >90minutes	1346	385	1.90 (1.40-2.78)
Harris et al [64]	2017	Advanced practice registered nurses	United States	Insufficient time for EHR documentation	333	69	3.72 (1.78-7.80)
Robertson et al [65]	2015	Primary care workers	United States	Extra time spent on EHR per week >6 hours	585	216	2.90 (1.90-4.40)

[a]OR: odds ratio.

[b]EHR: electronic health record.

Quality of Included Studies

The quality of the cross-sectional studies was assessed using the JBI checklist. Of the cross-sectional studies reviewed, only 16 had a response rate of more than 50%. In addition, 24 studies provided a clear and precise description of their inclusion and exclusion criteria for participants. Additionally, 32 cross-sectional studies provided a detailed and thorough statistical analysis of their data and results.

We used the NOS to assess the risk of bias and the overall quality of the case-control studies. In particular, one study failed to clarify its selection criteria for the control group and comparability with the exposed group, which resulted in a high risk of selection bias. Furthermore, none of the 5 case-control studies reported information on the nonresponse population, indicating a high risk of nonresponse bias. Overall, the risk of bias in the case-control studies was assessed as moderate. A full breakdown of the quality assessment for each study can be found in Multimedia Appendices 4 [6,30-60] and 5 [61-65].

In our meta-analysis, we examined 37 studies that focused on identifying the prevalence of burnout associated with EHR use, involving a total of 66,556 health care professionals. The internal heterogeneity of 37 cross-sectional studies was evident in all included cross-sectional studies (I^2=98.3%). Using random-effects models, we calculated the combined overall prevalence of EHR-related burnout of 40.4% (95% CI 37.6%-43.2%). Subgroup analysis showed that studies using the MBI-HSS reported a higher pooled prevalence of burnout (41.4%) than those using the mini-Z (35.1%) but lower than those using other instruments (43.2%). However, these differences were not statistically significant (Figures 2 and 3).

Figure 2. Forest plot of the pooled prevalence of burnout among health care professionals across cross-sectional studies [6,30-60]. IV: inverse variance methods.

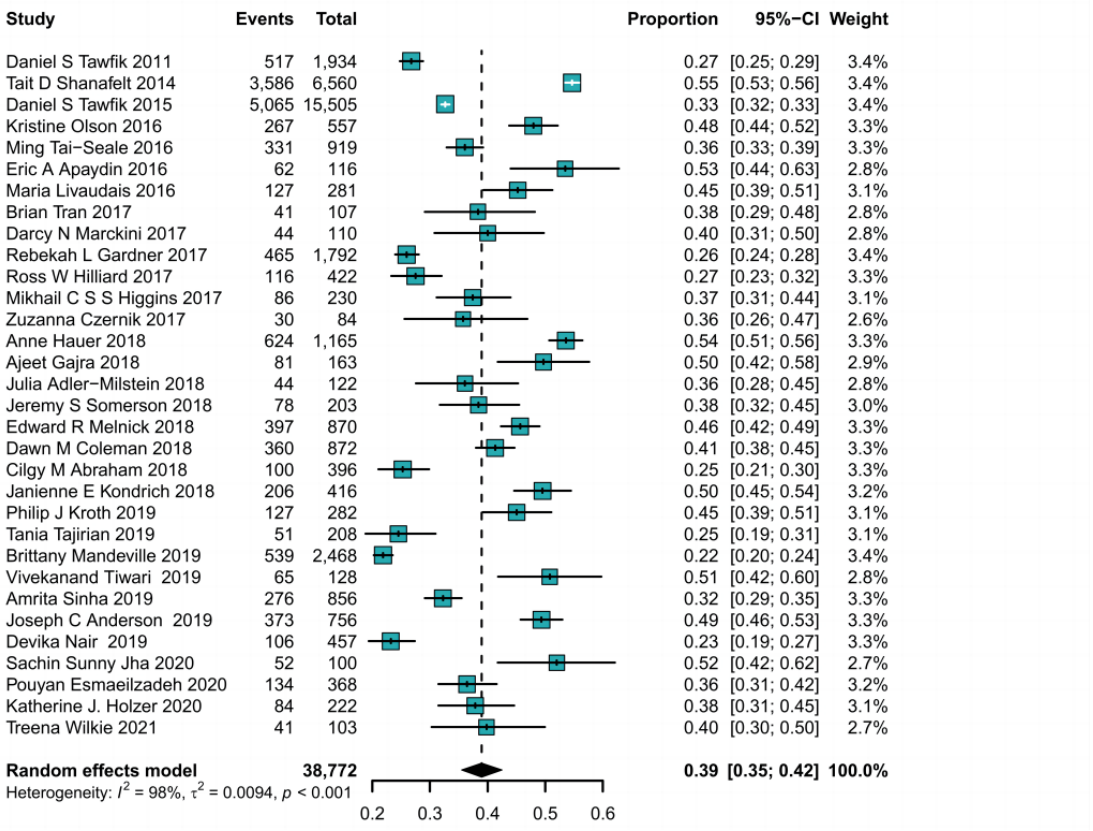
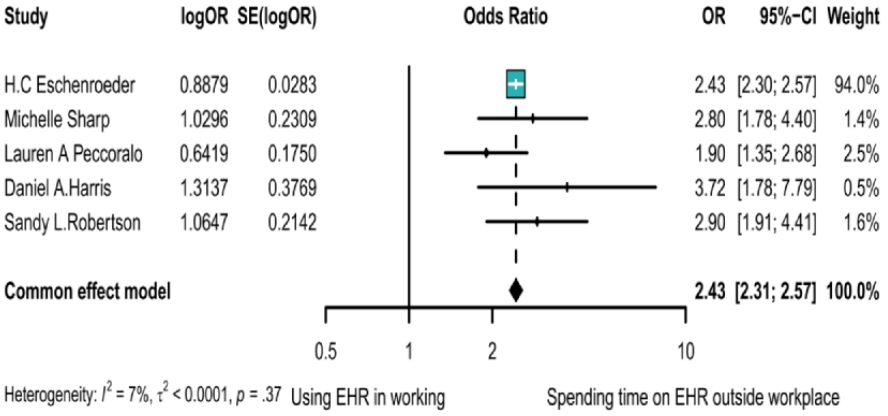


Figure 3. Measurement tool subgroup analysis of the pooled prevalence of burnout among health care professionals in cross-sectional studies [6,30-60]. IV: inverse variance methods; MBI: Maslach Burnout Inventory.



Publication Bias

The Egger test and the funnel plot were used to estimate the publication bias in the included studies (t=1.35, P=.18), indicating no significant publication bias. The distribution of the points in the funnel plot is symmetric. There was no statistical difference in publication bias. The results are available in Multimedia Appendices 8 and 9.

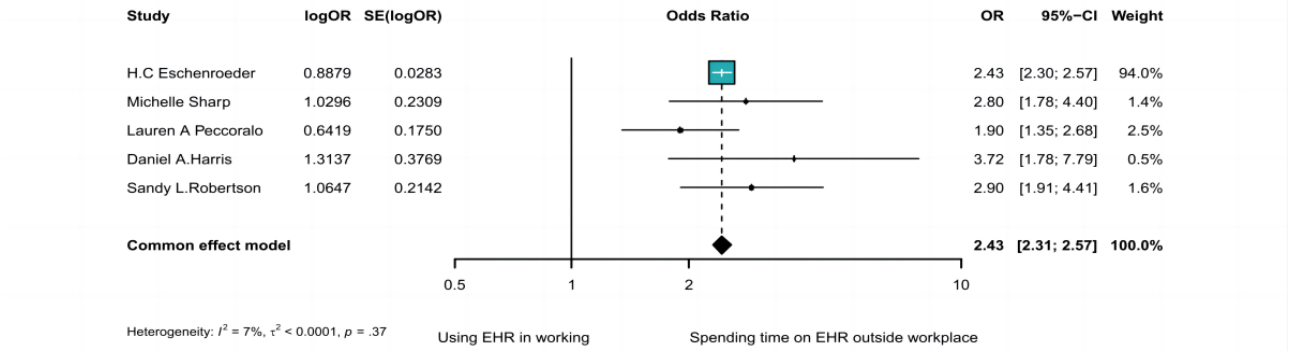
Sensitivity Analysis

Sensitivity analysis was performed using the individual omission method. The results showed that no single study had a significant effect on the pooled prevalence of burnout. The results of the sensitivity analysis indicated that the meta-analysis was robust.

The Association Between Time Spent on the EHR and Burnout

Data from 5 case-control studies with 27,784 participants were available for the meta-analysis of the time spent on EHR and burnout prevalence. There was no significant within-study heterogeneity (I^2=7.2%, P=.37), and a longer duration of EHR use was associated with a higher prevalence of burnout (OR 2.43, 95% CI 2.31-2.57) (Figure 4).

Figure 4. Forest plot of the association between the time spent using EHR and the risk of burnout [61-65]. EHR: electronic health record; IV: inverse variance methods; OR: odds ratio; SE: standard error of the TE; TE: take the logarithm of the effect value.



Main Causes of Burnout and Proposed Solutions

We have summarized the factors contributing to burnout among health care professionals in relation to EHR use in Table 3. Among these, challenges related to the design and availability of EHR systems were identified as the most significant contributors, as evidenced by 32 studies. Complaints from EHR users focused on several key issues: disruption to workflow [33]; cumbersome data entry (copy and paste) [59]; reduced direct communication with patients [38]; and annoyance with redundant, repetitive, or irrelevant alerts [52]. Poor EHR design has been shown to reduce work productivity and lead to prolonged EHR use [51,64]. This prolonged use has a negative impact on work-life balance of health care professionals and increasing burnout [42,57,65,66]. Workload factors, identified in 18 of the 32 studies, further exacerbate this problem. Specific aspects of workload that contribute to burnout include the number of hours worked per week [62-64], the frequency of night shifts [46,60], administrative documentation tasks [33,35,38,48,64], the volume of patient admissions [30,35,56], and the amount of information to manage in the EHR inbox [34,37,40]. Together, these factors exacerbate provider fatigue and increase the risk of burnout.

EHR usability, recognized as a contributing factor to burnout, relates to issues of accessibility and functionality of the system. This includes instances where the system is frequently unavailable due to maintenance, updates, or technical failures, as well as situations where the system is not user-friendly and requires excessive time to navigate and use effectively, potentially leading to burnout among health care professionals.

The factors contributing to burnout identified in the reviewed studies fall into 3 main categories: EHR use, work environment and organizational support, and the personal factors. Table 4 provides a summary of strategies to address these contributing factors. For example, the burden of medical clerical tasks imposed by EHR systems suggests the need to employ assistants or scribes to reduce the workload of health care professionals [31,67]. Evidence suggests that the EHR system itself can be improved by involving clinical staff in the design process [33], optimizing the user interface [39,64], minimizing the number of clicks required [52], and actively soliciting and incorporating user feedback [32]. In addition, some practitioners may not fully use health information technology in their roles and may be frustrated with EHR systems or similar systems [32]. To address this, health care organizations are advised to establish clear policies and procedures before implementing an EHR system and to provide ongoing health information technology education to reduce technology-related anxiety among users [32,52,68]. Finally, comprehensive and systematic initiatives are essential to effectively reduce burnout. Health care professionals are encouraged to work together to advocate for legislative and regulatory changes that ensure reasonable working hours, mandatory breaks, and safeguards against burnout [36,42,43,58,62].

Moreover, research also suggests that sociodemographic characteristics, interpersonal dynamics, and the work environment have a significant impact on the prevalence of burnout. In particular, factors such as being female, younger, and less experienced correlate with higher rates of burnout [34,48,55]. Conversely, high levels of satisfaction or positive perspectives on the use of EHR systems may reduce burnout [36,42,58].

Table 3. The influencing factors of burnout for studies.

Author	Design	Risk factors for burnout	Protective factors against burnout	Main EHR[a] factors influencing burnout
Tawfik et al [30]	Cross-sectional	NICU[b] with ≥10 weekly admissions, nursing care workload, and patient mortality	Burnout recognition education; implementation of burnout interventions at the individual and institutional level	Using EHR outside working or at home; time on using EHR
Shanafelt et al [31]	Cross-sectional	Using CPOE[c], female gender, emergency medicine, each additional hour per week	Assistant order entry; documentation support	Time spent on clerical tasks
Tawfik et al [32]	Cross-sectional	HIT[d] frustration, difficulty in falling asleep	Supplemental EHR training; scribes to assist documentation； team-based documentation and inbox management； automating data-entry tasks	Frustrated or stressed by EHR
Olson et al [33]	Cross-sectional	Poor control over workload, inefficient teamwork, lack of value alignment with leadership, and hectic-chaotic work atmosphere	Improve professional satisfaction; nonphysician order entry	Using EHR outside working or at home; insufficient documentation time
Tai-Seale et al [34]	Cross-sectional	Female gender and poor control over work schedule	Feeling highly valued; having good control over work schedule; working in a quiet or busy but reasonable environment; assist physician with email work; limit desktop medical work outside working hours (except in emergencies)	Using EHR outside working or at home; number of EHR system-generated in basket messages
Apaydin et al [35]	Cross-sectional	Managing unscheduled or same-day patients, lack of pharmacist support, administrative work, excessive overall workload, difficulty communicating with other professionals, inadequate care coordination, and answering patient emails	Interventions to facilitate provider-led quality improvement	Managing in-basket messages generated by EHR; responding to EHR alerts
Livaudais et al [36]	Cross-sectional	Negative perceptions of EHR	Perceiving positive effect of EHR in practice； technical support for EHR when using systems; EHR optimization program	Managing in-basket messages generated by EHR; poor EHR design; dealing with patient-call messages in systems
Tran et al [37]	Cross-sectional	Clinical full-time equivalents >0.9 and more incomplete messages in inbox	Perception positive attitudes about the effect of EHR or satisfied with EHR	Average additional 10 minutes spent on EHR after each visit; less efficient at completing EHR and inbox information
Marckini et al [38]	Cross-sectional	Female gender and dissatisfaction for clerical tasks	EHR optimization; improve physician efficiency; and job satisfaction	Managing in-basket messages generated by EHR; dissatisfaction with EHR
Gardner et al [39]	Cross-sectional	Primary care specialties, female gender, and reporting poor or marginal time for documentation	Perception positive attitudes about the effect of EHR or satisfied with EHR	Excessive data inputting in EHR; using EHR at home; frustrated with EHR
Hilliard et al [40]	Cross-sectional	High volume of patient call messages in the system and lack of control over workload	Copy and paste used in EHR documentation; assist with inbox tasks and create 2 administrative “desktops”	Using EHR outside working or at home; excessive data inputting in EHR; managing in-basket messages generated by EHR
Higgins et al [41]	Cross-sectional	Self-compassion, sleep disorder, lacking support from leaders, and poor control over schedules	Peer support, perceived appreciation and meaningfulness in work; maintaining values consistent with practice institution	Poor EHR usability; perception negative attitudes about the effect of EHR
Czernik et al [42]	Cross-sectional	Frustrated or stressed by EHR	Reducing the burden of documentation tasks; improving EHR usability; interventions to improve the EHR	Poor usability of EHR; information overload; degradation of medical documentation
Hauer et al [43]	Cross-sectional	Loss of practicing autonomy, female gender, frustrated with EHR, and increasing insurance and government regulation	Improve the functionality of EHR; enhance physician leadership and involvement; create a center for physician empowerment; create a physician health program	Using EHR outside workday

Author	Design	Risk factors for burnout	Protective factors against burnout	Main EHR[a] factors influencing burnout
Gajra et al [44]	Cross-sectional	Variable reimbursement models, interactions with payers, and increasing treating and caring demands	Use advanced practice providers; hire additional administrative staff; invest in information technology	Excessive data inputting in EHR; frustrated or stressed by EHR; using EHR outside workday
Adler-Milstein et al [45]	Cross-sectional	Poor self-rated EHR skills	Improve EHR design; scribe or team documentation; reduce documentation requirements	Using EHR outside working or at home; time spent on EHR; system-generated in-basket messages (>114) per week
Somerson et al [46]	Cross-sectional	Working >80 hours per week, verbal abuse from faculty, educational debt, “scut” work >10 hours per week	Nursing support; duty-hour restrictions; improve EHR functionality and efficiency; adequate, personalized training and support; adequate social work support	Time spent on EHR per week; used EHR >20 hours per week
Melnick et al [47]	Cross-sectional	Practice location (academic medical center) and medical specialty	Improve EHR usability	Using EHR outside working or at home; poor EHR usability
Coleman et al [48]	Cross-sectional	Work-related physical pain, work-home conflict, and younger age	Build personal resilience, enhance wellness; peer support; reduce administrative or EHR burden	Using EHR outside working or at home; increased EHR or documentation requirement
Abraham et al [49]	Cross-sectional	Intraorganizational factors	EHR with multifunctional; reduce high EHR workload; work with supportive colleagues; improve team communication	High EHR workload
Kondrich et al [50]	Cross-sectional	Feeling undervalued by patients, lacking superior support, little promotion chances, perceived unfair clinical working schedule, and nonacademic environment	Improve physician well-being	Feeling that the EHR detracts from patient care
Kroth et al [51]	Cross-sectional	Overall stress	Improve EHR design; clinician training; scribes to assist documentation; work at home boundaries; exercise, taking breaks	Information overloading; slow system response; excessive data inputting； fail to navigate quickly; note bloat; patient-clinician relationship interference; fear of missing something; billing oriented notes.
Tajirian et al [6]	Cross-sectional	Workflow issues	Reduce the administrative burden of EHR； improve EHR	Lower satisfaction and higher frustration with the EHR； poor intuitiveness and usability of EHR
Mandeville et al [52]	Cross-sectional	HIT-related stress and burnout and emergency medicine	Improved workflow	Daily frustration added by EHR； using EHR outside working or at home
Tiwari et al [53]	Cross-sectional	Lack of physical exercise and weekly working hours	Teamwork and working satisfaction; self-care training	Poor EHR usability; dissatisfaction with EHR
Sinha et al [54]	Cross-sectional	Interpersonal disengagement	Lower CLOC[e] ratio (total CLOC time to allocated appointment time); well-established personal resources	Using EHR outside working
Anderson et al [55]	Cross-sectional	Female gender, younger age, shorter practicing years, and having children at home	Taking 20 days or more of vacation time	Using EHR at home; ≥2-hour patient administration
Nair et al [56]	Cross-sectional	Working long hours, weekly number of nursing patients, practice environment, disinterested health systems, and dissatisfaction with remuneration	Caring for fewer patients per week	Using EHR outside working or at home; EHR requirements
Jha et al [57]	Cross-sectional	COVID-19 pandemic and in-house billing	Stay positive; improved EHR design	Documentation through EHR
Esmaeilzadeh and Mirzaei [58]	Cross-sectional	Less direct communication with patients, inadequate training for using HIT, and increasing computerization at work	Positive perceptions of EHR; more policy and legal interventions to ensure meaningful use of EHR	Poor EHR usability; time spent entering data
Holzer et al [59]	Cross-sectional	Receive COVID-19 patients	Using EHR to streamline clinical care activities; physician task relief	Using EHR outside work; increased EHR workload

Author	Design	Risk factors for burnout	Protective factors against burnout	Main EHR[a] factors influencing burnout
Wilkie et al [60]	Cross-sectional	High workload and insufficient resources	Good leadership; prioritize work-life balance	Poor EHR usability
Eschenroeder et al [61]	Case-control	Specialty	Organizational support for EHR	After-hours EHR charting time per week >6 hours; time-consuming data entry
Sharp et al [62]	Case-control	Working hours per week >70 hours	Report system to cover personal illness or emergency; access to mental health services; reduce EHR and clerical burden	>90 minutes on the EHR outside of the workday
Peccoralo et al [63]	Case-control	Clerical work time (>60 minutes/day) and poorer work-life integration	Reducing time spent on EHR and clerical tasks	Using EHR outside working (>90 minutes/day); EHR adds to daily work frustration
Harris et al [64]	Case-control	Insufficient time for documentation	Improve EHR usability; documentation practices optimization	Using EHR outside working or at home; EHR adding to daily frustration
Robertson et al [65]	Case-control	Dissatisfaction with work-life balance and female gender	EHR proficiency training	Extra time spent on EHR per week >6 hours

[a]EHR: electronic health record.

[b]NICU: neonatal intensive care unit.

[c]CPOE: computerized physician order entry.

[d]HIT: health information technology.

[e]CLOC: clinician logged-in outside clinic time.

Table 4. Proposed solutions for burnout mentioned.

Perspectives/solutions and suggestions	Measures
EHR[a]	
Improve EHR usability and performance	Enhance EHR user interface and design to reduce health care professionals to use
Institutions provide timely technical support during EHR use	Improving the effectiveness and efficiency of technological responses
Institutions should offer comprehensive training courses for EHR users	Ensure users master EHR skills to reduce burnout from technological issue
Working environment and organizational support	
Institutions introduce mechanisms for regular assessment of EHR efficacy	Regularly optimize and update the system based on user feedback
Establish a schedule, routine, and workflow	Design and optimize the workflow to ensure that the EHR aligns with the health care professional's actual work, reducing unnecessary steps and improving work efficiency
Enhance peer, managerial, and technical support	Provide appropriate human resources, such as medical assistants, scribes, and improving teamwork to distribute workload among health care professionals
Development of transparent policies and objectives	Establish clear policies and legislation to define the purpose, scope, and duration of EHR use, to delineate the responsibilities and obligations of health care professionals, and to reduce confusion and burnout
Personal	
Use of mental health resources and services	Counseling services and mindfulness meditation therapy help health care professionals better manage work stress and reduce their psychological distress
Encourage academic and career development	Plan career paths and training programs and create an environment for career development and learning

[a]EHR: electronic health record.

Discussion

Key Findings

This study explores the relationship between burnout and health care professionals. Our analysis revealed several key findings. First, the prevalence of burnout differs between assessment instruments, with the MBI-HSS indicating higher levels of burnout. However, this difference was not statistically significant. Second, there was a positive association between the average daily duration of EHR use and the risk of burnout. In particular, reducing the administrative burnout emerged as an effective strategy to reduce the risk of burnout [63]. Third, positive perceptions of the EHR and constructive work attitudes were correlated with the reduction in burnout.

The MBI-HSS is valued for its extensive validation and widespread acceptance as an essential tool for assessing burnout. Our findings suggest that the MBI-HSS may report higher rates of burnout due to several factors: sensitivity to burnout constructs—unlike self-report measures, which may rely predominantly on respondents' respondents' subjective feelings, the MBI-HSS comprehensively assesses burnout across multiple dimensions: emotional exhaustion, depersonalization, and personal accomplishment. This multidimensional assessment provides a nuanced perception of burnout, encompassing both its physical and psychological facets. These include the following: standardized cut-off scores—the MBI-HSS delineates specific cut-off scores for its dimensions, establishing clear criteria for identifying significant levels of burnout. This standardization promotes a consistent classification framework for burnout, which may contribute to the higher prevalence rates reported. Comprehensive assessment—the multidimensional approach of the MBI-HSS allows for a comprehensive assessment of burnout, including emotional exhaustion, depersonalization, and personal accomplishment. This thorough assessment is able to uncover more precise and detailed manifestations of burnout, thereby increasing detection rates. Benchmark for comparison—the MBI-HSS is often used as a benchmark for validating alternative burnout measures, and differences in results when compared with other instruments do not necessarily indicate a variance in prevalence. Rather, these differences underscore the accuracy of the MBI-HSS and the comprehensive scope of its assessment. The use of different instruments underlines the heterogeneity observed in our study results.

Solutions

This study demonstrates a robust relationship between workload, time spent using EHR, and burnout. Through a systematic review, we outline several pragmatic recommendations aimed at mitigating these problems.

Reduce Documentation and EHR Workload

A key strategy for alleviating workload concerns is to adopt a rational task allocation and effective teamwork model. Previous research highlights the effectiveness of this approach in reducing workload pressures [33,53]. By integrating medical assistants and scribes into the health care team, it is possible to distribute clerical tasks more evenly, thereby reducing the burden on health care professionals. This redistribution not only reduces workload but also increases overall operational efficiency [53,69,70]. In addition, the provision of targeted training is critical to improving teamwork dynamics, communication skills, and workflow efficiency. Such training efforts aim to cultivate a competent team capable of optimizing and streamlining workflow processes. The ultimate goal is to minimize documentation and EHR-related workloads, thereby making a significant contribution to reducing burnout among health care professionals [58,63].

Optimizing EHR and Training Courses

Continuous refinement of EHR systems through improved design, functionality, and integration of predesigned templates and phrases effectively increases system efficiency. The elimination of redundant steps and interactions further improves the user experience [32,71]. For example, customizing templates to include commonly used medical advice and alerts tailored to the specific needs of different departments significantly increases EHR efficiency [48,72]. Numerous studies have highlighted the critical role of improving user interaction with the EHR system. Developing a user-friendly interface that minimizes unnecessary clicks and reduces redundant and irrelevant data entry has been shown to significantly improve the user experience. Such improvements also significantly reduce the cognitive burden on health care professionals, resulting in a more streamlined and efficient health care delivery process [32,39,42]. In addition, comprehensive training and strong technical support are critical to improving the efficiency and effectiveness of EHR use. Systematic training aimed at promoting EHR proficiency among health care professionals can significantly improve operational efficiency and mitigate the effects of technology stress [46,58]. Research emphasizes the importance of training health care professionals to enhance EHR use and tailoring templates to specific clinical workflows.

Artificial Intelligence–Based Solutions

The integration of artificial intelligence (AI) into EHR systems represents a significant frontier for improvement. Innovations in machine learning, natural language processing (NLP), and large language models (LLMs) are poised to significantly increase the intelligence and automation capabilities of EHR systems [73,74]. Incorporating speech recognition and automated dictation or note-taking into hospital workflows can streamline the creation of medical documents, thereby increasing operational efficiency [75]. NLP is characterized by its ability to efficiently organize both unstructured and semistructured textual records, thereby facilitating a reduction in paperwork [76,77]. Recent research has highlighted the utility of LLMs, such as GPT-4, as powerful tools for medical documentation [78,79]. The use of technologies such as GPT-4 as a linguistic assistant or the use of intelligent templates can significantly speed up the medical documentation process for health care professionals, while improving the accuracy of documentation [79]. In addition, the researchers developed a data-driven method to generate recommendations for refining alert criteria through an explainable AI framework [80]. This advancement directly addresses the issue of overalerting in clinical decision support systems, which has been identified as a potential contributor to

burnout among health care professionals. By reducing unnecessary alerts, this approach promises to reduce the cognitive and operational workload of health care professionals, thereby improving both the quality of patient care and the work-life balance of health care staff. While AI technology could potentially help reduce burnout, it is important to recognize that the causes of burnout are complex and require further research.

Implications for Future Research

There is considerable evidence to support the need for comprehensive redesign of EHR systems to improve efficiency [32,51,53,81]. However, the literature reveals a paucity of published empirical research quantifying EHR limitations, user fatigue and burnout. While some studies have indirectly demonstrated the poor usability of EHR by measuring pupillary reflex and cognitive fatigue [82,83], claims of inefficiency are primarily based on subjective perceptions of users. Thus, there is a need for more studies that objectively assess usability and user experience. Future research should aim to quantitatively assess the usability of EHR systems and their impact on the physical and mental well-being of health care professionals.

Furthermore, the incorporation of AI, specifically LLMs, into EHR systems is an important future research direction to reduce burnout among health care professionals. Such research could include, but is not limited to, (1) reducing the amount of time health care professionals spend on nonclinical tasks by automating administrative tasks, including data entry, scheduling, and patient history taking; (2) using LLMs to efficiently generate and review medical documentation to ensure high quality and consistency of documentation while saving time; (3) improving the interpretability and transparency of clinical decision support to provide clinicians with trustworthy decision support to reduce their cognitive load; and (4) ensuring the ethical use of AI to guarantee that AI systems are used ethically and that algorithms are unbiased. The integration of AI into EHR systems must comply with strict privacy regulations to protect patient privacy [84]. Exploring the potential of AI could make a significant contribution to creating a more supportive and efficient health care ecosystem [73,79,85].

Limitations

This review has several limitations. First, it has a language bias by including only peer-reviewed literature published in English. This limitation may introduce information and selection bias by omitting non-English studies that may provide valuable insights or alternative viewpoints on the topic. Second, the internal heterogeneity of the included studies is remarkably high, with significant differences in methodology, participant demographics, and outcome measures between studies, which may bias the synthesis of findings. In addition, the geographical distribution of the selected studies is dominated by North American research, with only 1 study from Iran. This distribution may introduce regional bias, as health care practices and experiences in these areas may not accurately reflect global patterns.

In addition, the temporal scope of the study, covering the years 2020 to 2022, was significantly influenced by the COVID-19 pandemic. Data collected during this period may be subject to bias or inaccuracy due to the unprecedented impact of the pandemic on global health systems. Additionally, the pandemic introduced new stressors and challenges for health care professionals, which may have influenced the incidence and manifestation of their burnout. These factors should be carefully considered when interpreting the study results, as they may limit the generalizability and significance of the findings beyond the specific context and timeframe of the pandemic.

Conclusions

This review highlights the significant impact of the EHR and the workload of health care professionals on burnout and emphasizes the need for targeted solutions such as workflow optimization, improved training, and the use of medical scribes. It also identifies that the potential of AI to improve EHR efficiency is a promising direction. Despite these findings, there remains a critical need for empirical research to accurately quantify the challenges associated with EHR use and their impact on provider well-being. Future studies are encouraged to explore innovative solutions to foster a more supportive health care environment.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

J Liu, J Lin, and SL conceived and designed the study. YW, SL, MW, J Liu, and J Lin developed the methods. YW, SL, CW, J Lin, and J Liu developed the search strategy. All authors participated in drafting the manuscript. All authors have read and approved the final article. There was no funding for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 98 KB - medinform_v12i1e54811_app1.pdf]

Multimedia Appendix 2
PROSPERO registration.
[PDF File (Adobe PDF File), 1901 KB - medinform_v12i1e54811_app2.pdf]

Multimedia Appendix 3
Search strategy.
[DOCX File , 13 KB - medinform_v12i1e54811_app3.docx]

Multimedia Appendix 4
Joanna Briggs Institute checklist for the cross-sectional studies included.
[PDF File (Adobe PDF File), 58 KB - medinform_v12i1e54811_app4.pdf]

Multimedia Appendix 5
NOS results for the case-control studies included.
[PDF File (Adobe PDF File), 74 KB - medinform_v12i1e54811_app5.pdf]

Multimedia Appendix 6
Joanna Briggs Institute Prevalence Critical Appraisal Tool.
[DOCX File , 15 KB - medinform_v12i1e54811_app6.docx]

Multimedia Appendix 7
Basic characteristics of the studies included.
[PDF File (Adobe PDF File), 70 KB - medinform_v12i1e54811_app7.pdf]

Multimedia Appendix 8
Funnel plot for the studies included.
[PNG File , 136 KB - medinform_v12i1e54811_app8.png]

Multimedia Appendix 9
The results of the publication bias test.
[PNG File , 45 KB - medinform_v12i1e54811_app9.png]

References

1. Aldosari B. Patients' safety in the era of EMR/EHR automation. Inform Med Unlocked 2017;9:230-233 [FREE Full text] [doi: 10.1016/j.imu.2017.10.001]
2. Gatiti P, Ndirangu E, Mwangi J, Mwanzu A, Ramadhani T. Enhancing healthcare quality in hospitals through electronic health records: a systematic review. J Health Inform Dev Ctries 2021;15(2):1-25 [FREE Full text]
3. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun;30(1):e100704 [FREE Full text] [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
4. Li C, Parpia C, Sriharan A, Keefe DT. Electronic medical record-related burnout in healthcare providers: a scoping review of outcomes and interventions. BMJ Open 2022 Aug 19;12(8):e060865 [FREE Full text] [doi: 10.1136/bmjopen-2022-060865] [Medline: 35985785]
5. Heponiemi T, Gluschkoff K, Vehko T, Kaihlanen AM, Saranto K, Nissinen S, et al. Electronic health record implementations and insufficient training endanger nurses' well-being: cross-sectional survey study. J Med Internet Res 2021 Dec 23;23(12):e27096 [FREE Full text] [doi: 10.2196/27096] [Medline: 34941546]
6. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res 2020 Jul 15;22(7):e19274 [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]
7. Palojoki S, Saranto K, Reponen E, Skants N, Vakkuri A, Vuokko R. Classification of electronic health record-related patient safety incidents: development and validation study. JMIR Med Inform 2021 Aug 31;9(8):e30470 [FREE Full text] [doi: 10.2196/30470] [Medline: 34245558]
8. Tutty MA, Carlasare LE, Lloyd S, Sinsky CA. The complex case of EHRs: examining the factors impacting the EHR user experience. J Am Med Inform Assoc 2019 Jul 01;26(7):673-677 [FREE Full text] [doi: 10.1093/jamia/ocz021] [Medline: 30938754]
9. Jankovic I, Chen JH. Clinical decision support and implications for the clinician burnout crisis. Yearb Med Inform 2020 Aug;29(1):145-154 [FREE Full text] [doi: 10.1055/s-0040-1701986] [Medline: 32823308]

10. Thomas Craig KJ, Willis VC, Gruen D, Rhee K, Jackson GP. The burden of the digital environment: a systematic review on organization-directed workplace interventions to mitigate physician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):985-997 [FREE Full text] [doi: 10.1093/jamia/ocaa301] [Medline: 33463680]
11. Mohan V, Garrison C, Gold JA. Using a new model of electronic health record training to reduce physician burnout: a plan for action. JMIR Med Inform 2021 Sep 20;9(9):e29374 [FREE Full text] [doi: 10.2196/29374] [Medline: 34325400]
12. Melnick ER, Harry E, Sinsky CA, Dyrbye LN, Wang H, Trockel MT, et al. Perceived electronic health record usability as a predictor of task load and burnout among US physicians: mediation analysis. J Med Internet Res 2020 Dec 22;22(12):e23382 [FREE Full text] [doi: 10.2196/23382] [Medline: 33289493]
13. Mertz H. Electronic health record reform: an alternative response to physician burnout. Am J Med 2021;134(9):e498. [doi: 10.1016/j.amjmed.2021.04.022] [Medline: 34462089]
14. Dyrbye LN, Shanafelt TD, Sinsky CA, Cipriano PF, Bhatt J, Ommaya A, et al. Burnout among health care professionals: a call to explore and address this underrecognized threat to safe, high-quality care. NAM Perspectives 2017;7(7) [FREE Full text] [doi: 10.31478/201707b]
15. Kumar S. Burnout and doctors: prevalence, prevention and intervention. Healthcare (Basel) 2016;4(3):37 [FREE Full text] [doi: 10.3390/healthcare4030037] [Medline: 27417625]
16. Gesner E, Gazarian P, Dykes P. The burden and burnout in documenting patient care: an integrative literature review. Stud Health Technol Inform 2019;264:1194-1198 [FREE Full text] [doi: 10.3233/SHTI190415] [Medline: 31438114]
17. Kang C, Sarkar IN. Interventions to reduce electronic health record-related burnout: a systematic review. Appl Clin Inform 2024;15(1):10-25 [FREE Full text] [doi: 10.1055/a-2203-3787] [Medline: 37923381]
18. Johnson KB, Neuss MJ, Detmer DE. Electronic health records and clinician burnout: a story of three eras. J Am Med Inform Assoc 2021;28(5):967-973 [FREE Full text] [doi: 10.1093/jamia/ocaa274] [Medline: 33367815]
19. Williams MS. Misdiagnosis: burnout, moral injury, and implications for the electronic health record. J Am Med Inform Assoc 2021;28(5):1047-1050 [FREE Full text] [doi: 10.1093/jamia/ocaa244] [Medline: 33164089]
20. Baxter SL, Saseendrakumar BR, Cheung M, Savides TJ, Longhurst CA, Sinsky CA, et al. Association of electronic health record inbasket message characteristics with physician burnout. JAMA Netw Open 2022;5(11):e2244363 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.44363] [Medline: 36449288]
21. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151(4):W65-W94 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00136] [Medline: 19622512]
22. Maslach C, Jackson SE, Leiter MP. Maslach burnout inventory. In: Evaluating Stress. Lanham, MD: Scarecrow Education; 1997.
23. The Stanford Model of Professional Fulfillment™. Stanford Medicine: WellMD & WellPhD. 2016. URL: https://wellmd.stanford.edu/about/model-external.html [accessed 2024-05-01]
24. Shimotsu S, Poplau S, Linzer M. Validation of a brief clinician survey to reduce clinician burnout. J Gen Intern Med 2015;30(2 suppl):S79-S80.
25. Munn Z, Stone JC, Aromataris E, Klugar M, Sears K, Leonardi-Bee J, et al. Assessing the risk of bias of quantitative analytical studies: introducing the vision for critical appraisal within JBI systematic reviews. JBI Evid Synth 2023 Mar 01;21(3):467-471. [doi: 10.11124/JBIES-22-00224] [Medline: 36476419]
26. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603-605 [FREE Full text] [doi: 10.1007/s10654-010-9491-z] [Medline: 20652370]
27. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
28. Higgins JPT, Li T, Deeks JJ. Choosing effect measures and computing estimates of effect. In: Chandler J, Thomas J, Higgins JPT, Page MJ, Cumpston M, Li T, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions, Second Edition. Hoboken: Wiley; 2019:143-176.
29. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
30. Tawfik DS, Phibbs CS, Sexton JB, Kan P, Sharek PJ, Nisbet CC, et al. Factors associated with provider burnout in the NICU. Pediatrics 2017 May;139(5):e20164134 [FREE Full text] [doi: 10.1542/peds.2016-4134] [Medline: 28557756]
31. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
32. Tawfik DS, Sinha A, Bayati M, Adair KC, Shanafelt TD, Sexton JB, et al. Frustration with technology and its relation to emotional exhaustion among health care workers: cross-sectional observational study. J Med Internet Res 2021;23(7):e26817 [FREE Full text] [doi: 10.2196/26817] [Medline: 34255674]

33. Olson K, Sinsky C, Rinne ST, Long T, Vender R, Mukherjee S, et al. Cross-sectional survey of workplace stressors associated with physician burnout measured by the Mini-Z and the maslach burnout inventory. Stress Health 2019;35(2):157-175 [FREE Full text] [doi: 10.1002/smi.2849] [Medline: 30467949]
34. Tai-Seale M, Dillon EC, Yang Y, Nordgren R, Steinberg RL, Nauenberg T, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019;38(7):1073-1078 [FREE Full text] [doi: 10.1377/hlthaff.2018.05509] [Medline: 31260371]
35. Apaydin EA, Rose D, Meredith LS, McClean M, Dresselhaus T, Stockdale S. Association between difficulty with VA patient-centered medical home model components and provider emotional exhaustion and intent to remain in practice. J Gen Intern Med 2020;35(7):2069-2075 [FREE Full text] [doi: 10.1007/s11606-020-05780-8] [Medline: 32291716]
36. Livaudais M, Deng D, Frederick T, Grey-Theriot F, Kroth PJ. Perceived value of the electronic health record and its association with physician burnout. Appl Clin Inform 2022;13(4):778-784 [FREE Full text] [doi: 10.1055/s-0042-1755372] [Medline: 35981548]
37. Tran B, Lenhart A, Ross R, Dorr DA. Burnout and EHR use among academic primary care physicians with varied clinical workloads. AMIA Jt Summits Transl Sci Proc 2019;2019:136-144 [FREE Full text] [Medline: 31258965]
38. Marckini DN, Samuel BP, Parker JL, Cook SC. Electronic health record associated stress: a survey study of adult congenital heart disease specialists. Congenit Heart Dis 2019;14(3):356-361 [FREE Full text] [doi: 10.1111/chd.12745] [Medline: 30825270]
39. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
40. Hilliard RW, Haskell J, Gardner RL. Are specific elements of electronic health record use associated with clinician burnout more than others? J Am Med Inform Assoc 2020;27(9):1401-1410 [FREE Full text] [doi: 10.1093/jamia/ocaa092] [Medline: 32719859]
41. Higgins MCSS, Siddiqui AA, Kosowsky T, Unan L, Mete M, Rowe S, et al. Burnout, professional fulfillment, intention to leave, and sleep-related impairment among radiology trainees across the United States (US): a multisite epidemiologic study. Acad Radiol 2022;29(Suppl 5):S118-S125. [doi: 10.1016/j.acra.2022.01.022] [Medline: 35241358]
42. Czernik Z, Yu A, Pell J, Feinbloom D, Jones CD. Hospitalist perceptions of electronic health records: a multi-site survey. J Gen Intern Med 2022;37(1):269-271 [FREE Full text] [doi: 10.1007/s11606-020-06558-8] [Medline: 33479933]
43. Hauer A, Waukau HJ, Welch P. Physician burnout in Wisconsin: an alarming trend affecting physician wellness. WMJ 2018;117(5):194-200 [FREE Full text] [Medline: 30674095]
44. Gajra A, Bapat B, Jeune-Smith Y, Nabhan C, Klink AJ, Liassou D, et al. Frequency and causes of burnout in US community oncologists in the era of electronic health records. JCO Oncol Pract 2020;16(4):e357-e365 [FREE Full text] [doi: 10.1200/JOP.19.00542] [Medline: 32275848]
45. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
46. Somerson JS, Patton A, Ahmed AA, Ramey S, Holliday EB. Burnout among United States orthopaedic surgery residents. J Surg Educ 2020;77(4):961-968. [doi: 10.1016/j.jsurg.2020.02.019] [Medline: 32171748]
47. Melnick ER, Dyrbye LN, Sinsky CA, Trockel M, West CP, Nedelec L, et al. The association between perceived electronic health record usability and professional burnout among US physicians. Mayo Clin Proc 2020;95(3):476-487 [FREE Full text] [doi: 10.1016/j.mayocp.2019.09.024] [Medline: 31735343]
48. Coleman DM, Money SR, Meltzer AJ, Wohlauer M, Drudi LM, Freischlag JA, et al. Vascular surgeon wellness and burnout: a report from the Society for Vascular Surgery Wellness Task Force. J Vasc Surg 2021;73(6):1841-1850.e3 [FREE Full text] [doi: 10.1016/j.jvs.2020.10.065] [Medline: 33248123]
49. Abraham CM, Zheng K, Norful AA, Ghaffari A, Liu J, Topaz M, et al. Use of multifunctional electronic health records and burnout among primary care nurse practitioners. J Am Assoc Nurse Pract 2021;33(12):1182-1189 [FREE Full text] [doi: 10.1097/JXX.0000000000000533] [Medline: 33534286]
50. Kondrich JE, Han R, Clark S, Platt SL. Burnout in pediatric emergency medicine physicians: a predictive model. Pediatr Emerg Care 2022;38(2):e1003-e1008. [doi: 10.1097/PEC.0000000000002425] [Medline: 35100790]
51. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
52. Mandeville B, Cooper E, Haskell J, Viner-Brown S, Gardner RL. Use of health information technology by Rhode Island physicians and advanced practice providers, 2019. R I Med J (2013) 2020;103(1):21-24 [FREE Full text] [Medline: 32013299]
53. Tiwari V, Kavanaugh A, Martin G, Bergman M. High burden of burnout on rheumatology practitioners. J Rheumatol 2020;47(12):1831-1834 [FREE Full text] [doi: 10.3899/jrheum.191110] [Medline: 32238512]

54. Sinha A, Shanafelt TD, Trockel M, Wang H, Sharp C. Novel nonproprietary measures of ambulatory electronic health record use associated with physician work exhaustion. Appl Clin Inform 2021;12(3):637-646 [FREE Full text] [doi: 10.1055/s-0041-1731678] [Medline: 34261173]
55. Anderson JC, Bilal M, Burke CA, Gaidos JK, Lopez R, Oxentenko AS, et al. Burnout among US gastroenterologists and fellows in training: identifying contributing factors and offering solutions. J Clin Gastroenterol 2023;57(10):1063-1069 [FREE Full text] [doi: 10.1097/MCG.0000000000001781] [Medline: 36477385]
56. Nair D, Brereton L, Hoge C, Plantinga LC, Agrawal V, Soman SS, et al. Burnout among nephrologists in the United States: a survey study. Kidney Med 2022;4(3):100407 [FREE Full text] [doi: 10.1016/j.xkme.2022.100407] [Medline: 35386610]
57. Jha SS, Shah S, Calderon MD, Soin A, Manchikanti L. The effect of COVID-19 on interventional pain management practices: a physician burnout survey. Pain Physician 2020;23(4S):S271-S282 [FREE Full text] [Medline: 32942787]
58. Esmaeilzadeh P, Mirzaei T. Using electronic health records to mitigate workplace burnout among clinicians during the COVID-19 pandemic: field study in Iran. JMIR Med Inform 2021;9(6):e28497 [FREE Full text] [doi: 10.2196/28497] [Medline: 34033578]
59. Holzer KJ, Lou SS, Goss CW, Strickland J, Evanoff BA, Duncan JG, et al. Impact of changes in EHR use during COVID-19 on physician trainee mental health. Appl Clin Inform 2021;12(3):507-517 [FREE Full text] [doi: 10.1055/s-0041-1731000] [Medline: 34077972]
60. Wilkie T, Tajirian T, Thakur A, Mistry S, Islam F, Stergiopoulos V. Evolution of a physician wellness, engagement and excellence strategy: lessons learnt in a mental health setting. BMJ Lead 2023;7(3):182-188 [FREE Full text] [doi: 10.1136/leader-2022-000595] [Medline: 37200187]
61. Eschenroeder HC, Manzione LC, Adler-Milstein J, Bice C, Cash R, Duda C, et al. Associations of physician burnout with organizational electronic health record support and after-hours charting. J Am Med Inform Assoc 2021 Apr 23;28(5):960-966 [FREE Full text] [doi: 10.1093/jamia/ocab053] [Medline: 33880534]
62. Sharp M, Burkart KM, Adelman MH, Ashton RW, Biddison LD, Bosslet GT, et al. A national survey of burnout and depression among fellows training in pulmonary and critical care medicine: a special report by the association of pulmonary and critical care medicine program directors. Chest 2021;159(2):733-742 [FREE Full text] [doi: 10.1016/j.chest.2020.08.2117] [Medline: 32956717]
63. Peccoralo LA, Kaplan CA, Pietrzak RH, Charney DS, Ripp JA. The impact of time spent on the electronic health record after work and of clerical work on burnout among clinical faculty. J Am Med Inform Assoc 2021;28(5):938-947 [FREE Full text] [doi: 10.1093/jamia/ocaa349] [Medline: 33550392]
64. Harris DA, Haskell J, Cooper E, Crouse N, Gardner R. Estimating the association between burnout and electronic health record-related stress among advanced practice registered nurses. Appl Nurs Res 2018;43:36-41. [doi: 10.1016/j.apnr.2018.06.014] [Medline: 30220361]
65. Robertson SL, Robinson MD, Reid A. Electronic health record effects on work-life balance and burnout within the I population collaborative. J Grad Med Educ 2017;9(4):479-484 [FREE Full text] [doi: 10.4300/JGME-D-16-00123.1] [Medline: 28824762]
66. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012;172(18):1377-1385 [FREE Full text] [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
67. Shultz CG, Holmstrom HL. The use of medical scribes in health care settings: a systematic review and future directions. J Am Board Fam Med 2015;28(3):371-381 [FREE Full text] [doi: 10.3122/jabfm.2015.03.140224] [Medline: 25957370]
68. Green-McKenzie J, Somasundaram P, Lawler T, O'Hara E, Shofer FS. Prevalence of burnout in occupational and environmental medicine physicians in the United States. J Occup Environ Med 2020;62(9):680-685 [FREE Full text] [doi: 10.1097/JOM.0000000000001913] [Medline: 32890204]
69. Nguyen OT, Turner K, Charles D, Sprow O, Perkins R, Hong YR, et al. Implementing digital scribes to reduce electronic health record documentation burden among cancer care clinicians: a mixed-methods pilot study. JCO Clin Cancer Inform 2023;7:e2200166 [FREE Full text] [doi: 10.1200/CCI.22.00166] [Medline: 36972488]
70. Ghatnekar S, Faletsky A, Nambudiri VE. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. Health Technol (Berl) 2021;11(4):803-809 [FREE Full text] [doi: 10.1007/s12553-021-00568-0] [Medline: 34094806]
71. Lourie EM, Utidjian LH, Ricci MF, Webster L, Young C, Grenfell SM. Reducing electronic health record-related burnout in providers through a personalized efficiency improvement program. J Am Med Inform Assoc 2021;28(5):931-937 [FREE Full text] [doi: 10.1093/jamia/ocaa248] [Medline: 33166384]
72. DeWitt D, Harrison LE. The potential impact of scribes on medical school applicants and medical students with the new clinical documentation guidelines. J Gen Intern Med 2018;33(11):2002-2004 [FREE Full text] [doi: 10.1007/s11606-018-4582-8] [Medline: 30066114]
73. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. J Am Med Inform Assoc 2024:ocae052 [FREE Full text] [doi: 10.1093/jamia/ocae052] [Medline: 38497958]

74. Liu S, McCoy AB, Wright AP, Nelson SS, Huang SS, Ahmad HB, et al. Why do users override alerts? Utilizing large language model to summarize comments and optimize clinical decision support. J Am Med Inform Assoc 2024:ocae041 [FREE Full text] [doi: 10.1093/jamia/ocae041] [Medline: 38452289]
75. Payne TH, Alonso WD, Markiel JA, Lybarger K, Lordon R, Yetisgen M, et al. Using voice to create inpatient progress notes: effects on note timeliness, quality, and physician satisfaction. JAMIA Open 2018;1(2):218-226 [FREE Full text] [doi: 10.1093/jamiaopen/ooy036] [Medline: 31984334]
76. Wang JX, Sullivan DK, Wells AC, Chen JH. ClinicNet: machine learning for personalized clinical order set recommendations. JAMIA Open 2020;3(2):216-224 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa021] [Medline: 32734162]
77. Dymek C, Kim B, Melton GB, Payne TH, Singh H, Hsiao CJ. Building the evidence-base to reduce electronic health record-related clinician burden. J Am Med Inform Assoc 2021;28(5):1057-1061 [FREE Full text] [doi: 10.1093/jamia/ocaa238] [Medline: 33340326]
78. Truhn D, Loeffler CM, Müller-Franzes G, Nebelung S, Hewitt KJ, Brandner S, et al. Extracting structured information from unstructured histopathology reports using Generative Pre-Trained Transformer 4 (GPT-4). J Pathol 2024;262(3):310-319 [FREE Full text] [doi: 10.1002/path.6232] [Medline: 38098169]
79. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
80. Liu S, McCoy AB, Peterson JF, Lasko TA, Sittig DF, Nelson SD, et al. Leveraging explainable artificial intelligence to optimize clinical decision support. J Am Med Inform Assoc 2024;31(4):968-974 [FREE Full text] [doi: 10.1093/jamia/ocae019] [Medline: 38383050]
81. Atutxa A, Perez A, Casillas A, Atutxa A, Perez A, Casillas A. Machine learning approaches on diagnostic term encoding with the ICD for clinical documentation. IEEE J Biomed Health Inform 2018;22(4):1323-1329 [FREE Full text] [doi: 10.1109/JBHI.2017.2743824] [Medline: 28858819]
82. Khairat S, Coleman C, Ottmar P, Jayachander DI, Bice T, Carson SS. Association of electronic health record use with physician fatigue and efficiency. JAMA Netw Open 2020;3(6):e207385 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.7385] [Medline: 32515799]
83. Murphy DR, Satterly T, Giardina TD, Sittig DF, Singh H. Practicing clinicians' recommendations to reduce burden from the electronic health record inbox: a mixed-methods study. J Gen Intern Med 2019;34(9):1825-1832 [FREE Full text] [doi: 10.1007/s11606-019-05112-5] [Medline: 31292905]
84. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
85. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
EMR: electronic medical record
IV: inverse variation methods
JBI: Joanna Briggs Institute
LLM: large language model
MBI-HSS: Maslach Burnout Inventory-Human Services Survey instrument
NLP: natural language processing
NOS: Newcastle-Ottawa Scale
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Edited by C Lovis; submitted 27.11.23; peer-reviewed by JY Nam, J Wolff, I Mircheva, R Koppel; comments to author 14.01.24; revised version received 23.02.24; accepted 17.04.24; published 12.06.24.

Please cite as:
Wu Y, Wu M, Wang C, Lin J, Liu J, Liu S
Evaluating the Prevalence of Burnout Among Health Care Professionals Related to Electronic Health Record Use: Systematic Review and Meta-Analysis
JMIR Med Inform 2024;12:e54811
URL: https://medinform.jmir.org/2024/1/e54811
doi:10.2196/54811
PMID:38865188

Review

Transforming Health Care Through Chatbots for Medical History-Taking and Future Directions: Comprehensive Systematic Review

Michael Hindelang[1,2,3], MSc; Sebastian Sitaru[1], MD; Alexander Zink[1,4], MD, MPH, MBA, PhD

[1]Department of Dermatology and Allergy, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany

[2]Pettenkofer School of Public Health, Munich, Germany

[3]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilian University, LMU, Munich, Germany

[4]Division of Dermatology and Venereology, Department of Medicine Solna, Karolinska Institute, Stockholm, Sweden

Corresponding Author:
Michael Hindelang, MSc
Department of Dermatology and Allergy
TUM School of Medicine and Health
Technical University of Munich
Biedersteiner Straße 29
Munich, 80802
Germany
Phone: 49 894140 ext 3061
Email: michael.hindelang@tum.de

Abstract

Background: The integration of artificial intelligence and chatbot technology in health care has attracted significant attention due to its potential to improve patient care and streamline history-taking. As artificial intelligence–driven conversational agents, chatbots offer the opportunity to revolutionize history-taking, necessitating a comprehensive examination of their impact on medical practice.

Objective: This systematic review aims to assess the role, effectiveness, usability, and patient acceptance of chatbots in medical history–taking. It also examines potential challenges and future opportunities for integration into clinical practice.

Methods: A systematic search included PubMed, Embase, MEDLINE (via Ovid), CENTRAL, Scopus, and Open Science and covered studies through July 2024. The inclusion and exclusion criteria for the studies reviewed were based on the PICOS (participants, interventions, comparators, outcomes, and study design) framework. The population included individuals using health care chatbots for medical history–taking. Interventions focused on chatbots designed to facilitate medical history–taking. The outcomes of interest were the feasibility, acceptance, and usability of chatbot-based medical history–taking. Studies not reporting on these outcomes were excluded. All study designs except conference papers were eligible for inclusion. Only English-language studies were considered. There were no specific restrictions on study duration. Key search terms included "chatbot*," "conversational agent*," "virtual assistant," "artificial intelligence chatbot," "medical history," and "history-taking." The quality of observational studies was classified using the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) criteria (eg, sample size, design, data collection, and follow-up). The RoB 2 (Risk of Bias) tool assessed areas and the levels of bias in randomized controlled trials (RCTs).

Results: The review included 15 observational studies and 3 RCTs and synthesized evidence from different medical fields and populations. Chatbots systematically collect information through targeted queries and data retrieval, improving patient engagement and satisfaction. The results show that chatbots have great potential for history-taking and that the efficiency and accessibility of the health care system can be improved by 24/7 automated data collection. Bias assessments revealed that of the 15 observational studies, 5 (33%) studies were of high quality, 5 (33%) studies were of moderate quality, and 5 (33%) studies were of low quality. Of the RCTs, 2 had a low risk of bias, while 1 had a high risk.

Conclusions: This systematic review provides critical insights into the potential benefits and challenges of using chatbots for medical history–taking. The included studies showed that chatbots can increase patient engagement, streamline data collection, and improve health care decision-making. For effective integration into clinical practice, it is crucial to design user-friendly

interfaces, ensure robust data security, and maintain empathetic patient-physician interactions. Future research should focus on refining chatbot algorithms, improving their emotional intelligence, and extending their application to different health care settings to realize their full potential in modern medicine.

Trial Registration: PROSPERO CRD42023410312; www.crd.york.ac.uk/prospero

(JMIR Med Inform 2024;12:e56628) doi:10.2196/56628

KEYWORDS

medical history-taking; chatbots; artificial intelligence; natural language processing; health care data collection; patient engagement; clinical decision-making; usability; acceptability; systematic review; diagnostic accuracy; patient-doctor communication; cybersecurity; machine learning; conversational agents; health informatics

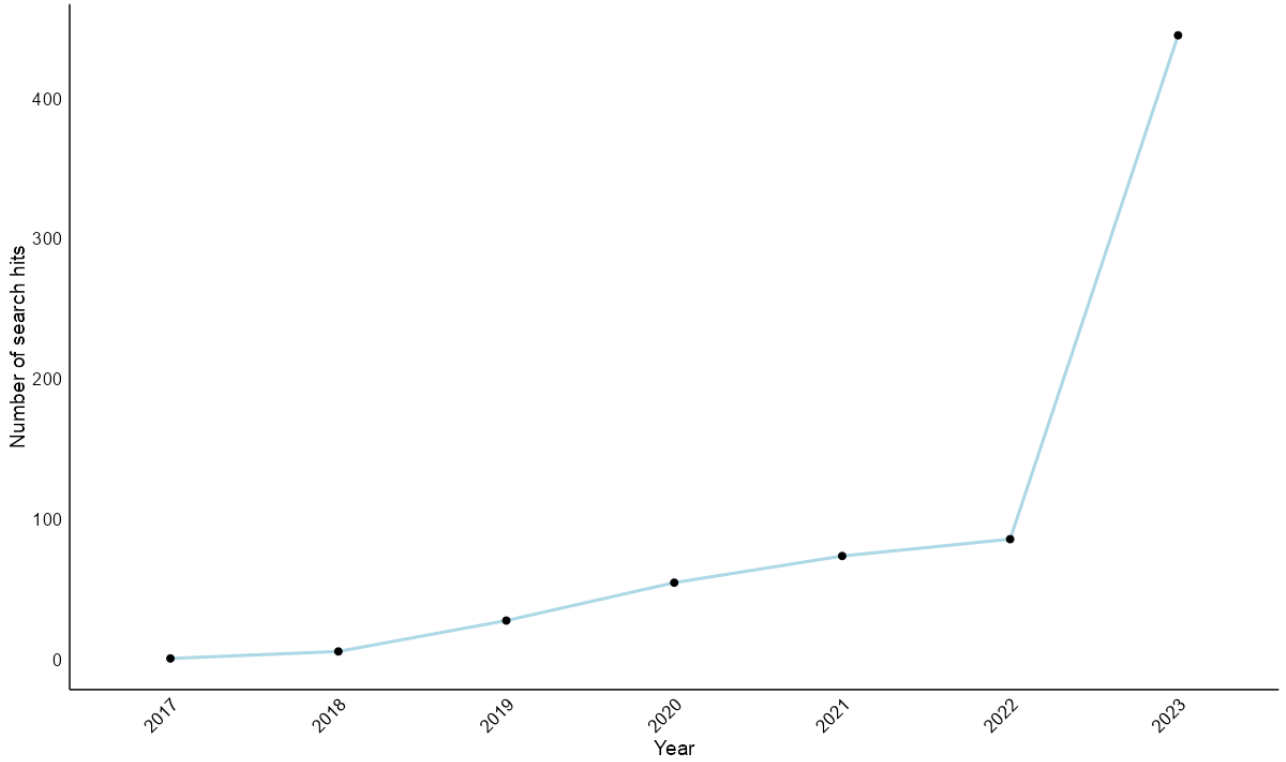
Introduction

Taking a patient's medical history is of central importance in the health care sector. Collecting comprehensive data is essential for accurate diagnosis and customized treatment [1]. Traditionally, clinicians have relied on interviews or questionnaires to gather this important information, but these methods can lack efficiency and accuracy, potentially leading to incomplete records and low patient engagement [2]. New technologies have brought about innovative solutions to streamline documentation, such as chatbots, with their ability to digitally transform data collection [3]. Chatbots can use artificial intelligence (AI) and natural language processing (NLP) to simulate conversations and minimize the limitations of paper-based processes [4-6]. The integration of chatbots promises significant improvements in care by enabling accurate, streamlined documentation that supports personalized, evidence-based clinical decision-making and greater patient engagement [7,8]. While chatbots are widely used in other areas, such as entertainment, customer service [9], security systems, and emergency communications [10-12], there is a lack of thorough research evaluating their effectiveness, usability, and acceptability of chatbots specifically for health care data collection. Research has focused on a narrow area without contextualizing the broader implications. To date, few people have had access to sophisticated AI due to its cost and complexity. However, new publicly available models, such as ChatGPT, are making these capabilities accessible to a wide audience by analyzing large amounts of literature and data in seconds to make time-critical decisions in a more data-driven and accurate way [13-17]. For interactions in the health care sector, specific and individual patient profiles can be addressed in order to improve documentation and the associated health outcomes. In addition, continued adoption will ensure that counseling by health care professionals remains widely accessible, especially in underserved communities [18]. In addition, their ability to work continuously and remotely can improve health care by ensuring that expert-level advice is always available, improving access to quality care, especially in underserved areas [18,19]. However, these benefits must be balanced by robust measures to ensure that the use of AI in health care improves, rather than undermines, patient care and trust [20].

Despite the promise of chatbots, important considerations are taken into account, particularly in health care. Cybersecurity is paramount, as chatbots handle sensitive medical information that must be protected from unauthorized access or data breaches [21,22]. Furthermore, despite the remarkable capabilities of chatbots in effectively processing and generating responses through predefined algorithms, they often lack the empathetic understanding and emotional intelligence inherent in human interactions [23]. This limitation can affect relationship-building and patient trust, especially during sensitive medical conversations [20].

Recent data highlighted the growing interest in the interplay between chatbots and medicine. An analysis of studies from the first study in 2017 to 2024 with the search query "chatbot*" AND "medicine" shows a significant increase, especially in 2022, with the trend rising from a single study in 2017 to 445 in 2023 (Figure 1).

Figure 1. Number of studies over recent years: "chatbot*" AND "medicine." This chart shows the increasing trend in publications on chatbots in medicine from 2017 to 2023. In 2022, there was an exponential increase in published studies, indicating a growing research interest and progress in chatbots in medicine.



Chatbots rely on advanced algorithms and AI-supported NLP for their technical function. These techniques enable chatbots to examine user input, provide applicable data in the form of feedback, and modify their interactions depending on context and user behavior, which can be refined through machine learning approaches, including information-driven learning and pattern recognition [24-26].

Considering the potential benefits and problems associated with chatbots, a thorough investigation is essential to assess their impact on the process of medical history–taking. While existing studies have examined the practicality and acceptability of chatbots in specific medical areas, such as psychological well-being or genetic counseling, a systematic literature review is needed for a complete understanding of chatbot-based history-taking [27-29].

The primary objective of this systematic review is to provide a comprehensive assessment of the role, effectiveness, usability, and patient acceptance of chatbots in medical history–taking. This systematic review also aims to explore the impact and future directions of integrating chatbots into clinical settings by assessing data accuracy, level of patient interaction, health care provider efficiency, and patient outcomes. Chatbots could transform the process of taking medical histories by supporting the accurate capture of patient information. In addition, this has the potential to increase productivity and improve the quality and delivery of health care services.

Methods

Overview

The systematic analysis was conducted in accordance with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for reporting systematic reviews to ensure transparency [30]. The protocol was registered under registration number CRD42023410312 in the PROSPERO database of the National Institute for Health Research [31].

Eligibility Criteria

Eligibility criteria for the studies were based on the PICOS (participants, interventions, comparators, outcomes, study design) framework for assessing participant demographics, types of interventions assessed, study designs, and outcome of interest [32]. We aimed to identify research investigating chatbots to facilitate medical history–taking to support physicians in diagnosis and treatment planning. The scope was limited to chatbots that facilitate patient disclosure of personal health information to improve accuracy and support clinical decision-making. In contrast, chatbots designed exclusively as "symptom-checkers," such as stand-alone apps providing rapid assessments and potential diagnoses, were excluded. This exclusion was made to focus on tools that facilitate comprehensive medical history–taking rather than immediate symptom-based advice. There were no limitations on the modality of chatbot input and output. The comparators were not subjected to any specific restrictions. The outcomes of interest included the feasibility, acceptability, and efficacy of chatbot-based history-taking interventions. There were no restrictions on study design, except for conference papers, which

were excluded to ensure the inclusion of studies with rigorous peer review and substantial data reporting. The review was limited to English-language studies because resources were limited.

Information Sources

PubMed, CENTRAL, Embase, MEDLINE (through Ovid), Scopus, and Open Science were searched to identify relevant studies. In addition, reference lists of relevant studies were screened manually.

Search Strategy

For each database, we developed a search strategy that included keywords, subject headings, mesh terms (in PubMed), filters, and restrictions to find relevant studies. The search terms focused on chatbots, anamnesis, history-taking, and related concepts: ("chatbot*" OR "conversational agent*" OR "chatterbot*" OR "virtual assistant" OR "intelligent virtual agent" OR "artificial intelligence chatbot" OR "AI chatbot" OR "conversational AI" OR "dialogue system") AND ("anamnesis" OR "medical history" OR "history-taking" OR "medical interview" OR "patient interview" OR "medical questionnaire" OR "patient questionnaire"). The last search was done in July 2024 (Multimedia Appendix 1). Additionally, a reference list search was conducted.

Selection Process

The selection process was done by 2 authors (MH and SS) independently screening the titles and abstracts of the identified studies based on the predetermined eligibility criteria. Potentially relevant studies were retrieved in full text and further assessed for eligibility. The full-text assessment was also performed independently (MH and SS). Any disagreements between the 2 authors were resolved through discussion, focusing on the eligibility criteria and study relevance. If consensus could not be reached, the involvement of a third author (AZ) was sought when necessary.

Data Collection Process

Data from the selected studies were extracted independently (MH and SS) using a data extraction form based on the PICO criteria (STROBE [Strengthening the Reporting of Observational Studies in Epidemiology]) [32,33]. The extracted data included information such as the first author, number of authors, country, year, title of the scientific journal, topics and type of journal, impact factor, and main results focused on history-taking (anamnesis). Additional data collected encompassed study design, setting, sample size, type of participants, female percentage, mean age (range), and results. Outcomes extracted focused on key aspects such as feasibility, acceptability, and efficacy. When full-text access was unavailable, the corresponding author was contacted by email. Data were visualized using the R-package for creating alluvial diagrams [34]. Any discrepancies in data extraction were resolved through a discussion between the 2 authors (MH and SS).

Quality Assessment

The methodological quality of the included observational studies was assessed using the STROBE criteria [33]. Each study was evaluated based on the fulfillment of the STROBE criteria. The studies were categorized into 3 categories: category A, if more than 80% of the STROBE criteria were fulfilled; category B, if 50%-80% were met; and category C if less than 50% of the criteria were fulfilled [35]. For example, category A studies provided comprehensive details on study objectives, participant selection, and statistical analysis. Category B had adequate but incomplete information. Category C studies frequently lacked critical details such as clear definitions of eligibility criteria or thorough data collection methods.

In addition, the RCTs included in this review were evaluated for risk of bias using the Risk of Bias tool and the robvis R-package [36,37]. The RoB 2 tool assesses various domains of bias, including randomization, allocation concealment, blinding, incomplete outcome data, selective reporting, and other potential sources of bias. The overall risk of bias score was determined for each study based on the number of criteria for high risk of bias met. Studies are considered to have a low risk of bias if no domains are rated as high risk and most domains are rated as low risk. Studies with some concerns in one or more domains but no high-risk ratings are considered to have some concerns. If any domain is rated as high risk, the study is considered to have a high risk of bias.

Software and Tools

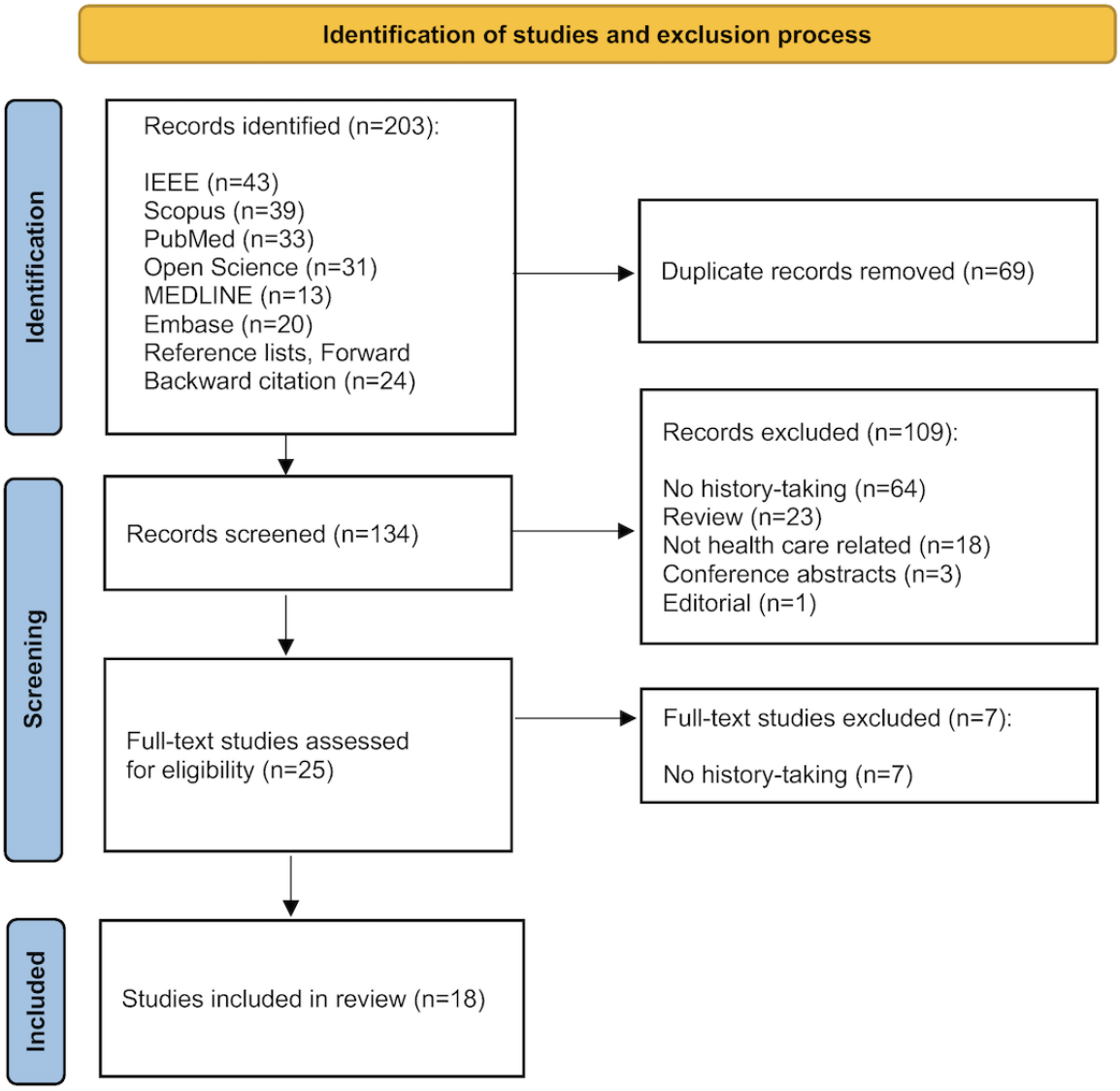
Data were managed and analyzed using R (version 4.2.1; The R Foundation). The ggplot2 package [38] was used for data visualization and the robvis R-package was used for risk of bias charts [37]. The alluvial R package [34] was used to create alluvial diagrams.

Results

Study Selection

The initial literature search yielded 203 records. After removing 69 duplicate studies, a total of 134 unique records were screened based on titles and abstracts. Of these, 109 studies did not meet the eligibility criteria and were excluded. Subsequently, 25 full-text studies were screened, resulting in 18 studies being included in the review (Figure 2).

Figure 2. Flowchart of the study search and inclusion. This flowchart details the systematic process of selecting studies for the review, starting from 203 records and narrowing down to 18 studies after removing duplicates and applying eligibility criteria. IEEE: Institute of Electrical and Electronic Engineers.



Study Characteristics

The studies investigated the use of chatbots for history-taking across diverse patient populations and sample sizes (range: n=5-61,070) and were mostly published in scientific health technology journals with varying impact factors (mean 4.52, SD 4.49; range: 0.14-14.71; Table 1). The studies used different research designs, including 9 cross-sectional studies, 3 case-control studies, 2 observational studies, and 3 RCTs (Multimedia Appendix 1 and Tables 1-3).

Table 1. General characteristics of the included studies. This table summarizes the number of authors, countries, and journal topics of the studies, showing most research from Germany and the United States, and a focus on Health Informatics and Technology.

	Count, n (%)
Numbers of authors	
1-3	4 (22)
4-6	8 (44)
>6	6 (33)
Countries	
Germany	6 (33)
United States	6 (33)
Switzerland	3 (17)
Australia	2 (11)
New Zealand	1 (6)
Scientific journals	
Topics of scientific journals	
Health Informatics and Technology	12 (67)
Medical Imaging and Radiology	2 (11)
Genetics and Genetic Counseling	2 (11)
Surgical Procedures and Techniques	1 (6)
Mental Health and Psychology	1 (6)

Table 2. Study characteristics. This table details study characteristics, including author, year, design, sample size, participant type, and key findings, highlighting diverse participant demographics and study outcomes.

Reference		Participants				Methods and result	
Authors (year)	Study design	n	Type of participants	Female (%)	Mean age (years)	Type of measurement	Relevant results
Denecke et al (2018) [39]	Cross-sectional study	22	Music therapy patients	41	39 (range 19-73)	Usability test of the tool and corresponding questionnaire	CUI[a]-based self-anamnesis app well-received, potential for collecting anamnesis data.
Denecke et al (2022) [40]	Cross-sectional study	5	Radiology patients	40	39.2 (range 17-73)	System usability scale	Digital medical interview assistant with good usability.
Faqar-Uz-Zaman et al (2022) [41]	RCT[b]	450	Patients with abdominal pain in ER[c]	52.2	44 (range 18-97)	Accuracy of diagnosis by ER doctor and Ada app according to the final diagnosis	Classic patient-physician interaction superior to AI[d]-based tool, but AI benefits diagnostic efficacy.
Frick et al (2021) [42]	Cross-sectional study	148	German participants	53	33.32 (SD 12.59)	Scales for disclosure and concealment of medical information	Patients prefer disclosing to physicians over chatbots. No significant difference in concealment.
Gashi et al (2021) [43]	Cross-sectional study	N/A[e]	N/A	N/A	N/A	N/A	AnCha chatbot improves patient-doctor communication, enhances diagnostic process.
Ghosh et al (2018) [44]	Case-control study	30 scenarios	Not specified	N/A	N/A	True positives and false positives, precision	Medical chatbot helps with automated patient preassessment.
Heald et al (2021) [27]	Feasibility study	506	Various types of care	58	56.6 (SD 12.5)	Colon cancer risk assessment tool	Chatbot feasible for increasing genetic screening in at-risk individuals.
Hennemann et al (2022) [45]	Observational study	49	Adult patients from an outpatient psychotherapy clinic	61	33.41 (SD 12.79)	Interviews, questionnaires, diagnostic software	Chatbot shows moderate to good accuracy for condition suggestions.
Hong et al (2022) [46]	Cross-sectional study	20	Primary care patients	60	50	Web-based survey	Patients believe chatbot helps clinicians better understand their health.
Ireland et al (2021) [28]	Cross-sectional study	83	Adults who had whole exome sequencing for genetic condition diagnosis	53	range 23.2-80.4	Transcript analysis	Chatbot enhances genetic counseling by providing genomic information.
Jungmann et al (2019) [47]	Case-control study	6	Psychotherapists, psychology students, and laypersons	50	40 (therapists) 22 (students)	Case vignettes, health app comparison	Chatbot shows moderate diagnostic agreement, improvement needed for childhood disorders.
Nazareth et al (2021) [48]	Retrospective, observational study	61,070	Women's health	96	N/A	Genetic testing results	Chatbot helps identify patients at high risk for hereditary cancer syndromes.
Ni et al (2017) [49]	Cross-sectional study or proof-of-concept	11	Patients with chest pain, respiratory infections, headaches, and dizziness	N/A	N/A	Question accuracy, prediction accuracy	Chatbot generates medical reports with varying accuracy based on disease category.

Reference		Participants				Methods and result	
Authors (year)	Study design	n	Type of participants	Female (%)	Mean age (years)	Type of measurement	Relevant results
Ponathil et al (2020) [50]	Cross-sectional study	50	Adults	50	N/A	NASA Task Load Index workload instrument IBM Usability Questionnaire Technology Acceptance Model Questionnaire	Chatbot interface saves time, preferred for collecting family health history.
Reis et al (2020) [51]	Case-control study	16	Physicians	35	35.51	N/A	Failure of cognitive agent highlights need for managing resistance and transparency.
Schneider et al (2023) [52]	RCT	30	Hymenoptera venom allergic patients	N/A	38.93 (SD 12.56)	Standardized questionnaire	Chatbot-supported anamnesis saves time, potential for allergology assessments.
Wang et al (2015) [29]	RCT, hospital	70	Majority of patients from underserved populations (low-income families, elders, people with disabilities, and immigrants)	60	Majority in age group 45-54	Interview, questions	Technological support for documenting family history risks is accepted and feasible.
Welch et al (2020) [53]	Cross-sectional study	3204	General population	100	49.4 (SD 7.1)	Standardized questionnaire	Chatbot engages users, potential for gathering family health history at population level.

[a]CUI: conversational user interface.

[b]RCT: randomized controlled trial.

[c]ER: emergency room.

[d]AI: artificial intelligence.

[e]Not applicable.

Table 3. Chatbot characteristics. This table outlines the chatbots used in the studies, including their name, goal, modality, techniques, outcomes, user preferences, and challenges, showcasing varied applications and technological approaches in health care. Table format based on Schachner et al [54].

Authors (year)	Name	Goal	Modality	Techniques	Main outcomes	User preference	Challenges
Denecke et al (2018) [39]	Ana	Collect medical history for music therapy	Mobile app: Text input	AIML[a], rule-based	Comprehensive data collection, usability	Engaging, intuitive	Integration, diverse interactions, data completeness
Denecke et al (2022) [40]	Not specified	Improve radiological diagnostics	Telegram CUI[b]	RiveScript (rule-based)	Enhanced knowledgeability, diagnostic quality	User-friendly	Clinical workflow integration, data security
Faqar-Uz-Zaman et al (2022) [41]	Ada	Evaluate diagnostic accuracy in ER[c]	iPad app	AI[d] questionnaire, ML[e]	Increased diagnostic accuracy	Not specified	Physician integration, diagnostic variability
Frick et al (2021) [42]	Not specified	Elicit truthful medical disclosure	Digital survey	Common CA[f] technologies	Disclosure versus concealment	Prefer physicians	Information accuracy, privacy
Gashi et al (2021) [43]	AnCha	Collect previsit medical history	IBM Watson, web-based	Rule-based tree	Efficient data collection	Reduces previsit anxiety	Clinical integration, data security
Ghosh et al (2018) [44]	Quro	User symptom check, personalized assessments	Web interface	NLP[g], ML	Precision in condition prediction	High engagement	Data complexity, accurate predictions
Heald et al (2021) [27]	Not specified	Screen for heritable cancer syndromes	Web-based, text-based	AI conversation, NLP	Efficient risk assessment, facilitated testing	High engagement, completion rates	Workflow integration, genetic risk understanding
Hennemann et al (2022) [45]	Ada	Diagnose mental disorders	App-based symptom checker	AI analysis, NLP	Moderate diagnostic accuracy	Mixed preferences	Diagnostic performance, user input dependency
Hong et al (2022) [46]	Genie	Collect detailed medical histories	Web-based, AI speech-to-text	AI, NLP	Improved history collection	Helpful for PCPs[h]	Ease of use, AI use concerns
Ireland et al (2021) [28]	Edna	Support genomic findings decision-making	Mobile, tablet, PC	NLP, Sentiment Analysis	Enhanced patient agency, informed decisions	Ease of access, supports consent	Empathy, complex interactions, data privacy
Jungmann et al (2019) [47]	Ada	Diagnose mental disorders	Mobile app	AI symptom analysis	Moderate diagnostic agreement	Not specified	Accuracy for complex cases
Nazareth et al (2021) [48]	Gia	Hereditary cancer risk triage	Web-based, mobile	NLP	Automated risk triage, educational interactions	High engagement	Workflow integration, privacy, diverse needs
Ni et al (2017) [49]	Mandy	Automate patient intake	Mobile app	NLP, data-driven analysis	Reduced staff workload, privacy maintenance	Improves physician efficiency	Full clinical integration, privacy, diverse interactions
Ponathil et al (2020) [50]	VCA	Collect family health history	Web-based chat	Not specified	Higher satisfaction, lower workload	Preferred by most users	Multiple clicks, extensive interaction
Reis et al (2020) [51]	Cognitive Agent	Automate anamnesis-diagnosis-treatment	Voice-based AI chatbot	ML, NLP, speech recognition	Reduced documentation time	Reduces nonbillable activities	Physician resistance, legal concerns, oversimplification
Schneider et al (2023) [52]	Not specified	Standardize allergy history-taking	HTML-based, digital	HTML, Java scripting	Time-efficient, accurate history-taking	High satisfaction	Question clarity, specificity
Wang et al (2015) [29]	VICKY	Collect family health histories	Touchscreen tablet	Speech recognition, decision trees	High satisfaction, effective identification	Easy to use, recommended	Data entry issues, complex questions
Welch et al (2020) [53]	It Runs In My Family	Assess hereditary cancer risk	Web-based chatbot	NLP	High engagement, thorough assessments	Prefer chatbot to web forms	Data accuracy, interface design, demographic reach

[a]AIML: artificial intelligence markup language.

[b]CUI: conversational user interface.

[c]ER: emergency room.

[d]AI: artificial intelligence.

[e]ML: machine learning.

[f]CA: conversational agent.

[g]NLP: natural language processing.

[h]PCP: primary care physician.

The alluvial diagram (Figure 3 [27,29,39-53]) shows an overview of the literature over time, indicating the year, the country of origin, and the medical area of focus for each study. The included studies were published from 2015 to 2023. Most of the studies were published in 2020 and 2022. The included studies (Figures 3 [27,29,39-53] and 4) were conducted in Switzerland [39,40,43], Germany [41,42,45,47,51,52], the United States [27,29,46,48,50,53], Australia [28,44], and New Zealand [49]. The studies cover a diverse range of medical areas: general medicine [42-44,49,51] genetics [28,29,48,50] cancer research [27,53], family medicine [46], mental health [45,47], radiology [40], surgery [41], allergy [52], and music therapy [39].

Figure 3. Alluvial diagram of the publication date, country, and area of studies. The alluvial diagram illustrates the distribution of studies by year, country, and medical area from 2015 to 2023, highlighting increased publications in 2020 and 2022, with contributions from Germany, the United States, and Switzerland across various medical fields.

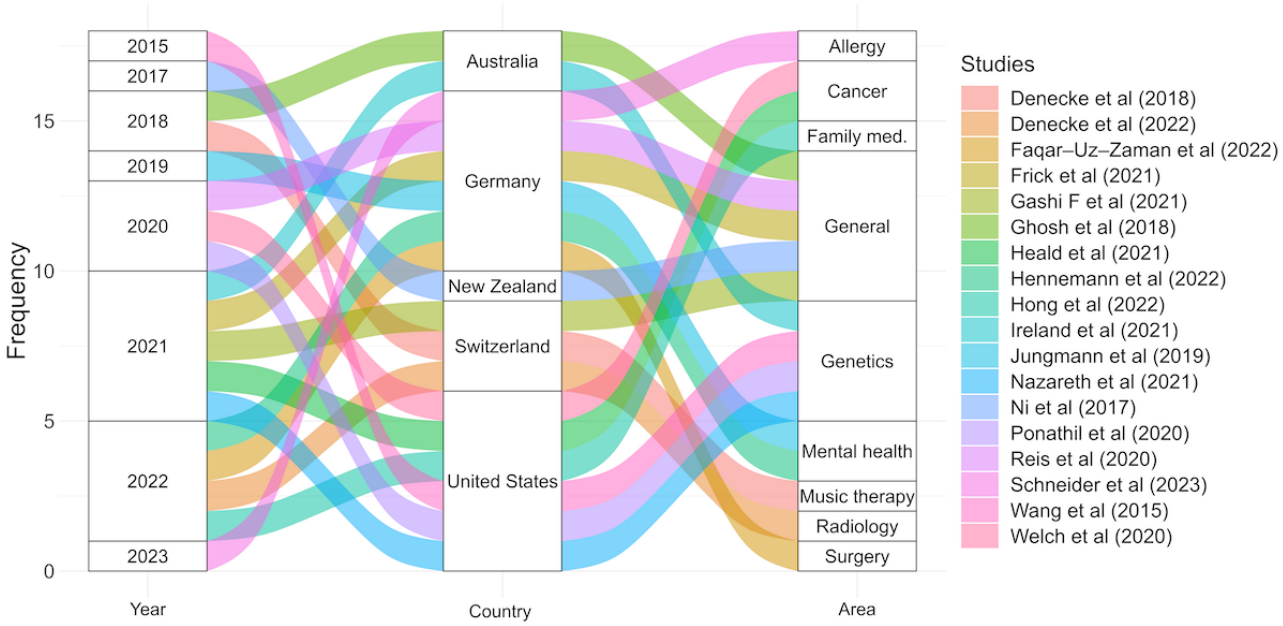
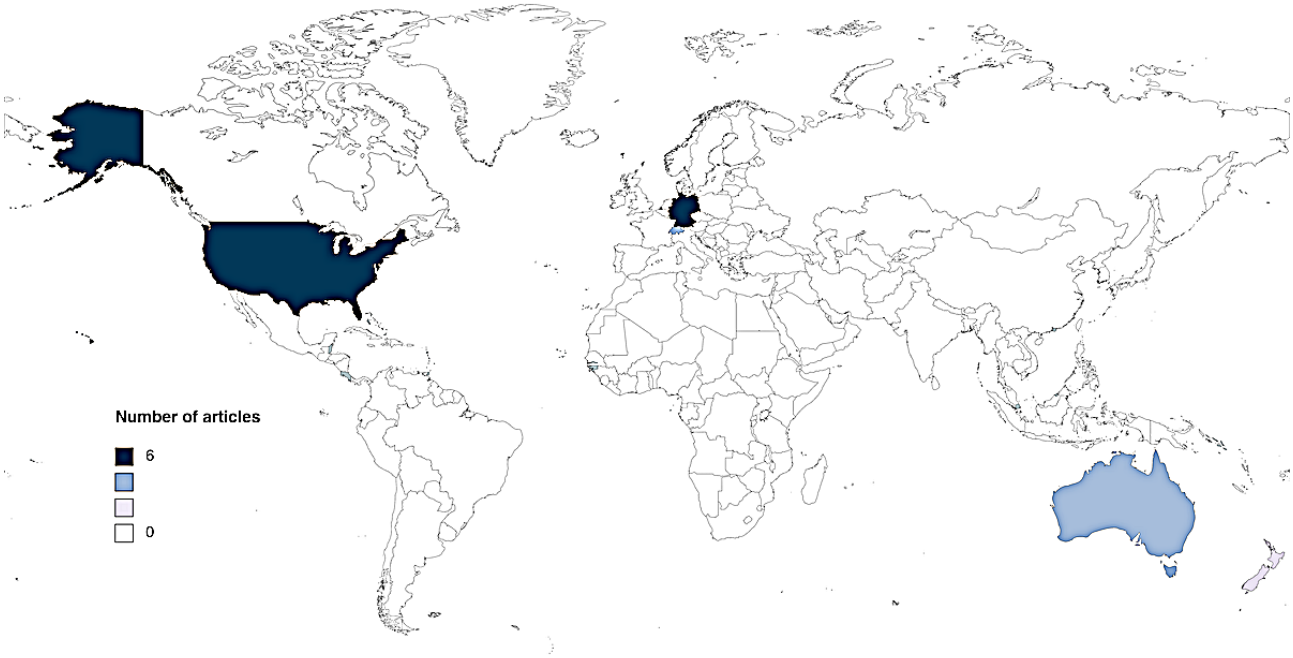


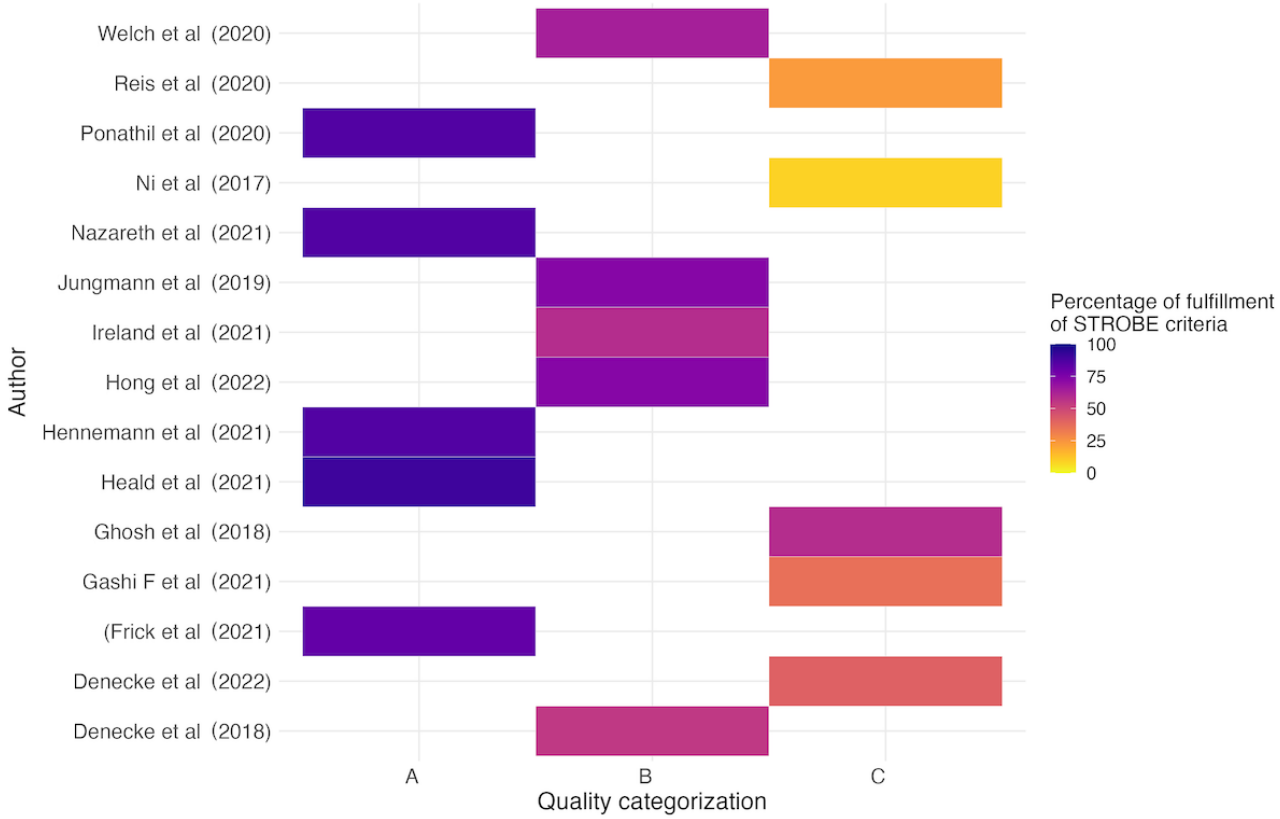
Figure 4. World map showing the number of studies published in each country. This map shows the geographical distribution of the studies, with most research originating from Germany and the United States. Created with MapChart [55].



Quality Appraisal of the Included Studies

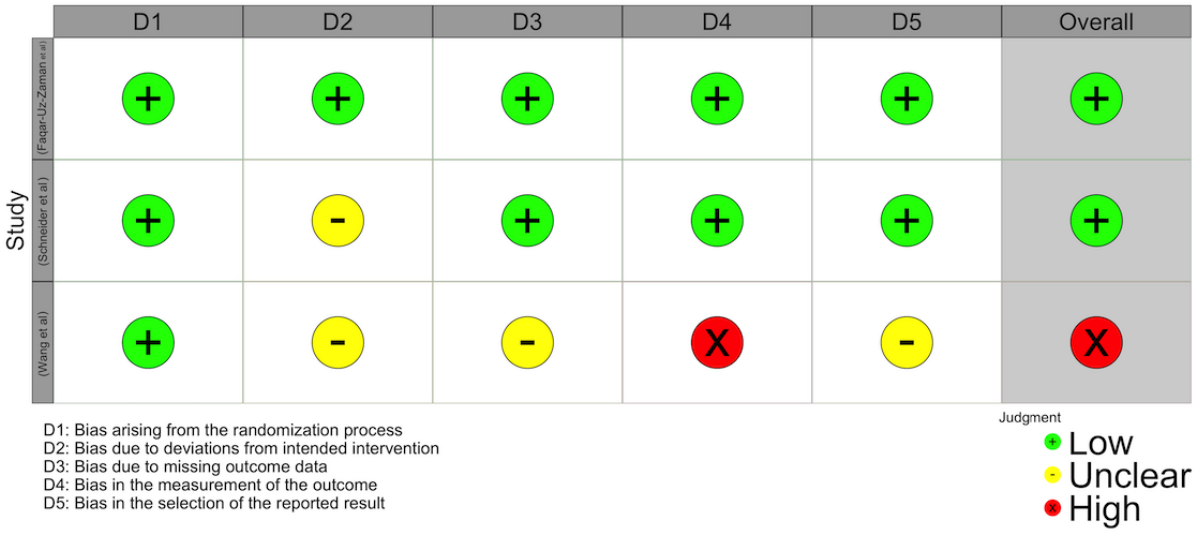
Among the 16 observational studies, 6 (38%) studies were classified as category A [27,42,45,48,50], indicating high methodological quality with more than 80% of the STROBE criteria fulfilled (Multimedia Appendix 1). A total of 5 (31%) studies were classified as category B [28,39,46,47,53], meeting 50%-80% of the STROBE criteria, and 5 (31%) studies were classified as category C [40,43,44,49,51], meeting less than 50% of the STROBE criteria (Figure 5 [27,28,39,40,42-51,53]). The lack of adherence to STROBE criteria in observational studies can have a significant impact on the quality. Missing elements, such as clear definitions of eligibility criteria or participants or detailed methods, lead to biases that reduce validity and reliability. For example, the study of Denecke et al [40] showed a high risk of selection bias due to a small, nonrepresentative sample and lack of eligibility criteria, limiting the generalizability of their findings. Gashi et al [43] faced biases from the absence of a control group and unclear eligibility criteria. This could impact the validity of the effectiveness results. Ghosh et al [44] showed high bias from simulated scenarios without real patient interactions. This could lead to overestimated accuracy and applicability in real-world settings.

Figure 5. Fulfillment of STROBE criteria and categorization. This bar chart categorizes observational studies by their adherence to STROBE criteria, showing 37.5% of high-quality (category A), and an even split between moderate (category B) and lower quality (category C). STROBE: Strengthening the Reporting of Observational Studies in Epidemiology.



The studies by Schneider et al [52] and Faqar-Uz-Zaman et al [41] showed a low risk of bias according to the RoB tool, with detailed methodology and statistical analysis. In contrast, the study by Wang et al [29] showed a risk of bias due to the absence of intention-to-treat analysis and participants being aware of the intervention (Multimedia Appendix 1 and Figure 6), which could skew results by excluding noncompleters and altering participant behavior.

Figure 6. Risk of bias domains (RoB-tool) for randomized controlled trials.



Summary of Statistical Analyses

The studies included in this systematic review used a variety of statistical methods. Descriptive statistics summarized demographics and usability ratings. Comparative analyses used 2-tailed *t* tests and chi-square tests to compare diagnostic accuracy and user engagement. κ statistics measured agreement between chatbot and expert diagnoses. Precision and accuracy metrics were assessed using precision, recall, and F_1-scores. Nonparametric tests, such as the Mann-Whitney *U* test showed significant reductions in anamnesis duration. CIs and *P* values were reported where relevant to clarify the strength of the evidence.

Usability and User Experience of Chatbots

Five studies focused on the usability and user experience of chatbots in history-taking (Tables 2 and 3). Denecke et al [39,40] found that chatbots were well-received by participants and showed potential for history-taking. Usability scores were high, between 90 and 100 (average 96). Ponathil et al [50] found that using a voice-controlled assistant interface for taking family health history significantly reduced history-taking duration. Ghosh et al [44] implemented a medical chatbot that assists with automated patient preassessment through symptom analysis, demonstrating the possibility of avoiding form-based data entry. The chatbot correctly identified at least one of the top three conditions in 83% (n=25) of cases and two out of three conditions in 67% (n=20) of cases. Welch et al [53] found high engagement and interest in chatbots, suggesting the potential for gathering family health history information at the population level in the United States. Of the over 14,000 who participated in the assessment of the study, 54.4% (n=7616) of users went beyond the consent step, and 22.7% (n=3178) of users completed the full assessment.

Chatbots and Patient-Doctor Communication

One study highlighted the potential of chatbots to improve patient-doctor communication. Gashi et al [43] reported that using a chatbot could reduce patient nervousness, allow patients to respond more thoughtfully, and give physicians a more comprehensive picture of the patient's condition.

Diagnostic Accuracy and Efficacy of Chatbots

Nazareth et al [48] found that a chatbot can help identify high-risk patients for hereditary cancer syndromes. A total of 27.2% (n=14,850) of the chatbot users met the criteria for genetic testing, and 5.6% (n=73) of the chatbot users had a pathogenic variant. Ni et al [49] reported that Mandy, a chatbot, automates history-taking, understands symptoms expressed in natural language, and generates comprehensive reports for further medical investigations, with varying degrees of accuracy depending on the disease category. Hennemann et al [45] reported that the app-based symptom checker with an AI chatbot showed agreement with therapist diagnoses in 51% (n=25) of cases for the first condition suggestion and in 69% (n=34) of cases for the top five condition suggestions. Jungmann et al [47] tested a health app's diagnostic agreement with case vignettes for mental disorders, pointing to the need for improvement in diagnostic accuracy, especially for mental disorders in childhood and adolescence.

Patient Perceptions and Acceptance of Chatbots

Hong et al [46] reported that most primary care patients believed that chatbots could help clinicians better understand their health and identify health risks. Ireland et al [28] found that the development of the Edna tool, an AI-based chatbot that interacts with patients via speech-to-text, signifies progress toward creating digital health processes that are accessible, acceptable, and well-supported, enabling patients to make informed decisions about additional findings. Heald et al [27] highlighted the feasibility of using chatbots for increasing genetic screening and testing in individuals at risk of hereditary colorectal cancer syndromes.

Challenges and Limitations of Chatbots

Reis et al [51] noted the importance of managing user resistance and fostering realistic expectations when implementing AI-based history-taking tools. Frick et al [42] found that patients preferred to disclose medical information to a physician rather than a conversational agent.

Effectiveness on Chatbots

Faqar-Uz-Zaman et al [41] found that classic patient-physician interaction was superior to an AI-based diagnostic tool applied by patients. However, they also noted that AI tools can benefit clinicians' diagnostic efficacy and improve the quality of care. Schneider et al [52] found that a chatbot-supported anamnesis could save significant time by 57.3%, in assessing Hymenoptera venom allergies with high completeness (73.3%) and patient satisfaction (75%). Wang et al [29] demonstrated that technological support for documenting family history risks can be highly accepted, feasible, and effective.

Discussion

Principal Results

This systematic review highlights that the use of chatbots can improve medical history–taking. Results of the included studies have shown that chatbots can facilitate data collection while increasing patient engagement and satisfaction [39,49]. Chatbots show value, especially in collecting structured data such as family history [29,50,53]. As highlighted, the collection of family history benefits significantly from chatbot automation due to the simple nature of their queries, which typically require binary responses. This area contrasts with the challenges of collecting data on undiagnosed symptoms, where patient responses are inherently more nuanced and variable. The inherent abilities of chatbots to handle yes or no questions efficiently and without misinterpretation make them particularly valuable in this context, minimizing human error and optimizing the data collection process. Several studies have highlighted that chatbots provide a more engaging patient interaction, often perceived as less intimidating than traditional face-to-face conversations [27,46]. This interaction is crucial as it motivates patients to disclose more comprehensive health information, which can lead to better health outcomes. While chatbots excel at retrieving and conveying information through interactions that require limited context, their capabilities remain limited when it comes to more nuanced understanding and complex emotions. Research has shown that specific sensitive topics are best-discussed face-to-face with a human, where building trust is paramount [42]. Chatbots, on the other hand, offer relief through constant availability and allow patients to share details from any location and at any time, which can expand access—especially for urgent needs that require quick access to medical history [41,53]. This expanded access aims to improve care, especially in cases where timely data can make the difference between outcomes. In addition, chatbots support overburdened care providers by systematically presenting summarized patient data, potentially enabling faster and more accurate decisions [43,52]. Such support is invaluable in high-pressure situations requiring rapid action based on comprehensive information. These findings are consistent with

previous research that emphasizes the ability of chatbots to capture patient reports in a structured, comprehensive way [3,22]. Their conversational design facilitates higher engagement and satisfaction through interactive discussions [4,50]. This contributes to improved documentation of patient histories. Furthermore, automated information capture has been confirmed to increase both the efficiency and accessibility of health care by simplifying reporting processes [21,39].

While chatbots already promise success in supporting diagnostic processes, the required level of accuracy must be achieved for complex medical scenarios that require in-depth understanding and sound clinical judgment. The limitations of current systems are highlighted in the studies by Hennemann et al [45] and Jungmann et al [47], highlighting the need to improve the algorithms and decision-making processes to manage complex health conditions.

While the seamless integration of conversational agents into clinical workflows requires robust data infrastructures and user-friendly interfaces, such integration can drive adoption among care providers and patients if done in a secure manner [48]. Customized chatbots are required to serve different patient audiences and different facilities. Addressing these needs can increase patient engagement and satisfaction [48,50].

However, the development of such technologies requires careful consideration [56]. Rushing to release chatbots without thorough refinement and validation can lead to inaccuracies and potentially detrimental outcomes. These hastily deployed chatbots run the risk of failing to understand complex medical situations and recommending incorrect diagnoses or treatments. The use of chatbots requires caution and rigorous testing or validation to minimize the risks [57-59].

Limitations

Although this systematic review provided useful insights, certain limitations must be acknowledged. As we only considered papers published in English, we may have overlooked important work published in other languages. In the future, a more comprehensive review that includes multilingual research could promote a more complete understanding of chatbots worldwide. The variability of study designs, patient groups, and health care contexts makes it difficult to draw definitive conclusions. Different studies, such as those by Denecke et al [39] and Faqar-Uz-Zaman et al [41], focused on different settings and patient groups, which influenced the results. Cross-sectional studies provide snapshots of usability, while RCTs provide robust evidence. Heterogeneity in demographics and health status also affects generalizability, as seen in the studies by Welch et al [53] and Wang et al [29]. Bias assessment frequently showed unmet STROBE criteria. Clear eligibility criteria and detailed methods could influence reliability. For example, Gashi et al [43] lacked defined selection criteria, and Jungmann et al [47] had a selection bias. Inconsistent reporting and lack of blinding in some RCTs, such as Wang et al [29], impaired internal validity.

The methodological quality of the included studies varied. At the same time, most observational studies demonstrated satisfactory quality, and a significant proportion fulfilled only some of the STROBE criteria. Additionally, the risk of bias assessment of the RCTs revealed a high risk of bias in one of the studies [41]. It is important to consider these limitations when interpreting the data and trying to understand how they relate to clinical practice. In addition, only published research has been included in this systematic review, which may lead to publication bias as studies with positive results are more likely to be published [41].

Future Directions

Based on the findings and limitations of this systematic review, future research should focus on conducting more standardized and well-designed studies in this field. Emphasizing rigorous study designs, such as RCTs, with larger sample sizes and standardized outcome measures will enhance the scientific validity of the research and provide more substantial evidence of the effectiveness of chatbots in history-taking. Standardized outcome measures between studies are crucial for better comparability. Future studies should use measures such as diagnostic accuracy, patient satisfaction, engagement, and usability ratings. Instruments, such as the system usability scale or the technology acceptance model, could be used. Further investigation is needed to explore the specific contexts and patient populations where chatbots for history-taking may be most effective [29,50,53]. Different medical areas and health situations may present special considerations and challenges that could influence the implementation and acceptance of chatbot-based systems for taking medical histories, such as in the case of older people due to a more limited technical affinity or long medical histories in people with chronic illnesses.

Moreover, future research should address the challenges and limitations identified in this review. Efforts should be made to minimize bias and improve the methodological quality of studies. Conducting studies with more homogeneous patient populations and using consistent outcome measures would enhance the comparability and generalizability of the findings [39].

Finally, it would be valuable to explore the integration of chatbots with other technologies or interventions to optimize the history-taking process. The integration of chatbots with modern technologies, such as NLP, machine learning algorithms, and decision support systems, has the potential to significantly improve history-taking [21,46,51]. NLP could improve the ability to understand and interpret patient responses to the chatbot. The interactions will be more fluid and intuitive. Machine learning algorithms can be used to continuously improve chatbot responses based on patient interactions. This could lead to more accurate and personalized information. The integration of decision support systems can provide health care providers with real-time evidence-based recommendations. Research designs to investigate these integrations could include comparative studies for measuring differences in diagnostic accuracy, patient satisfaction, and efficiency between 2 groups. One group could use a simple chatbot, and another group could use an advanced chatbot with integrated NLP and machine learning.

Conclusions

The systematic review provides an insightful overview of the use of chatbots in medical history–taking. The results show that chatbots can increase data completeness and user satisfaction. This can encourage patient engagement, and more accurate assessment can be achieved in a reduced timeframe. Chatbots can be used in primary care before the face-to-face visit. This would not only reduce the workload of medical staff but also enable more targeted interaction between patients and physicians. Future research should focus on different areas to improve the use of chatbots for medical history–taking. Larger studies and RCTs are essential for adequate validation. The use of chatbots needs to be investigated in different health care settings and with different patient groups, for example, in patients with chronic diseases, mental illness, or older patients and in people who are not tech-savvy. Another area that needs to be considered is analyzing the impact of chatbots on workflows in clinics or practices and the change in the doctor-patient relationship. In addition, data protection and security issues must be clarified to ensure the protection of patient data, especially considering the latest developments in AI models. These offer new opportunities for more precise and personalized interactions. Research should optimize these models for history-taking and integrate them into decision support systems for real-time evidence-based recommendations. If these areas are addressed, chatbots can significantly transform health care by improving efficiency, accuracy, and patient engagement, especially for underserved patient populations, as well as chronic disease management and real-time symptom assessment.

Acknowledgments

This systematic review was funded by the Department of Dermatology and Allergology of the Technical University of Munich, Germany. Funding did not influence the review process or results.

Data Availability

All data generated or analyzed during this study are included in this published article. All aggregate data collected for this review are available from the corresponding author upon reasonable request.

Authors' Contributions

MH conceptualized and designed the analysis, collected the data, performed the screening and analysis, and was the primary author of the article. SS served as the second reviewer for screening and quality appraisal. AZ critically reviewed and provided feedback on the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategies conducted, overview of studies, quality assessment of included studies.
[PDF File (Adobe PDF File), 324 KB - medinform_v12i1e56628_app1.pdf]

Multimedia Appendix 2
PRISMA Checklist.
[PDF File (Adobe PDF File), 20 KB - medinform_v12i1e56628_app2.pdf]

References

1. Fowler FJ, Levin CA, Sepucha KR. Informing and involving patients to improve the quality of medical decisions. Health Aff (Millwood) 2011;30(4):699-706. [doi: 10.1377/hlthaff.2011.0003] [Medline: 21471491]
2. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J 1975;2(5969):486-489 [FREE Full text] [doi: 10.1136/bmj.2.5969.486] [Medline: 1148666]
3. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
4. Denecke K, May R, Deng Y. Towards emotion-sensitive conversational user interfaces in healthcare applications. Stud Health Technol Inform 2019;264:1164-1168. [doi: 10.3233/SHTI190409] [Medline: 31438108]
5. Hess GI, Fricker G, Denecke K. Improving and evaluating eMMA's communication skills: a chatbot for managing medication. Stud Health Technol Inform 2019;259:101-104. [Medline: 30923283]
6. Marietto MDGB, Aguiar RV, Barbosa GDO, Botelho WT, Pimentel E, Franca RDS, et al. Artificial intelligence markup language: a brief tutorial. Int J Comp Sci Eng 2013;4(3):1-20. [doi: 10.5121/ijcses.2013.4301]

7. Rebelo N, Sanders L, Li K, Chow JCL. Learning the treatment process in radiotherapy using an artificial intelligence-assisted chatbot: development study. JMIR Form Res 2022;6(12):e39443 [FREE Full text] [doi: 10.2196/39443] [Medline: 36327383]
8. Chew HSJ. The use of artificial intelligence-based conversational agents (Chatbots) for weight loss: scoping review and practical recommendations. JMIR Med Inform 2022;10(4):e32578 [FREE Full text] [doi: 10.2196/32578] [Medline: 35416791]
9. Xu Y, Zhang J, Deng G. Enhancing customer satisfaction with chatbots: the influence of communication styles and consumer attachment anxiety. Front Psychol 2022;13:902782 [FREE Full text] [doi: 10.3389/fpsyg.2022.902782] [Medline: 35936304]
10. Amiri P, Karahanna E. Chatbot use cases in the COVID-19 public health response. J Am Med Inform Assoc 2022;29(5):1000-1010 [FREE Full text] [doi: 10.1093/jamia/ocac014] [Medline: 35137107]
11. Almalki M, Azeez F. Health chatbots for fighting COVID-19: a scoping review. Acta Inform Med 2020;28(4):241-247 [FREE Full text] [doi: 10.5455/aim.2020.28.241-247] [Medline: 33627924]
12. Judson TJ, Odisho AY, Young JJ, Bigazzi O, Steuer D, Gonzales R, et al. Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic. J Am Med Inform Assoc 2020;27(9):1450-1455 [FREE Full text] [doi: 10.1093/jamia/ocaa130] [Medline: 32531066]
13. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]
14. Someya T, Amagai M. Toward a new generation of smart skins. Nat Biotechnol 2019;37(4):382-388. [doi: 10.1038/s41587-019-0079-1] [Medline: 30940942]
15. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health 2023;5(3):e102 [FREE Full text] [doi: 10.1016/S2589-7500(23)00023-7] [Medline: 36754723]
16. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
17. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
18. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024;3(1):12 [FREE Full text] [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
19. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595 [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
20. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020;22(6):e15154 [FREE Full text] [doi: 10.2196/15154] [Medline: 32558657]
21. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 2021;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
22. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
23. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or pandora's box? JAMA Intern Med 2023;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]
24. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018;5(4):e64 [FREE Full text] [doi: 10.2196/mental.9782] [Medline: 30545815]
25. Oh YJ, Zhang J, Fang M, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. Int J Behav Nutr Phys Act 2021;18(1):160 [FREE Full text] [doi: 10.1186/s12966-021-01224-6] [Medline: 34895247]
26. Bickmore TW, Silliman RA, Nelson K, Cheng DM, Winter M, Henault L, et al. A randomized controlled trial of an automated exercise coach for older adults. J Am Geriatr Soc 2013;61(10):1676-1683. [doi: 10.1111/jgs.12449] [Medline: 24001030]
27. Heald B, Keel E, Marquard J, Burke CA, Kalady MF, Church JM, et al. Using chatbots to screen for heritable cancer syndromes in patients undergoing routine colonoscopy. J Med Genet 2021;58(12):807-814. [doi: 10.1136/jmedgenet-2020-107294] [Medline: 33168571]
28. Ireland D, Bradford D, Szepe E, Lynch E, Martyn M, Hansen D, et al. Introducing Edna: a trainee chatbot designed to support communication about additional (secondary) genomic findings. Patient Educ Couns 2021;104(4):739-749. [doi: 10.1016/j.pec.2020.11.007] [Medline: 33234441]

29. Wang C, Bickmore T, Bowen DJ, Norkunas T, Campion M, Cabral H, et al. Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories. Genet Med 2015;17(10):822-830 [FREE Full text] [doi: 10.1038/gim.2014.198] [Medline: 25590980]
30. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
31. PROSPERO—International prospective register of systematic reviews. NIHR. URL: https://www.crd.york.ac.uk/prospero/ [accessed 2023-04-02]
32. Miller SA, Forrest JL. Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions. J Evid Based Dent Pract 2001;1(2):136-141. [doi: 10.1016/s1532-3382(01)70024-3]
33. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008] [Medline: 18313558]
34. Bojanowski M, Edwards R. alluvial: R package for creating alluvial diagrams. R package version: 0.1-2. 2016. URL: https://cran.r-project.org/web/packages/alluvial/citation.html [accessed 2024-08-01]
35. Mendy A, Gasana J, Vieira ER, Forno E, Patel J, Kadam P, et al. Endotoxin exposure and childhood wheeze and asthma: a meta-analysis of observational studies. J Asthma 2011;48(7):685-693. [doi: 10.3109/02770903.2011.594140] [Medline: 21732750]
36. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
37. McGuinness LA, Higgins JPT. Risk-of-bias visualization (robvis): an R package and shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]
38. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd Edition. New York: Springer International Publishing; 2016.
39. Denecke K, Hochreutener S, Pöpel A, May R. Self-anamnesis with a conversational user interface: concept and usability study. Methods Inf Med 2018;57(5-06):243-252. [doi: 10.1055/s-0038-1675822] [Medline: 30875703]
40. Denecke K, Lombardo P, Nairz K. Digital medical interview assistant for radiology: opportunities and challenges. Stud Health Technol Inform 2022;293:39-46 [FREE Full text] [doi: 10.3233/SHTI220345] [Medline: 35592958]
41. Faqar-Uz-Zaman SF, Anantharajah L, Baumartz P, Sobotta P, Filmann N, Zmuc D, et al. The diagnostic efficacy of an app-based diagnostic health care application in the emergency room: eRadaR-Trial. A prospective, double-blinded, observational study. Ann Surg 2022;276(5):935-942. [doi: 10.1097/SLA.0000000000005614] [Medline: 35925755]
42. Frick NR, Brünker F, Ross B, Stieglitz S. Comparison of disclosure/concealment of medical information given to conversational agents or to physicians. Health Informatics J 2021;27(1):1460458221994861 [FREE Full text] [doi: 10.1177/1460458221994861] [Medline: 33779384]
43. Gashi F, Regli SF, May R, Tschopp P, Denecke K. Developing intelligent interviewers to collect the medical history: lessons learned and guidelines. Stud Health Technol Inform 2021;279:18-25. [doi: 10.3233/SHTI210083] [Medline: 33965913]
44. Ghosh S, Bhatia S, Bhatia A. Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system. Stud Health Technol Inform 2018;252:51-56. [Medline: 30040682]
45. Hennemann S, Kuhn S, Witthöft M, Jungmann SM. Diagnostic performance of an app-based symptom checker in mental disorders: comparative study in psychotherapy outpatients. JMIR Ment Health 2022;9(1):e32832 [FREE Full text] [doi: 10.2196/32832] [Medline: 35099395]
46. Hong G, Smith M, Lin S. The AI will see you now: feasibility and acceptability of a conversational AI medical interviewing system. JMIR Form Res 2022;6(6):e37028 [FREE Full text] [doi: 10.2196/37028] [Medline: 35759326]
47. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res 2019;3(4):e13863 [FREE Full text] [doi: 10.2196/13863] [Medline: 31663858]
48. Nazareth S, Hayward L, Simmons E, Snir M, Hatchell KE, Rojahn S, et al. Hereditary cancer risk using a genetic chatbot before routine care visits. Obstet Gynecol 2021;138(6):860-870 [FREE Full text] [doi: 10.1097/AOG.0000000000004596] [Medline: 34735417]
49. Ni L, Lu C, Liu N, Liu J. MANDY: towards a smart primary care chatbot application. Commun Comput Inf Sci 2017:38-52. [doi: 10.1007/978-981-10-6989-5_4]
50. Ponathil A, Ozkan F, Welch B, Bertrand J, Madathil KC. Family health history collected by virtual conversational agents: an empirical study to investigate the efficacy of this approach. J Genet Couns 2020;29(6):1081-1092. [doi: 10.1002/jgc4.1239] [Medline: 32125052]
51. Reis L, Maier C, Mattke J, Creutzenberg M, Weitzel T. Addressing user resistance would have prevented a healthcare AI project failure. MIS Q Exec 2020;19(4):8 [FREE Full text] [doi: 10.17705/2msqe.00038]
52. Schneider S, Gasteiger C, Wecker H, Höbenreich J, Biedermann T, Brockow K, et al. Successful usage of a chatbot to standardize and automate history taking in Hymenoptera venom allergy. Allergy 2023;78(9):2526-2528. [doi: 10.1111/all.15720] [Medline: 36946258]

53. Welch BM, Allen CG, Ritchie JB, Morrison H, Hughes-Halbert C, Schiffman JD. Using a chatbot to assess hereditary cancer risk. JCO Clin Cancer Inform 2020;4:787-793 [FREE Full text] [doi: 10.1200/CCI.20.00014] [Medline: 32897737]
54. Schachner T, Keller R, Wangenheim FV. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. J Med Internet Res 2020;22(9):e20701 [FREE Full text] [doi: 10.2196/20701] [Medline: 32924957]
55. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/
56. Ni Z, Peng ML, Balakrishnan V, Tee V, Azwa I, Saifi R, et al. Implementation of chatbot technology in health care: protocol for a bibliometric analysis. JMIR Res Protoc 2024;13:e54349 [FREE Full text] [doi: 10.2196/54349] [Medline: 38228575]
57. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
58. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
59. Wilson L, Marasoiu M. The development and use of chatbots in public health: scoping review. JMIR Hum Factors 2022;9(4):e35882 [FREE Full text] [doi: 10.2196/35882] [Medline: 36197708]

Abbreviations

AI: artificial intelligence
NLP: natural language processing
PICOS: participants, interventions, comparators, outcomes, and study design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Castonguay; submitted 22.01.24; peer-reviewed by T Agresta, S Sakilay, H Aghayan Golkashani; comments to author 04.05.24; revised version received 08.05.24; accepted 11.07.24; published 29.08.24.

Please cite as:
Hindelang M, Sitaru S, Zink A
Transforming Health Care Through Chatbots for Medical History-Taking and Future Directions: Comprehensive Systematic Review
JMIR Med Inform 2024;12:e56628
URL: https://medinform.jmir.org/2024/1/e56628
doi:10.2196/56628
PMID:

Review

Automated Identification of Postoperative Infections to Allow Prediction and Surveillance Based on Electronic Health Record Data: Scoping Review

Siri Lise van der Meijden[1,2], MSc; Anna M van Boekel[1], MD; Harry van Goor[3], MD, PhD; Rob GHH Nelissen[4], MD, PhD; Jan W Schoones[5], MA; Ewout W Steyerberg[6], PhD; Bart F Geerts[2], MD, MBA, PhD; Mark GJ de Boer[7], MD, PhD; M Sesmu Arbous[1], MD, PhD

[1]Intensive Care Unit, Leiden University Medical Center, Leiden, Netherlands

[2]Healthplus.ai BV, Amsterdam, Netherlands

[3]General Surgery Department, Radboud University Medical Center, Nijmegen, Netherlands

[4]Department of Orthopedics, Leiden University Medical Center, Leiden, Netherlands

[5]Directorate of Research Policy, Leiden University Medical Center, Leiden, Netherlands

[6]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

[7]Department of Infectious Diseases, Leiden University Medical Center, Leiden, Netherlands

Corresponding Author:
Siri Lise van der Meijden, MSc
Intensive Care Unit
Leiden University Medical Center
Albinusdreef 2
Leiden, 2333 ZA
Netherlands
Phone: 31 526 9111
Email: S.L.van_der_meijden@lumc.nl

Abstract

Background: Postoperative infections remain a crucial challenge in health care, resulting in high morbidity, mortality, and costs. Accurate identification and labeling of patients with postoperative bacterial infections is crucial for developing prediction models, validating biomarkers, and implementing surveillance systems in clinical practice.

Objective: This scoping review aimed to explore methods for identifying patients with postoperative infections using electronic health record (EHR) data to go beyond the reference standard of manual chart review.

Methods: We performed a systematic search strategy across PubMed, Embase, Web of Science (Core Collection), the Cochrane Library, and Emcare (Ovid), targeting studies addressing the prediction and fully automated surveillance (ie, without manual check) of diverse bacterial infections in the postoperative setting. For prediction modeling studies, we assessed the labeling methods used, categorizing them as either manual or automated. We evaluated the different types of EHR data needed for the surveillance and labeling of postoperative infections, as well as the performance of fully automated surveillance systems compared with manual chart review.

Results: We identified 75 different methods and definitions used to identify patients with postoperative infections in studies published between 2003 and 2023. Manual labeling was the predominant method in prediction modeling research, 65% (49/75) of the identified methods use structured data, and 45% (34/75) use free text and clinical notes as one of their data sources. Fully automated surveillance systems should be used with caution because the reported positive predictive values are between 0.31 and 0.76.

Conclusions: There is currently no evidence to support fully automated labeling and identification of patients with infections based solely on structured EHR data. Future research should focus on defining uniform definitions, as well as prioritizing the development of more scalable, automated methods for infection detection using structured EHR data.

(JMIR Med Inform 2024;12:e57195) doi:10.2196/57195

KEYWORDS

postoperative infections; surveillance; prediction; surgery; artificial intelligence; chart review; electronic health record; scoping review; postoperative; surgical; infection; infections; predictions; predict; predictive; bacterial; machine learning; record; records; EHR; EHRs; synthesis; review methods; review methodology; search; searches; searching; scoping

Introduction

Postoperative bacterial infections, including deep or superficial surgical site infections (SSIs), urinary tract infections (UTIs), and pneumonia, are the most frequent complications after surgery. Postoperative infections can be categorized into subtypes, usually based on location or severity according to the Clavien-Dindo classification [1]. The overall incidence of postoperative infections within 30 days of surgery varies between 6.5% and 25% [2-4]. Considering the 313 million patients undergoing surgery globally each year, these postoperative infections have an enormous impact on population health and overall health care costs [5]. Effective postoperative infection prevention and management require early detection of high-risk patients through prediction and data-driven surveillance. It is imperative for developing and validating prediction and surveillance systems to be able to accurately identify patients who have postoperative infections. Machine learning modeling practices use the term “labeling” for the identification of patients with the outcome of interest. Labeling and surveillance are both challenges due to underreporting in (hospital) complication registries, ranging from 38% to 77% when compared with a manual chart review [6,7]. Consequently, the current reference standard for identifying patients with postoperative infections relies on labor-intensive manual chart review, with an estimated 1.5 full-time equivalents per 10,000 admissions [8,9]. Furthermore, manual surveillance and labeling are prone to interobserver variability [10,11] and human errors [12], highlighting the need for more robust methods to address this devastating postoperative problem.

To achieve a more objective, cost-effective, and resource-efficient identification of patients with postoperative infections, it is imperative to leverage the electronic health records (EHRs) to automatically detect patients with infections without human checking on high-risk patients based on readily available EHR data. Different types of data are present within the EHR, including structured, tabular, and free-text records in which diagnoses and clinical symptoms are reported. A previously performed systematic review identified semiautomated and fully automated surveillance methods for hospital-acquired infections (HAIs) [13]. As more than 90% of the included systems required manual checking of infectious cases, it was concluded that fully automated surveillance of HAIs cannot be routinely used yet in health care settings.

To go beyond manual labeling and manual surveillance and to explore the current methods and criteria used in prediction modeling studies, the aim of this study was to perform a scoping review on available labeling methods for postoperative infections and fully automated surveillance systems (ie, not requiring manual checking). We aimed to (1) evaluate the current methods and criteria used to label patients with postoperative infections in prediction modeling and biomarker validation studies, (2) explore available automated surveillance methods and their performance (sensitivity, specificity, positive predictive value [PPV], and negative predictive value [NPV]) in comparison with reference standard manual chart review, and (3) determine the necessary data types and sources needed to perform automated detection of postoperative infections.

Methods

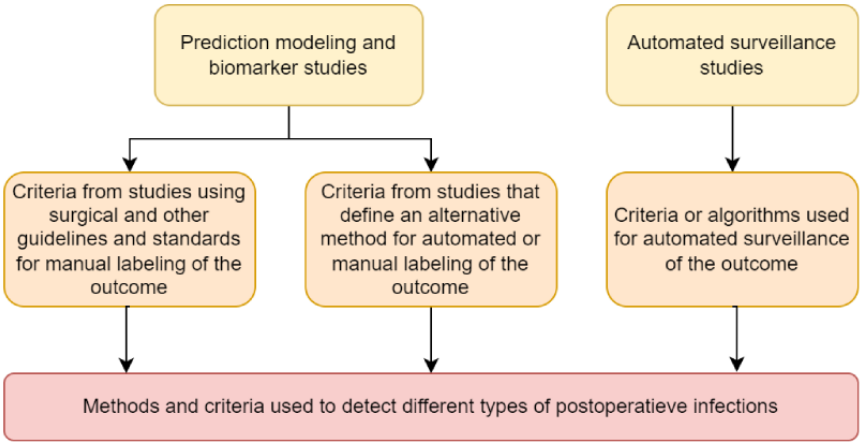
Overview

This scoping review combined 2 literature searches to evaluate current methods used by prediction modeling and biomarker validation studies to label patients with postoperative infections and the use of automated surveillance systems to identify patients with postoperative infections based on EHR data. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist was used. The protocol was registered on Open Science Framework [14].

Search Strategy

First, prediction modeling validation studies using machine learning methods, statistical models, and biomarkers to predict postoperative infections were identified. Second, a separate search was performed to identify studies on automated surveillance for postoperative and other hospital-acquired infections (Figure 1). Surveillance studies focusing on surgical populations often only investigate SSIs. As we aimed to study all bacterial infections that may occur after surgery, surveillance studies in a hospital-wide setting were also included. Both searches were performed in PubMed, Embase (OVID), Web of Science (Core Collection), the Cochrane Library, and Emcare (OVID). Studies were included from inception (ie, 1966) to August 1, 2023. The search queries were generated with help from an information specialist (JWS) from the Leiden University Medical Center. The details of the search queries are provided in Appendix A of Multimedia Appendix 1.

Figure 1. Data sources from the literature for identifying infections in prediction modeling or biomarker studies and automated surveillance studies.



Selection Criteria

The selection of studies was performed in Covidence, a program used to manage systematic literature searches. The inclusion and exclusion criteria used are presented in Table S1 in Multimedia Appendix 1. All titles and abstracts were screened by 2 independent reviewers (AMVB and BFG for prediction models; SLVDM and BFG for surveillance studies). The full texts of all potentially relevant studies were retrieved and assessed by 2 reviewers (SLVDM and BFG) for eligibility. Any disagreement on the inclusion or exclusion of studies was resolved through reassessment and discussion with a third reviewer (MSA). The data from the different reports were collected by 1 researcher (AMVB or SLVDM), and inconsistencies were checked for by a second researcher (BFG).

Data Extraction and Definitions

The following data were extracted for the prediction modeling studies: name of the prediction tool, type of prediction tool (machine learning, biomarker, and statistical model), surgical subpopulations, type of postoperative infection predicted, and criteria and guidelines used to manually or automatically label patients with infections. Manual labeling involves individuals conducting EHR chart reviews and applying specific criteria, often derived from surgical guidelines, to determine the presence or absence of infections in patient records. The criteria for diagnosing patients with an infection, for example, from a reference guideline from the literature, were identified and extracted.

For automated surveillance studies, the population, study design, years of data collection, type of infection surveyed, type of algorithm used, definition used to automatically detect infections, reference standard used to compare the automated method with, type of validation performed, and performance metrics reported compared with the reference standard were collected. The main metrics used to assess performance were the method's sensitivity, specificity, PPV, and NPV. Other metrics extracted are presented in Table S17 in Multimedia Appendix 1, including the area under the receiver operating characteristic curve, accuracy, F_1-score, κ score, Pearson correlation coefficient, and agreement percentage. Only performance metrics were assessed for surveillance studies, as for prediction modeling studies, and no accuracy of the labeling method compared with a reference standard was determined.

Data Synthesis

For each method to identify and label patients with infections, the data type categories needed from the EHR were assessed to identify infections based on the definition used. These could be structured EHR data (type A), including tabular information stored, such as complication registries, medication information, and vital signs; free-text clinical notes (type B), including all clinical information stored in free-text, such as discharge letters and daily reports; microbiology results (type C), which is seen as a separate category, as it differs per hospital how well-structured this information is stored [15]; and an additional interpretation layer (whether the results are positive) is needed to use this information; or imaging results (type D), or a combination of these categories. The definitions were further differentiated based on the data types and criteria needed to adhere to the definitions in Appendix E in Multimedia Appendix 1.

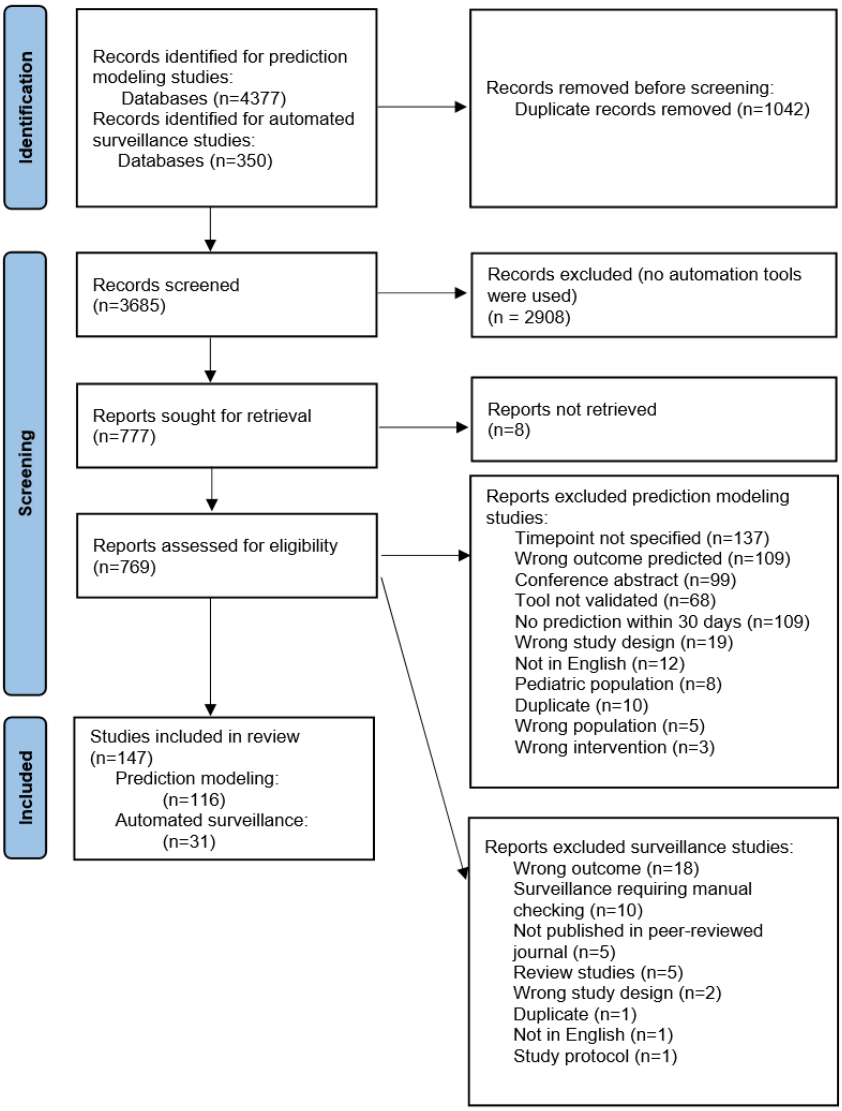
Some prediction models and surveillance systems are focused on predicting or detecting all severity types of bacterial infections, while others focus only on infections requiring pharmacological or surgical treatment. For example, some definitions include the prescription of antibiotics as one of the criteria, while others base their criteria on clinical symptoms only. As the severity of the infections surveyed or predicted influences the intended use case and number of infections identified, we classified the definitions according to the Clavien-Dindo scale [1]. Finally, the performance of the automated infection surveillance systems compared with that of the reference standard manual review was visualized per subtype of infection. The results were grouped according to the type of infection, such as HAI (type not further specified), SSI, pneumonia, anastomotic leakage and abdominal infections, bloodstream infections (including central venous catheter-related infections and sepsis), and UTIs. Infections that did not belong to one of these groups were categorized as "other."

Results

Overview

We included a total of 147 studies published between 2003 and 2023 (Figure 2). Of these, 116 studies focused on the prediction of postoperative infections; either the development and validation of prediction models or a predictive biomarker were performed, or validation was performed of preexisting risk scores. These included the American College of Surgeons National Surgical Quality Improvement Program surgical risk calculator (ACS NSQIP; 33/116 studies), the National Nosocomial Infection Surveillance System (NNIS; 4/116 studies), and the Surgical Risk Preoperative Assessment System (SURPAS; 5/116 studies).

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram.



Out of 116 studies, 4 did not report the methodology used to determine which patients had the outcome of interest (ie, postoperative infection). In 83% (97/116) of prediction modeling studies, manual labeling based on diagnostic guidelines was performed, or a publicly available, manually labeled database was used, such as the participant use data file from the ACS NSQIP program (Table 1). A total of 13% (15/116) of studies used an alternative, non–guideline-based method to label patients with infections, 11 of whom used manual labeling, 3 of whom did not explicitly mention manual or automatic labeling, and 1 of whom used automatic labeling. In total, 93% (108/116) of the prediction modeling studies used manual labeling to determine the outcome of interest or a manually labeled, publicly available data set to perform their research.

Table 1. Definitions of patients with bacterial infections.

Type of infection and reference	Origin of definition	Type A (structured)[a]	Type B (free-text)[b]	Type C (microbiology results)	Type D (imaging results)	Minimum Clavien-Dindo
HAI[c]						
WHO[d] [16]	Diagnostic guidelines	✓	✓	✓	✓	1
ECDC[e] [17]	Diagnostic guidelines	✓	✓			1
Ehrentraut et al [18]	Automated surveillance		✓			1
Sakji et al [19]	Automated surveillance		✓			1
Tvardik et al [20]	Automated surveillance		✓			1
Pneumonia						
ECDC/ASC NSQIP[f] [21]	Diagnostic guidelines	✓	✓	✓	✓	1
Kinlin et al [22]	Prediction modeling	✓	✓	✓	✓	1
Blacky et al [23]	Automated surveillance	✓		✓		1
Bouzbid et al [24]	Automated surveillance	✓		✓		2
Cato et al [25]	Automated surveillance	✓		✓		1
FitzHenry et al [26]	Automated surveillance		✓			1
Tvardik et al [20]	Automated surveillance		✓			1
Colborn et al [27]	Automated surveillance	✓				1
Stern et al [28]	Automated surveillance	✓			✓	1
SSI[g]						
CDC[h]/ASC NSQIP [29]	Diagnostic guidelines	✓	✓	✓	✓	1-3a
WHO [30]	Diagnostic guidelines	✓	✓	✓	✓	1-3a
Daneman et al [31]	Prediction modeling	✓				1
Weller et al [32]	Prediction modeling	✓				2
Crispin et al [33]	Prediction modeling	✓				3a
Martin et al [34]	Prediction modeling	✓	✓			2
Campillo-Gimenez et al [35]	Automated surveillance	✓	✓			1
Cato et al [25]	Automated surveillance	✓		✓		1
FitzHenry et al [26]	Automated surveillance		✓			1

Type of infection and reference	Origin of definition	Type A (structured)[a]	Type B (free-text)[b]	Type C (microbiology results)	Type D (imaging results)	Minimum Clavien-Dindo
Leclère et al [36]	Automated surveillance	✓		✓		1
Leth et al [37]	Automated surveillance	✓		✓		2
Suzuki et al [38]	Automated surveillance	✓	✓	✓		2
Tvardik et al [20]	Automated surveillance		✓			1
Thirukumaran et al [39]	Automated surveillance		✓			1
Colborn et al [27]	Automated surveillance	✓				1
Abdominal and AL[i]						
Rahbari et al [40]	Diagnostic guidelines	✓	✓		✓	1
Stidham et al [41]	Prediction modeling	✓				3a
Miyakita et al [42]	Prediction modeling	✓				3b
Mckenna et al [43]	Prediction modeling	✓				2
Nudel et al [44]	Prediction modeling	✓	✓			3a
Kawai et al [45]	Prediction modeling		✓		✓	2
Lin et al [46]	Prediction modeling		✓		✓	2
Shi et al [47]	Prediction modeling				✓	1
van Kooten et al [48]	Prediction modeling		✓		✓	1
UTI[j]						
ECDC/ASC NSQIP [49]	Diagnostic guidelines	✓	✓	✓		2
Cheng et al [50]	Prediction modeling		✓	✓		1
Bouam et al [51]	Automated surveillance			✓		1
Bouzbid et al [24]	Automated surveillance	✓		✓		2
Branch-Elliman et al [52]	Automated surveillance	✓	✓			1
Cato et al [25]	Automated surveillance	✓		✓		1
Choudhuri et al [53]	Automated surveillance	✓		✓		1
FitzHenry et al [26]	Automated surveillance		✓			1
Leth et al [37]	Automated surveillance	✓		✓		2

Type of infection and reference	Origin of definition	Type A (structured)[a]	Type B (free-text)[b]	Type C (microbiology results)	Type D (imaging results)	Minimum Clavien-Dindo
Redder et al [54]	Automated surveillance	✓		✓		2
Tvardik et al [20]	Automated surveillance		✓			1
van der Werff et al [55]	Automated surveillance	✓	✓	✓		2
Venable and Dissanaike [56]	Automated surveillance	✓	✓			—[k]
Wald et al [57]	Automated surveillance	✓		✓		1
Colborn et al [27]	Automated surveillance	✓				1
Bloodstream infections						
Moore et al [58]	Diagnostic guidelines	✓		✓		1
Singer et al [59] (sepsis-3 criteria)	Diagnostic guidelines	✓				2
Blacky et al [23]	Automated surveillance	✓		✓		1
Bouam et al [51]	Automated surveillance			✓		1
Bouzbid et al [24]	Automated surveillance	✓		✓		2
Cato et al [25]	Automated surveillance	✓		✓		1
FitzHenry et al [26]	Automated surveillance		✓			1
Leal et al [60]	Automated surveillance			✓		1
Leal et al [61]	Automated surveillance			✓		1
Lin et al [62]	Automated surveillance			✓		1
Redder et al [54]	Automated surveillance	✓		✓		2
Tvardik et al [20]	Automated surveillance		✓			1
Valik et al [63]	Automated surveillance	✓		✓		2
Venable and Dissanaike [56]	Automated surveillance	✓	✓			—[k]
Woeltje et al [64]	Automated surveillance	✓		✓		1
Colborn et al [27]	Automated surveillance	✓				1
Clostirdrium difficile						
Dubberke et al [65]	Automated surveillance			✓		1
Clostridium difficile						
van der Werff et al [66]	Automated surveillance			✓		1

Type of infection and reference	Origin of definition	Type A (structured)[a]	Type B (free-text)[b]	Type C (microbiology results)	Type D (imaging results)	Minimum Clavien-Dindo
External ventricular and lumbar drain-related meningitis						
van Mourik et al [67]	Automated surveillance	✓		✓		1
MRSA[l]						
Peterson et al [68]	Automated surveillance			✓		1
PJI[m]						
Fu et al [69]	Automated surveillance		✓			1
Neurological						
Cheng et al [70]	Prediction modeling	✓	✓	✓		1

[a]Type A (structured): structured electronic health record data, including tabular information stored such as complication registries, medication information, and vital signs.

[b]Type B (free-text): free-text clinical notes, including all clinical information stored in free-text such as discharge letters and daily reports.

[c]HAI: hospital-acquired infections.

[d]WHO: World Health Organization.

[e]ECDC: European Centre for Disease Prevention and Control.

[f]ACS NSQIP: American College of Surgeons National Surgical Quality Improvement Program.

[g]SSI: surgical site infection.

[h]CDC: Centers for Disease Control and Prevention.

[i]AL: anastomotic leakage.

[j]UTI: urinary tract infection.

[k]Not applicable.

[l]MRSA: Methicillin-resistant *Staphylococcus aureus*.

[m]PJI: prosthetic joint infection.

Automated Surveillance

We included 31 automated surveillance studies for bacterial infections. Surveillance was performed and reported per patient, admission, procedure, patient days, or culture. Different types of surveillance systems were studied, and some studies have reported on more than 1 method. Most often (21/31, 68%), a set of criteria or rules was defined to automatically detect infections based on EHR data, followed by natural language processing (NLP) algorithms for free-text from the EHR (7/31, 23%) and other classification algorithms such as logistic regression (3/31, 10%). Except for one study [25], all the studies validated their automated surveillance algorithms against a reference standard (manual chart review, often according to one of the established diagnostic guidelines). Comparing the automated surveillance algorithm to manual chart review according to the established guidelines resulted in a range of sensitivity (0.79-0.96), specificity (0.81-0.96), PPV (0.31-0.76), and NPV (0.96-1.00) estimates for the different types of infection (Figure 3). The performance of all the combinations of postoperative infection data needed to run the automated surveillance algorithm varied (Figure 4). Reported performance per surveillance algorithm is provided in Table S17 in Multimedia Appendix 1.

Figure 3. Performance of automated surveillance of postoperative infections compared with manual reference standard chart review. Panel A is the sensitivity, B is the specificity, C is the PPV, and D is the NPV. HAI: hospital-acquired infection; NPV: negative predictive value; PPV: positive predictive value; SSI: surgical site infection; UTI: urinary tract infection.

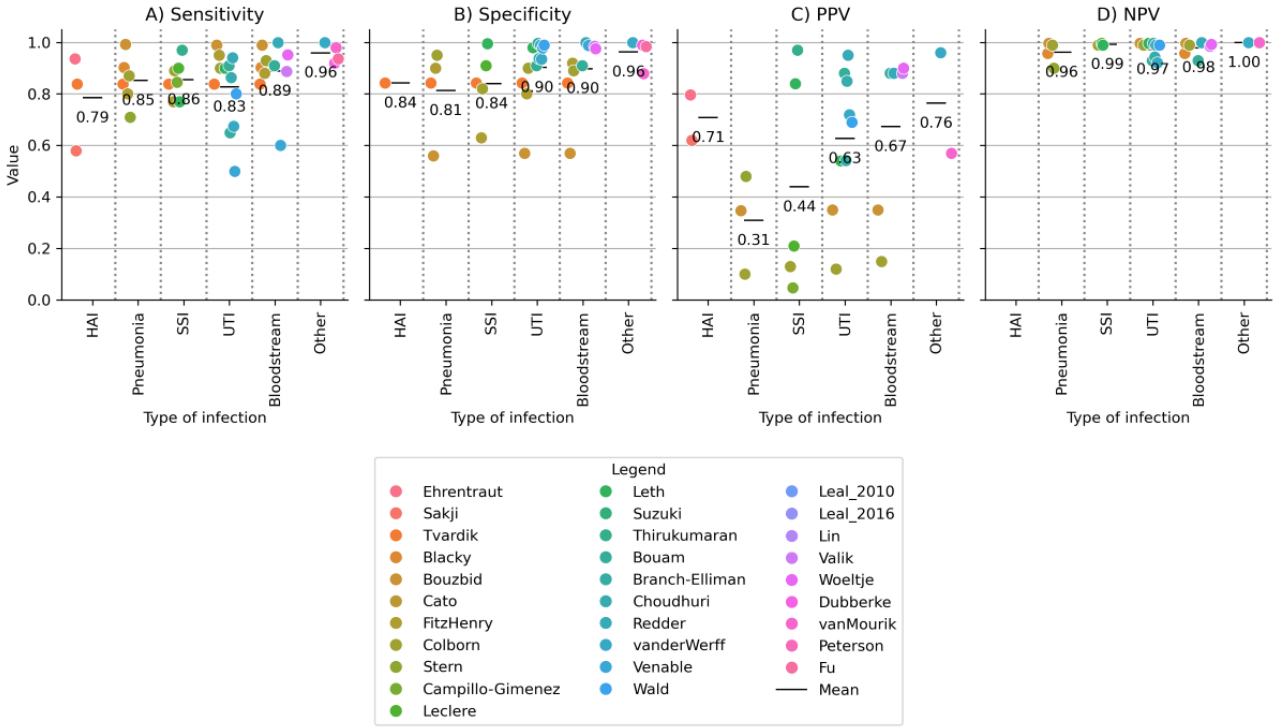
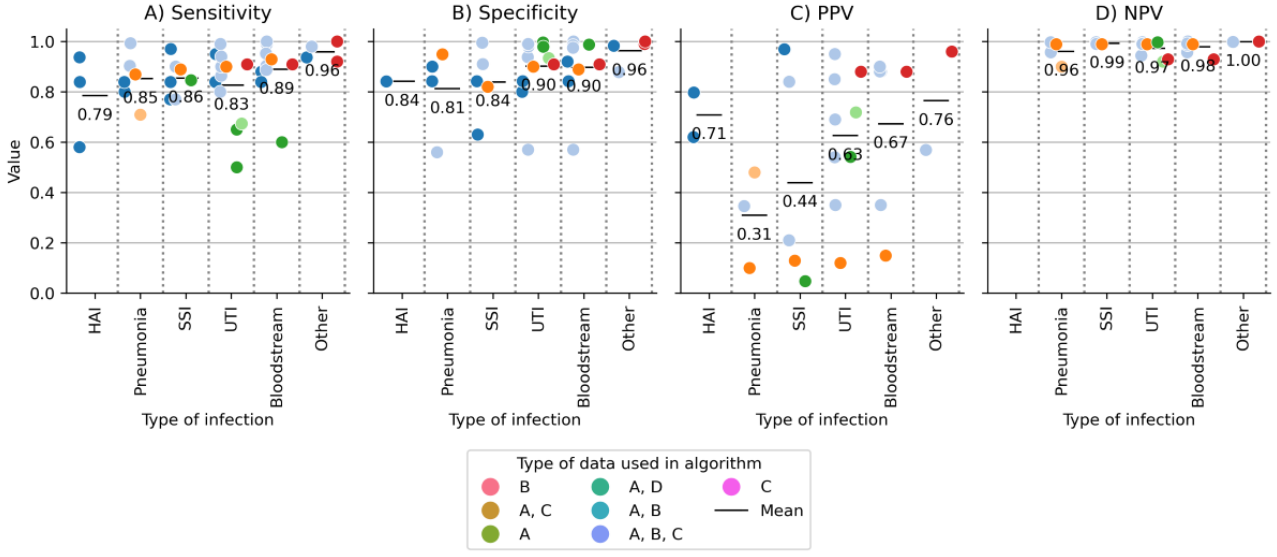


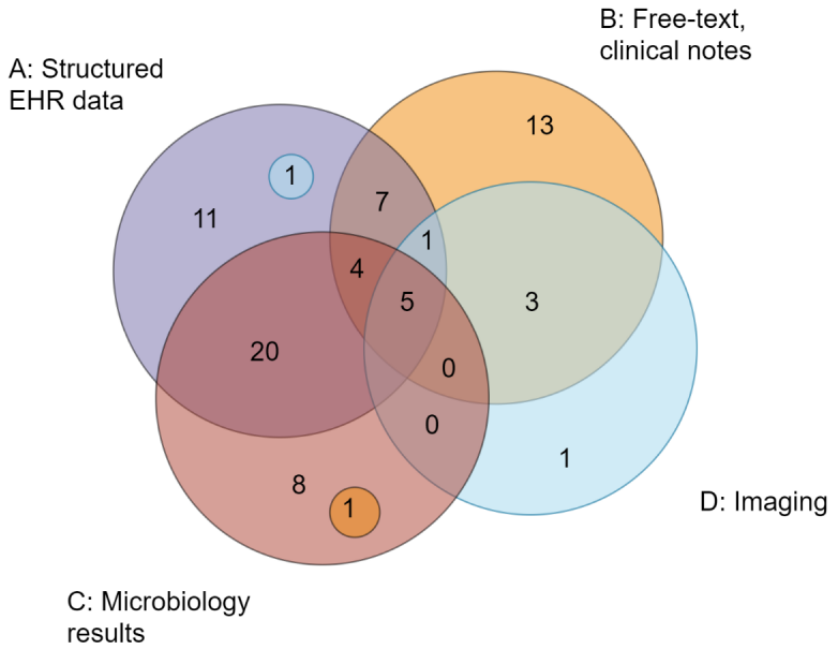
Figure 4. Performance per data type category used in automated surveillance algorithms. A=Structured electronic health record data only (eg, registrations and medication), B=Free-text clinical notes, C=microbiology results. Panel A is the sensitivity, B is the specificity, C is the PPV and D is the NPV. HAI: hospital-acquired infection; NPV: negative predictive value; PPV: positive predictive value; SSI: surgical site infection; UTI: urinary tract infection.



Electronic Health Record Data for Automated Identification and Surveillance

In the 147 included studies, 75 different methods and definitions were used to identify different types of bacterial infections. A total of 56% (42/75) used 2 or more datatypes to label, diagnose, or surveil infections, and 45% (34/75) required free-text and clinical notes as at least one of their data sources. In Table 1, the different types of data from the EHR needed to automatically detect patients with an infection are specified for each diagnostic guideline or infection definition used in the different prediction modeling studies or automated surveillance methods. Figure 5 shows the total number of methods used to identify patients with bacterial infections and the different data categories used. Most frequently (20/75, 27%), a combination of microbiology results and structured EHR data was used, followed by free-text (13/75, 17%) and structured EHR data (11/75, 15%). In total, 45% (34/75) of the identified methods used free text and clinical notes as one of their data sources.

Figure 5. Venn diagram of the types of data used to identify bacterial infections in the included studies and guidelines. The data were divided into structured electronic health record data, free-text and clinical notes, microbiology results, and imaging. In total, 75 unique definitions were identified for different types of bacterial infections. EHR: electronic health record.



For hospital-acquired infections (no specification of subtype), free-text information was needed for all definitions and methods, limiting the ability to detect patients with an infection based on structured EHR data. For pneumonia, some automated surveillance studies have identified patients without the need for free-text information [24,25], but they did include culture results in their definition. For SSIs and UTIs, a wide range of criteria were used compared with other types of infections. Abdominal surgery-related anastomotic leakage and abdominal infections were identified based on antibiotic treatment or surgical reinterventions supplemented with free-text data or imaging results. Bacterial culture data, in combination with structured EHR parameters, are used in most methods for detecting bloodstream infections. For *Clostridium difficile* infections, cerebral extraventricular and lumbar drain-related meningitis, methicillin-resistant *Staphylococcus aureus*, and prosthetic joint infection, the authors used a maximum of 2 criteria from different categories to define infection. Prediction modeling studies that did not use manual chart review for labeling patients in the data set relied on the registration of infections or the performance of surgical interventions, sometimes in combination with antibiotic administration [32].

When assessing infection severity according to the different Clavien-Dindo definitions, most (64%, 48/75) were based on identifying infections according to a Clavien-Dindo score of 1 or more. This indicates that, based on the registration of infection or clinical criteria only, infections were surveyed and predicted. In 23% of definitions (17/75), the prescription of antibiotic therapy or surgical intervention was included as the criterion, resulting in a Clavien-Dindo score of 2 or higher.

Discussion

This scoping review assessed the methods and criteria used for identifying postoperative bacterial infections in prediction modeling and fully automated surveillance studies. We identified a total of 75 different methods and definitions from 147 included studies to identify patients with different types of bacterial infections. We found that 45% (34/75) used unstructured free-text and clinical notes as at least one of their data sources. Furthermore, out of 116 postoperative infection prediction studies, 108 (93%) used manual labeling based on self-defined criteria or diagnostic guidelines or used publicly available manually labeled databases. In addition, among the 31 automated surveillance studies, various methods, such as NLP, classification algorithms, and predefined criteria or rules on structured data, were used to automatically detect infections. Compared with manual chart review, automated surveillance systems have reported sensitivities for different types of infections ranging from 0.79 to 0.96, specificities from 0.81 to 0.96, PPVs from 0.31 to 0.76, and NPVs from 0.96 to 1.00. Finally, we found that different criteria were used among both prediction and surveillance studies to identify patients with infections, indicating that there is no uniform definition being used. Given the current use of different types of criteria and data used in prediction and surveillance studies, we were not able to identify or formulate a uniform and reliable method to automatically label patients with infections based on structured EHR data.

Prediction and surveillance of postoperative infections are crucial for early detection and assessment of the impact of preventative interventions but are currently hindered because the labeling of these cases is performed by resource-intensive manual chart review. In contrast to a previous study on semiautomated surveillance where high-risk patients were manually checked [13], we included only fully automated surveillance systems that were built to avoid requiring any human intervention. However, human intervention might still be required to incorporate the systems as well as to clean and

preprocess the EHR data. Furthermore, we broadened the scope by assessing current labeling methods for prediction modeling studies, which, with some exceptions, were based on manual labeling according to established guidelines. In line with our findings, the predominant use of manual labeling was reported in a meta-analysis on the predictive performance of machine learning algorithms for SSI prediction [71]. Although manual labeling based on chart review is still the predominant method and is considered the reference standard, it must be noted that manual labeling may be flawed due to human errors and interobserver variability [9,10,12]. Furthermore, validating models only on national registries and databases limits the generalizability of developed prediction models and surveillance systems to other settings [72].

We extensively researched different definitions and methods from prediction modeling studies, guidelines, and surveillance studies to identify patients with bacterial infections that may occur after surgery and summarized different types of data needed to adhere to the different definitions. This study has several limitations. First, heterogeneity between studies (eg, differences in study design) prevented a meta-analysis, making it difficult to draw generalizable conclusions on optimal labeling methods. However, combining different types of studies allowed us to generate insight into the current methods of labeling and identifying patients with infections. Second, the distinction between structured and unstructured data may differ according to hospital data set and region (eg, microbiology results can be registered as free-text or tabular data). Despite these limitations, we could identify a lack of uniform definitions for labeling of postoperative infections exists, and that manual labeling is currently the predominant method. Third, pre-existing infections could have impacted the performance of surveillance algorithms and prediction models as well as label reliability [73]. This could explain the relatively lower PPVs and warrants further research before reliable implementation of automated surveillance systems.

Different types of data were used among the definitions and methods, including structured tabular data, microbiological data, free-text data, and imaging results. The importance of reliable, high-quality outcome data is essential for the reliable use of artificial intelligence and surveillance systems [74]. Using structured EHR data is preferable, as free text is often subject to misinterpretation and contains personal patient-specific data that conflict with privacy legislation and thus have restrictions on data use [75]. By extracting free-text information, NLP shows promise in uncovering postoperative infections from free-text data. However, challenges remain with respect to generalizability [76], transparency, reliability, and potential biases, including concerns about accuracy or unintended errors [77,78]. Furthermore, NLP methods can be computationally expensive, depend on the quality of the input data, and are influenced by nuances in language, dialects, and medical jargon. Considering that NLP methods can vary significantly in complexity, ranging from simple string searches to advanced neural networks, future research should investigate whether increased complexity leads to improved surveillance accuracy. The use of microbiology results in definitions is prevalent, despite their occasional unreliability due to the possibility of false negatives or positives, causing under- or overreporting of infections [13], and heterogeneous storage practices. This reliance on microbiology results could lead to errors or inconsistencies in infection identification.

Accurately identifying patients with infections based on an automated analysis of EHR data remains a challenge, and validation is difficult owing to the limitations of manual chart review, which until now has remained the reference standard for postoperative infections and other relevant patient outcomes. Manual labeling based on manual EHR chart review is unfeasible when scaling artificial intelligence–based or statistical prediction models to more than one hospital, with 100,000 patient records each. In some of the included studies, alternative approaches were identified that relied on treatments and other structured data sources [27,31-33,41-43]. For future prediction model development and surveillance, alternative approaches to identifying patients with infection should be explored, such as focusing on pharmacological and interventional treatments performed by clinicians, as these approaches are often stored in a structured format in the EHR system [27]. Emphasis should be placed on the consensus on the definition and whether it is worse to miss infections that do not require treatment compared with those that do. Compared with sensitivity, specificity, and NPV, automated surveillance systems have a lower PPV where heterogeneity is observed between the different types of infections. The PPV to detect pneumonia and SSIs is lower compared with other types of infections. This could be due to variations in clinical presentation, differences in diagnostic criteria, or the inherent complexity and variability of these particular infections. A lower PPV in general could be due to the use of low classification cutoffs to not miss any cases, but it could also indicate that the reference standard manual labeling may have resulted in erroneous labels and that the systems found infections where the human annotator did not [79]. In addition to detecting individual patients with infections, automated surveillance systems hold promise for assessing hospital incidence rates, predicting rates of complications, and evaluating the effectiveness of quality improvement initiatives, where the emphasis may shift from high PPVs to broader statistical insights.

In conclusion, there is currently no evidence to support fully automated labeling and identification of patients with infections based solely on structured EHR data. This is due to the diverse definitions of postoperative infection and the need for unstructured data types, such as free text and clinical notes, which were required as data sources in nearly half of the instances to assess an infection. Furthermore, manual labeling was still the predominant method in prediction modeling studies. Fully automatic surveillance methods may result in overreporting due to a relatively low PPV and heavy reliance on free-text data. Future research must focus on defining uniform or globally accepted definitions of postoperative infection that use criteria that can be extracted from the EHR, as well as prioritizing the development of more scalable automated methods for infection detection using EHR data.

Acknowledgments

This research was partially funded by the Recovery Assistance for Cohesion and the Territories of Europe grant provided by the European Regional Development Fund (ERDF; grant KVW-00351).

Authors' Contributions

SLVDM, AMVB, MSA, BFG, RGHHN, EWS, MGJDB, and HVG contributed to the conception of the study. SLVDM, BFG, and MSA designed the study. JWS assisted in conducting the literature searches. SLVDM, BFG, AMVB, and MSA conducted the literature search and selection of the studies. SLVDM and MSA performed the data synthesis and data analyses. SLVDM wrote the initial draft of the paper. AMVB, MSA, BFG, RGHHN, EWS, MGJDB, and HVG reviewed and corrected the paper. All the authors read and approved the final paper.

Conflicts of Interest

BFG is currently the chief executive officer and majority shareholder of healthplus.ai BV and its subsidiaries. BFG has also consulted for and received research grants from Philips NV and Edwards Lifesciences LLC. SLVDM works as a data scientist and PhD candidate at Healthplus.ai and LUMC. SLVDM owns share options in Healthplus.ai. The rest of the authors declare no competing interests.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 449 KB - medinform_v12i1e57195_app1.docx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 102 KB - medinform_v12i1e57195_app2.pdf]

References

1. Clavien PA, Barkun J, de Oliveira ML, Vauthey JN, Dindo D, Schulick RD, et al. The Clavien-Dindo classification of surgical complications: five-year experience. Ann Surg 2009;250(2):187-196. [doi: 10.1097/SLA.0b013e3181b13ca2] [Medline: 19638912]
2. International Surgical Outcomes Study group. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. Br J Anaesth 2016;117(5):601-609 [FREE Full text] [doi: 10.1093/bja/aew316] [Medline: 27799174]
3. Gillespie BM, Harbeck E, Rattray M, Liang R, Walker R, Latimer S, et al. Worldwide incidence of surgical site infections in general surgical patients: a systematic review and meta-analysis of 488,594 patients. Int J Surg 2021;95:106136 [FREE Full text] [doi: 10.1016/j.ijsu.2021.106136] [Medline: 34655800]
4. Wan YI, Patel A, Achary C, Hewson R, Phull M, Pearse RM, International Surgical Outcomes Study (ISOS) Group. Postoperative infection and mortality following elective surgery in the international surgical outcomes study (ISOS). Br J Surg 2021;108(2):220-227. [doi: 10.1093/bjs/znaa075] [Medline: 33711143]
5. Weiser TG, Haynes AB, Molina G, Lipsitz SR, Esquivel MM, Uribe-Leitz T, et al. Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. Lancet 2015;385 Suppl 2:S11. [doi: 10.1016/S0140-6736(15)60806-6] [Medline: 26313057]
6. Ubbink DT, Visser A, Gouma DJ, Goslings JC. Registration of surgical adverse outcomes: a reliability study in a university hospital. BMJ Open 2012;2(3):e000891 [FREE Full text] [doi: 10.1136/bmjopen-2012-000891] [Medline: 22637372]
7. Veen EJ, Janssen-Heijnen MLG, Bosma E, de Jongh MAC, Roukema JA. The accuracy of complications documented in a prospective complication registry. J Surg Res 2012;173(1):54-59. [doi: 10.1016/j.jss.2010.08.042] [Medline: 20934713]
8. Du M, Xing Y, Suo J, Liu B, Jia N, Huo R, et al. Real-time automatic hospital-wide surveillance of nosocomial infections and outbreaks in a large Chinese tertiary hospital. BMC Med Inform Decis Mak 2014;14:9 [FREE Full text] [doi: 10.1186/1472-6947-14-9] [Medline: 24475790]
9. Brossette SE, Hacek DM, Gavin PJ, Kamdar MA, Gadbois KD, Fisher AG, et al. A laboratory-based, hospital-wide, electronic marker for nosocomial infection: the future of infection control surveillance? Am J Clin Pathol 2006;125(1):34-39. [Medline: 16482989]
10. Klompas M. Interobserver variability in ventilator-associated pneumonia surveillance. Am J Infect Control 2010;38(3):237-293. [doi: 10.1016/j.ajic.2009.10.003] [Medline: 20171757]
11. Tokars JI, Richards C, Andrus M, Klevens M, Curtis A, Horan T, et al. The changing face of surveillance for health care-associated infections. Clin Infect Dis 2004;39(9):1347-1352. [doi: 10.1086/425000] [Medline: 15494912]

12. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
13. Streefkerk HRA, Verkooijen RP, Bramer WM, Verbrugh HA. Electronically assisted surveillance systems of healthcare-associated infections: a systematic review. Euro Surveill 2020;25(2):1900321 [FREE Full text] [doi: 10.2807/1560-7917.ES.2020.25.2.1900321] [Medline: 31964462]
14. van der Meijden SL. Scoping Review Protocol - Automated Detection of Postoperative Infections to Allow Prediction and Surveillance Based on EHR Data - a Scoping Review. 2022. URL: https://osf.io/4cuge/ [accessed 2023-11-29]
15. Rhoads DD, Sintchenko V, Rauch CA, Pantanowitz L. Clinical microbiology informatics. Clin Microbiol Rev 2014;27(4):1025-1047 [FREE Full text] [doi: 10.1128/CMR.00049-14] [Medline: 25278581]
16. Prevention of hospital-acquired infections. World Health Organization. 2002. URL: https://iris.who.int/bitstream/handle/10665/67350/WHO_CDS_CSR_EPH_2002.12.pdf [accessed 2024-08-14]
17. CDC/NHSN surveillance definitions for specific types of infections. National Healthcare Safety Network. 2024. URL: https://www.cdc.gov/nhsn/pdfs/pscmanual/17pscnosinfdef_current.pdf [accessed 2024-01-01]
18. Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalianis H. Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. Health Informatics J 2018;24(1):24-42 [FREE Full text] [doi: 10.1177/1460458216656471] [Medline: 27496862]
19. Sakji S, Gicquel Q, Pereira S, Kergourlay I, Proux D, Darmoni S, et al. Evaluation of a french medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. Stud Health Technol Inform 2010;160(Pt 1):252-256. [Medline: 20841688]
20. Tvardik N, Kergourlay I, Bittar A, Segond F, Darmoni S, Metzger MH. Accuracy of using natural language processing methods for identifying healthcare-associated infections. Int J Med Inform 2018;117:96-102. [doi: 10.1016/j.ijmedinf.2018.06.002] [Medline: 30032970]
21. Pneumonia (Ventilator-associated [VAP] and non-ventilatorassociated pneumonia [PNEU]) event. National Healthcare Safety Network. 2024. URL: https://www.cdc.gov/nhsn/pdfs/pscmanual/6pscvapcurrent.pdf [accessed 2024-01-01]
22. Kinlin LM, Kirchner C, Zhang H, Daley J, Fisman DN. Derivation and validation of a clinical prediction rule for nosocomial pneumonia after coronary artery bypass graft surgery. Clin Infect Dis 2010;50(4):493-501. [doi: 10.1086/649925] [Medline: 20085462]
23. Blacky A, Mandl H, Adlassnig KP, Koller W. Fully automated surveillance of healthcare-associated infections with MONI-ICU: a breakthrough in clinical infection surveillance. Appl Clin Inform 2011;2(3):365-372 [FREE Full text] [doi: 10.4338/ACI-2011-03-RA-0022] [Medline: 23616883]
24. Bouzbid S, Gicquel Q, Gerbier S, Chomarat M, Pradat E, Fabry J, et al. Automated detection of nosocomial infections: evaluation of different strategies in an intensive care unit 2000-2006. J Hosp Infect 2011;79(1):38-43. [doi: 10.1016/j.jhin.2011.05.006] [Medline: 21742413]
25. Cato KD, Liu J, Cohen B, Larson E. Electronic surveillance of surgical site infections. Surg Infect (Larchmt) 2017;18(4):498-502 [FREE Full text] [doi: 10.1089/sur.2016.262] [Medline: 28402721]
26. FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Med Care 2013;51(6):509-516 [FREE Full text] [doi: 10.1097/MLR.0b013e31828d1210] [Medline: 23673394]
27. Colborn KL, Zhuang Y, Dyas AR, Henderson WG, Madsen HJ, Bronsert MR, et al. Development and validation of models for detection of postoperative infections using structured electronic health records data and machine learning. Surgery 2023;173(2):464-471 [FREE Full text] [doi: 10.1016/j.surg.2022.10.026] [Medline: 36470694]
28. Stern SE, Christensen MA, Nevers MR, Ying J, McKenna C, Munro S, et al. Electronic surveillance criteria for non-ventilator-associated hospital-acquired pneumonia: assessment of reliability and validity. Infect Control Hosp Epidemiol 2023:1-7. [doi: 10.1017/ice.2022.302] [Medline: 36920040]
29. Surgical site infection event (SSI). National Healthcare Safety Network. 2024. URL: https://www.cdc.gov/nhsn/pdfs/pscmanual/9pscssicurrent.pdf [accessed 2024-01-01]
30. Global guidelines for the prevention of surgical site infection. World Health Organization. 2018. URL: https://www.who.int/publications/i/item/9789241550475 [accessed 2023-12-01]
31. Daneman N, Simor AE, Redelmeier DA. Validation of a modified version of the national nosocomial infections surveillance system risk index for health services research. Infect Control Hosp Epidemiol 2009;30(6):563-569. [doi: 10.1086/597523] [Medline: 19415966]
32. Weller GB, Lovely J, Larson DW, Earnshaw BA, Huebner M. Leveraging electronic health records for predictive modeling of post-surgical complications. Stat Methods Med Res 2018;27(11):3271-3285. [doi: 10.1177/0962280217696115] [Medline: 29298612]
33. Crispin A, Klinger C, Rieger A, Strahwald B, Lehmann K, Buhr HJ, et al. The DGAV risk calculator: development and validation of statistical models for a web-based instrument predicting complications of colorectal cancer surgery. Int J Colorectal Dis 2017;32(10):1385-1397. [doi: 10.1007/s00384-017-2869-6] [Medline: 28799112]

34. Martin S, Turner E, Nguyen A, Thornton B, Nazerali RS. An evaluation of the utility of the Breast Reconstruction Risk Assessment score risk model in prepectoral tissue expander breast reconstruction. Ann Plast Surg 2020;84(5S Suppl 4):S318-S322. [doi: 10.1097/SAP.0000000000002320] [Medline: 32187065]
35. Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. Stud Health Technol Inform 2013;192:572-575. [Medline: 23920620]
36. Leclère B, Lasserre C, Bourigault C, Juvin ME, Chaillet MP, Mauduit N, et al. Matching bacteriological and medico-administrative databases is efficient for a computer-enhanced surveillance of surgical site infections: retrospective analysis of 4,400 surgical procedures in a French university hospital. Infect Control Hosp Epidemiol 2014;35(11):1330-1335. [doi: 10.1086/678422] [Medline: 25333426]
37. Leth RA, Nørgaard M, Uldbjerg N, Thomsen RW, Møller JK. Surveillance of selected post-caesarean infections based on electronic registries: validation study including post-discharge infections. J Hosp Infect 2010;75(3):200-204. [doi: 10.1016/j.jhin.2009.11.018] [Medline: 20381909]
38. Suzuki H, Clore GS, Perencevich EN, Hockett-Sherlock SM, Goto M, Nair R, et al. Development of a fully automated surgical site infection detection algorithm for use in cardiac and orthopedic surgery research. Infect Control Hosp Epidemiol 2021;42(10):1215-1220 [FREE Full text] [doi: 10.1017/ice.2020.1387] [Medline: 33618788]
39. Thirukumaran CP, Zaman A, Rubery PT, Calabria C, Li Y, Ricciardi BF, et al. Natural language processing for the identification of surgical site infections in orthopaedics. J Bone Joint Surg Am 2019;101(24):2167-2174 [FREE Full text] [doi: 10.2106/JBJS.19.00661] [Medline: 31596819]
40. Rahbari NN, Weitz J, Hohenberger W, Heald RJ, Moran B, Ulrich A, et al. Definition and grading of anastomotic leakage following anterior resection of the rectum: a proposal by the international study group of rectal cancer. Surgery 2010;147(3):339-351. [doi: 10.1016/j.surg.2009.10.012] [Medline: 20004450]
41. Stidham RW, Waljee AK, Day NM, Bergmans CL, Zahn KM, Higgins PDR, et al. Body fat composition assessment using analytic morphomics predicts infectious complications after bowel resection in Crohn's disease. Inflamm Bowel Dis 2015;21(6):1306-1313 [FREE Full text] [doi: 10.1097/MIB.0000000000000360] [Medline: 25822011]
42. Miyakita H, Sadahiro S, Saito G, Okada K, Tanaka A, Suzuki T. Risk scores as useful predictors of perioperative complications in patients with rectal cancer who received radical surgery. Int J Clin Oncol 2017;22(2):324-331. [doi: 10.1007/s10147-016-1054-1] [Medline: 27783239]
43. McKenna NP, Bews KA, Cima RR, Crowson CS, Habermann EB. Development of a risk score to predict anastomotic leak after left-sided colectomy: which patients warrant diversion? J Gastrointest Surg 2020;24(1):132-143 [FREE Full text] [doi: 10.1007/s11605-019-04293-y] [Medline: 31250368]
44. Nudel J, Bishara AM, de Geus SWL, Patil P, Srinivasan J, Hess DT, et al. Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database. Surg Endosc 2021;35(1):182-191 [FREE Full text] [doi: 10.1007/s00464-020-07378-x] [Medline: 31953733]
45. Kawai K, Hirakawa S, Tachimori H, Oshikiri T, Miyata H, Kakeji Y, et al. Updating the predictive models for mortality and morbidity after low anterior resection based on the National Clinical Database. Dig Surg 2023;40(3-4):130-142. [doi: 10.1159/000531370] [Medline: 37311436]
46. Lin V, Tsouchnika A, Allakhverdiiev E, Rosen AW, Gögenur M, Clausen JSR, et al. Training prediction models for individual risk assessment of postoperative complications after surgery for colorectal cancer. Tech Coloproctol 2022;26(8):665-675. [doi: 10.1007/s10151-022-02624-x] [Medline: 35593971]
47. Shi J, Wu Z, Wu X, Shan F, Zhang Y, Ying X, et al. Early diagnosis of anastomotic leakage after colorectal cancer surgery using an inflammatory factors-based score system. BJS Open 2022;6(3):zrac069 [FREE Full text] [doi: 10.1093/bjsopen/zrac069] [Medline: 35657137]
48. van Kooten RT, Bahadoer RR, Ter Buurkes de Vries B, Wouters MWJM, Tollenaar RAEM, Hartgrink HH, et al. Conventional regression analysis and machine learning in prediction of anastomotic leakage and pulmonary complications after esophagogastric cancer surgery. J Surg Oncol 2022;126(3):490-501 [FREE Full text] [doi: 10.1002/jso.26910] [Medline: 35503455]
49. Urinary Tract Infection (Catheter-Associated Urinary Tract Infection [CAUTI] and Non-Catheter-Associated Urinary Tract Infection [UTI]) Events. National Healthcare Safety Network. 2024. URL: https://www.cdc.gov/nhsn/pdfs/pscmanual/7psccauticurrent.pdf [accessed 2024-01-01]
50. Cheng X, Liu Y, Wang W, Yan J, Lei X, Wu H, et al. Preoperative risk factor analysis and dynamic online nomogram development for early infections following primary hip arthroplasty in geriatric patients with hip fracture. Clin Interv Aging 2022;17:1873-1883 [FREE Full text] [doi: 10.2147/CIA.S392393] [Medline: 36575659]
51. Bouam S, Girou E, Brun-Buisson C, Karadimas H, Lepage E. An intranet-based automated system for the surveillance of nosocomial infections: prospective validation compared with physicians' self-reports. Infect Control Hosp Epidemiol 2003;24(1):51-55. [doi: 10.1086/502115] [Medline: 12558236]
52. Branch-Elliman W, Strymish J, Kudesia V, Rosen AK, Gupta K. Natural language processing for real-time catheter-associated urinary tract infection surveillance: results of a pilot implementation trial. Infect Control Hosp Epidemiol 2015;36(9):1004-1010. [doi: 10.1017/ice.2015.122] [Medline: 26022228]

53. Choudhuri JA, Pergamit RF, Chan JD, Schreuder AB, McNamara E, Lynch JB, et al. An electronic catheter-associated urinary tract infection surveillance tool. Infect Control Hosp Epidemiol 2011;32(8):757-762. [doi: 10.1086/661103] [Medline: 21768758]
54. Redder JD, Leth RA, Møller JK. Incidence rates of hospital-acquired urinary tract and bloodstream infections generated by automated compilation of electronically available healthcare data. J Hosp Infect 2015;91(3):231-236. [doi: 10.1016/j.jhin.2015.05.011] [Medline: 26162918]
55. van der Werff SD, Thiman E, Tanushi H, Valik JK, Henriksson A, Ul Alam MU, et al. The accuracy of fully automated algorithms for surveillance of healthcare-associated urinary tract infections in hospitalized patients. J Hosp Infect 2021;110:139-147 [FREE Full text] [doi: 10.1016/j.jhin.2021.01.023] [Medline: 33548370]
56. Venable A, Dissanaike S. Is automated electronic surveillance for healthcare-associated infections accurate in the burn unit? J Burn Care Res 2013;34(6):591-597. [doi: 10.1097/BCR.0b013e3182a2aa0f] [Medline: 24121803]
57. Wald HL, Bandle B, Richard A, Min S. Accuracy of electronic surveillance of catheter-associated urinary tract infection at an academic medical center. Infect Control Hosp Epidemiol 2014;35(6):685-691. [doi: 10.1086/676429] [Medline: 24799645]
58. Moore LJ, Moore FA, Todd SR, Jones SL, Turner KL, Bass BL. Sepsis in general surgery: the 2005-2007 National Surgical Quality Improvement Program perspective. Arch Surg 2010;145(7):695-700. [doi: 10.1001/archsurg.2010.107] [Medline: 20644134]
59. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
60. Leal J, Gregson DB, Ross T, Flemons WW, Church DL, Laupland KB. Development of a novel electronic surveillance system for monitoring of bloodstream infections. Infect Control Hosp Epidemiol 2010;31(7):740-747. [doi: 10.1086/653207] [Medline: 20470039]
61. Leal JR, Gregson DB, Church DL, Henderson EA, Ross T, Laupland KB. The validation of a novel surveillance system for monitoring bloodstream infections in the Calgary Zone. Can J Infect Dis Med Microbiol 2016;2016:2935870 [FREE Full text] [doi: 10.1155/2016/2935870] [Medline: 27375749]
62. Lin MY, Woeltje KF, Khan YM, Hota B, Doherty JA, Borlawsky TB, Centers for Disease ControlPrevention Epicenter Program. Multicenter evaluation of computer automated versus traditional surveillance of hospital-acquired bloodstream infections. Infect Control Hosp Epidemiol 2014;35(12):1483-1490 [FREE Full text] [doi: 10.1086/678602] [Medline: 25419770]
63. Valik JK, Ward L, Tanushi H, Müllersdorf K, Ternhag A, Aufwerber E, et al. Validation of automated sepsis surveillance based on the Sepsis-3 clinical criteria against physician record review in a general hospital population: observational study using electronic health records data. BMJ Qual Saf 2020;29(9):735-745 [FREE Full text] [doi: 10.1136/bmjqs-2019-010123] [Medline: 32029574]
64. Woeltje KF, McMullen KM, Butler AM, Goris AJ, Doherty JA. Electronic surveillance for healthcare-associated central line-associated bloodstream infections outside the intensive care unit. Infect Control Hosp Epidemiol 2011;32(11):1086-1090. [doi: 10.1086/662181] [Medline: 22011535]
65. Dubberke ER, Nyazee HA, Yokoe DS, Mayer J, Stevenson KB, Mangino JE, et al. Implementing automated surveillance for tracking Clostridium difficile infection at multiple healthcare facilities. Infect Control Hosp Epidemiol 2012;33(3):305-308 [FREE Full text] [doi: 10.1086/664052] [Medline: 22314071]
66. van der Werff SD, Fritzing M, Tanushi H, Henriksson A, Dalianis H, Ternhag A, et al. The accuracy of fully automated algorithms for surveillance of healthcare-onset infections in hospitalized patients. Antimicrob Steward Healthc Epidemiol 2022;2(1):e43 [FREE Full text] [doi: 10.1017/ash.2022.32] [Medline: 36310782]
67. van Mourik MSM, Groenwold RHH, Berkelbach van der Sprenkel JW, van Solinge WW, Troelstra A, Bonten MJM. Automated detection of external ventricular and lumbar drain-related meningitis using laboratory and microbiology results and medication data. PLoS One 2011;6(8):e22846 [FREE Full text] [doi: 10.1371/journal.pone.0022846] [Medline: 21829659]
68. Peterson KE, Hacek DM, Robicsek A, Thomson Jr RB, Peterson LR. Electronic surveillance for infectious disease trend analysis following a quality improvement intervention. Infect Control Hosp Epidemiol 2012;33(8):790-795. [doi: 10.1086/666625] [Medline: 22759546]
69. Fu S, Wyles CC, Osmon DR, Carvour ML, Sagheb E, Ramazanian T, et al. Automated detection of periprosthetic joint infections and data elements using natural language processing. J Arthroplasty 2021;36(2):688-692 [FREE Full text] [doi: 10.1016/j.arth.2020.07.076] [Medline: 32854996]
70. Cheng L, Bai W, Song P, Zhou L, Li Z, Gao L, et al. Development and validation of a nomograph model for post-operative central nervous system infection after craniocerebral surgery. Diagnostics (Basel) 2023;13(13):2207 [FREE Full text] [doi: 10.3390/diagnostics13132207] [Medline: 37443601]
71. Wu G, Khair S, Yang F, Cheligeer C, Southern D, Zhang Z, et al. Performance of machine learning algorithms for surgical site infection case detection and prediction: a systematic review and meta-analysis. Ann Med Surg (Lond) 2022;84:104956 [FREE Full text] [doi: 10.1016/j.amsu.2022.104956] [Medline: 36582918]

72. de Hond AAH, Shah VB, Kant IMJ, Van Calster B, Steyerberg EW, Hernandez-Boussard T. Perspectives on validation of clinical predictive algorithms. NPJ Digit Med 2023;6(1):86 [FREE Full text] [doi: 10.1038/s41746-023-00832-9] [Medline: 37149704]
73. van Rooden SM, Tacconelli E, Pujol M, Gomila A, Kluytmans JAJW, Romme J, et al. A framework to develop semiautomated surveillance of surgical site infections: an international multicenter study. Infect Control Hosp Epidemiol 2020;41(2):194-201 [FREE Full text] [doi: 10.1017/ice.2019.321] [Medline: 31884977]
74. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
75. Humbert-Droz M, Mukherjee P, Gevaert O. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: case study for the extraction of symptoms from clinical notes. JMIR Med Inform 2022;10(3):e32903 [FREE Full text] [doi: 10.2196/32903] [Medline: 35285805]
76. Khambete MP, Su W, Garcia JC, Badgeley MA. Quantification of BERT diagnosis generalizability across medical specialties using semantic dataset distance. AMIA Jt Summits Transl Sci Proc 2021;2021:345-354 [FREE Full text] [Medline: 34457149]
77. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
78. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
79. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. J Educ Eval Health Prof 2013;10:12 [FREE Full text] [doi: 10.3352/jeehp.2013.10.12] [Medline: 24324853]

Abbreviations

ACS NSQIP: American College of Surgeons National Surgical Quality Improvement Program
EHR: electronic health record
HAI: hospital-acquired Infections
NNIS: National Nosocomial Infections Surveillance System
NLP: natural language processing
NPV: negative predictive value
PPV: positive predictive value
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SSI: surgical site infection
SURPAS: Surgical Risk Preoperative Assessment System
UTI: urinary tract infection

Edited by J Klann; submitted 07.02.24; peer-reviewed by O Endrich, G Laynor, K Weber, K Colborn; comments to author 28.06.24; revised version received 12.07.24; accepted 16.07.24; published 10.09.24.

Please cite as:
van der Meijden SL, van Boekel AM, van Goor H, Nelissen RGHH, Schoones JW, Steyerberg EW, Geerts BF, de Boer MGJ, Arbous MS
Automated Identification of Postoperative Infections to Allow Prediction and Surveillance Based on Electronic Health Record Data: Scoping Review
JMIR Med Inform 2024;12:e57195
URL: https://medinform.jmir.org/2024/1/e57195
doi:10.2196/57195
PMID:39255011

Review

State-of-the-Art Fast Healthcare Interoperability Resources (FHIR)–Based Data Model and Structure Implementations: Systematic Scoping Review

Parinaz Tabari[1], MSc; Gennaro Costagliola[1], Prof Dr; Mattia De Rosa[1], PhD; Martin Boeker[2], Prof Dr

[1]Department of Informatics, University of Salerno, Fisciano, Italy

[2]Institute for Artificial Intelligence and Informatics in Medicine, Medical Center rechts der Isar, School of Medicine and Health, Technical University of Munich, Munich, Germany

Corresponding Author:
Parinaz Tabari, MSc
Department of Informatics
University of Salerno
Via Giovanni Paolo II, 132
Fisciano, 84084
Italy
Phone: 39 089 963319
Email: ptabari@unisa.it

Abstract

Background: Data models are crucial for clinical research as they enable researchers to fully use the vast amount of clinical data stored in medical systems. Standardized data and well-defined relationships between data points are necessary to guarantee semantic interoperability. Using the Fast Healthcare Interoperability Resources (FHIR) standard for clinical data representation would be a practical methodology to enhance and accelerate interoperability and data availability for research.

Objective: This research aims to provide a comprehensive overview of the state-of-the-art and current landscape in FHIR-based data models and structures. In addition, we intend to identify and discuss the tools, resources, limitations, and other critical aspects mentioned in the selected research papers.

Methods: To ensure the extraction of reliable results, we followed the instructions of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. We analyzed the indexed articles in PubMed, Scopus, Web of Science, IEEE Xplore, the ACM Digital Library, and Google Scholar. After identifying, extracting, and assessing the quality and relevance of the articles, we synthesized the extracted data to identify common patterns, themes, and variations in the use of FHIR-based data models and structures across different studies.

Results: On the basis of the reviewed articles, we could identify 2 main themes: dynamic (pipeline-based) and static data models. The articles were also categorized into health care use cases, including chronic diseases, COVID-19 and infectious diseases, cancer research, acute or intensive care, random and general medical notes, and other conditions. Furthermore, we summarized the important or common tools and approaches of the selected papers. These items included FHIR-based tools and frameworks, machine learning approaches, and data storage and security. The most common resource was “Observation” followed by “Condition” and “Patient.” The limitations and challenges of developing data models were categorized based on the issues of data integration, interoperability, standardization, performance, and scalability or generalizability.

Conclusions: FHIR serves as a highly promising interoperability standard for developing real-world health care apps. The implementation of FHIR modeling for electronic health record data facilitates the integration, transmission, and analysis of data while also advancing translational research and phenotyping. Generally, FHIR-based exports of local data repositories improve data interoperability for systems and data warehouses across different settings. However, ongoing efforts to address existing limitations and challenges are essential for the successful implementation and integration of FHIR data models.

(JMIR Med Inform 2024;12:e58445) doi:10.2196/58445

KEYWORDS

data model; Fast Healthcare Interoperability Resources; FHIR; interoperability; modeling; PRISMA

Introduction

Background

In informatics, operations and data structures can be described by a set of concepts called data models. Because structures and data points need to be connected to represent connections, data modeling offers a visual representation of the system, in a whole or in some parts. For instance, one of the most used conceptual data models is the entity relationship model which is generally linked to a relational database [1]. Data modeling is a process that defines how the data should be maintained in a database. Data types, constraints, relationships, and metadata definitions are among the features specified by a data model [2]. Data models are also crucial for clinical research as they enable researchers to fully use the vast amount of clinical data stored in medical systems. Standardized data and well-defined relationships between data points are necessary to "guarantee reproducible research findings" [3].

Furthermore, data modeling can facilitate interoperability between medical systems. Interoperability refers to the ability to exchange information between computer systems, which is essential in various fields, such as artificial intelligence (AI), big data research and analytics, medical communication, and multinational collaboration. In the medical field, interoperable systems can reduce errors and documentation workload, empower patients, and facilitate information retrieval. In research, real-world information can be collected and used for data mining and AI to generate new hypotheses [4]. The management board of the Healthcare Information and Management Systems Society (HIMSS) defined 3 levels of interoperability: fundamental, structural, and semantic. Fundamental interoperability refers to the communication method between IT firms and devices, while structural interoperability is the format and structure of data being communicated. Semantic interoperability, by contrast, involves the ability of disparate and heterogeneous systems to not only exchange information but also interpret and use it autonomously [5]. Developing a data model would enhance structural and semantic interoperability between medical information systems. Furthermore, efficient data exchange contributes to the reduction of time and financial resources [6].

Health Level 7 (HL7) is a standard-developing organization focused on enhancing information exchange among health care systems. These standards are fundamental in the adoption of electronic health records (EHRs). Fast Healthcare Interoperability Resources (FHIR) is the most recent interoperability standard, preceded by HL7 version 2 and HL7 version 3 [7]. FHIR aims to advance messaging standards to enhance semantic interoperability [8]. Using this standard for clinical data representation is a practical methodology to enhance and accelerate data availability for research. These models can also have the potential to be transformed into other models for analytics purposes [9]. FHIR mapping is the process of identifying the corresponding FHIR resources to real-world data elements. This is an essential step in the FHIR data modeling procedure [10]. When the objective is to maintain semantic interoperability with legacy applications, performing manual data transformations and mappings is necessary to guarantee that the exchanged data are interpreted properly and as expected by all end points [8].

Because not all health care information is structured, there is a need to use other approaches for mapping and FHIR modeling. Natural language processing (NLP) is a branch of AI that deals with the computerized interpretation, representation, and analysis of natural (human) language. In the health care domain, this technology is widely used to interpret and analyze unstructured health data, such as diagnostic reports, medical notes, and prescriptions [11]. The extracted information can then be represented in a structured format, such as a FHIR-based model. In general, it is possible to formalize and integrate unstructured and structured EHR data through a FHIR-based framework [12].

FHIR-based data normalization pipelines are valuable tools in data capture and EHR phenotyping [13]. For instance, a pipeline called NLP2FHIR standardizes unstructured EHR data [14]. Concerning the big data domain, workflows of data harmonization pipelines integrated with FHIR would present a scalable data modeling of large data sets [15]. It is also feasible to use FHIR data models to standardize heterogeneous annotation corpora [16]. All the mentioned potentials will lead to better semantic interoperability between medical systems. To the best of our knowledge, no research has been done so far to comprehensively assess the practical implementations of FHIR-based models and infrastructures. Thus, in this research, we aim to review recent advancements in this field, focusing on the functional data model or structure implementations using this standard. More specifically, this scoping review focused on addressing the question, "What insights can be gained from analyzing the state-of-the-art FHIR-based data modeling approaches considering technological advancements, application in the medical domain, and potential limitations?"

Objectives

The research objectives are as follows: (1) to provide a comprehensive overview of FHIR-based data models in the context of interoperability, structure, and functionality and summarize the state of the art for developing FHIR-based data models and (2) to highlight limitations, challenges, advantages, and opportunities brought about by FHIR-based data models

Methods

Overview

This review was conducted according to the instructions of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [17]. This checklist aims to facilitate the development of a deeper comprehension of pertinent terminology, fundamental concepts, and essential items to report for scoping reviews [17]. The checklist is available in Multimedia Appendix 1.

Study Protocol

We used the PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols) 2015 checklist

to formulate and draft the review protocol. Protocols for systematic reviews facilitate the organization and recording of review procedures, ensuring the reproducibility of research. In addition, they serve as a safeguard against indecisive judgment during the review process and allow the readers to determine whether selective reporting has been applied [18]. The full checklist and the review protocol are available in Multimedia Appendix 2.

Eligibility Criteria

To select the papers, we considered the articles that encompass the FHIR standard in the data model development or infrastructure design. The inclusion and exclusion criteria were defined in more detail in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Original articles and case studies from journals and conferences
- Articles related to the Fast Healthcare Interoperability Resources (FHIR)-based data models and structures focusing on a health care condition or using real-world patient data, registries, or data sets
- Articles with high-quality and detailed workflow processes with at least one architecture or data model diagram
- Articles that discuss the barriers, challenges, or limitations of developing FHIR-based data models and infrastructures in a health care domain

Exclusion criteria

- Not written in English
- Not accessible
- Letter to the editors, reviews, editorials, commentary articles, short papers without detailed implementation information, posters, and preprint articles
- Not relevant to research questions and objectives; in other words, articles not focusing on FHIR-based data model development or not providing practical and detailed insights into the development or use of FHIR-based data models by a schematic approach
- Papers lacking specific use cases or real-world data sources (practical implementations) or without discussion of limitations and challenges

Information Sources and Search Strategy

We searched academic databases, such as PubMed, Scopus, Web of Science (standard selection of databases—Web of Science Core Collection), IEEE Xplore, and the ACM Digital Library in May 2023.

The search was conducted using database-specific variants of the basic search term ([“fhir”] AND [“data model” OR “modelling” OR “minimum data set” OR “data element”]) with their synonyms, variations, and full forms.

It is worth mentioning that no time limit was applied to the search to obtain a comprehensive overview of all published articles in this field. We should clarify that the initial pages of Google Scholar (9-10 pages) were investigated as a supplement to the mentioned academic libraries to retrieve additional papers. Full searches are available in Multimedia Appendix 3.

Study Selection

In a stepwise process, 2 coauthors (PT and MDR) independently screened the retrieved articles and selected the initial studies by applying the inclusion and exclusion criteria to the titles or abstracts or, in some cases, full texts (by rapid skimming). Inconsistencies in the selection were discussed with other coauthors until a consensus was reached. EndNote (Endnote X9; Bld 12062) software was used for article screening and investigation in each step. The full texts of the initially selected articles were assessed in the next phase to check compliance with the eligibility criteria. PT thoroughly reviewed the articles and then discussed with other authors about inclusions. Disagreements were resolved after group discussions.

Each selected study was thoroughly investigated for the appropriateness and clarity of the research methodology and design. We also assessed them to ensure alignment with the study objectives. The rigor of the methods, tools, and techniques used for FHIR-based architectural design was considered in this phase. The presentation of results and the coherence of model interpretation were also closely examined.

Data Charting Process and Data Items

Two coauthors (PT and MDR) extracted and analyzed the selected articles and charted the data. The final analysis was thoroughly reviewed and confirmed by other coauthors to ensure reliability and rigor. The collaborative review process among coauthors further enhanced the robustness of the results’ interpretation, ensuring a comprehensive and well-rounded analysis of the gathered evidence. The following information was extracted and collected in a spreadsheet: (1) bibliographic information, such as title, authors, and year of publication; (2) data sources; (3) FHIR resources; (4) data transformation and mapping; (5) standards, tools, terminologies, and models; (6) data validation and evaluation; (7) use case.

Synthesis of the Results

After extraction, we assessed the information to find themes or categories. Subsequently, we performed a general analysis of the papers, based on the overall technical themes and the medical domains. In addition, any important technologies used most in the included articles were comprehensively presented and

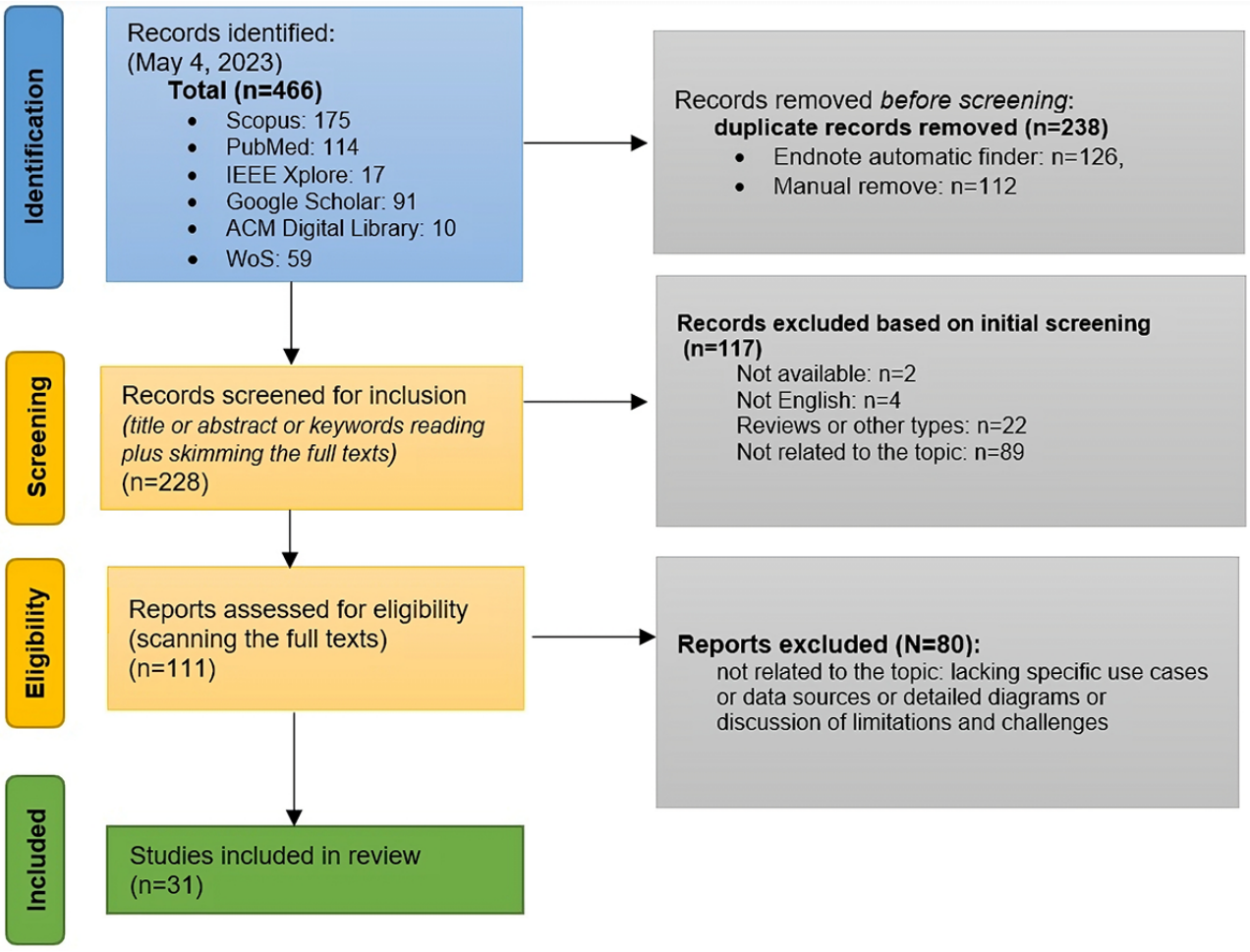
discussed afterward. Resource frequency analysis was performed via the investigation and counting of FHIR resources used in each data model and infrastructure to find out which resources were more common in system developments. One of the most important aims of our research was to extract and categorize the implementation limitations mentioned by the researchers. Therefore, these aspects were also addressed subsequently to provide a thorough viewpoint of challenges that future scientists may face.

Results

Selection of Sources of Evidence

Of the overall 466 articles found during the comprehensive search, 238 (51.1%) studies were duplicates. Of the remaining 228 articles, 117 (51.3%) were excluded based on reading titles or abstracts or skimming some full texts. Of the remaining 111 articles for the next phase (full-text assessment), 31 (27.9%) articles were eventually selected to be included in this review. Figure 1 illustrates the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) chart of this study.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection. WoS: Web of Science.



Structural Categorization

After analyzing the full texts of the 31 articles, we categorized them based on 2 models: dynamic (pipeline-based) and static data models.

Dynamic Data Models

Data pipelines are chains of functions and activities that lead the input to the output in an attempt for the flow of data to be smooth and automated from source to destination [19]. Dynamic or pipeline-based data models deal with moving, transforming, and analyzing the data using the FHIR standard in their approach. In this category, the FHIR standard has been used as a canonical data model to develop dynamic models. This group encompasses the articles with processes that go beyond static representation and include the movement and transformation of data. Of the 31 articles included, 25 (81%) were related to the development of dynamic data models using the FHIR standard. Table 1 summarizes the extracted information about this category.

Table 1. Dynamic models.

Study	FHIR[a] resources	Data source	Data transformation and mapping	Standards, tools, terminologies, and models	Validation and evaluation	Use case
Lenert et al [9]	Patient, Encounter, Condition, Procedure, MedicationRequest, MedicationAdministration, and Observation	Epic EHR[b], large academic institution	Flat file to FHIR, FHIR to OMOP[c], FHIR to PCORnet[d]	HL7[e] version 2.X, OMOP, PCORnet, Flat file, FHIR CDR[f], and CDM[g]	By published quality assessment tools	Health care research, especially in the context of COVID-19
Wen et al [20]	Composition and ValueSet	i2b2[h] Obesity Challenge data set, MIMIC[i] III obesity discharge summaries	FHIR-based NLP[j] extensions to CQL[k], FHIR extensions to NLP2FHIR pipeline	NLP2FHIR pipeline, CQL, NLP engines (cTAKES[l], MEDXN[m], MedTime) and PheKB[n]	Phenotype algorithms in PheKB and obesity phenotyping algorithm plus 2 obesity data sets	Obesity
Hong et al [13]	Composition, Condition, MedicationStatement, Procedure and FamilyMemberHistory	i2b2 obesity challenge (discharge summaries)	EHR data to FHIR resources	NLP tools (cTAKES, MedXN, MedTime) and NLP2FHIR pipeline, machine learning algorithms (logistic regression, support vector machine, decision tree, and random forest)	Using MIMIC-III obesity data set as a second data set, evaluation measures (precision, recall, and F_1-score) for performance evaluation	Obesity
Hong et al [14]	Composition, Condition, Observation, Procedure, MedicationStatement, Medication and FamilyMemberHistory	Mayo Clinic's unstructured EHR data	Unstructured and structured EHR data to FHIR resources	NLP2FHIR pipeline, UIMA[o] clinical NLP tools (cTAKES, MedXN, MedTime), LOINC[p], SNOMED CT[q], RxNorm[r], and ATC[s]	Reusing annotation corpora, standardizing annotation corpora, NLP2FHIR performance evaluation by precision, recall, and F_1-score	Random notes from EHR
Zong et al [21]	Questionnaire and QuestionnaireResponse	Mayo Clinic patients with colorectal cancer and ACP[t]	Unstructured reports to structured reports and synoptic report to ACP, and ACP FHIR model to CRF[u]	NLP tools and UDP[v] data sources	Precision, recall, F_1-score, and accuracy	Colorectal cancer
Hong et al [12]	MedicationStatement	Medication data from Mayo Clinic's EHR	Unstructured EHR data to FHIR, structured data to FHIR resource, and FHIR resources to annotation schemas	NLP tools (cTAKES, MedXN, MedTime), rule-based approach, SNOMED CT, CAS[w], RxNorm, UIMA, and protégé	Precision, recall, and F_1-score	Random notes from EHR
Williams et al [15]	Patient, Encounter, Observation, Procedure, MedicationRequest, MedicationAdministration, and Condition	MIMIC-IV database for validation	Raw hospital records to AI[x]-friendly and harmonized representation, and database tables to FHIR standard	ETL[y] framework and Postgres	Openly available MIMIC-IV database to test FHIR-DHP[z] and syntactic validation of FHIR mapping	Intensive care
Fischer et al [22]	Patient, Encounter, and Observation	German Pulmonary Hypertension registry	CSV file to FHIR bundle collection, source filed names to standard terminology (SNOMED CT, LOINC, ATC, and *ICD*[aa]*-10*) and source data to OMOP schema	ETL process, XSLT[ab], XPath, OMOP CDM, SNOMED CT, LOINC, ATC, and *ICD-10*	Feasibility assessment by computation time and source data coverage in the target CDM	Pulmonary Hypertension registry
Pfaff et al [23]	Patient, Encounter, Condition, Procedure, Observation, MedicationRequest, and Practitioner	i2b2	CDM to FHIR	CDM and FHIR PIT[ac]	Comparison of generated data by the pipeline and equivalent clinical data of CDWH[ad] warehouse	Asthma

Study	FHIR[a] resources	Data source	Data transformation and mapping	Standards, tools, terminologies, and models	Validation and evaluation	Use case
Rosenau et al [24]	Condition, Observation, Procedure, MedicationStatement, Immunization, DiagnosticReport, and Specimen	GECCO[ae]	Clinical data to FHIR, structured query to FHIR search, and CQL requests	SNOMED CT, LOINC, *ICD-10*-GM[af], ATC, CQL, and ETL processes	Create test patients and automated and manual test	COVID-19
Zong et al [25]	Observation, Condition, Medication, FamilyMemberHistory, and Patient	Mayo Clinic clinical data warehouse	Clinical entries to FHIR resources, FHIR to RDF[ag]	RDF, classification, machine learning and deep learning, cTAKES, MedXN, MedTime, NLP2FHIR, bag of features, Node2vec, *ICD-9*, RxNorm, and LOINC	Conventional 10-fold cross-validation, AUROC[ah], and AUPRC[ai]	Cancer
Bennett et al [26]	CodeSystem, ValueSet, MedicationRequest, MedicationDispense, and MedicationAdministration	MIMIC-IV	MIMIC-IV to FHIR	FSH[aj], py mimic FHIR package, PostgreSQL, and SNOMED CT	Validation by open-source FHIR server (HAPI[ak] FHIR) by bundles	Intensive care (ED[al] data)
El-Sappagh et al [27]	Patient, Practitioner, RelatedPerson, Observation, Condition, AdverseEvent, AllergyIntolerance, Location, FamilyMemberHistory, CarePlan, Goal, NutritionOrder, Medication, MedicationRequest, MedicationStatement, Device, Encounter, EpisodeOfCare, CareTeam, and Procedure	WBAN[am], patient profiles in EHR, and manual data sent by patients	RDB[an] to FHIR, FHIR to RDB, EHR data to FHIR, and direct mapping of historical data to FASTO[ao] ontology	FASTO ontology (using FHIR, SSN[ap], BFO[aq], and CPG[ar]), OWL[as] 2, WBAN, CDSS[at], Protégé, PHR[au], ISO IEEE 11073, LOINC, SNOMED CT, UoM[av], FHRBase database, FHIR RESTful[aw], OAuth2[ax], SPARQL[ay], D2RQ platform, Jena API[az], and Pallet and HermiT reasoners	Ontology is evaluated (assessment of correctness, consistency, and completeness of ontology knowledge) and manual evaluation by experts	Type 1 diabetes mellitus
Zong et al [28]	DiagnosticReport and Observation	ACP and clinical records of Mayo Clinic's patients	Structured and unstructured data to FHIR-based data profile and directly-inherited data element mapping	CRF, DMM[ba] model, ETL process, and topic modeling	Precision, recall, and F_1-score	Cancer clinical trials-colorectal
Hong et al [29]	Patient, Observation, Condition, and Procedure	Ovarian cancer database, laboratory test database, and CDM database	Local code to standard code, laboratory test codes to LOINC codes, and mapping between local identifiers and FHIR resource identifiers	Shiny web framework, Shiny apps library, R packages for FHIR data visualization, HAPI FHIR API, LOINC, ICD, and CPT[bb]	Feasibility and adaptability test using public FHIR servers	Ovarian cancer
Hong et al [16]	Condition, FamilyMemberHistory, Procedure, Observation, MedicationStatement, and Medication	Three annotated corpora from SHARPn project, MedXN project, and active Mayo's clinical NLP project (Family History NLP Project)	Source annotation schemas and FHIR annotation schema	UMLS[bc], SNOMED CT, LOINC, RxNorm, NLP tasks, support vector machine, annotation tools (Knowtator and Anafora), Protégé ontology editor, and HAPI FHIR API	Evaluation with annotation corpora, calculated precision, recall, and F_1-score	Annotated clinical notes
Marteau et al [30]	Not defined	SHC[bd] data repositories and Synthea Patient Generator	Map OMOP CDM concepts to FHIR resources by OMOP-on-FHIR (a novel clinical infrastructure)	ETL processes, OMOP CDM, OMOP-on-FHIR, PostgreSQL, psql[be], SMART[bf] on FHIR, and Synthea Patient Generator	Qualitative feedback collection and SUS[bg]	Pediatric musculoskeletal disorders
Ismail et al [31]	Patient, Observation, Condition, and Practitioner	MCHHJ[bh] and CRMHIS[bi]	Data elements to FHIR resources	MongoDB, FHIR RESTful web services, DAO[bj], Google's REST console app	User study and questionnaires, generate requests and view responses	Maternal health

Study	FHIR[a] resources	Data source	Data transformation and mapping	Standards, tools, terminologies, and models	Validation and evaluation	Use case
Guinez-Molinos et al [32]	Patient, Specimen, DiagnosticReport, and Observation	UC Christus laboratory	Minimum data set fields to FHIR	HAPI FHIR libraries, BPMN[bk], Cawemo, clinFHIR graphBuilder, JWT[bl], and MySQL	Performance evaluation (response time, throughput, process management time, main memory storage, secondary storage), and usability test	PCR[bm] SARS-CoV-2 tests
Burkhardt et al [33]	Patient, Organization, Communication Consent, Questionnaire, QuestionnaireResponse, and CarePlan	Requirement analysis outputs (undergraduate students were surveyed)	Data elements to FHIR	FHIR RESTful API, FHIR Search API, Google's Flutter, Keycloak, HAPI FHIR, AWS[bn], Docker, Postgres DB, JWT, and Apache web server	Not stated	COVID-19 symptom tracking
De et al [34]	Patient, Practitioner, RelatedPerson, Organization, HealthcareService, Appointment, Device, Encounter, DocumentReference, AllergyIntolerance, AdverseEvent, BodyStructure, Specimen, Procedure, FamilyMemberHistory, Observation, Condition, Medication, Immunization, CarePlan, ExplanationOfBenefit, and Account	The web-based patient portal at the Mayo Clinic Rochester	Biomedical text to UMLS and patient secure messages to hidden microconcepts	MetaMap, LDA[bo], multipurpose Annotation Environment, and FHIR definitions	F_1-score to check the consistency between annotators	Random samples of secure patient messages
Liu et al [35]	Condition, Procedure, MedicationStatement, FamilyMemberHistory, Composition, and Bundle	i2b2 2008 obesity data set and MIMIC III data set	Clinical text to FHIR bundle	Deep learning models (text GCN[bp], GRU[bq], and CNN[br]), scikit-learn, TensorFlow, Keras, text classification, NLP2FHIR pipeline, cTAKES, and SNOMED CT codes	Accuracy and macroaveraged precision, F_1-score, and recall	Obesity and random notes from discharge summaries
Zong et al [36]	Observation, Condition	Mayo Clinic's UDP (a clinical data warehouse)	Mappings of report data to 3 data elements (patient clinic number, name, and date of birth) and mappings between elements of PheWAS[bs] profile and FHIR	UML[bt], (*ICD-9* and *ICD-10*) codes, LOINC, phecode, Forge editor, FHIR profiling, cross- validation, chi-square distribution associated allelic *P* value, and KS[bu] test	Cross-validation and FHIR specifications and IGs[bv]	Cancer
Xiao et al [37]	Patient, Encounter, Location, Condition, MedicationStatement, Observation, Procedure, Practitioner, and ConceptMap	MIMIC-III data set (OMOP CDM-based)	OMOP to RDF mappings and OMOP-FHIR mappings	OWL, Protégé, FHIR ShEx[bw], FHIR RDF, VKG[bx] (also known as OBDA[by]), MIMIC-OMOP ETL tool, OMOP CDM, Ontop toolkits, SQL, and SPARQL	Using OMOP CDM-based MIMIC-III data set for system evaluation and comparing patient counts identified over MIMIC database and virtual CKG[bz]	Intensive care

Study	FHIR[a] resources	Data source	Data transformation and mapping	Standards, tools, terminologies, and models	Validation and evaluation	Use case
Kukhareva et al [38]	Patient, Encounter, Observation, Procedure, and Related Person	Epic EHR	Local codes to LOINC, local codes to standard codes, and QUICK[ca] to different FHIR versions and profiles	EHR web services, FHIR services, Authorization services, SMART-on-FHIR, native EHR FHIR APIs, SNOMED, and LOINC	Feasibility check by clinicians	Neonatal bilirubin management

[a]FHIR: Fast Healthcare Interoperability Resources.
[b]EHR: electronic health record.
[c]OMOP: Observational Medical Outcomes Partnership.
[d]PCORnet: Patient-Centered Outcomes Research Network.
[e]HL7: Health Level 7.
[f]CDR: Clinical Data Repositories.
[g]CDM: Common Data Model.
[h]i2b2: informatics for integrating biology and the bedside.
[i]MIMIC: Medical Information Mart for Intensive Care.
[j]NLP: natural language processing.
[k]CQL: Clinical Quality Language.
[l]cTAKES: clinical Text Analysis and Knowledge Extraction System.
[m]MedXN: Medication Extraction and Normalization.
[n]PheKB: Phenotype Knowledge Base.
[o]UIMA: Unstructured Information Management Architecture.
[p]LOINC: Logical Observation Identifiers Names and Codes.
[q]SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.
[r]RxNorm: medical prescription normalized.
[s]ATC: Anatomical Therapeutic Chemical.
[t]ACP: Australian Colorectal Cancer Profile.
[u]CRF: case report form.
[v]UPD: Unified Data Platform.
[w]CAS: Common Analysis System.
[x]AI: artificial intelligence.
[y]ETL: Extract, Transform, and Load.
[z]DHP: Data Harmonization Pipeline.
[aa]ICD: International Classification of Diseases.
[ab]XSLT: Extensible Stylesheet Language Transformations.
[ac]PIT: Patient data Integration Tool.
[ad]CDWH: Carolina Data Warehouse for Health.
[ae]GECCO: German Corona Consensus Dataset.
[af]ICD-10-GM: International Classification of Diseases–German Modification.
[ag]RDF: Resource Description Framework.
[ah]AUROC: Area Under the Receiver Operating Characteristic Curve.
[ai]AUPRC: Area Under the Precision-Recall Curve.
[aj]FSH: FHIR Short Hand.
[ak]HAPI: HL7 application programming interface.
[al]ED: emergency department.
[am]WBAN: Wireless Body Area Network.
[an]RDB: Relational Database.
[ao]FASTO: FHIR And Semantic Sensor Network based Type 1 diabetes Ontology.
[ap]SSN: Semantic Sensor Network.
[aq]BFO: Basic Formal Ontology
[ar]CPG: clinical practice guideline.
[as]OWL: Web Ontology Language.
[at]CDSS: Clinical Decision Support System.

[au]PHR: personal health record.
[av]UoM: units of measurement.
[aw]REST: Representational State Transfer.
[ax]OAuth: open authorization.
[ay]SPARQL: SPARQL Protocol and RDF Query Language.
[az]API: application programming interface.
[ba]DMM: Dirichlet multinomial mixture.
[bb]CPT: Current Procedural Terminology.
[bc]UMLS: Unified Medical Language System.
[bd]SHC: Shriner's Children.
[be]psql: a terminal-based front end to PostgreSQL.
[bf]SMART: Substitutable Medical Apps and Reusable Technology.
[bg]SUS: System Usability Scale.
[bh]MCHHJ: Maternal and Child Health Handbook in Japan.
[bi]CRMHIS: Common Requirements for Maternal Health Information Systems.
[bj]DAO: Data Access Objects.
[bk]BPMN: Business Process Model and Notation.
[bl]JWT: JSON Web Token.
[bm]PCR: polymerase chain reaction.
[bn]AWS: Amazon Web Service.
[bo]LDA: latent Dirichlet allocation.
[bp]GCN: graph convolutional network.
[bq]GRU: Gated Recurrent Unit.
[br]CNN: Convolutional Neural Network.
[bs]PheWAS: Phenome-Wide Association Studies.
[bt]UML: Unified Modeling Language.
[bu]KS: Kolmogorov–Smirnov.
[bv]IG: implementation guide.
[bw]ShEx: Shape Expressions Language.
[bx]VKG: virtual knowledge graph.
[by]OBDA: Ontology-Based Data Access.
[bz]CKG: Clinical Knowledge Graph.
[ca]QUICK: Quality Improvement and Clinical Knowledge.

Static Data Models

Static models do not follow a sequential or linear flow of data processing; instead, they capture and integrate data in broader aspects and mainly consider data mappings rather than the flow of data. These models are more likely to focus on capturing relationships between variables. They focus on the representation and organization of data within the FHIR standard without necessarily addressing the dynamic aspects of data flow or processing. Out of 31 included articles, 6 (19%) studies were related to the development of static data models. Table 2 summarizes the important information of the articles that presented these data models.

Table 2. Static models.

Study	FHIR[a] resources	Data source	Data transformation and mapping	Standards, tools, terminologies, and models	Validation and evaluation	Use case
González-Castro et al [10]	Observation, Device, Questionnaire, QuestionnaireResponse, FamilyMemberHistory, AllergyIntolerance, Patient, Procedure, MedicationStatement, Condition, and Encounter	Patient medical records, PGD[b]	Map data elements to FHIR resources	SNOMED[c], and LOINC[d]	Mapping possibilities check	Cancer survivorship (colon and breast cancers)
Montazeri et al [39]	Patient, Observation, Condition, Medication, ServiceRequest, and Practitioner	CPOE[e] systems, Shafa Hospital (Kerman, Iran)	Data elements to FHIR	CPOE, DigiSurvey platform	Expert panel	Cardiovascular
Shivers et al [40]	AllergyIntolerance, Appointment, CarePlan, Communication, Condition, Consent, CoverageEncounter, HealthcareService, Medication, MedicationAdministration, MedicationStatement, Observation, Patient, Practitioner, Procedure, and ServiceRequest	DAK[f] data dictionaries that contain core data elements for recommendations about family planning and sexually transmitted infections	Data mappings to FHIR and semantic terminologies (*ICD*[g]*-10*, SNOMED CT[h], LOINC, and RxNorm[i])	*ICD-10*, SNOMED CT, LOINC, RxNorm, IG[j], UMLS[k], and IPS[l]	Iterative validation of mappings to identify discrepancies gaps, and errors	Family planning and sexually transmitted infections
Lambarki et al [41]	Patient, Organization, Condition, ClinicalImpression, ServiceRequest, Encounter, Observation, Procedure, and MedicationRequest	DKTK[m]	FHIR data elements to corresponding ADT[n] and ISO standard (11179-3 fields)	*ICD-10*, *ICD*-O-3[o], TNM[p], Forge, Simplifier, FHIR validator, clinFHIR, LOINC, ADT/GEKID schema, and OID[q]	FHIR validator to validate FHIR profiles	Oncology
Lichtner et al [42]	Composition, EvidenceVariable, PlanDefinition, ActivityDefinition, Citation, ArtifactAssessment, Evidence, and Group	Members of the COVID-19 evidence ecosystem project (CEOsys)	Model's items to FHIR resources, information model to EBMonFHIR[r] resources	EBMonFHIR, CPG[s]-on-FHIR, FSH[t], SUSHI[u], HL7[v] FHIR IG Publisher tool, FHIR core artifacts, GRADE EtD[w] framework, PICO[x] framework, Cochrane PICO ontology, SNOMED CT, LOINC, *ICD-10*, ATC[y], UCUM[z], CEOsys, FEvIR[aa] platform	Implementation of a recent COVID-19 guideline recommendation to evaluate EBMonFHIR-based guideline representation	Evidence-based CPG recommendations, COVID-19 intensive care patients' guideline (evaluation phase)
Khalifa et al [43]	Patient, Practitioner, PractitionerRole, Organization, RiskAssessment, Task, ServiceRequest, MedicationRequest, CarePlan, DeviceRequest, NutritionOrder, SupplyRequest, Questionnaire	Sample reports from ARUP laboratory portal	Genetic laboratory test reports to KDEs[ab]- KDEs to FHIR specification	FHIR profiling, (FHIR CG IG STU1[ac])	Not mentioned	Genetic laboratory tests

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]PGD: patient-generated data.

[c]SNOMED: Systemized Nomenclature of Medicine.

[d]LOINC: Logical Observation Identifiers, Names, and Codes.

[e]CPOE: computerized physician order entry.

[f]DAK: Digital Adaptation Kit.

[g]ICD: International Classification of Diseases.

[h]SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.
[i]RxNorm: medical prescription normalized.
[j]IG: implementation guide.
[k]UMLS: Unified Medical Language System.
[l]IPS: International Patient Summary.
[m]DKTK: German Cancer Consortium.
[n]ADT: Association of Comprehensive Cancer Centres (German).
[o]ICD-O: International Classification of Diseases for Oncology.
[p]TNM: Tumor, Node, Metastasis.
[q]OID: object identifier.
[r]EBMonFHIR: Evidence-Based Medicine on Fast Healthcare Interoperability Resources.
[s]CPG: clinical practice guideline.
[t]FSH: FHIR Short Hand.
[u]SUSHI: SUSHI Unshortens Short Hand Inputs.
[v]Hl7: Health Level 7.
[w]GRADE EtD: Grading of Recommendations Assessment, Development and Evaluation Evidence to Decision.
[x]PICO: Population, Intervention, Comparison and Outcome.
[y]ATC: Anatomical Therapeutic Chemical.
[z]UCUM: Unified Code for Units of Measure.
[aa]FEvIR: Fast Evidence Interoperability Resources.
[ab]KDE: Key Data Elements.
[ac]FHIR CG IG STU1: FHIR Clinical Genomics Implementation Guide–Release 1.

Medical Use Case–Specific Summary of Papers

In this phase, we tried to maintain the medical domain consistency in summarizing the articles, and there may be some overlaps between the categories of each article's health care domain. In the following sections, the included papers are summarized and ordered by specific medical use cases and health care applications.

Chronic Diseases

A standard-driven methodology called Clinical Quality Language (CQL) 4NLP was developed to integrate a collection of NLP extensions represented in the HL7 FHIR standard, into the CQL to enhance EHR-driven phenotyping. Using the FHIR standard, specifically the FHIRPath system, enhanced metadata handling and querying by allowing the integration of NLP-derived metadata (such as hypotheticals and negation) into queries. The use case of this research was obesity comorbidities [20]. Another study in the obesity domain used a normalization pipeline to automatically analyze and understand the information in medical records. This FHIR-based approach could detect different sections of medical records and identify important concepts and states of obesity using discharge summaries. The methodology enhanced precise data extraction and portable EHR phenotyping [13]. A similar approach was followed to conduct a case study with obesity data sets. The objective was to predict this condition and the related comorbidities. The sample of adults was categorized into 2 groups called obesity and nonobesity considering their BMI. The design allowed the sharing of deidentified data because only higher-level concepts from knowledge bases and clinical ontologies were included in the FHIR components [35].

In another study, heterogeneous data from a pulmonary hypertension registry were integrated into the Observational Medical Outcomes Partnership–Common Data Model (OMOP CDM) data standard. Common parameters were first identified and mapped to Logical Observation Identifiers Names and Codes (LOINC) and Systemized Nomenclature of Medicine–Clinical Terms (SNOMED CT) as standard terminologies. Extracted data in the form of FHIR bundles were then transformed to OMOP CDM using the Extensible Stylesheet Language Transformations (XSLT). The researchers claimed that FHIR bundles and XSLT can be efficiently and simply used as components of an Extract, Transform, and Load (ETL) process, which can eventually increase data interoperability and applicability [22]. The goal of another research in this area was to map source variables and the value sets to FHIR data elements. The researchers developed a tool called Clinical Asset Mapping Program for FHIR to read Common Data Models (such as informatics for integrating biology and the bedside and Patient-Centered Outcomes Research Network data models) and map the items to FHIR. Using FHIR as a Common Data Model can enhance collaboration, interoperability, and data sharing among health care centers. The clinical use case in the mentioned study was "Asthma" [23].

OMOP-on-FHIR is a technology to convert data elements in OMOP CDM format to the FHIR standard. The researchers used this framework to implement 2 apps to facilitate cohort administration in the context of pediatric musculoskeletal disease research. Accordingly, FHIR can facilitate data access from OMOP CDM databases, support practical integration into health care systems, and enable the development of interoperable clinical applications [30]. For type 1 diabetes mellitus, research presented an ontology-based Clinical Decision Support System based on FHIR and Semantic Sensor Network-Based Type 1 Diabetes Ontology (FASTO). The researchers integrated the FHIR standard, clinical practice guidelines (CPGs), Basic

Formal Ontology, and Semantic Sensor Network and implemented a cloud-based interoperable mobile health system for monitoring and managing patients with this condition. Broader adoption and seamless integration within existing EHRs can be achieved through using FHIR and ontology semantics [27]. A multimethod approach involving the development of a Minimum Data Set for cardiovascular computerized physician order entry was presented in another study. The researchers identified and classified critical data elements by reviewing the content of medical records and then mapped them to the FHIR standard. The FHIR standard was used to maintain interoperability between EHR and computerized physician order entry, which can eventually avoid duplicate data entries and redundancies [39].

COVID-19 and Infectious Diseases

In the context of COVID-19, clinical data across sites were federated by maintaining a single master patient identifier and consistent demographic information. In addition, this proposed methodology was used to distribute data across networks and maintain common data elements, such as mortality status and social determinants of health data. In the aforementioned approach, the data were loaded into an FHIR Clinical Data Repository, which finally produced real-time linked repositories, including FHIR, OMOP, and Patient-Centered Outcomes Research Network. The researchers found that using FHIR as the initial canonical data model and FHIR subscription protocols for transformation and synchronization of multiple data models has potential benefits for health care research, including the automated creation of research data marts for COVID-19 research [9]. An interoperable platform based on the FHIR standard was developed for convenient reporting and sharing of the polymerase chain reaction SARS-CoV-2 tests across countries. The aim was to create a Minimum Data Set for the tests, followed by modeling associated processes and end points. Implementation continued with standards and interoperability design, software development, testing, and implementation [32]. Another COVID-19–related tool called StayHome was developed for collecting patient-reported outcomes. This reusable mobile app was designed to collect COVID-19 symptoms and share them with health care organizations. The FHIR standard was used to ensure interoperability [33]. In another study on COVID-19, the automatic generation of research ontologies through a terminology server and FHIR profiles was analyzed. The researchers also investigated the process of translating user inputs into FHIR queries. On the basis of the results, it is possible to automatically generate mapping files and ontologies for FHIR-based data and profiles [24]. FHIR-based and evidence-based CPG recommendations for patients with COVID-19 were outlined in another approach. Iterative consensus-based mapping of model elements and links to FHIR correspondences along with modeling of recommendations were covered in the mentioned framework. According to the CPG-on-FHIR architecture, the generated guideline recommendations were represented using FHIR profiles. Using this FHIR-based architecture facilitates the creation of computerized guidelines and their seamless integration into EHR systems [42]. In the fields of family planning and sexually transmitted infections, the researchers structured data dictionaries to improve the mapping procedures to FHIR and multiple terminologies, such as the International Classification of Diseases 10th Revision. The corresponding FHIR resources and codes were then identified and mapped to each data dictionary term. The goal was to prepare inputs (mappings and data dictionaries) for an implementation guide (IG) generation tool and enhance the creation of machine-interpretable guidelines [40]. To clarify, FHIR IG is a collection of guidelines and rules designed to facilitate the adaptation of profiles to align with specific care contexts and promote the standardization of information exchange [44].

Cancer Research

In the context of research in cancer clinical trials, FHIR-based pipelines can be used to automatically populate the case report forms (CRFs). The Electronic Data Capture framework was developed in a study to model colorectal cancer trials as a case study. With this strategy, real-world trials can be supported using EHR data [21]. Classification of cancer types and prediction of cancers from unknown primaries were the aims of another research in this field. In the mentioned study [25], genetic data elements (from the oncology reports of patients with cancer) and the associated phenotyping data (from an EHR) were extracted. Researchers presented a network-based infrastructure that modeled the EHR and genetic data with FHIR and Resource Description Framework (RDF) to enhance cancer prediction. In this respect, the performance of different machine learning and deep learning techniques was compared and analyzed [25]. In a paper related to colorectal cancer, data elements were extracted from the CRFs of cancer clinical trials using a data population application. The information was then mapped to an equivalent element in the FHIR cancer profile [28]. An interactive statistics and analysis platform called Shiny FHIR was implemented for ovarian cancer. The system included related R packages (R Foundation for Statistical Computing), FHIR resources, and Shiny (a web application framework). In the FHIR data modeling phase, the ovarian cancer data elements were mapped to corresponding FHIR resources. On the basis of the findings, Shiny can be used in parallel with FHIR to perform interactive analysis [29]. Another interoperable data model called Cancer Survivorship Interoperable Data Elements (CASIDE) was developed in the context of cancer survivorship. The researchers defined data elements and then mapped them to the corresponding FHIR resources. Patient information was illustrated by a collection of FHIR resources to enhance secondary use and sharing of medical data. The research declared the benefits of using FHIR-based models in conjunction with machine learning techniques. In addition, data entry tools can be seamlessly integrated with FHIR-based EHRs [10]. A harmonized data model was also developed in the context of cancer research based on FHIR. German cancer care providers are generally required to report patient data to cancer registries using a specific schema called ADT/GEKID. Therefore, in the mentioned research, the XML representation was compared to the extended version in the German Cancer Consortium (DKTK), and a codification of the cancer life cycle was created. The DKTK FHIR-based data model was represented, and the FHIR resources were identified. Other oncology FHIR profiling efforts were analyzed for reuse in DKTK. It was proved that

multiple health care domains can be efficiently modeled using the FHIR standard and that using embedded mapping annotations, FHIR can be smoothly integrated with other standards [41]. The integration of genetic data from heterogeneous sources, including EHR data and genetic reports, was provided using another FHIR-based data model. The objective was to enable the validation of the Phenome-Wide Association Studies results across different institutions using the FHIR-based data profile. The researchers used the developed model to identify cancer genotype-phenotype associations, followed by validation of the associations according to a literature review [36].

Random and General Medical Notes

The modeling capability of a data normalization pipeline (NLP2FHIR) was assessed in a study focusing on core clinical resources and unstructured EHR data. The researchers attempted to integrate the unstructured elements to develop an FHIR-based model that successfully standardized the annotated corpora [14]. Another framework was designed to integrate unstructured and structured data into an interoperable format by implementing an NLP-based pipeline using the FHIR-type system. On the basis of the results, the model facilitates the integration of NLP-driven EHR data into a standard FHIR format, supports diverse NLP tools, and provides strong extension capacities [12]. A framework presented in another research for standardizing heterogeneous annotation corpora included 2 main modules (automatic schema mapping module and expert-based annotation and verification module). The system used annotated clinical notes and proved that using FHIR with this kind of heterogeneous data can enhance data reuse as well as integration in medical NLP research [16]. A data model in the context of secure patient messages was developed based on FHIR concepts (related to base, foundation, clinical, and financial categories). The objective was to define significant information contained in these sources. After annotating the sentences and creating a huge corpus, the researchers extracted hidden topics related to 3 microconcepts (fatigue, patient visit, and prednisone as highly discussed topics) through topic modeling. The presented data model could distinguish critical concepts in messages and can be used to identify other narratives on multiple platforms [34].

Acute or Intensive Care

In the field of intensive care, researchers aimed to convert the Medical Information Mart for Intensive Care (MIMIC)-IV database elements to FHIR. This database contains patient data from intensive care departments. To support the use of MIMIC-IV on FHIR, a resource demo and a FHIR IG were also created. The benefits of using the FHIR data model are claimed to be its extensive details, which facilitate mappings and conversions of data elements [26]. A FHIR Data Harmonization Pipeline was developed in another study based on an ETL framework. The harmonization of EHR data was performed in 5 phases, including querying the hospital database, mapping the retrieved data to the FHIR format, validating the mapping, transferring the FHIR resources to the patient model database, and exporting the data to the JSON format. Consequently, raw clinical records were transformed into AI-friendly and harmonized representations of data because the hierarchical structure of FHIR may not be sufficiently accessible and standard for AI frameworks. The data could then provide the fast and generic integration of cohort identification methods, facilitating big data processing [15]. In an application for the management of bilirubin in neonates, custom FHIR interfaces were included. After extensive intrainstitutional use, several strategies were explored to modify the app for cross-institutional transfer. Adapting the app for cross-institutional dissemination included clinician-specific implementation using custom FHIR application programming interfaces (APIs), gathering user feedback, differentiating functionality based on FHIR capabilities, implementing gradual replacement with native FHIR interfaces, and using the HL7 Quality Improvement and Clinical Knowledge (QUICK) logical data model for mapping to different FHIR versions and profiles [38]. The QUICK model encapsulates specific details of FHIR (eg, the differences between elements and extensions), enabling a more focused approach to the attributes and classes. This allows for the logical data model specifications to be identified with greater clarity [45]. Another research focused on knowledge graphs (KGs) and semantic modeling. In the mentioned research, the relational databases of the OMOP were used to develop the FHIR-Ontop-OMOP system. The aim was to generate virtual KGs from the databases. The generated KGs were evaluated for the accuracy of data transformation and compatibility with FHIR RDF using an intensive care data set (including medications, vital signs, observations, survival data, and so on). This semantic system could fully represent an OMOP database as an FHIR-compliant representation using KGs, thus enhancing the interoperability of OMOP CDM and FHIR [37].

Other Conditions

A study aimed at implementing a maternal health record system with a data access model based on RESTful web services. In the proposed data model, important data elements were mapped to FHIR resources. Maintaining the related data as FHIR resources enhanced interoperability, efficient data exchange, and evidence-based decision-making [31]. Another article dealt with genetic laboratory tests. The researchers aimed to map the test elements to FHIR format based on an IG. FHIR clinical genomic IG is a beneficial and almost comprehensive tool for sharing genetic test results [43].

Technical Approaches

Overview

Concerning developing data models or infrastructures using the FHIR standard, several tools have been used in the reviewed research articles. This section summarizes the important or common tools and approaches. These items include FHIR-based tools and frameworks, machine learning approaches, and data storage and security.

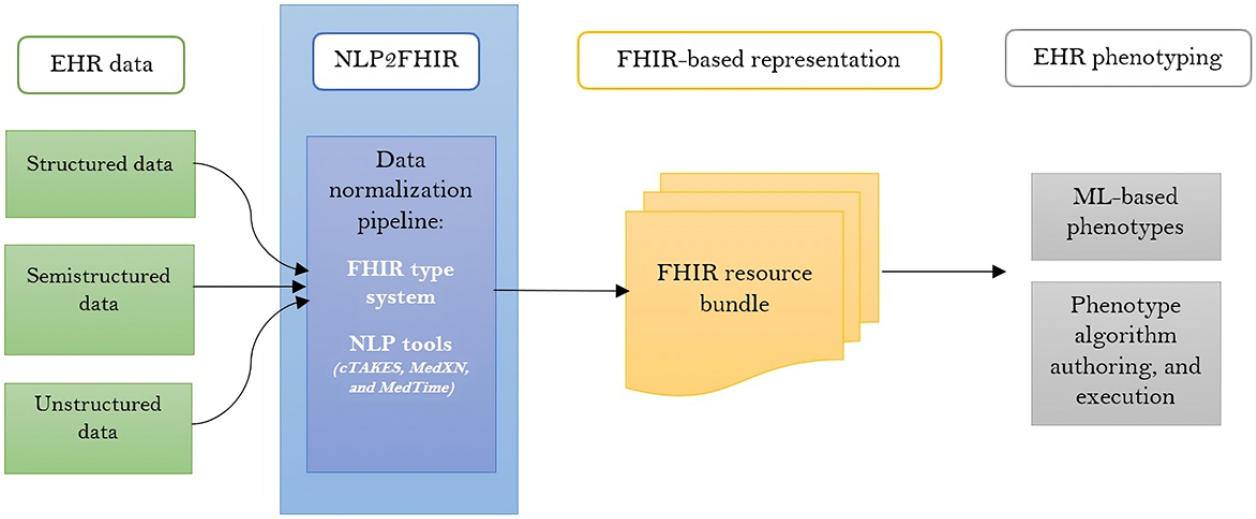
FHIR-Based Tools and Frameworks

NLP2FHIR Pipeline

In the field of NLP, there is a FHIR-related clinical data normalization pipeline called NLP2FHIR for EHR data modeling. This pipeline can be used to standardize and integrate structured and unstructured data stored in EHRs. In other words,

it can make unstructured EHR data consistent and integrate it with structured data. This procedure facilitates portable EHR-driven phenotyping and large-scale data-driven analytics. Several studies used NLP tools as part of the data model's implementation. The NLP2FHIR pipeline was used in 5 articles [13,14,20,25,35]. As shown in Figure 2, this pipeline receives the EHR data in various formats (structured, semistructured, and unstructured) as input. The pipeline itself uses the FHIR-type system as well as NLP tools. The binaries required to run this pipeline are MedTagger, clinical Text Analysis and Knowledge Extraction System (cTAKES), MedTime, Medication Extraction and Normalization (MedXN), and Unified Medical Language System Vocabulary and Terminology Service. The raw clinical data are then transformed into FHIR bundles. Phenotypes can be created based on FHIR bundles, and finally, the FHIR-based data are easily integrated into EHR systems [46].

Figure 2. NLP2FHIR data normalization pipeline and its applications [46]. cTAKES: Clinical Text Analysis and Knowledge Extraction System; EHR: Electronic Health Record; FHIR: Fast Healthcare Interoperability Resources; MedXN: Medication Extraction and Normalization; ML: Machine Learning; NLP: Natural Language Processing.



To elaborate more, MedTagger is an open-source NLP pipeline that underpins the implementation and handling of unstructured clinical data. The system differentiates between general NLP processes and task-specific NLP knowledge, enabling experts to directly encode clinical information using relevant terms and phrases [47]. cTAKES is another open-source NLP system that extracts free-text and narrative information from EHRs and enables semantic processing of this information. It is developed based on the OpenNLP toolkit and a framework called Unstructured Information Management Architecture [48]. MedTime is a hybrid framework containing both machine learning and rule-based approaches for the extraction of temporal information from unstructured clinical text. This system has a high performance in recognizing temporal expressions and clinical incidents [49]. MedXN is a system for extracting pharmaceutical information from clinical notes, making it compatible with RxNorm representation with high accuracy [50]. Unified Medical Language System Vocabulary and Terminology Service enables the interaction of the Unified Medical Language System and the different source vocabularies [51].

Using this pipeline, the data contained in discharge summaries can be transformed into FHIR resources [13]. In addition, normalization and mapping rules as well as NLP-based FHIR extensions can be implemented through NLP2FHIR. It is proven that this pipeline can be a practical tool for modeling unstructured data to eventually integrate the structured elements into models [14]. When NLP-derived artifacts are stored as FHIR extension metadata fields through NLP2FHIR, these elements can be seamlessly incorporated into queries. This integration supports more comprehensive and precise querying by including clinically relevant metadata extracted from unstructured data [20]. In a study conducted by Zong et al [25], each entry in family history records was processed by the NLP2FHIR pipeline, which involved identifying and normalizing medical concepts with MedXN, cTAKES, and MedTime tools. Liu et al [35] followed a workflow of tokenizing documents from 2 data sets and improved the embedding performance by preprocessing (eg, removing less frequent words as well as stop words). The JSON-formatted FHIR resources from the NLP2FHIR pipeline were then transferred into token-like representations categorized into FHIR resources and bundles. cTAKES was also used for concept normalization. The researchers compared the performance of models based on the information in this pipeline with models with original texts.

Substitutable Medical Apps, Reusable Technology–on-FHIR

This specification can be used for data and security requirements for health-related applications. Substitutable Medical Apps, Reusable Technology (SMART)-on-FHIR defines a workflow of secure requests for data access, as well as receiving and using that data [52]. In other words, this specification is a framework that includes web standards that are used to define health applications based on the FHIR-based data stored in an FHIR server. Marteau et al [30] developed a SMART-on-FHIR application, including a query and an upload page to enhance data organization and accessibility. The research highlights that clinicians and health care professionals can query health care applications through FHIR APIs [30]. The applications

containing SMART-on-FHIR can interact and integrate with EHR systems through APIs and provide efficient "plug and play interoperability." Kukhareva et al [38] discussed the balance between portability and functionality of SMART on the FHIR applications and how the developers should consider this balance. A comprehensive approach with the integration of user-centered and technical methods is needed to optimize this balance.

Evidence-Based Medicine–on-FHIR and CPG-on-FHIR

Evidence-Based Medicine–on-FHIR (EBMonFHIR) is a knowledge asset project on FHIR resources for EBM. The objective of EBMonFHIR is to offer interoperability for people who generate, analyze, synthesize, disseminate, and implement clinical evidence and CPGs [53]. CPG-on-FHIR is an IG that uses FHIR resources to build computable and interoperable representations of clinical guideline contents [54]. Lichtner et al [42] developed an IG that used the resources developed by EBMonFHIR to represent primary evidence and the evidence-to-decision process. These resources were eventually integrated into the CPG-on-FHIR framework. Both EBMonFHIR and CPG-on-FHIR are supported by the HL7 Clinical Decision Support staff and represent different aspects of evidence-based guideline recommendations. The former focuses on the justification aspects of the recommendations, while the latter focuses on the implementation aspects of the recommendations.

clinFHIR

clinFHIR is a web-based, open-source educational environment that also allows developers to create or search FHIR-based resources [55]. ClinFHIR graphBuilder is used to model the relationships between resources. This tool also assembles resource instances into a graph with related resources to specify a scenario using FHIR [32]. Accordingly, the structure of models can be visualized using clinFHIR software [41].

HL7 Application Programming Interface FHIR

HL7 API (HAPI) FHIR is a comprehensive implementation of FHIR in the Java language [56]. The API is available for both FHIR clients and servers [57]. Several studies used HAPI FHIR in the data model implementation process. Bennett et al [26] used the HAPI FHIR server in the process of validation, bulk export, and writing data to NDJSON files. Hong et al [29] used the API to put ovarian cancer data into FHIR resources. They also used the client API to upload structured FHIR data elements to the FHIR server. HAPI was one of the test servers that was used to assess data quality and server stability. The API can also be used in the NLP domain. Hong et al [16] used HAPI FHIR for annotation serialization; they converted the annotations to FHIR XML and JSON formats that were eventually represented in an FHIR-consistent format. The HAPI FHIR resource validator API was also used to validate the resources for compliance with the FHIR specification. In the model presented by Guinez-Molinos et al [32], the HAPI FHIR database was used to store resources, and the HAPI server was responsible for the interoperability layer of the model. The HAPI libraries were also used to construct resources, messages, and end points. For persistent storage of FHIR-based data and as an API server, Burkhardt et al [33] used HAPI FHIR V4.2.0. HAPI generally offers standard functionalities, such as create, read, update, and delete APIs, along with specialized domain-specific tools, including CQL. This capability enables developers to concentrate more on the specific needs of their app.

Machine Learning Approaches

Apart from the use of the NLP2FHIR pipeline discussed in the previous section, some other articles used simple NLP tools and algorithms to convert unstructured data into structured data elements adhering to a specific schema for better data description [21]. Hong et al [12] used Unstructured Information Management Architecture NLP tools such as MedXN and MedTime in the normalization phase to enhance interoperability. MedXN was used to extract medication extraction concepts, and MedTime was used to extract FHIR-defined temporal elements. Separate NLP extraction modules were developed to extract information directly from free text for those entities that cannot be extracted by current NLP tools. In a classification system developed by Hong et al [13], 4 machine learning algorithms, including support vector machine (SVM), random forest, logistic regression, and decision tree, were implemented to train the classifiers of the disease prediction module; the features that were used by the system were extracted from FHIR resources as well as terminology extensions. Among all methods, the random forest approach had the best performance. Zong et al [25] analyzed some deep learning and machine learning backbone models to compare the performance of cancer prediction. The bag of features (or bag of words) was used in their research based on the values of attributes in the FHIR model. A graph embedding method called Node2vec was used to learn the patient's features (a vector). Generally, 3 methods of feature generation were compared, including the bag of features, Node2vec, and the bag of features combined with Node2vec. Moreover, 7 classification algorithms were analyzed and compared (random forest, logistic regression, naive Bayes, deep neural network, SVM, graph convolutional network [GCN], and convolutional neural network). Node2vec+bag of features and random forest classifier showed the best performance. To analyze the potential of integrating the unstructured FHIR data representations into deep learning methods, Liu et al [35] used Gated Recurrent Unit, CNN, Text GCN on NLP2FHIR inputs, and raw text. The results highlighted that the best performance was achieved by using the Text GCN classifier in NLP2FHIR input. Therefore, this combination can enhance interoperable EHR phenotyping.

In the data model presented by Zong et al [28], NLP tools were used to provide structured data for the ETL process from unstructured data (such as surgical reports). To cluster each patient in the patient subgrouping process, a model called Dirichlet Multinomial Mixture was used. In the Dirichlet Multinomial Mixture model, one document represents a single topic, which makes it suitable for clustering short texts. The genetic relationship extractor that was developed by Hong et al [16] used SVM as a learning model; the goal was to extract the "FamilyMemberHistory.relationship" FHIR element. Eventually, the NLP performance of the corpora was analyzed. On the basis of the results, an NLP engine can be developed on a pooled corpora that offers enough annotations to train a model. To learn the concealed topics of patient messages, De et al [34] used

latent Dirichlet allocation, an unsupervised topic-learning model. It was claimed that latent Dirichlet allocation is effective in finding common topics with well-known terms but, at the same time, tends to overlook less frequent yet important topics in patient messages.

Data Storage and Security

Several studies used PostgreSQL (also known as Postgres) as the database management system [15,26,30,33]. This system is an SQL-based open-source relational database management system that is compatible with JSON document storage. In the study of Williams et al [15], data storage was based on FHIR resource type, and each resource was mapped to a separate JSON structure. Bennett et al [26] used the MIMIC-IV database; the data contained in the data source was loaded into Postgres and the HAPI FHIR server. The data elements were then mapped to JSON within that system. The research indicated a substantial increase in storage requirements when data is converted to FHIR and further when inserted into HAPI FHIR. Specifically, the HAPI FHIR format required significantly more storage space compared to the basic relational structure. To store the "OMOP CDM database" generated by the Synthea synthetic patient generator, Marteau et al [30] used Postgres. The database was then modified to incorporate additional data needed for OMOP-on-FHIR. A PostgreSQL client application (psql) was subsequently used to interact with the database.

Burkhardt et al [33] also used this system along with the Apache web server in their proposed architecture. By contrast, Ismail et al [31] used MongoDB (NoSQL, or nonrelational data storage) for efficient data record manipulation processes. MongoDB can conveniently handle JSON structure, which is the format of the FHIR resources sent and received by servers. Using MongoDB provides a straightforward transformation of JSON objects into JSON documents, making storage and management more efficient. The database can handle FHIR resource searches based on specified criteria. This is facilitated by the MongoDB Data Access Object component that is responsible for validating the JSON strings received from clients.

Some researchers implemented OAuth for security purposes [27] and used Keycloak as an identity provider [33]. JSON Web Token was also used in other studies [32,33] to provide authentication services. This token securely shares information between end points by a JSON object.

Resource Frequencies

Table 3 illustrates the frequency of each resource in the included articles. It should be mentioned that in this section, we only discuss the official base FHIR resources, and the items mainly considered as profiles, extensions, or domain-specific resources are not illustrated here (eg, "LabTest, Imaging, Referral, Risk, CoverageEligibility, ClaimPayment, ProcedureRequest, Dosage, and DeviceUseStatement" [27,34,38,40]).

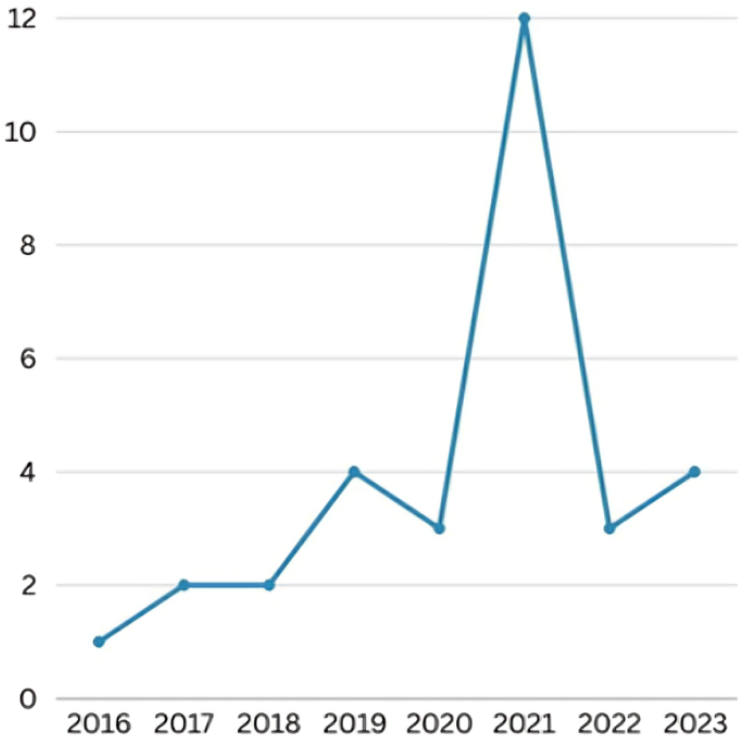
Table 3. Frequency of Fast Healthcare Interoperability Resources (FHIR) resources mentions in the included articles (n=31).

FHIR resources	Articles, n (%)
Observation	21 (68)
Condition	19 (61)
Patient	18 (58)
Procedure	16 (52)
Encounter	11 (35)
MedicationStatement	10 (32)
FamilyMemberHistory and Practitioner	8 (26)
MedicationRequest and Medication	7 (23)
Composition and CarePlan	5 (16)
MedicationAdministration, Questionnaire, AllergyIntolerance, Organization, and ServiceRequest	4 (13)
QuestionnaireResponse, DiagnosticReport, Specimen, RelatedPerson, Device	3 (10)
ValueSet, Immunization, AdverseEvent, Location, NutritionOrder, Communication, Consent, HealthcareService, and Appointment	2 (6)
CodeSystem, MedicationDispense, Goal, EpisodeOfCare, CareTeam, DocumentReference, BodyStructure, ExplanationOfBenefit, Account, Coverage, ClinicalImpression, EvidenceVariable, PlanDefinition, ActivityDefinition, Citation, ArtifactAssessment, Evidence, Group, Bundle, ConceptMap, PractitionerRole, RiskAssessment, Task, DeviceRequest, and SupplyRequest	1 (3)

Publication Distribution

Figure 3 illustrates the distribution of included studies according to the publication year. As shown in the figure, the year 2021 encompassed the highest number of publications.

Figure 3. Yearly distribution of reviewed articles.



Limitations and Challenges of Developing Data Models

Overview

Data model development can present many challenges and constraints arising from issues of data integration, interoperability, standardization, performance, scalability, generalizability, etc. In this section, we discuss the significant challenges and limitations identified in the reviewed papers, which researchers should take into account when developing FHIR-based data models and architectures. It is worth noting that there would be overlaps in the categorization of limitations and challenges.

Data Integration

It is the process of combining data from multiple sources and creating a unified data set. The initial stage of working on data analysis, reporting, and forecasting is data integration [69].

Regarding data integration, the problems mentioned in the analyzed papers include requiring some manual ETL processes [9], manual download of FHIR resources [12], reproducibility issues [20], maintaining robustness [9,20], using only 2 data sets and information loss [13], challenging content normalization [14], single CRF with limited data elements and inadequate questions [21], quality and completeness of the database documentation and nonautomatic concept recognition and considerable data preparation process [15], using incomplete synthetic database [30], no data curation during transformation, bias in database [26], storage space costs [26,33], maintaining inadequate aspects of data [27], synchronization issues and hard-coded mapping [22], privacy and confidentiality issues and limited corpora reuse [16], ignoring continuous changing of values over time [25], lack of comprehensive use of health care records due to lack of education and awareness [31], mapping rules based on only 2 use cases and not including information about generic data [10], manual FHIR mapping and reviewing data of only 1 setting [39], the requirement to implement a structured information model to an existing data dictionary [40], issues with ADT data set as a national reference (completeness and accuracy) [41], and no robust mapping [37].

Interoperability

Interoperability issues are related to the limitations in the seamless transfer and exchange of information between medical systems or applications [70].

Interoperability issues mentioned in the papers include reproducibility concerns [20], organizational interoperability issues [24], privacy and confidentiality issues [16], synchronization issues and hard-coded mapping [22], no evidence to attain a balance of functionality and portability and dissemination barriers due to development and integration costs [38], and extra mapping effort, which affects flexibility and adaptability of the framework [36].

Data Standardization

It is a crucial step in transforming data into a uniform format to enable the shared use of advanced tools and techniques, large-scale analytics, and collaborative research [9,71]. Therefore, standardization issues are related to challenges in attaining standard and consistent data representation across different medical systems.

The issues that could be categorized into the data standardization challenges include difficult rule-changing in compiled java code

for data transformation [9]; semantic gaps between NLP system's data model and FHIR specification [14]; no adequate standardization [12]; mapping accuracy issues [15]; handling valid source system data with no match in FHIR [23]; SNOMED CT postcoordination limitations [24]; not mapping to the US Core as standard ontology, not mapping other databases linked to MIMIC-IV, and not covering some clinical modifiers and qualifiers by FHIR redefinitions [16]; SNOMED coverage restrictions [10]; manual FHIR mapping [39]; duplication of the mapping terms and the necessity to assess the need for a new FHIR profile versus continuing with the existing one [40]; LOINC codes for some observations (SNOMED could be used instead), no available code systems for many value sets, and lack of ubiquitous adoption of FHIR profiles due to the issues with SNOMED license [41]; requiring constant synchronization to the updates because the model was based on EBMonFHIR resources (have low maturity level and subject to changes) and impossibility of showing all guideline information in the FHIR resource format [42]; no textual structure due to lack of gold standard labels, not using syntax-based features for semantic representation, and elimination of some contextual information [35]; translation issues (from OMOP to preferred code systems of FHIR) [37].

Performance

Performance issues are related to obstacles in the efficient processing and retrieval of data that can compromise system performance, for example, the data are not processed within an acceptable response time [72].

According to the analyzed papers, performance issues include integration speed limitations due to transactional EHR [9], performance limitations [13], lack of sophisticated evaluation method [21], limited implementation assessment [30], performance validation issues and no validation for real questions [28], technical challenges [29], model's limited functionality and lack of comprehensive specification [10], no evidence to attain a balance of functionality and portability [38], reduced response rate due to using a web-based questionnaire [39], no execution engine available for representation format [42], evaluation and validation [34], performance rate lower than others and low F_1-score [35], evaluation issues (one instance of MIMIC-III OMOP CDM, no rigorous evaluation) and no comprehensive assessment [43].

Scalability

Scalability limitations can be considered as the data model's weakness in handling increasing data volume or workload.

Generalizability

Generalizability problems are challenges in the applicability of the data model to other aspects or settings.

Considering the selected papers, scalability and generalizability limitations include few resources being used [15]; limited corpora reuse [16]; not enough compatibility and generalizability experiments [15]; not using a real environment [27]; possible bias when conducting a similar study and challenging generalizability for other types of a disease (in this case other cancers) [28]; restriction in the adaptability of the best-performing prediction model and requiring more complex methods to empower prediction and cover diversity [25]; the generalizability issues of the platform [32]; no broad adoption of the app due to issues related to resources and expertise, the best performance for specific programs [33]; few data models are used with no exhaustive evaluation [35]; low contribution to the medical field, failure to distinguish differences in genetic data, low resources for evaluation, and lack of comprehensive data modeling comparisons [36]; and not studying other test types and small sample size [43].

Discussion

Principal Findings

In this review, we aimed to provide a comprehensive PRISMA-based overview of data models using FHIR in the context of interoperability, structure, and functionality and summarize the state of the art for developing FHIR-based data models. In addition, we highlighted limitations, challenges, advantages, and opportunities brought about by FHIR data models. On the basis of the reviewed papers, the most common resources were from the "Clinical" (Observation and Condition) and "Base" (Patient resource) categories of FHIR resources. To develop the models, researchers focused more on the use cases, such as chronic diseases, cancer, COVID-19 and infectious diseases, and intensive care. The reason may be the availability of data in these fields. For instance, Mayo Clinic's clinical data warehouses provide cancer data for researchers. Moreover, i2b2 and MIMIC offer health care data sets about chronic diseases such as obesity. MIMIC-IV on FHIR is also accessible for research in critical care, which provides deidentified FHIR-based data [26].

In terms of limitations, data integration issues are among the most significant challenges in developing data models. The necessity for manual ETL processes, the potential for information loss, and the use of constrained and incomplete data sets can impede the data integration process. Furthermore, organizational differences and hard-coded mappings complicate the seamless exchange of data, affecting interoperability. Issues, such as speed limitations and a lack of robust evaluation metrics, negatively impact performance. In addition, scalability and generalizability are further hindered by limited resources, insufficient compatibility experiments, and small sample sizes.

However, apart from the limitations and challenges, there are numerous advantages to using FHIR-based data models. This standard uses a set of resources and attributes (either common or unique) that enhance data modeling procedures. Constraining the attributes based on an adaptation of clinically relevant ontologies, such as *International Classification of Diseases, Ninth Revision (ICD-9)*, and *International Classification of Diseases, Tenth Revision (ICD-10)*, LOINC, and SNOMED CT, is done through common data types (eg, codeable concepts and string). FHIR can be integrated with other data models, such as RDF, to provide a network-based model for disease prediction. It also supports feature generation and network population in these frameworks [25].

The use of the FHIR models provides the potential to significantly enhance the efficiency and effectiveness of health care research [9,15,29]. CRFs can be automatically populated with FHIR-based EHR data [21], and this automation can identify patient subgroups by topic modeling [28]. EHR data can also be harmonized and mapped to FHIR elements to enhance interoperability and quality of care [15]. Therefore, accessing health-related data for research would be more efficiently achieved when the data are in the FHIR format. Researchers and practitioners can access FHIR app galleries through FHIR APIs and SMART-on-FHIR applications, which can promote health care research and quality of care. FHIR also enables health care professionals to create use case–specific and customized applications [30]. However, implementing SMART-on-FHIR apps poses multiple dissemination challenges and barriers because FHIR-based APIs are not generally considered in the initial stages of implementation in some EHRs [38]. By contrast, as more EHRs choose FHIR as a data exchange standard, nonacademic health care settings will also tend to produce FHIR-formatted data with their EHR systems using FHIR APIs [23]. Maintaining health care data in the format of FHIR resources and using RESTful APIs provide more efficient data transmission compared to conventional record-keeping methods [31].

Another possibility is to use patient messages in web-based portals for health care research. In this respect, a FHIR data model can be developed to extract essential information and concepts from this type of data; FHIR is a beneficial option because it encapsulates modular actions and concepts in health information sharing [34]. Moreover, FHIR exports of local data repositories increase data interoperability for systems and data warehouses [22]. Using FHIR ensures standardized data representation, supports data quality through validation tools (eg, IGs and FHIR specifications), offers flexible adaptation, and benefits from strong community support [35]. The transition of traditional medical guidelines to machine-readable FHIR IGs is a sophisticated and advancing process and needs validation approaches. These use case–specific IGs can eventually enhance real-world application and interoperability of the clinical guidelines [40], considering the constant feedback and inputs from related health care communities [43]. Furthermore, it is crucial to thoroughly follow domain-specific FHIR IGs to gain optimal semantic interoperability. However, even with this guidance provided by IGs, developers still have to make numerous representation and implementation decisions, which may not always be ideal [33].

Evidence-based and computer-interpretable guidelines can be developed using structured data from frameworks' evidence and reviews, followed by mapping the derived items to EBM-on-FHIR resources. This approach aligns with the CPG-on-FHIR framework and includes FHIR profiles and IGs [42].

On the basis of the results of a research paper on using the FHIR standard in oncology, specific health care domains can also be modeled with minimal gaps between FHIR and other standards using annotations of embedded mappings [41]. In addition, in the field of EHR phenotyping and data capture, using FHIR-based data normalization pipelines is considered valuable and beneficial [13]. Using FHIR-based NLP extensions and FHIR composition resources represents NLP components in phenotyping algorithms [20]. Inherently, as mentioned earlier, FHIR resources have granular and atomic characteristics that enable them to share only the required elements for specific use cases and purposes rather than a wide range of elements. This feature is useful for developing specialized AI platforms and interpreting machine learning algorithms [10]. By contrast, this multilayer and complex structure of FHIR may cause some accessibility issues for AI algorithms. To be useful for AI, FHIR data often need to be transformed into a simpler format with a higher level of abstraction, making it more compatible with typical data preprocessing tools. This transformation seems to be necessary to make the data more manageable and easier to analyze by AI systems [15].

Using machine learning algorithms and NLP models is advantageous when integrating with FHIR data models and structures [12] in cases such as standardizing heterogeneous corpora [16]. FHIR modeling for EHR data enhances data integration, data transmission, translational research, and phenotyping [12]. Similarly, the NLP2FHIR pipeline can be used to enhance the standardization of unstructured EHR data [14]. The researchers illustrated how deep learning models can be effectively transferred and used across different settings or systems when dealing with data that have been processed using NLP2FHIR representations. This pipeline performed better in text classification in comparison with models using original texts [35].

Limitations

This review has some limitations. First, we focused only on the articles that dealt with specific use cases with real-world data. By contrast, this approach enabled us to gain insight into the practical applications of the subject matter in real-world contexts rather than merely theoretical ones. In this review, we may have some generalizability issues or biases, especially in the resources and methodologies used and with ontological, general, and theoretical data models. Second, we did not analyze articles published in languages other than English, so we potentially missed some articles due to this criterion. Third, because we did not thoroughly analyze the gray literature and preprints, and due to the relatively small number of included articles, some results may not be generalizable to the entire field of FHIR data modeling.

Comparison With Prior Work

There are some valuable review studies considering the FHIR standard (Table 4). In a scoping review, Balch et al [58] investigated machine learning–based clinical information systems that used the FHIR standard. The focus of the review was to analyze analytics and data management platforms, CDSSs, and APIs and assess the systems' functionalities as well as strengths and weaknesses. Then, the researchers proposed a clinical structure that integrated FHIR and machine learning techniques.

Table 4. The summary of previous reviews considering FHIR[a].

Study, year	Title	Items included in the review
Balch et al [58], 2023	Machine Learning–Enabled Clinical Information Systems Using Fast Healthcare Interoperability Resources Data Standards: Scoping Review	• Investigation of FHIR-based systems using machine learning methods focusing on decision support, data analytics, and APIs[b] • PRISMA-ScR[c] guideline's steps • Categorization of the articles based on functionalities, limitations, and strengths • Proposing a machine learning–based system using FHIR
Nan and Xu [59], 2023	Designing Interoperable Health Care Services Based on Fast Healthcare Interoperability Resources: Literature Review	• Reviewing FHIR-based studies about interoperable health services • Study year and country distributions and charts • Flowchart of paper selection • Clinical categorization of studies and corresponding FHIR resources • Data model migrations to FHIR • Data management methods • Data integration modes • Presenting a FHIR practice design and its development architecture • Commonly used tools
Pimenta et al [60], 2023	Interoperability of Clinical Data through FHIR: A review	• Some selected examples and applications of FHIR (data standards, analysis, API implementations, and research) • PRISMA[d] chart
Pavão et al [61], 2023	The Fast Health Interoperability Resources (FHIR) Standard and Homecare, a Scoping Review	• Home care research studies focusing on FHIR • Screening and inclusion details • FHIR resources • Technological tools in the implementation phase • Privacy and security measures
Ayaz et al [62], 2021	The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities	• Focusing on FHIR and EHR[e]; all articles dealt with FHIR related to research questions • Screening and inclusion details • Study year, type, and country distributions • Primary subject categories over time • FHIR resource list • Types of applications that leverage FHIR • Data mappings form or to FHIR • Objectives of using FHIR • Challenges of using FHIR
Vorisek et al [63], 2022	Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review	• FHIR-based implementations in health care research • PRISMA flowchart for article inclusion • Study year distribution and coauthorship network chart • Research domains • FHIR applications, international standards, and medical domain • FHIR resources • Items mapped to FHIR • Objectives of using FHIR • Limitations of using FHIR
Duda et al [64], 2022	HL7[f] FHIR-based tools and initiatives to support clinical research: a scoping review	• Trends and gaps in using FHIR in health care research • PRISMA flowchart for article and project inclusion • The expansion of Marquis-Gravel categorization [43] of FHIR-based projects contributing to research, in the categories of preparation, prestudy, study setup, recruitment, study conduct, and poststudy activities • Gaps in using FHIR in clinical research
Lehne et al [65], 2019	The Use of FHIR in Digital Health - A Review of the Scientific Literature	• Investigation of using FHIR in digital health care • Article selection flowchart • Study year and article category distribution charts • Abstract text mining for most frequent words
Yogesh and Karthikeyan [66], 2022	Health Informatics: Engaging Modern Healthcare Units: A Brief Overview	• FHIR architecture in health care units • Narrative explanation of FHIR definitions, FHIR data layers and resources, and workflow relations • Health informatics challenges, some related to FHIR

Study, year	Title	Items included in the review
Schweitzer et al [67], 2022	Data Exchange Standards in Teleophthalmology: Current and Future Developments	• Interoperability standards in the field of store-and-forward ophthalmology • Reviewing IHE[g], HL7 standards, DICOM[h], and health care terminologies
Torab-Miandoab et al [68], 2023	Interoperability of heterogeneous health information systems: a systematic literature review	• Interoperability in heterogeneous health care systems • PRISMA flowchart for article selection • Charts and figures for frequencies and trends of interoperability articles • Summary and categorization of interoperability standards and architecture components • Word cloud figures for frequent standards and platforms • Interoperability levels

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]API: application programming interface.

[c]PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews.

[d]PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

[e]EHR: electronic health record.

[f]HL7: Health Level 7.

[g]IHE: Integrating the Healthcare Enterprise.

[h]DICOM: Digital Imaging and Communications in Medicine.

Nan and Xu [59] reviewed FHIR-related papers on designing and building interoperable health care services with a focus on data standardization, management, and integration. The researchers analyzed detailed processes and techniques for each group, resulting in a comprehensive FHIR practice guideline. Similar to our research, Nan and Xu [59] reviewed important techniques and FHIR resources for developing health care services. The difference between this research and our study is the focus of the research; they [59] investigated a broader range of FHIR-based studies and services, while we focused more on data models and infrastructures.

Pavão et al [61] reviewed research articles using the FHIR standard in home care services. The researchers aimed to analyze FHIR resources, types of home care applications, privacy and security considerations, and deployment tools. Pimenta et al [60] reviewed interoperability with FHIR and summarized some important points and examples. The researchers selected some articles and extracted FHIR applications in each study. Ayaz et al [62] in 2021 reviewed all aspects of FHIR in the articles published from January 2012 to December 2019. Our study also considered more recent articles from 2020 to 2023. The main objective of their study was to analyze the articles according to the implementation, challenges, future applications, and opportunities of this standard. The researchers reviewed articles that focused on all categories, including apps, FHIR implementation models, FHIR resources, FHIR framework, mapping framework and data model, challenges, and FHIR goals. The researchers also summarized the resources used in the included articles; "Observation" and "Patient" resources, respectively, were the most commonly used resources in the included articles. We also performed this analysis and had close results; as we mentioned earlier in our study, "Observation," "Condition," and "Patient" resources were used more frequently. The mentioned researchers also discussed the mapping approaches from other techniques or methods to FHIR. The focus of the study by Vorisek et al [63] was to review the FHIR standard from a "health research" perspective. The researchers analyzed the studies that used FHIR in any aspect of the research process, such as data collection, recruitment, data standardization, analysis, and consent management. In addition, they categorized the articles with generic or specific clinical specialty approaches. We also categorized our articles based on the medical field. In the mentioned research, it was reported that most studies used other terminologies and data models besides FHIR, which included SNOMED CT, LOINC, International Classification of Diseases 10th Revision, OMOP CDM, and more. It was reported that among "data capture–related" studies, the "Questionnaire" resource was used more frequently, as expected. In addition to the scientific aspects, the limitations of using FHIR were similarly discussed. They highlighted that the limitations may include the evolving contents of FHIR resources, legal issues, safety, and the need to have an FHIR server. In our study, by contrast, we categorized the limitations into other aspects.

Regarding the medical research aspects of FHIR, Duda et al [64] also presented a literature review. The study extended the "Marquis-Gravel categories" [73], in which it is possible to categorize the way each project contributes to research tasks. The FHIR projects focused on research were investigated, which included the activities of preparation (eg, mapping to and from FHIR), prestudy (eg, defining or refining of cohorts), study design (eg, data collection for research), recruitment (eg, including screening criteria in EHR), study conduct (eg, patient data collection), and poststudy (eg, data sharing). Most projects focused on "general research preparation" (eg, infrastructure and data pipeline development). Lehne et al [65] reviewed the application of FHIR in digital health. On the basis of their research, the reviewed articles were mostly related to data models, mobile or web applications, and medical devices. Yogesh and Karthikeyan [66] reviewed the FHIR architectural specifications, such as the linkages, workflow state, health informatics, and public health safety approaches using this standard. The researchers also highlighted the likely challenges with health care data standards, including coding speed and accuracy issues, code mappings, compatibility issues between

new and legacy systems, and communication concerns between EHRs and patients.

Some other articles reviewed general interoperability and data exchange standards. In the research conducted by Schweitzer et al [67], the researchers narratively described and compared exchange approaches, such as Digital Imaging and Communications in Medicine (DICOM), the Integrating the Health Care Enterprise initiative, and clinical terminologies (such as SNOMED CT) as well as FHIR in the field of teleophthalmology. In their research, the ophthalmology-related FHIR resource, which is "VisionPrescription," as well as the current proposal of the related IG were discussed. Torab-Miandoab et al [68] reviewed interoperability approaches and requirements for semantic interoperability between heterogeneous health information systems. It was found that FHIR, Clinical Document Architecture (CDA), Service-Oriented Architecture, Reference Information Model, Health Insurance Portability and Accountability Act security act, SNOMED CT, XML, JAVA, SQL, and API can be considered the most important requirements to implement semantic interoperability. On the basis of the results, a summary of interoperability standards in the context of terminology, content, transport, and security was also presented. The researchers highlighted the categorization of interoperability architecture components with the main categories of service-oriented architecture, archetype-based, web-based, client-server, multiagent, blockchain-based, XML-based, cloud-based, ontology-based, object-oriented, and local network.

Future Directions and Recommendations

In the course of this study, we encountered some ideas and recommendations for future research. These included the following: (1). a comparison of health care data models with the use of FHIR and other standards, including earlier versions of HL7 interoperability standards (such as HL7-version 2 and version 3 and CDA), OpenEHR, and OMOP CDM. The aim would be to provide a detailed analysis of the models created with these standards, focusing on the methodological aspects, limitations, strengths, and maintenance of interoperability. (2). an examination of the ontological aspects of data models and a discussion of how they represent medical terminologies and concepts.

Conclusions

FHIR serves as a highly promising interoperability standard for developing real-world health care applications. The integration of FHIR with other data models facilitates the development of more interoperable domain-specific solutions and improves research efficiency. In addition, the implementation of FHIR modeling for EHR data facilitates the integration, transmission, and analysis of data while also advancing translational research and phenotyping. Several FHIR data models have been developed to enhance the extraction of essential information and concepts from unstructured data such as patient summaries retrieved from EHRs. Generally, FHIR-based exports of local data repositories improve data interoperability for systems and data warehouses across different settings. However, ongoing efforts to address existing limitations and challenges are essential for the successful implementation and integration of FHIR data models.

Acknowledgments

Partial funding by the German Federal Ministry of Education and Research within the Medical Informatics Initiative DIFUTURE FKZ 01ZZ2304A.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 548 KB - medinform_v12i1e58445_app1.pdf]

Multimedia Appendix 2
PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analyses Protocols) 2015 checklist and review protocol.
[PDF File (Adobe PDF File), 147 KB - medinform_v12i1e58445_app2.pdf]

Multimedia Appendix 3
Search strategy.
[PDF File (Adobe PDF File), 132 KB - medinform_v12i1e58445_app3.pdf]

References

1. Carvalho G, Mykolyshyn S, Cabral B, Bernardino J, Pereira V. Comparative analysis of data modeling design tools. IEEE Access 2022;10:3351-3365. [doi: 10.1109/access.2021.3139071]
2. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care 2012 Jul;50 Suppl:S60-S67 [FREE Full text] [doi: 10.1097/MLR.0b013e318259bff4] [Medline: 22692260]

3. Danese MD, Halperin M, Duryea J, Duryea R. The generalized data model for clinical research. BMC Med Inform Decis Mak 2019 Jun 24;19(1):117 [FREE Full text] [doi: 10.1186/s12911-019-0837-5] [Medline: 31234921]
4. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
5. Stan O, Miclea L. Local EHR management based on FHIR. In: Proceedings of the 2018 IEEE International Conference on Automation, Quality and Testing, Robotics. 2018 Presented at: AQTR '18; May 24-26, 2018; Cluj-Napoca, Romania p. 1-5 URL: https://ieeexplore.ieee.org/document/8402719 [doi: 10.1109/aqtr.2018.8402719]
6. Borisov V, Minin A, Basko V, Syskov A. FHIR data model for intelligent multimodal interface. In: Proceedings of the 26th Telecommunications Forum. 2018 Presented at: TELFOR '18; November 20-21, 2018; Belgrade, Serbia p. 420-425 URL: https://ieeexplore.ieee.org/document/8611918 [doi: 10.1109/telfor.2018.8611918]
7. Noumeir R. Active learning of the HL7 medical standard. J Digit Imaging 2019 Jun 23;32(3):354-361 [FREE Full text] [doi: 10.1007/s10278-018-0134-3] [Medline: 30353411]
8. Amar F, April A, Abran A. Electronic health record and semantic issues using fast healthcare interoperability resources: systematic mapping review. J Med Internet Res 2024 Jan 30;26:e45209 [FREE Full text] [doi: 10.2196/45209] [Medline: 38289660]
9. Lenert LA, Ilatovskiy AV, Agnew J, Rudisill P, Jacobs J, Weatherston D, et al. Automated production of research data marts from a canonical fast healthcare interoperability resource data repository: applications to COVID-19 research. J Am Med Inform Assoc 2021 Jul 30;28(8):1605-1611 [FREE Full text] [doi: 10.1093/jamia/ocab108] [Medline: 33993254]
10. González-Castro L, Cal-González VM, Del Fiol G, López-Nores M. CASIDE: a data model for interoperable cancer survivorship information based on FHIR. J Biomed Inform 2021 Dec;124:103953 [FREE Full text] [doi: 10.1016/j.jbi.2021.103953] [Medline: 34781009]
11. Zhou B, Yang G, Shi Z, Ma S. Natural language processing for smart healthcare. IEEE Rev Biomed Eng 2024;17:4-18. [doi: 10.1109/rbme.2022.3210270]
12. Hong N, Wen A, Shen F, Sohn S, Liu S, Liu H, et al. Integrating structured and unstructured EHR data using an FHIR-based type system: a case study with medication data. AMIA Jt Summits Transl Sci Proc 2018;2017:74-83 [FREE Full text] [Medline: 29888045]
13. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmussen LV, et al. Developing a FHIR-based EHR phenotyping framework: a case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform 2019 Nov;99:103310 [FREE Full text] [doi: 10.1016/j.jbi.2019.103310] [Medline: 31622801]
14. Hong N, Wen A, Shen F, Sohn S, Wang C, Liu H, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open 2019 Dec;2(4):570-579 [FREE Full text] [doi: 10.1093/jamiaopen/ooz056] [Medline: 32025655]
15. Williams E, Kienast M, Medawar E, Reinelt J, Merola A, Klopfenstein SA, et al. A standardized clinical data harmonization pipeline for scalable ai application deployment (FHIR-DHP): validation and usability study. JMIR Med Inform 2023 Mar 21;11:e43847 [FREE Full text] [doi: 10.2196/43847] [Medline: 36943344]
16. Hong N, Wen A, Mojarad MR, Sohn S, Liu H, Jiang G. Standardizing heterogeneous annotation corpora using HL7 FHIR for facilitating their reuse and integration in clinical NLP. AMIA Annu Symp Proc 2018;2018:574-583 [FREE Full text] [Medline: 30815098]
17. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
18. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015 Jan 02;349:g7647 [FREE Full text] [Medline: 25555855]
19. Raj A, Bosch J, Olsson HH, Wang TJ. Modelling data Pipelines. In: Proceedings of the 46th Euromicro Conference on Software Engineering and Advanced Applications. 2020 Presented at: SEAA '20; August 26-28, 2020; Portoroz, Slovenia p. 13-20 URL: https://ieeexplore.ieee.org/document/9226314 [doi: 10.1109/seaa51224.2020.00014]
20. Wen A, Rasmussen LV, Stone D, Liu S, Kiefer R, Adekkanattu P, et al. CQL4NLP: development and integration of FHIR NLP extensions in clinical quality language for EHR-driven phenotyping. AMIA Jt Summits Transl Sci Proc 2021;2021:624-633 [FREE Full text] [Medline: 34457178]
21. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-based computational pipeline for automatic population of case report forms for colorectal cancer clinical trials using electronic health records. JCO Clin Cancer Inform 2020 Mar;4:201-209 [FREE Full text] [doi: 10.1200/CCI.19.00116] [Medline: 32134686]
22. Fischer P, Stöhr MR, Gall H, Michel-Backofen A, Majeed RW. Data Integration into OMOP CDM for heterogeneous clinical data collections via HL7 FHIR bundles and XSLT. Stud Health Technol Inform 2020 Jun 16;270:138-142. [doi: 10.3233/SHTI200138] [Medline: 32570362]
23. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, et al. Fast Healthcare Interoperability Resources (FHIR) as a meta model to integrate common data models: development of a tool and quantitative validation study. JMIR Med Inform 2019 Oct 16;7(4):e15199 [FREE Full text] [doi: 10.2196/15199] [Medline: 31621639]

24. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, et al. Generation of a fast healthcare interoperability resources (FHIR)-based ontology for federated feasibility queries in the context of COVID-19: feasibility study. JMIR Med Inform 2022 Apr 27;10(4):e35789 [FREE Full text] [doi: 10.2196/35789] [Medline: 35380548]
25. Zong N, Ngo V, Stone DJ, Wen A, Zhao Y, Yu Y, et al. Leveraging genetic reports and electronic health records for the prediction of primary cancers: algorithm development and validation study. JMIR Med Inform 2021 May 25;9(5):e23586 [FREE Full text] [doi: 10.2196/23586] [Medline: 34032581]
26. Bennett AM, Ulrich H, van Damme P, Wiedekopf J, Johnson AE. MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format. J Am Med Inform Assoc 2023 Mar 16;30(4):718-725 [FREE Full text] [doi: 10.1093/jamia/ocad002] [Medline: 36688534]
27. El-Sappagh S, Ali F, Hendawi A, Jang J, Kwak K. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. BMC Med Inform Decis Mak 2019 May 10;19(1):97 [FREE Full text] [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
28. Zong N, Stone DJ, Sharma DK, Wen A, Wang C, Yu Y, et al. Modeling cancer clinical trials using HL7 FHIR to support downstream applications: a case study with colorectal cancer data. Int J Med Inform 2021 Jan;145:104308 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104308] [Medline: 33160272]
29. Hong N, Prodduturi N, Wang C, Jiang G. Shiny FHIR: an integrated framework leveraging shiny R and HL7 FHIR to empower standards-based clinical data applications. Stud Health Technol Inform 2017;245:868-872 [FREE Full text] [Medline: 29295223]
30. Marteau BL, Zhu Y, Giuste F, Shi W, Carpenter A, Hilton C, et al. Accelerating multi-site health informatics with streamlined data infrastructure using OMOP-on-FHIR. Annu Int Conf IEEE Eng Med Biol Soc 2022 Jul;2022:4687-4690. [doi: 10.1109/EMBC48229.2022.9871865] [Medline: 36085809]
31. Ismail S, Alshmari M, Qamar U, Butt WH, Latif K, Ahmad HF. HL7 FHIR compliant data access model for maternal health information system. In: Proceedings of the IEEE 16th International Conference on Bioinformatics and Bioengineering. 2016 Presented at: BIBE '16; October 31-November 2, 2016; aichung, Taiwan p. 51-56 URL: https://ieeexplore.ieee.org/document/7789959 [doi: 10.1109/bibe.2016.9]
32. Guinez-Molinos S, Andrade JM, Medina Negrete A, Espinoza Vidal S, Rios E. Interoperable platform to report polymerase chain reaction SARS-CoV-2 tests from laboratories to the Chilean government: development and implementation study. JMIR Med Inform 2021 Jan 20;9(1):e25149 [FREE Full text] [doi: 10.2196/25149] [Medline: 33417587]
33. Burkhardt H, Brandt P, Lee J, Karras S, Bugni P, Cvitkovic I, et al. StayHome: a FHIR-native mobile COVID-19 symptom tracker and public health reporting tool. Online J Public Health Inform 2021 Mar 21;13(1):e2 [FREE Full text] [doi: 10.5210/ojphi.v13i1.11462] [Medline: 33936522]
34. De A, Huang M, Feng T, Yue X, Yao L. Analyzing patient secure messages using a fast health care interoperability resources (FIHR)-based data model: development and topic modeling study. J Med Internet Res 2021 Jul 30;23(7):e26770 [FREE Full text] [doi: 10.2196/26770] [Medline: 34328444]
35. Liu S, Luo Y, Stone D, Zong N, Wen A, Yu Y, et al. Integration of NLP2FHIR representation with deep learning models for EHR phenotyping: a pilot study on obesity datasets. AMIA Jt Summits Transl Sci Proc 2021;2021:410-419 [FREE Full text] [Medline: 34457156]
36. Zong N, Sharma DK, Yu Y, Egan JB, Davila JI, Wang C, et al. Developing a FHIR-based framework for phenome wide association studies: a case study with a pan-cancer cohort. AMIA Jt Summits Transl Sci Proc 2020;2020:750-759 [FREE Full text] [Medline: 32477698]
37. Xiao G, Pfaff E, Prud'hommeaux E, Booth D, Sharma DK, Huo N, et al. FHIR-Ontop-OMOP: building clinical knowledge graphs in FHIR RDF with the OMOP common data model. J Biomed Inform 2022 Oct;134:104201 [FREE Full text] [doi: 10.1016/j.jbi.2022.104201] [Medline: 36089199]
38. Kukhareva P, Warner P, Rodriguez S, Kramer H, Weir C, Nanjo C, et al. Balancing functionality versus portability for SMART on FHIR applications: case study for a neonatal bilirubin management application. AMIA Annu Symp Proc 2019;2019:562-571 [FREE Full text] [Medline: 32308850]
39. Montazeri M, Khajouei R, Mahdavi A, Ahmadian L. Developing a minimum data set for cardiovascular Computerized Physician Order Entry (CPOE) and mapping the data set to FHIR: a multi-method approach. J Med Syst 2023 Apr 14;47(1):47. [doi: 10.1007/s10916-023-01943-2] [Medline: 37058148]
40. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: authoring FHIR implementation guides based on WHO recommendations. J Biomed Inform 2021 Oct;122:103891 [FREE Full text] [doi: 10.1016/j.jbi.2021.103891] [Medline: 34450285]
41. Lambarki M, Kern J, Croft D, Engels C, Deppenwiese N, Kerscher A, et al. Oncology on FHIR: a data model for distributed cancer research. Stud Health Technol Inform 2021 May 24;278:203-210. [doi: 10.3233/SHTI210070] [Medline: 34042895]
42. Lichtner G, Alper BS, Jurth C, Spies C, Boeker M, Meerpohl JJ, et al. Representation of evidence-based clinical practice guideline recommendations on FHIR. J Biomed Inform 2023 Mar;139:104305 [FREE Full text] [doi: 10.1016/j.jbi.2023.104305] [Medline: 36738871]

43. Khalifa A, Mason CC, Garvin JH, Williams MS, Del Fiol G, Jackson BR, et al. Interoperable genetic lab test reports: mapping key data elements to HL7 FHIR specifications and professional reporting guidelines. J Am Med Inform Assoc 2021 Nov 25;28(12):2617-2625 [FREE Full text] [doi: 10.1093/jamia/ocab201] [Medline: 34569596]
44. Dos Santos Leandro G, Moro CM, Cruz-Correia RJ, Portela Santos EA. FHIR implementation guide for stroke: a dual focus on the patient's clinical pathway and value-based healthcare. Int J Med Inform 2024 Oct;190:105525. [doi: 10.1016/j.ijmedinf.2024.105525] [Medline: 39033722]
45. QUICK data model. HL7 International. URL: https://www.hl7.org/fhir/us/qicore/2018jan/quick/help.html [accessed 2024-04-29]
46. BD2KOnFHIR / NLP2FHIR. GitHub. URL: https://github.com/BD2KOnFHIR/NLP2FHIR [accessed 2024-04-29]
47. Fu S, Leung LY, Wang Y, Raulli A, Kallmes DF, Kinsman KA, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. JMIR Med Inform 2019 Apr 21;7(2):e12109 [FREE Full text] [doi: 10.2196/12109] [Medline: 31066686]
48. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
49. Lin YK, Chen H, Brown RA. MedTime: a temporal information extraction system for clinical narratives. J Biomed Inform 2013 Dec;46 Suppl:S20-S28 [FREE Full text] [doi: 10.1016/j.jbi.2013.07.012] [Medline: 23911344]
50. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 2014;21(5):858-865 [FREE Full text] [doi: 10.1136/amiajnl-2013-002190] [Medline: 24637954]
51. OHNLPIR / UMLS_VTS. GitHub. URL: https://github.com/OHNLPIR/UMLS_VTS [accessed 2024-04-29]
52. SMART on FHIR: introduction. smile Digital Health. URL: https://smilecdr.com/docs/smartu [accessed 2024-04-29]
53. EBMonFHIR. Confluence. URL: https://confluence.hl7.org/display/CDS/EBMonFHIR [accessed 2024-04-29]
54. CPGonFHIR. Confluence. URL: https://confluence.hl7.org/display/CDS/CPGonFHIR [accessed 2024-04-29]
55. ClinFHIR. eCQI Resource Center. URL: https://ecqi.healthit.gov/tool/clinfhir [accessed 2024-04-29]
56. A free and open source global good: powering interoperability around the world for. HAPI FHIR. URL: https://hapifhir.io/ [accessed 2024-04-29]
57. hapifhir / hapi-fhir. GitHub. URL: https://github.com/hapifhir/hapi-fhir [accessed 2024-04-29]
58. Balch JA, Ruppert MM, Loftus TJ, Guan Z, Ren Y, Upchurch GR, et al. Machine learning-enabled clinical information systems using fast healthcare interoperability resources data standards: scoping review. JMIR Med Inform 2023 Aug 24;11:e48297 [FREE Full text] [doi: 10.2196/48297] [Medline: 37646309]
59. Nan J, Xu LQ. Designing interoperable health care services based on fast healthcare interoperability resources: literature review. JMIR Med Inform 2023 Aug 21;11:e44842 [FREE Full text] [doi: 10.2196/44842] [Medline: 37603388]
60. Pimenta N, Chaves A, Sousa R, Abelha A, Peixoto H. Interoperability of clinical data through FHIR: a review. Procedia Comput Sci 2023;220:856-861. [doi: 10.1016/j.procs.2023.03.115]
61. Pavão J, Bastardo R, Santos M, Rocha NP. The Fast Health Interoperability Resources (FHIR) standard and homecare, a scoping review. Procedia Comput Sci 2023;219:1249-1256. [doi: 10.1016/j.procs.2023.01.408]
62. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
63. Vorisek CN, Lehne M, Klopfenstein SA, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
64. Duda SN, Kennedy N, Conway D, Cheng AC, Nguyen V, Zayas-Cabán T, et al. HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. J Am Med Inform Assoc 2022 Aug 16;29(9):1642-1653 [FREE Full text] [doi: 10.1093/jamia/ocac105] [Medline: 35818340]
65. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - a review of the scientific literature. Stud Health Technol Inform 2019 Sep 03;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]
66. Yogesh MJ, Karthikeyan J. Health informatics: engaging modern healthcare units: a brief overview. Front Public Health 2022 Apr 29;10:854688 [FREE Full text] [doi: 10.3389/fpubh.2022.854688] [Medline: 35570921]
67. Schweitzer M, Steger B, Hoerbst A, Augustin M, Pfeifer B, Hausmann U, et al. Data exchange standards in teleophthalmology: current and future developments. Stud Health Technol Inform 2022 May 16;293:270-277. [doi: 10.3233/SHTI220380] [Medline: 35592993]
68. Torab-Miandoab A, Samad-Soltani T, Jodati A, Rezaei-Hachesu P. Interoperability of heterogeneous health information systems: a systematic literature review. BMC Med Inform Decis Mak 2023 Jan 24;23(1):18 [FREE Full text] [doi: 10.1186/s12911-023-02115-5] [Medline: 36694161]
69. 11 most common issues with data integration [solved]. DataToBiz. URL: https://tinyurl.com/4ybmfxau [accessed 2024-04-29]
70. Lenhardt WC. Data interoperability: infrastructure, historical artifact, or science meme? Examples from the geosciences. In: Companion: Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social

Computing. 2019 Presented at: CSCW '19; November 9-13, 2019; Austin, TX p. 272-277 URL: https://dl.acm.org/doi/10.1145/3311957.3359458 [doi: 10.1145/3311957.3359458]
71. Standardized data: the OMOP common data model. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/data-standardization/ [accessed 2024-04-29]
72. Supranovich R, Newmyer R. Performance issues. In: Supranovich R, Newmyer R, editors. Bringing Micro to the Macro: Adapting Clinical Interventions for Supervision and Management. New York, NY: Routledge; 2020:71-79.
73. Marquis-Gravel G, Roe MT, Turakhia MP, Boden W, Temple R, Sharma A, et al. Technology-enabled clinical trials: transforming medical evidence generation. Circulation 2019 Oct 22;140(17):1426-1436. [doi: 10.1161/CIRCULATIONAHA.119.040798] [Medline: 31634011]

Abbreviations

AI: artificial intelligence
API: application programming interface
CASIDE: Cancer Survivorship Interoperable Data Elements
CPG: clinical practice guideline
CQL: Clinical Quality Language
CRF: case report form
cTAKES: Clinical Text Analysis and Knowledge Extraction System
DICOM: Digital Imaging and Communications in Medicine
DKTK: German Cancer Consortium (From German)
EBMonFHIR: Evidence-Based Medicine–on-FHIR
EHR: electronic health record
ETL: Extract, Transform, and Load
FASTO: Fast Healthcare Interoperability Resources and Semantic Sensor Network-Based Type 1 Diabetes Ontology
FHIR: Fast Healthcare Interoperability Resources
GCN: graph convolutional network
HAPI: Health Level 7 application programming interface
HIMSS: Healthcare Information and Management Systems Society
HL7: Health Level 7
IG: implementation guide
KG: knowledge graph
LOINC: Logical Observation Identifiers Names and Codes
MedXN: Medication Extraction and Normalization
MIMIC: Medical Information Mart for Intensive Care
NLP: natural language processing
OMOP CDM: Observational Medical Outcomes Partnership−Common Data Model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-P: Preferred Reporting Items for Systematic review and Meta-Analysis Protocols
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
QUICK: Quality Improvement and Clinical Knowledge
RDF: Resource Description Framework
SMART: Substitutable Medical Apps, Reusable Technology
SNOMED CT: Systemized Nomenclature of Medicine−Clinical Terms
SVM: support vector machine

Edited by C Lovis; submitted 17.03.24; peer-reviewed by C Vorisek; comments to author 03.06.24; revised version received 28.07.24; accepted 17.08.24; published 24.09.24.

Please cite as:
Tabari P, Costagliola G, De Rosa M, Boeker M
State-of-the-Art Fast Healthcare Interoperability Resources (FHIR)–Based Data Model and Structure Implementations: Systematic Scoping Review
JMIR Med Inform 2024;12:e58445
URL: https://medinform.jmir.org/2024/1/e58445
doi:10.2196/58445
PMID:39316433

Review

Use of SNOMED CT in Large Language Models: Scoping Review

Eunsuk Chang[1], MD, MPH, PhD; Sumi Sung[2], PhD

[1]Republic of Korea Air Force Aerospace Medical Center, Cheongju, Republic of Korea

[2]Department of Nursing Science, Research Institute of Nursing Science, Chungbuk National University, Cheongju, Republic of Korea

Corresponding Author:
Sumi Sung, PhD
Department of Nursing Science
Research Institute of Nursing Science
Chungbuk National University
1 Chungdae-ro
Seowon-gu
Cheongju, 28644
Republic of Korea
Phone: 82 43 249 1731
Fax: 82 43 266 1710
Email: sumisung@cbnu.ac.kr

Abstract

Background: Large language models (LLMs) have substantially advanced natural language processing (NLP) capabilities but often struggle with knowledge-driven tasks in specialized domains such as biomedicine. Integrating biomedical knowledge sources such as SNOMED CT into LLMs may enhance their performance on biomedical tasks. However, the methodologies and effectiveness of incorporating SNOMED CT into LLMs have not been systematically reviewed.

Objective: This scoping review aims to examine how SNOMED CT is integrated into LLMs, focusing on (1) the types and components of LLMs being integrated with SNOMED CT, (2) which contents of SNOMED CT are being integrated, and (3) whether this integration improves LLM performance on NLP tasks.

Methods: Following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, we searched ACM Digital Library, ACL Anthology, IEEE Xplore, PubMed, and Embase for relevant studies published from 2018 to 2023. Studies were included if they incorporated SNOMED CT into LLM pipelines for natural language understanding or generation tasks. Data on LLM types, SNOMED CT integration methods, end tasks, and performance metrics were extracted and synthesized.

Results: The review included 37 studies. Bidirectional Encoder Representations from Transformers and its biomedical variants were the most commonly used LLMs. Three main approaches for integrating SNOMED CT were identified: (1) incorporating SNOMED CT into LLM inputs (28/37, 76%), primarily using concept descriptions to expand training corpora; (2) integrating SNOMED CT into additional fusion modules (5/37, 14%); and (3) using SNOMED CT as an external knowledge retriever during inference (5/37, 14%). The most frequent end task was medical concept normalization (15/37, 41%), followed by entity extraction or typing and classification. While most studies (17/19, 89%) reported performance improvements after SNOMED CT integration, only a small fraction (19/37, 51%) provided direct comparisons. The reported gains varied widely across different metrics and tasks, ranging from 0.87% to 131.66%. However, some studies showed either no improvement or a decline in certain performance metrics.

Conclusions: This review demonstrates diverse approaches for integrating SNOMED CT into LLMs, with a focus on using concept descriptions to enhance biomedical language understanding and generation. While the results suggest potential benefits of SNOMED CT integration, the lack of standardized evaluation methods and comprehensive performance reporting hinders definitive conclusions about its effectiveness. Future research should prioritize consistent reporting of performance comparisons and explore more sophisticated methods for incorporating SNOMED CT's relational structure into LLMs. In addition, the biomedical NLP community should develop standardized evaluation frameworks to better assess the impact of ontology integration on LLM performance.

(JMIR Med Inform 2024;12:e62924) doi:10.2196/62924

KEYWORDS

SNOMED CT; ontology; knowledge graph; large language models; natural language processing; language models

Introduction

Background

The recent emergence of large language models (LLMs), exemplified by Bidirectional Encoder Representations from Transformers (BERT) [1] and GPT [2], has significantly advanced the capabilities of machines in natural language understanding (NLU) and natural language generation (NLG). Despite achieving state-of-the-art performance on a range of natural language processing (NLP) tasks, LLMs exhibit a deficiency in knowledge when confronted with knowledge-driven tasks [3]. These models acquire factual information from extensive text corpora during training, embedding this knowledge implicitly within their numerous parameters and consequently posing challenges in terms of verification and manipulation [4]. Moreover, numerous studies have demonstrated that LLMs struggle to recall facts and frequently encounter hallucinations, generating factually inaccurate statements [5,6]. This poses a significant obstacle to the effective application of LLMs in critical scenarios, such as medical diagnosis and legal judgment [7].

Efforts have been made to address the black box nature of LLMs and mitigate potential hallucination problems. Approaches include enhancing language model (LM) veracity through strategies such as retrieval chain-of-thought prompting [8] and retrieval-augmented generation [9]. Another significant avenue involves integrating knowledge graphs (KGs) or ontologies into LMs using triple relations or KG subgraphs [7,10]. KGs, renowned for their excellence in representing knowledge within a domain, can provide answers when combined with LMs [11], making them valuable for common sense–based reasoning and fact-checking models [12]. However, LLMs often face challenges when trained and tested predominantly on general-domain datasets or KGs, such as Wikipedia and WordNet [13], making it difficult to gauge their performance on datasets containing biomedical texts. The differing word distributions in general and biomedical corpora pose challenges for biomedical text mining models [14].

Biomedicine-specific KGs may be a potential solution to the abovementioned problems. In the biomedical domain, KGs, also known as ontologies, are relatively abundant, with the Unified Medical Language System (UMLS) [15] being one of the most frequently used ontologies [16]. The UMLS serves as a thesaurus for biomedical terminology systems such as the Medical Subject Headings, International Classification of Diseases, Gene Ontology, Human Phenotype Ontology, and SNOMED CT, all curated and managed by the United States National Library of Medicine.

Among UMLS member terminologies, SNOMED CT stands out as the most comprehensive biomedical ontology, encompassing a wide range of biomedical and clinical entities, including signs, symptoms, diseases, procedures, and social contexts [17]. These entities are represented by concepts (clinical ideas), descriptions (human-readable terms linked to concepts), and relations (comprising hierarchical *is-a* relations and horizontal attribute relations). As SNOMED CT is increasingly integrated into electronic health record (EHR) systems, as required by the Fast Healthcare Interoperability Resource (FHIR) to ensure interoperability among health care institutions [18], terminology servers supporting SNOMED CT have become ubiquitous. With its ready availability across health care institutions, SNOMED CT has gained attention as a knowledge source or ontology for representing biomedical and clinical knowledge [17]. In this case, the abstract model of SNOMED CT is used to describe and store biomedical facts in a hierarchical and structured manner, readily available across health care institutions.

Integrating SNOMED CT into LLMs holds significant potential for advancing various aspects of health care and biomedical research. By incorporating the comprehensive and structured biomedical knowledge from SNOMED CT, LLMs can better understand medical terminology, relationships between clinical concepts, and domain-specific context, potentially reducing errors and hallucinations when understanding or generating biomedical texts. This integration could enhance clinical decision support systems, improve the accuracy of automated coding and billing processes, facilitate more precise information retrieval from medical literature, and support the development of personalized medicine approaches. Furthermore, it may enable more accurate NLP of clinical notes and medical records, potentially leading to improved patient care and outcomes through better data analysis and insights.

Objectives

This scoping review aimed to examine the use of SNOMED CT as a knowledge source to be incorporated into LLMs, specifically focusing on the methodology of integrating these 2 modalities. This review sought to answer the following research questions: (1) What are the dominant types and components of LLMs being integrated with SNOMED CT? (2) Which contents of SNOMED CT (ie, descriptions, relations, or entity classes) are being integrated into LLMs? and (3) Does the integration of SNOMED CT into LLMs improve the performance on NLP tasks in terms of NLU and NLG? Answers to these questions could suggest future methodological approaches for more effectively integrating human-engineered knowledge into LLMs.

Methods

This scoping review was guided by the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) framework, which outlines the recommended steps and reporting standards for conducting scoping reviews (Multimedia Appendix 1) [19].

Study Identification

We defined LLMs as transformer-based LMs pretrained on large-scale corpora [20] (Multimedia Appendix 2). Given that transformer-based models currently dominate in the field and

are likely to continue doing so in the coming years, reviewing other LMs, such as recurrent neural networks and more conventional statistical models, does not hold scientific significance for current and future applications. Therefore, focusing on transformer-based models allows a more cohesive and in-depth analysis of the most relevant and cutting-edge techniques in the field.

To explore scientific literature describing transformer-based models, we conducted our literature search on ACM Digital Library, ACL Anthology, IEEE Xplore, PubMed, and Embase on March 12, 2024, using the following query terms: (1) ("language *model" OR "pretrained *model" OR "language processing" OR "embedding") AND ("SNOMED" OR "Unified Medical Language System" OR "UMLS" OR "*medical") AND ("knowledge graph" OR "ontolog*" OR "knowledge*base" OR "knowledge infusion") and (2) ("SNOMED") AND ("large language model" OR "BERT" OR "GPT"). Queries were modified according to the bibliographic databases when necessary. Queries were designed to search for articles published from 2018 to 2023. The start date of the query was set to 2018 when BERT, the first transformer-based LM to gain widespread adoption, was introduced, marking the beginning of significant research into transformer-based LLMs.

Study Selection

Articles were extracted from ACM Digital Library, ACL Anthology, IEEE Xplore, PubMed, and Embase. Duplicates were removed, and 2 authors (SS and EC) examined the full text of the retrieved articles for the presence of the term "SNOMED." We prioritized a full-text search first before title and abstract review because many potentially eligible papers do not explicitly mention "SNOMED" in their titles or abstracts. To be eligible for our review, articles had to have SNOMED CT incorporated into NLP pipelines, which encompass processes from text cleansing through pretraining and inference to model evaluation, specifically for tasks involving NLU and NLG. We then further excluded studies that met ≥1 of the following criteria: (1) published in languages other than English; (2) categorized as reviews, surveys, keynotes, or editorial articles; (3) did not incorporate SNOMED CT at any stage of the NLP pipeline; (4) aimed to create, develop, enrich, or enhance ontologies or graphs; (5) did not involve the processing of natural language (NL) text; or (6) solely used SNOMED CT codes for retrieving patients of interest from EHRs or for annotating instances with SNOMED CT codes as gold-standard target labels for LM training.

Result Synthesis

Through discussions and qualitative assessments, we analyzed the included articles according to the following characteristics: chronological and geographic publication trends, baseline LLM and its output, dataset used for training and testing the model, methods for integrating SNOMED CT into the LLM, and the model's end task and performance (Textbox 1).

Textbox 1. Methods for synthesizing the review.

Synthesis of results

- Chronological and geographic publication trends
- Baseline large language model (LLM) and its output
- Dataset used for training and testing the model
- Methods for integrating SNOMED CT into the LLM (methodologies for knowledge graph [KG]–enhanced LLMs [7])
  - KG-enhanced LLM pretraining: works that apply KGs during the pretraining stage and improve the knowledge expression of LLMs
  - KG-enhanced LLM interpretability: works that use KGs to understand the knowledge learned by LLMs and interpret the reasoning process of LLMs
  - KG-enhanced LLM inference: research that uses KGs during the inference stage of LLMs, which enables LLMs to access the latest knowledge without retraining
- End task and performance
  - End task natural language understanding: entity recognition or typing, entity or relation extraction, document classification, question answering (multiple choice), and inference End task natural language generation: text summarization, question answering (short or essay answers), translation, and dialogue generation Performance analysis: nominal percentage gains in performance after SNOMED CT integration

We elucidated the methodology for incorporating SNOMED CT into NLP pipelines following the categorization methods previously outlined by Pan et al [7]. These methods categorized methodologies for KG-enhanced LLMs into three distinctive types: (1) KG-enhanced LLM pretraining, (2) KG-enhanced LLM interpretability, and (3) KG-enhanced LLM inference. The end tasks of LLMs after SNOMED CT integration included NLU and NLG. Regarding the performance analysis, we presented the nominal percentage gains in performance after SNOMED CT integration without analyzing their statistical significance, as most studies did not perform statistical significance testing. We refrained from conducting direct study-to-study comparisons due to concerns about the heterogeneity of testing corpora and evaluation metrics across different studies.

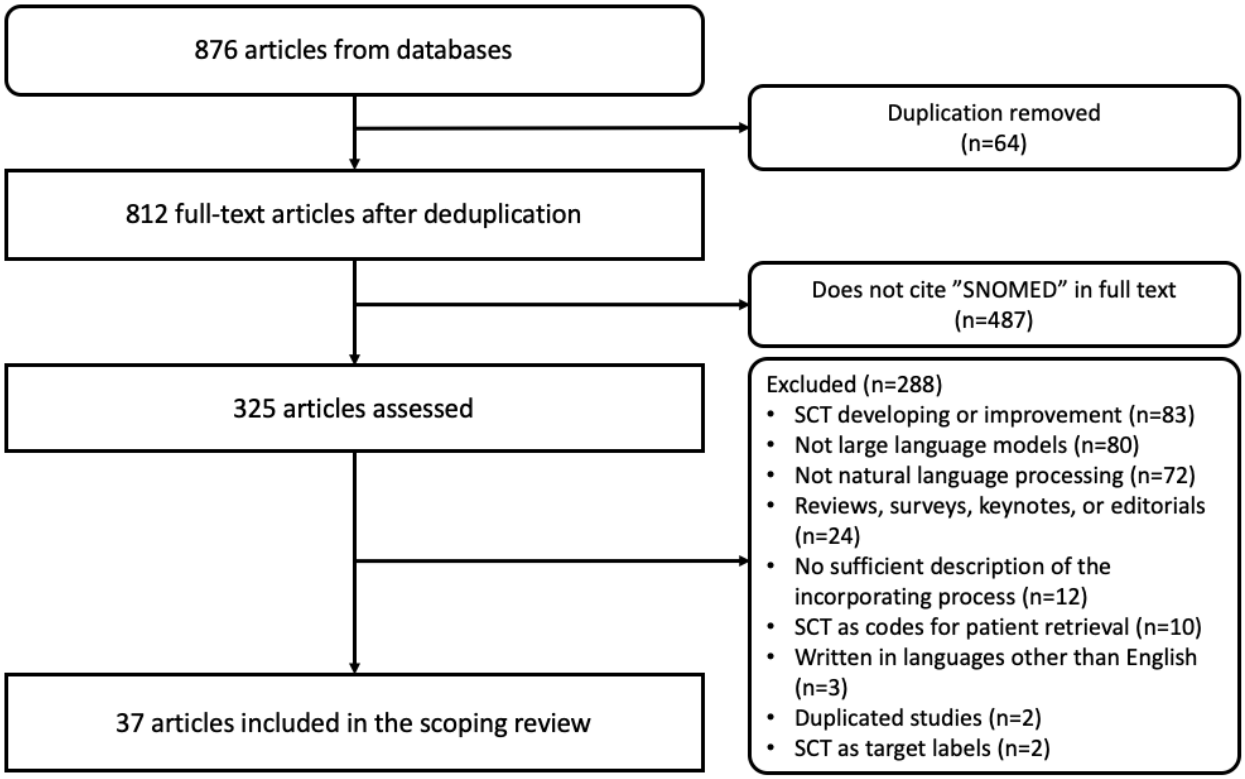
Results

Selected Papers

The query yielded 876 articles from the 5 bibliographic databases, with 634 (72.4%) obtained from the first query and

242 (27.6%) from the second query (Figure 1). After the removal of duplicates, 812 (92.7%) articles were reviewed to check whether the term "SNOMED" was mentioned in their full texts. A total of 325 (37.1%) articles were then reviewed according to the inclusion and exclusion criteria. Consequently, 37 (4.2%) publications were finally selected for the scoping review (Figure 1). The characteristics of the individual papers and other features, including the language of used datasets and SNOMED CT descriptions, other ontologies used, and the types of entities represented by SNOMED CT, are detailed in Multimedia Appendix 3.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of article selection. SCT: SNOMED CT.



Chronological and Geographic Publication Trends

Table 1 presents the publication trends noted in the review. Although our literature search covered publications from 2018 onward, no studies published in 2018 were included in the final review. The largest volume of studies was published in 2022 (13/37, 35%), followed by those published in 2020 (10/37, 27%).

When the number of countries was counted according to the first authors' institutional affiliations, the largest number of studies was noted to originate from the United States (10/37, 27%). While most of the studies (26/37, 70%) were conducted in countries that are members of SNOMED International, some were performed in nonmember countries such as Bulgaria and China, where separate license fees and in-house translation of SNOMED CT descriptions to the local language were required.

Table 1. Chronological and geographic publication trends among the included studies.

Study characteristics	Studies
Publication year	
2019	[21-23]
2020	[24-33]
2021	[34-36]
2022	[37-49]
2023	[50-57]
Countries	
Australia	[26,35]
Bulgaria	[34,52]
Canada	[55]
China (including Hong Kong)	[28,38,39,41,43,45,48,50,56]
Germany	[47,51]
India	[22,31,32]
Israel	[53]
Spain	[21,29,30,37,40]
United Kingdom	[54,57]
United States	[23-25,27,33,36,42,44,46,49]
Publication type	
Journal paper	[23-26,36,42-46,50,55-57]
Conference paper	[21,22,27-35,37-41,45,47-49,51-54]

Baseline LLMs and Their Outputs

Most of the included studies (27/37, 73%) used BERT and its variants as the baseline LLMs for NLU and NLG tasks. Variants such as RoBERTa [58] and ALBERT [59] were also used to address BERT's relatively small training corpora and long training time [31,37,38,50,53]. To overcome the limited applicability of these general-purpose LLMs to biomedical texts, many studies (13/37, 35%) used LLMs trained on large-scale biomedical corpora, such as BioBERT [14] and PubMedBERT [60], which were trained on PubMed articles, and ClinicalBERT [61] and EHRBERT [23], which were trained on clinical notes. SapBERT [62], initialized by PubMedBERT, was further fine-tuned using contrastive learning with UMLS synonyms to better accommodate SNOMED CT synonym descriptions [44,47]. To support biomedical NLP tasks in languages other than English, LLMs trained on corpora in those languages were also adopted, such as medBERT.de [63], designed specifically for the German medical domain [51], and ERNIE-health, pretrained from Chinese medical records [41]. Aside from these BERT-based models, GPT emerged as a new baseline LLM since 2023. Makhervaks et al [53] used BioGPT [64], whose decoder was pretrained on biomedical corpora, to enhance the generation of artificial sentences. In addition, Xu et al [55] used GPT-3.5 for ranking suggested annotation terms in their study (Table 2).

A primary assertive role of LLMs was representing biomedical entities from text data. While most proposed methods produced embedding vectors to convey contextual information about the biomedical entities that appeared in texts, Kalyan and Sangeetha [31] introduced a Siamese RoBERTa model to generate concept vectors from synonym relationships defined by SNOMED CT. These basic outputs of LLMs might undergo additional task-specific layers to perform desired end tasks, which will be discussed later. Beyond producing embedding representations of entities, some studies required LLMs to perform classification or ranking tasks after fine-tuning, predicting the most likely relevant standard concepts [23,24,26,34,41,55], entity types [35,38,51], sentences [49,53], or matched foreign language words, enabling machine translation [28-30,39]. LLMs with encoder-decoder architectures, such as BART [65], were used for dedicated NLG tasks [32,57].

Table 2. Large language models used in the included studies.

Base and fine-tuned models	Studies
BERT[a]	
Vanilla BERT	[22,24,26,27,33,40,42-44,50,53,54,56,57]
RoBERTa	[31,37,38,50]
ALBERT	[53]
ELECTRA	[53]
DeBERTa	[53]
mBERT	[37,45]
BioBERT	[27,33,34,46,48,49,52]
ClinicalBERT	[25,33,35,36]
PubMedBERT	[45,46]
SAPBERT	[44,47]
EHRBERT	[23]
SciBERT	[46]
BioELECTRA	[53]
German BERT models	[51]
GPT	
GPT-3.5	[55]
BioGPT	[53]
BART	[57]
Transformer neural networks	
Transformer NMT[b] model	[21,28-30,39]
Denoising autoencoder	[32]
ERNIE[c]	
ERNIE-health	[41]

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]NMT: neural machine translation.

[c]ERNIE: Enhanced Language Representation with Informative Entities.

Data for Training and Testing Models

When using general-domain LLMs, authors deployed additional fine-tuning or pretraining on biomedical corpora to better adapt their models for biomedical NLP tasks. The pretraining corpora included PubMed or MEDLINE articles [28,30,38,39,46] and other publicly available datasets, such as Wikipedia articles [29] and tweets [37] related to biomedical topics. Synthetic sentences were also used to address data scarcity, which was generated based on SNOMED CT descriptions or relations [21,29].

While some studies (8/37, 22%) used real-world clinical narrative records [21,30,48,52] or customized (ie, manually annotated by researchers) data [25,27,41,56] for testing their models, most of the studies (29/37, 78%) used publicly available datasets, especially when researchers were participating in shared task competitions or dealing with English texts. CADEC [66] and PsySTAR [67], open datasets built from drug review posts in which concept mentions were mapped to SNOMED CT concepts, were used for validating and testing concept normalization models [31,45]. The Medical Concept Normalization (MCN) corpus, drawn from discharge summaries annotated using SNOMED CT and RxNorm concepts, was experimented on by concept normalization models [24,26]. The WMT corpora, provided by the annual Conference on Machine Translation Shared Tasks, were used to test multilingual machine translation tasks by participating researchers [28,29,39]. Makhervaks et al [53] and Chopra et al [22] used sentence pairs in the MedNLI corpus [68], annotated by medical doctors into 3 categories—contradictory, entailing, and neutral—for NL inference tasks. The MedMentions corpus [69] identifies >350,000 mentions from >4000 PubMed abstracts, linking them to the UMLS concepts; it was used in the studies by Zotova et al [40] and Dong et al [54], in which SNOMED CT was loaded onto the UMLS. The ShARe/CLEF 2013 corpus [70] consists of deidentified clinical notes annotated with disease mentions using the SNOMED CT subset of the UMLS; it was used for testing concept normalization tasks [44,54].

SNOMED CT Content Integration Into NLP Pipelines

Overview

While the categorization methods by Pan et al [7] pertained to the integration of LLMs with general-purpose KGs, we treated SNOMED CT as a specified form of KG. Their third category—KG-enhanced LLM interpretability—was omitted due to the lack of relevant studies in our review. In addition, we found no studies that fit into the subcategories "Integrating KGs into Training Objectives" (under "KG-enhanced LLM pretraining") and "Dynamic Knowledge Fusion" (under "SNOMED CT–enhanced LLM inference"). The overarching categorization of all included methods is shown in Textbox 2.

Textbox 2. Summarized categorizations of SNOMED CT–incorporated large language model (LLM) methods (allowed duplicated counting of studies).

Category and subcategory

- SNOMED CT–enhanced LLM pretraining
  - Integrating SNOMED CT into LLM inputs (n=28, 76%)
  - Integrating SNOMED CT into additional fusion modules (n=5, 14%)
- SNOMED CT–enhanced LLM inference
  - Retrieval-augmented knowledge fusion (n=5, 14%)

Integration of SNOMED CT Into LLM Inputs

Overview

Research in this area concentrated on developing new training objectives for LLMs that incorporate knowledge awareness. More specifically, this line of research aimed to incorporate relevant portions or subsets of SNOMED CT as additional input to LLMs during training. Because a disproportionately large number of included studies (28/37, 76%) fell into this category, we analyzed the methodology by two additional themes: (1) the content of SNOMED CT that was integrated into an LLM and (2) the part of the NLP pipeline into which the aforementioned content was incorporated. After qualitative analysis of the included articles and heuristic discussions among reviewers, we categorized the former theme into descriptions (including descriptions of synonyms), relations, and entity types (classes) and the latter theme into encoders and training data. SNOMED CT contents could be incorporated into LLM encoders either as embedding vectors or as annotations or tags when incorporated into the training corpus.

Table 3 shows the distribution of models across SNOMED CT contents and NLP pipelines, allowing for duplicated counting of a single study if it adopted ≥2 methods.

Table 3. Distributions of models across SNOMED CT contents and natural language processing (NLP) pipelines.

SNOMED CT content integrated into the NLP pipeline	Part of the NLP pipeline where SNOMED CT contents were integrated into	
	Encoder (as vector embedding)	Training corpora (as annotated text)
Description	[31,35,41,43,44,54]	[21,23,24,28-30,32,34,39,40,47-50,52,54,57]
Relation	[31,45]	[21,34,40,52,53]
Entity type (class)	—[a]	[25,38,42,51]

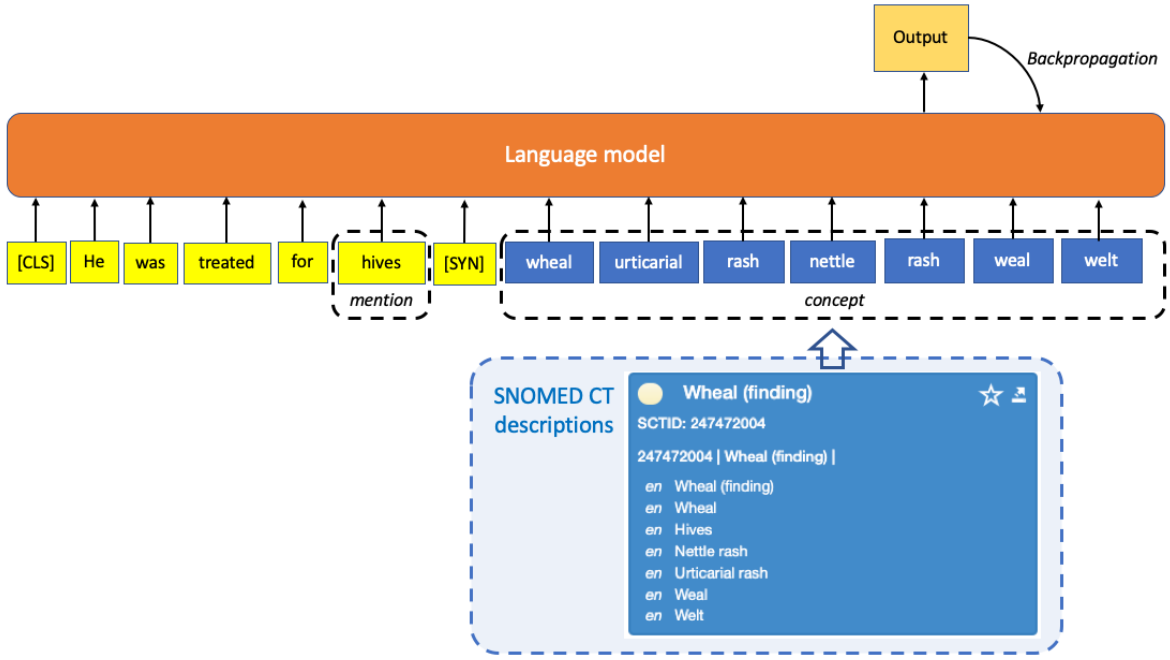
[a]Not available.

Integration of SNOMED CT Descriptions

Vector representations of SNOMED CT concept descriptions were created to facilitate seamless fusion into LLM encoders. The vectors for SNOMED CT description embeddings were used to calculate cosine similarity between the original mentions and SNOMED CT descriptions for concept normalization tasks [35,41,43,54].

Instead of transforming text descriptions into vector embeddings, NL description texts were directly added to training corpora to expand the size of in-domain vocabulary (Figure 2). The description texts of synonyms were either concatenated in the training corpora before being input into an LLM for pretraining [24,47,49,54,57] or they replaced the original entity mentions in the text with standardized terms [32,48]. The descriptions of SNOMED CT codes were also prepended to the word sequences as classifier tokens for LLM pretraining [23]. The multilingual feature of SNOMED CT descriptions was exploited to address the limited availability of training datasets in foreign languages by adding the translated SNOMED CT descriptions into the training corpora [28-30,39,50].

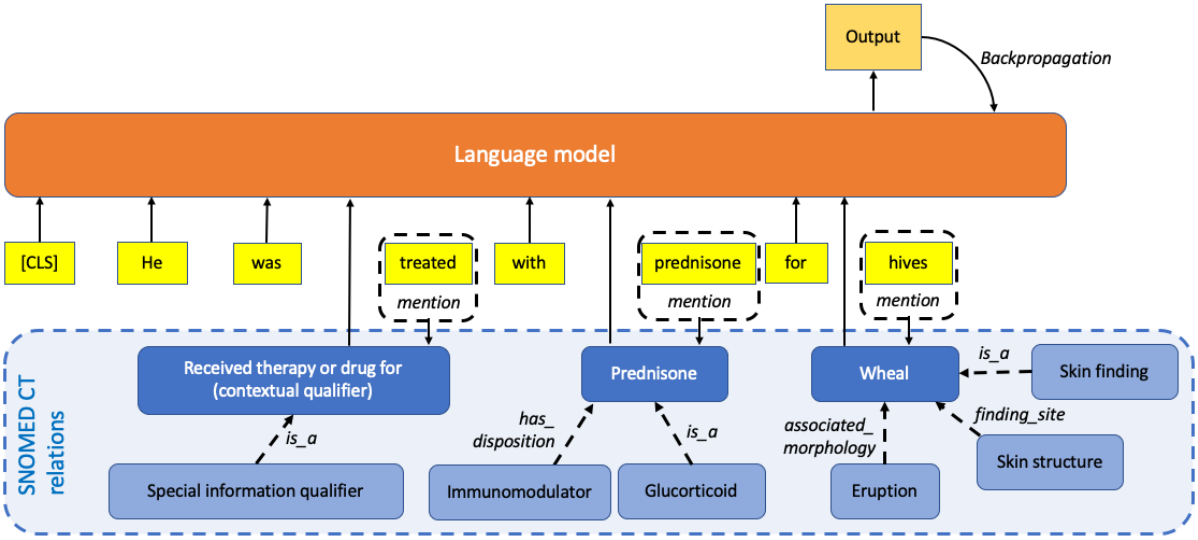
Figure 2. Integrating SNOMED CT descriptions into large language models. CLS: classification; SYN: synonym.



Integration of SNOMED CT Relations

This line of research introduced relevant subgraph information of SNOMED CT, representing SNOMED CT relations as graph edges, into LLMs (Figure 3). Kalyan and Sangeetha [31] encoded SNOMED CT concept descriptions to generate concept embedding vectors and learn representation vectors of concept mentions in the text, further improving the representations by retrofitting the target concept vectors with SNOMED CT synonym relations. CODER [45] used KG embedding methods such as DistMult and ANALOGY [71] to learn relational knowledge from SNOMED CT, enabling the quantification of term-relation-term similarity as well as term-term similarity.

Figure 3. Integrating SNOMED CT relations into large language models. CLS: classification.

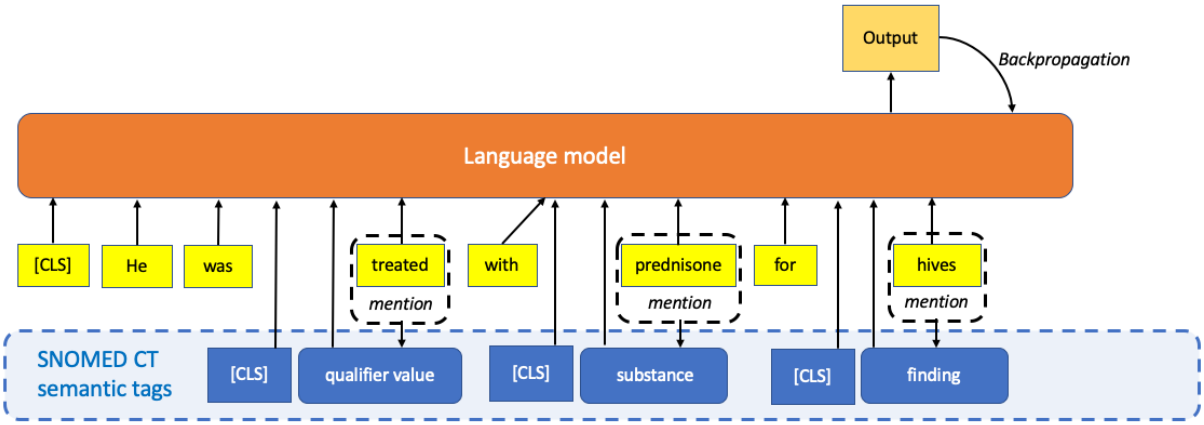


A different approach was taken to introduce textual relation triplets defined by SNOMED CT to expand the size of training corpora. Soto et al [21] exploited the relations defined in SNOMED CT, such as *is_a* and *occurs_in*, to generate synthetic training corpora. Relations defined in SNOMED CT were also used to apply weak supervision to sentence pairs extracted from PubMed to establish contradiction labels in the dataset [53]. Other authors exploited the existing mappings to other ontologies (eg, International Classification of Diseases-10 and UMLS) to enrich the training corpus with the description texts from the linked ontology concepts [34,40,52].

Integration of SNOMED CT Entity Types

The type of entities was incorporated into training corpora by distantly labeling the identified entities with SNOMED CT semantic tags (eg, diseases and chemicals; Figure 4) [25,38]. In other studies, training corpora were annotated with SNOMED CT top-level hierarchies [51] or subclasses of top-level hierarchies [42] to label sentences per their respective tasks.

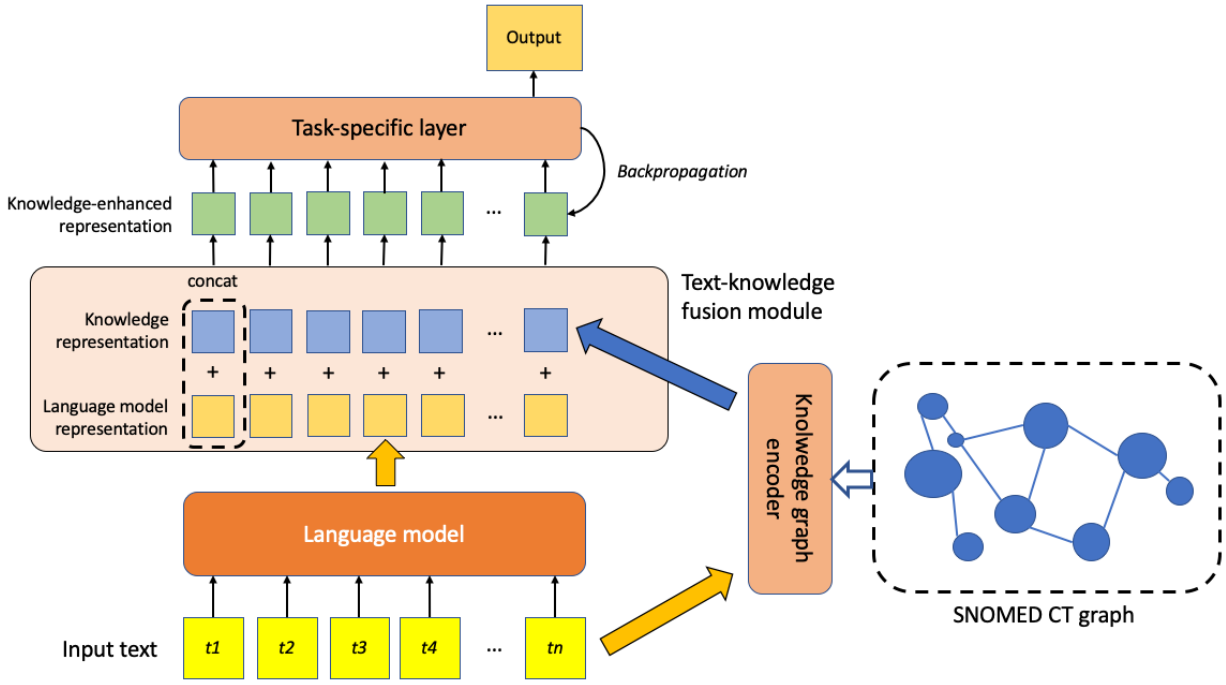
Figure 4. Integrating SNOMED CT entity type information into large language models. CLS: classification.



Integration of SNOMED CT Into Additional Fusion Modules

In this approach, concept information was processed separately before being concatenated and fused with the LLM embedding output (Figure 5). Authors created knowledge-directed embeddings using SNOMED CT graphs, where concepts were represented as nodes and relations as edges, and concatenated them with the LLM contextual embeddings. The merged representations of text and graph embeddings were then passed through a task-specific knowledge fusion module to achieve end tasks such as semantic similarity measurement [36,46], classification [22,27], and question answering [33,46]. To represent the graph information of SNOMED CT concepts, Chang et al [36] used a graph convolutional network [72] for encoding node features and edges. Chopra et al [22] proposed the Bio-MTDDN model, which introduced the shortest path information between corresponding SNOMED CT concepts into knowledge-directed embeddings.

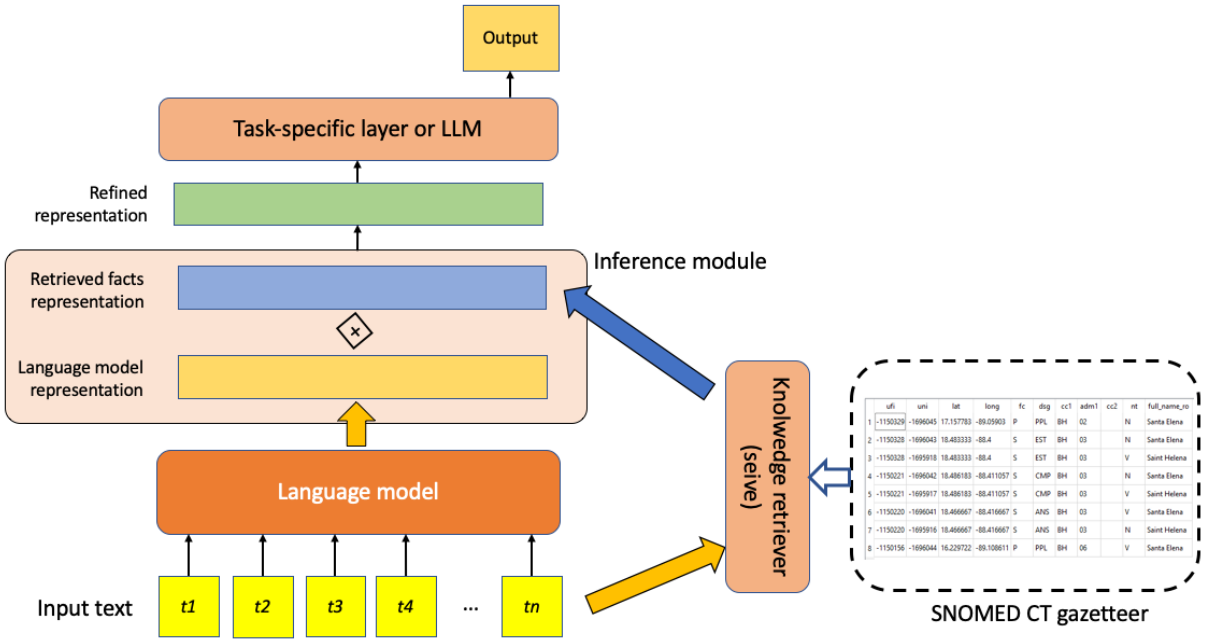
Figure 5. Integrating SNOMED CT into additional fusion modules.



Retrieval-Augmented Knowledge Fusion

In this approach, SNOMED CT was located outside LLMs as a fact-consulting knowledge base, injecting knowledge during inference (Figure 6). The module functioned as a gazetteer (dictionary), matching mentions in texts against the dictionary of SNOMED CT descriptions to filter out irrelevant entities from the models and map textual mentions to the most likely SNOMED CT concepts [24,26,37,55,56]. These methods primarily concentrated on entity recognition and question answering, capturing both textual semantic meanings and up-to-date real-world knowledge.

Figure 6. Retrieval-augmented knowledge fusion. LLM: large language model.



End Task and Performance Gain After SNOMED CT Integration

Overview

Most of the included studies (30/37, 81%) focused on NLU tasks, such as entity typing and classification. NLG tasks, including translation and summarization, were also attempted by a substantial number of studies (9/37, 24%), often involving various NLU pipelines before producing the final text output. Therefore, notably, works on NLU may also appear in the NLG category. Herein, we also compared the performance of models integrated with SNOMED CT to that of their counterparts without SNOMED CT integration.

NLU Tasks

Entity Extraction and Typing

Entity typing or named entity recognition tasks aim to detect specific types of entities by identifying the spans of their mentions in the text. These can be regarded as multiclassification tasks, where the number of classes is arbitrarily chosen by researchers. To fine-tune LLMs for type classification, authors annotated entities in texts by matching domain gazetteer strings (eg, "BIO" tagging scheme) [37,38,49] or using off-the-shelf automatic concept extractors [27]. The identified entities were then classified into human-annotated entity types [37,38] or topmost nodes in the SNOMED CT hierarchies [27,51]. In addition to typing individual entities, extraction and typing of relations between 2 entities were also attempted to align the detected entities with FHIR resources [25], such as protein to chemical and gene to disease [46] as well as disease to inflicted family members [35].

Many researchers did not conduct a comparative performance analysis of their SNOMED CT–integrated models against out-of-domain vanilla models. Among the few researchers who reported such comparisons, Jha and Zhang [46] demonstrated a gain in the F_1-score after the integration of SNOMED CT, while Montañés-Salas et al [37] found a positive impact only on recall (Table 4).

Table 4. Percentage performance gain in biomedical entity typing tasks after SNOMED CT integration into large language models.

Studies	F_1-score gain (%)	Precision gain (%)	Recall gain (%)	AUC[a] gain (%)
Montañés-Salas et al [37] (Best 2 model)	−0.11 (0.899→0.898)	−7.97 (0.928→0.854)	+8.60 (0.872→0.947)	—[b]
Jha and Zhang [46] (PubMedBERT on BC2GM)	+4.08 (0.80982→0.84287)	—	—	—

[a]AUC: area under the receiver operating characteristic curve.

[b]Not available.

Classification

We defined classification tasks as occurring at the sentence or document level, rather than at the word, entity, or phrase level. When classification tasks were implemented, semantic similarity [36] or the conditional probability of a positive case [22,33,53] was calculated, and the case was categorized as positive if the probability exceeded a threshold. Binary classification was performed to determine whether a sentence pair was entailed [33], contradictory [22,53], or similar [36]. Multilabel classification was conducted to categorize utterances by clinical

encounter components, such as symptoms, complaints, and medications [27]; social determinants of health [42]; or narrators' intent [48].

Table 5 shows the percentage performance gain after SNOMED CT integration in classification tasks. While Yadav et al [33] and Zhang et al [48] estimated the performance of their models based on the F_1-score, precision, and recall, Khosla et al [27] and Makhervaks et al [53] measured performance in terms of the area under the receiver operating characteristic curve, which improved by 0.87% to 14.83% after the integration of SNOMED CT. Chang et al [36] reported the Pearson correlation to assess clinical semantic textual similarity, and the incorporation of SNOMED CT into ClinicalBERT improved the performance of the model by 1.77% and 2.36% using cui2vec [73] and KG embeddings, respectively.

Table 5. Percentage performance gain in classification tasks after SNOMED CT integration into large language models.

Studies	F_1-score gain (%)	Precision gain (%)	Recall gain (%)	AUC[a] gain (%)	Accuracy gain (%)
Chopra et al [22]	—[b]	—	—	—	+0.99
Yadav et al [33]	+26.05 (0.4718→0.5947)	+36.87 (0.4616→0.6318)	+16.41 (0.4826→0.5618)	—	+17.27 (0.4790→0.5617)
Khosla et al [27]	—	—	—	+0.85 (0.468→0.472)	—
Zhang et al [48]					
BioBERT for intent detection	+1.15 (0.701→0.693)	—	—	—	—
Semantic matching for content recognition	—	−0.90 (1.000→0.991)	+12.15 (0.724→0.812)	—	—
Makhervaks et al [53]					
BERT based on MedNLI-General	—	—	—	+14.83 (0.661→0.759)	—
Bio-GPT on MedNLI-General	—	—	—	+10.34 (0.725→0.800)	—

[a]AUC: area under the receiver operating characteristic curve.

[b]Not available.

MCN Tasks

The most prominent end task in NLU was MCN, with 15 studies involved. MCN, the task of linking textual mentions to concepts in an ontology, provides a solution for unifying different ways of referring to the same concept. All the studies approached concept recognition as a multilabel classification task involving entity extraction and entity typing from words, phrases, or sentences. Models were trained on corpora annotated with SNOMED CT concepts and semantic types to identify concept mentions and generate a list of candidate SNOMED CT concepts that best match those mentions from testing texts. When training from annotated corpora was not available, MetaMap [74] was used to extract biomedical entities mentioned in free texts and map them to ontology concepts [25,26,35,50]. When candidate concepts were ranked, representation vectors of mentions and concept descriptions were generated, and their similarity was calculated using cosine similarity [31,35,44,45,54], linear transformation such as support vector classifiers [52], or softmax function [23,41,43]. In a more rule-oriented approach, Borchert and Schapranow [47] calculated weights based on semantic type and preferred term status from a gazetteer to reorder candidate lists. In other studies [24,26,50], sieve-based multipass entity linking systems [75] were used to rank the most likely concepts and achieved superior performance compared to neural classifiers.

Most of the studies observed positive gains in accuracy in MCN tasks after SNOMED CT integration (Table 6). Two authors reported the pre- and postintegration F_1-scores, recall values, and precision values and observed inconsistent results, with one reporting positive gains in the F_1-score and precision value and the other demonstrating a loss in the F_1-score and precision value after the integration of SNOMED CT.

Table 6. Percentage performance gain in medical concept normalization tasks after SNOMED CT integration into large language models.

Studies	F_1-score gain (%)	Precision gain (%)	Recall gain (%)	Accuracy gain (%)
Peterson et al [25]	−1.05 (0.95→0.94)	−1.04 (0.96→0.95)	0 (0.94→0.94)	—[a]
Wang et al [26] (vs training data dictionary with exact match, ignore order "yes")[b]	—	—	—	+27.36 (0.6013→0.7658)
Hristov et al [34]	—	—	—	+73.21 (0.56→0.97)
Dai et al (2021) [35]	—	—	—	+45.08 (0.417→0.605)
Xu and Miller [44] (on ShARe/CLEF 2013)	—	—	—	+0.68 (0.8333→0.8277)
Dong et al [54] (BLINKout on ShARe/CLEF 2013)	+5.87 (0.818→0.866)	+15.11 (0.741→0.853)	−3.62 (0.912→0.879)	+10.68 (0.777→0.860)

[a]Not available.

[b]The training data dictionary was constructed based on the Medical Concept Normalization corpus data. The SNOMED CT dictionary included the RxNorm dictionary.

NLG Tasks

Machine Translation

Several studies that participated in the WMT Biomedical Shared Task [76] described their methods for translating biomedical texts from various foreign languages, such as Spanish, French, German, and Chinese, as well as less-resourced languages, such as Basque, into English or vice versa. Transformer-based multilingual neural machine translation systems were the mainstream architectures, which were trained on dictionaries derived from SNOMED CT [28,30,39] or clinical notes artificially generated from SNOMED CT terminology contents [21,29].

The translation performance was reported using the Bilingual Evaluation Understudy (BLEU) score [77]. While most studies (4/5, 80%) presented improved BLEU scores by up to 131.66% [21] compared to their out-of-domain models, some studies (1/5, 20%) reported nonsuperior results [30] (Table 7).

Table 7. Performance comparison of biomedical translation tasks with and without SNOMED CT integration into large language models (LLMs).

Studies and translation direction	Performance on test data without SNOMED CT integration into an LLM (BLEU[a] score)	Performance on test data with SNOMED CT integration into an LLM (BLEU score)	BLEU score gain after SNOMED CT integration into an LLM (%)
Soto et al [21]			
Basque to Spanish	10.55	24.44	+131.66
Soto et al [30]			
Spanish to English	57.25	56.89	−0.63
English to Spanish	47.19	47.15	−0.08
Corral and Saralegi [29]			
English to Basque	12.85	13.61	+5.91
Peng et al [28]			
English to French	38.98	41.66	+6.88
French to English	38.31	38.44	+0.34
Wang et al [39]			
English to Italian	33.53	42.17	+25.77
Italian to English	36.43	43.72	+20.01
English to Portuguese	38.73	50.12	+29.41
Portuguese to English	41.84	54.74	+30.83
English to Russian	25.25	36.25	+43.56
Russian to English	39.76	47.09	+18.44

[a]BLEU: Bilingual Evaluation Understudy.

Text Summarization

For medical text summarization, encoder-decoder LLMs were used to process input embeddings and produce simplified texts. Pattisapu et al [32] primarily focused on the simplification of verbose sentences. They substituted biomedical mentions with UMLS-preferred names and tokenized them at the subword level to produce noisy input sentences for training. In contrast, Searle et al [57] summarized entire hospital encounters into a few sentences by ranking the most salient ones to constitute the summary. To address the hallucination problem arising from LLMs, authors used SNOMED CT semantic tags of the extracted biomedical terms to configure guidance signals for clinical problems and interventions.

Recall-Oriented Understudy for Gisting Evaluation recall [78] measures how many n-grams in the source text appear in the summarization. Pattisapu et al [32] reported no gain in ROUGE recall when incorporating SNOMED CT into NLP pipelines. Searle et al [57] presented ROUGE-F_1, a harmonized measure of the recall and precision for ROUGE, and observed improvements by 3.6% (from 11.1 to 11.5) and 48.84% (from 8.6 to 12.8) on the Medical Information Mart for Intensive Care III and King's College Hospital corpora, respectively, after incorporating SNOMED CT.

Question Answering and Generation

Generating answers for short-answer or essay questions, as opposed to multiple-choice questions, can be classified as NLG. The task of question answering may involve preliminary NLU pipelines, such as intent and content recognition. Zhang et al [48] developed a clinical communication training dialogue system incorporated with SNOMED CT synonyms for the augmentation of textual data and BioBERT for intent recognition. They qualitatively evaluated the performance of the conversation system using scales rated by physicians from 29 training records, which indicated a comparable precision as clinical experts.

Discussion

LLMs and SNOMED CT

In this scoping review, we observed that BERT was the mainstream LLM integrated with SNOMED CT. Considering the significant time required to publish state-of-the-art methodologies, especially in peer-reviewed journals [79], it is unsurprising that more recent inventions, such as GPT-3.5 and BART, were less prevalent in articles published from 2018 to 2023. Researchers in this field exploited biomedically oriented BERT variants, such as BioBERT and PubMedBERT, reflecting the need for biomedical tasks to be trained or fine-tuned on specialized corpora [16]. However, due to privacy and confidentiality concerns, there is a dearth of clinical documents and patient notes, making it difficult to sufficiently train biomedical LLMs to an extent comparable to those in the general domain [80]. SNOMED CT can supplement or even substitute biomedical pretraining corpora, addressing the chronic shortage, as noted in this review. A substantial number of studies included in this review used SNOMED CT to expand pretraining corpora by concatenating synonyms or relations in documents or generating synthetic texts based on SNOMED CT descriptions or relations.

We identified 3 approaches to incorporating SNOMED CT into LLMs: LLM input, additional fusion modules, and knowledge retriever, with the former 2 intervening in the pretraining process of LLMs. While either lexical or graph information from SNOMED CT could be incorporated into the pretraining stage, the lexicon of SNOMED CT descriptions was the predominant form of integration. This underscores that SNOMED CT chiefly introduces synonym information to LLMs, yet relation information remains underused in NLP research. The advantage of SNOMED CT in defining relations between biomedical entities through semantic networks needs to be adopted for more sophisticated tasks such as knowledge inference and validation and highlighted within the biomedical NLP research community.

End Tasks and Performance Reports

A significant number of studies included in this review engaged in the concept recognition process from free texts, whether as the final task or an intermediate step for subsequent tasks. Recognizing and extracting SNOMED CT concepts from the unstructured sections of EHRs is becoming crucial in clinical settings, where substantial patient information, such as social history and socioeconomic status, remains untapped in free-text clinical notes [81]. Leveraging previously unrepresented SNOMED CT concepts from free-text clinical data holds great potential in significantly enhancing clinical care and research, especially in the era of smart applications where patient-generated data can be integrated into EHRs through the representation of patient-authored texts with SNOMED CT concepts [82].

Only a small fraction of the included models disclosed performance comparisons before and after SNOMED CT integration. For example, only 6 (40%) out of 15 studies on MCN tasks provided information about the gain in the F_1-scores or accuracy after SNOMED CT incorporation. This suggests that many biomedical NLP researchers do not focus on the role of SNOMED CT or other ontologies in improving their models. Moreover, some authors chose to demonstrate only selected metrics, potentially leading to publication bias that favors improved performance at first glance. In our review, we identified 7 studies that presented only 1 metric without disclosing others (excluding those that reported only the BLEU score, which is widely recognized as the best metric for measuring translation performance). This focus on a single metric may encourage researchers to optimize their models for that metric, potentially leading to underperformance in other areas. The NLP community needs to propose standardized methods for presenting performance and, if possible, develop new metrics that better reflect the specifics of NLU and NLG tasks performed by LLMs.

Implications for Future Endeavors

The knowledge-intensive approaches to enhancing LMs, which are often renounced by those favoring deep learning–based approaches, still comprise a small portion of the artificial intelligence research community. However, in the face of immense computational power and the availability of data

required by LLMs and deep learning–based systems, an increasing number of researchers now advocate the harmonization of the 2 approaches [83], and a plethora of KG-enhanced LLMs is developed in the general domain [10,84]. In addition to improving the performance of artificial intelligence models, ontologies and human-curated knowledge bases can address the explainability and controllability of artificial intelligence, probing facts within the human-interpretable form of system architectures [85]. Exploring the trade-offs in combining the 2 approaches is anticipated to contribute toward trustworthy and reliable artificial intelligence.

Among various biomedical terminology systems and ontologies, SNOMED CT was the primary focus in this review as a KG integrated with LLMs. Although the UMLS continues to dominate NLP research in the biomedical domain [16], SNOMED CT has the potential to expand its influence, given its governance over the health care industry. Consequently, the use of SNOMED CT as a reliable knowledge source becomes more feasible, considering its presence in various EHR systems or common data models. While this review did not identify real-world SNOMED CT–incorporated LLM applications directly tied to EHR systems, SNOMED CT is implicitly expected to support these systems as a standardized terminology system bound to syntactic interoperability structures such as FHIR and OpenEHR. In addition, medical institutions already implementing SNOMED CT in their EHR systems are anticipated to incorporate LLM applications and use SNOMED CT at the point of care [86]. Explicit descriptions of SNOMED CT in technical specifications or scientific papers by developers of these applications would have been valuable to include in this review.

Limitations

One of the limitations of this scoping review is that we examined LLMs that accepted SNOMED CT only as a working ontology, leaving other biomedical ontologies out of our scope. To the best of our knowledge, however, there is no comprehensive review of the use of other biomedical ontologies within LLMs. The queries used in this review, especially the first one, retrieved articles that used a variety of biomedical ontologies, such as the UMLS, Medical Subject Headings, Gene Ontology, and Medical Wikidata. We chose to limit the scope of our review to SNOMED CT due to the heterogeneity of components among different ontology systems and the difficulty in delineating the contributions of each ontology in a standardized way. A more consolidated analysis of different ontologies used within LLMs awaits more comprehensive work.

A significant proportion of the included studies (23/37, 62%) were retrieved from conference proceedings. While we excluded short abstract articles and included only those that provided sufficient information to be categorized by our preset features, interested readers might find it challenging to delve into detailed methodologies from these proceedings articles. However, many of these papers refer to additional materials, such as GitHub (GitHub, Inc) repositories, to provide raw data and source codes; for example, Khosla et al [27] provided the source code of their system on GitHub [87]. We encourage more studies to share additional materials on open developer platforms to enhance methodology transparency and accelerate NLP research.

Another limitation of this review is that we could not conclude on how the integration of SNOMED CT improved the performance of LLMs. While most of the studies (14/18, 78%) observed a positive impact on performance after SNOMED CT integration, their statistical significance was not indicated. Moreover, the diversity of evaluation methods prevented us from performing a meta-analysis across all the included studies. While we examined whether SNOMED CT integration improved LLM performance by presenting percentage gains across various metrics, these results are prone to being misleading due to potential publication bias and the insufficient number of included studies. Nevertheless, this before-and-after comparison method, often adopted for comparative studies, effectively measures the effect of interventions (SNOMED CT in our case) within a single group or entity [88]. To control for confounding factors, we excluded models whose performance differences could be attributable to modalities other than SNOMED CT integration. For example, we excluded the study by Zotova et al [40] from our analysis because their performance might have been affected by the use of a different testing corpus. An evenhanded testing bed, such as a shared task competition under a single testing method requiring all participants to report performance differences before and after KG integration, could provide a controlled evaluation to reliably and objectively measure the contributions of KGs.

Conclusions

In conclusion, this scoping review explored the methodologies and effectiveness of integrating SNOMED CT into LLMs. The predominant approach involved using SNOMED CT concept descriptions or graph embeddings as inputs for LM encoders, many of which were involved in MCN tasks. The endeavor to identify and extract SNOMED CT concepts from free texts was proven to be instrumental in enhancing the understanding and generation of NL texts for downstream tasks in the biomedical realm. However, our study revealed both a lack of standardized methods for assessing KG integration into LLMs and a scarcity of explicit performance reporting in existing research, highlighting significant gaps in current evaluation practices. These findings underline the need for more consistent reporting and evaluation practices in this field of research. Future research is anticipated to be more aware of the advantage of SNOMED CT when incorporating it into LLMs and to report findings in a manner that facilitates comparison across different works.

Acknowledgments

This work was supported by the National Research Foundation of Korea grant funded by the Republic of Korea government (Ministry of Science and Information and Communication Technology; RS-2024-00354718).

Data Availability

The data analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 134 KB - medinform_v12i1e62924_app1.pdf]

Multimedia Appendix 2
Brief introduction to large language models.
[PDF File (Adobe PDF File), 412 KB - medinform_v12i1e62924_app2.pdf]

Multimedia Appendix 3
Summary of the included studies.
[XLSX File (Microsoft Excel File), 61 KB - medinform_v12i1e62924_app3.xlsx]

References

1. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018. [doi: 10.48550/arXiv.1810.04805]
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online on May 28, 2020. [doi: 10.48550/arXiv.2005.14165]
3. Chen Q, Li FL, Xu G, Yan M, Zhang J, Zhang Y. DictBERT: dictionary description knowledge enhanced language model pre-training via contrastive learning. arXiv Preprint posted online on August 1, 2022. [doi: 10.48550/arXiv.2208.00635]
4. Hou Y, Jiao W, Liu M, Allen C, Tu Z, Sachan M. Adapters for enhanced modeling of multilingual knowledge and text. arXiv Preprint posted online on October 24, 2022. [doi: 10.48550/arXiv.2210.13617]
5. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Mar 03;55(12):1-38. [doi: 10.1145/3571730]
6. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv Preprint posted online on November 9, 2023. [doi: 10.48550/arxiv.2311.05232]
7. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. IEEE Trans Knowl Data Eng 2024 Jul;36(7):3580-3599. [doi: 10.1109/tkde.2024.3352100]
8. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv Preprint posted online on January 28, 2022. [doi: 10.48550/arXiv.2201.11903]
9. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020 Presented at: NIPS'20; December 6-12, 2020; Vancouver, BC.
10. Hu L, Liu Z, Zhao Z, Hou L, Nie L, Li J. A survey of knowledge enhanced pre-trained language models. IEEE Trans Knowl Data Eng 2024 Apr;36(4):1413-1430. [doi: 10.1109/tkde.2023.3310002]
11. Lawrence P. Knowledge graphs + large language models = the ability for users to ask their own questions? Medium. 2023 Mar 31. URL: https://medium.com/@peter.lawrence_47665/knowledge-graphs-large-language-models-the-ability-for-users-to-ask-their-own-questions-e4afc348fa72 [accessed 2023-12-30]
12. Anand V, Ramesh R, Jin B, Wang Z, Lei X, Lin CY. MultiModal language modelling on knowledge graphs for deep video understanding. In: Proceedings of the 29th ACM International Conference on Multimedia. 2021 Presented at: MM '21; October 20-24, 2021; Virtual Event, China. [doi: 10.1145/3474085.3479220]
13. Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press; 1998.
14. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
15. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
16. Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z, et al. Pre-trained language models in biomedical domain: a systematic survey. ACM Comput Surv 2023 Oct 05;56(3):1-52. [doi: 10.1145/3611651]

17. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. J Am Med Inform Assoc 2021 Aug 13;28(9):2017-2026 [FREE Full text] [doi: 10.1093/jamia/ocab084] [Medline: 34151978]
18. Posnack S, Barker W. The heat is on: US caught FHIR in 2019. Health IT Buzz. 2021 Jul 29. URL: https://www.healthit.gov/buzz-blog/health-it/the-heat-is-on-us-caught-fhir-in-2019 [accessed 2023-12-30]
19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
20. Min B, Ross H, Sulem E, Veyseh AP, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv 2023 Sep 14;56(2):1-40. [doi: 10.1145/3605943]
21. Soto X, Perez-De-Vinaspre O, Oronoz M, Labaka G. Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. In: Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation. 2019 Presented at: MomenT@MTSummit 2019; August 19-23, 2019; Dublin, Ireland. [doi: 10.1093/jamia/ocz110]
22. Chopra S, Gupta A, Kaushik A. MSIT_SRIB at MEDIQA 2019: knowledge directed multi-task framework for natural language inference in clinical domain. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP@ACL 2019; August 1, 2019; Florence, Italy. [doi: 10.18653/v1/w19-5052]
23. Li F, Jin Y, Liu W, Rawat BP, Cai P, Yu H. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
24. Xu D, Gopale M, Zhang J, Brown K, Begoli E, Bethard S. Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)-based ranking for concept normalization. J Am Med Inform Assoc 2020 Oct 01;27(10):1510-1519 [FREE Full text] [doi: 10.1093/jamia/ocaa080] [Medline: 32719838]
25. Peterson KJ, Jiang G, Liu H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. J Biomed Inform 2020 Oct;110:103541 [FREE Full text] [doi: 10.1016/j.jbi.2020.103541] [Medline: 32814201]
26. Wang Y, Hur B, Verspoor K, Baldwin T. A multi-pass sieve for clinical concept normalization. Traitement Automatique Des Langues 2020;61(2) [FREE Full text]
27. Khosla S, Vashishth S, Lehman JF, Rose C. MedFilter: improving extraction of task-relevant utterances through integration of discourse structure and ontological knowledge. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP 2020; November 16-20, 2020; Online. [doi: 10.18653/v1/2020.emnlp-main.626]
28. Peng W, Liu J, Wang M, Li L, Meng X, Yang H, et al. Huawei's submissions to the WMT20 biomedical translation task. In: Proceedings of the Fifth Conference on Machine Translation. 2020 Presented at: WMT@EMNLP 2020; November 19-20, 2020; Online.
29. Corral A, Saralegi X. Elhuyar submission to the biomedical translation task 2020 on terminology and abstracts translation. In: Proceedings of the Fifth Conference on Machine Translation. 2020 Presented at: WMT@EMNLP 2020; November 19-20, 2020; Online.
30. Soto X, Perez-de-Vinaspre O, Labaka G, Oronoz M. Ixamed's submission description for WMT20 Biomedical shared task: benefits and limitations of using terminologies for domain adaptation. In: Proceedings of the Fifth Conference on Machine Translation. 2020 Presented at: WMT@EMNLP 2020; November 19-20, 2020; Online.
31. Kalyan KS, Sangeetha S. Target concept guided medical concept normalization in noisy user-generated texts. In: Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. 2020 Presented at: DeeLIO 2020; November 19-20, 2020; Online. [doi: 10.18653/v1/2020.deelio-1.8]
32. Pattisapu N, Prabhu N, Bhati S, Varma V. Leveraging social media for medical text simplification. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020 Presented at: SIGIR '20; July 25-30, 2020; Virtual Event. [doi: 10.1145/3397271.3401105]
33. Yadav S, Pallagani V, Sheth A. Medical knowledge-enriched textual entailment framework. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020 Presented at: COLING 2020; December 8-13, 2020; Online. [doi: 10.18653/v1/2020.coling-main.161]
34. Hristov A, Tahchiev A, Papazov H, Tulechki N, Primov T, Boytcheva S. Application of deep learning methods to SNOMED CT encoding of clinical texts: from data collection to extreme multi-label text-based classification. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. 2021 Presented at: RANLP 2021; September 1-3, 2021; Online. [doi: 10.26615/978-954-452-072-4_063]
35. Dai X, Rybinski M, Karimi S. SearchEHR: a family history search system for clinical decision support. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021 Presented at: CIKM '21; November 1-5, 2021; Virtual Event. [doi: 10.1145/3459637.3481986]
36. Chang D, Lin E, Brandt C, Taylor RA. Incorporating domain knowledge into language models by using graph convolutional networks for assessing semantic textual similarity: model development and performance comparison. JMIR Med Inform 2021 Nov 26;9(11):e23101 [FREE Full text] [doi: 10.2196/23101] [Medline: 34842531]

37. Montañés-Salas RM, López-Bosque I, García-Garcés L, del-Hoyo-Alonso R. ITAINNOVA at SocialDisNER: a transformers cocktail for disease identification in social media in Spanish. In: Proceedings of the 29th International Conference on Computational Linguistic. 2022 Presented at: COLING 2022; October 12-17, 2022; Gyeongju, South Korea.
38. Ying H, Luo S, Dang T, Yu S. Label refinement via contrastive learning for distantly-supervised named entity recognition. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2022 Presented at: NAACL 2022; July 10-15, 2022; Seattle, WA. [doi: 10.18653/v1/2022.findings-naacl.203]
39. Wang W, Meng X, Yan S, Tian Y, Peng W. Huawei BabelTar NMT at WMT22 biomedical translation task: how we further improve domain-specific NMT. In: Proceedings of the Seventh Conference on Machine Translation. 2022 Presented at: WMT 2022; December 7-8, 2022; Abu Dhabi, United Arab Emirates.
40. Zotova E, Cuadros M, Rigau G. ClinIDMap: towards a clinical IDs mapping for data interoperability. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022 Presented at: LREC 2022; June 20-25, 2022; Marseille, France.
41. Tang G, Liu T, Cai X, Gao S, Fu L. Standardization of clinical terminology based on hybrid recall and Ernie. In: Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences. 2022 Presented at: ISAIMS '22; October 13-15, 2022; Amsterdam, The Netherlands. [doi: 10.1145/3570773.3570782]
42. Han S, Zhang RF, Shi L, Richie R, Liu H, Tseng A, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. J Biomed Inform 2022 Mar;127:103984 [FREE Full text] [doi: 10.1016/j.jbi.2021.103984] [Medline: 35007754]
43. Chen Y, Hu D, Li M, Duan H, Lu X. Automatic SNOMED CT coding of Chinese clinical terms via attention-based semantic matching. Int J Med Inform 2022 Mar;159:104676. [doi: 10.1016/j.ijmedinf.2021.104676] [Medline: 34990940]
44. Xu D, Miller T. A simple neural vector space model for medical concept normalization using concept embeddings. J Biomed Inform 2022 Jun;130:104080 [FREE Full text] [doi: 10.1016/j.jbi.2022.104080] [Medline: 35472514]
45. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform 2022 Feb;126:103983 [FREE Full text] [doi: 10.1016/j.jbi.2021.103983] [Medline: 34990838]
46. Jha K, Zhang A. Continual knowledge infusion into pre-trained biomedical language models. Bioinformatics 2022 Jan 03;38(2):494-502. [doi: 10.1093/bioinformatics/btab671] [Medline: 34554186]
47. Borchert F, Schapranow MP. HPI-DHC @ BioASQ DisTEMIST: Spanish biomedical entity linking with pre-trained transformers and cross-lingual candidate retrieval. In: Proceedings of the Conference and Labs of the Evaluation Forum. 2022 Presented at: CLEF 2022; September 5-8, 2022; Bologna, Italy.
48. Zhang X, Yu BX, Liu Y, Chen G, Wing-Yiu Ng G, Chia NH, et al. Conversational system for clinical communication training supporting user-defined tasks. In: Proceedings of the IEEE International Conference on Teaching, Assessment and Learning for Engineering. 2022 Presented at: TALE 2022; December 4-7, 2022; Hung Hom, Hong Kong. [doi: 10.1109/tale54877.2022.00071]
49. Morine MJ, Priami C, Coronado E, Haber J, Kaput J. A comprehensive and holistic health database. In: Proceedings of the IEEE International Conference on Digital Health. 2022 Presented at: ICDH 2022; July 10-16, 2022; Barcelona, Spain. [doi: 10.1109/icdh55609.2022.00039]
50. Li L, Zhai Y, Gao J, Wang L, Hou L, Zhao J. Stacking-BERT model for Chinese medical procedure entity normalization. Math Biosci Eng 2023 Jan;20(1):1018-1036 [FREE Full text] [doi: 10.3934/mbe.2023047] [Medline: 36650800]
51. Llorca I, Borchert F, Schapranow MP. A meta-dataset of german medical corpora: harmonization of annotations and cross-corpus NER evaluation. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. 2023 Presented at: ClinicalNLP@ACL 2023; July 14, 2023; Toronto, ON. [doi: 10.18653/v1/2023.clinicalnlp-1.23]
52. Hristov A, Ivanov P, Aksenova A, Asamov T, Gyurov P, Primov T, et al. Clinical text classification to SNOMED CT codes using transformers trained on linked open medical ontologies. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. 2023 Presented at: RANLP 2023; September 4-6, 2023; Varna, Bulgaria. [doi: 10.26615/978-954-452-092-2_057]
53. Makhervaks D, Gillis P, Radinsky K. Clinical contradiction detection. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023 Presented at: EMNLP 2023; December 6-10, 2023; Singapore, Singapore. [doi: 10.18653/v1/2023.emnlp-main.80]
54. Dong H, Chen J, He Y, Liu Y, Horrocks I. Reveal the unknown: out-of-knowledge-base mention discovery with entity linking. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023 Presented at: CIKM '23; October 21-25, 2023; Birmingham, UK. [doi: 10.1145/3583780.3615036]
55. Xu J, Mazwi M, Johnson AE. AnnoDash, a clinical terminology annotation dashboard. JAMIA Open 2023 Jul 08;6(3):ooad046 [FREE Full text] [doi: 10.1093/jamiaopen/ooad046] [Medline: 37425489]
56. Liu F, Liu M, Li M, Xin Y, Gao D, Wu J, et al. Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction. Quant Imaging Med Surg 2023 Jun 01;13(6):3873-3890 [FREE Full text] [doi: 10.21037/qims-22-1158] [Medline: 37284084]

57. Searle T, Ibrahim Z, Teo J, Dobson RJ. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. J Biomed Inform 2023 May;141:104358 [FREE Full text] [doi: 10.1016/j.jbi.2023.104358] [Medline: 37023846]
58. Liu Z, Lin W, Shi Y, Zhao J. A robustly optimized BERT pre-training approach with post-training. In: Proceedings of the 20th China National Conference on Chinese Computational Linguistics. 2021 Presented at: CCL 2021; August 13-15, 2021; Hohhot, China. [doi: 10.1007/978-3-030-84186-7_31]
59. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv Preprint posted online on September 26, 2019. [doi: 10.48550/arXiv.1909.11942]
60. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
61. Alsentzer E, Murph J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP 2019; June 7, 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1909]
62. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021 Presented at: NAACL-HLT 2021; June 6-11, 2021; Online. [doi: 10.18653/v1/2021.naacl-main.334]
63. Bressem KK, Papaioannou JM, Grundmann P, Borchert F, Adams LC, Liu L, et al. medBERT.de: a comprehensive German BERT model for the medical domain. Expert Syst Appl 2024 Mar 01;237:121598. [doi: 10.1016/j.eswa.2023.121598]
64. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022 Nov 19;23(6):bbac409. [doi: 10.1093/bib/bbac409] [Medline: 36156661]
65. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: ACL 2020; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.703]
66. Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: a corpus of adverse drug event annotations. J Biomed Inform 2015 Jun;55:73-81 [FREE Full text] [doi: 10.1016/j.jbi.2015.03.010] [Medline: 25817970]
67. Zolnoori M, Fung KW, Patrick TB, Fontelo P, Kharrazi H, Faiola A, et al. The PsyTAR dataset: from patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. Data Brief 2019 Mar 15;24:103838 [FREE Full text] [doi: 10.1016/j.dib.2019.103838] [Medline: 31065579]
68. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP 2018; October 31-November 4, 2015; Brussels, Belgium. [doi: 10.18653/v1/d18-1187]
69. Mohan S, Li D. MedMentions: a large biomedical corpus annotated with UMLS concepts. arXiv Preprint posted online on February 25, 2019. [doi: 10.48550/arxiv.1902.09476]
70. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: Proceedings of the 4th International Conference of the CLEF Initiative onInformation Access Evaluation. Multilinguality, Multimodality, and Visualization. 2013 Presented at: CLEF 2013; September 23-26, 2013; Valencia, Spain. [doi: 10.1007/978-3-642-40802-1_24]
71. Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings. arXiv Preprint posted online on May 6, 2017. [doi: 10.48550/arXiv.1705.02426]
72. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv Preprint posted online on September 9, 2016. [doi: 10.48550/arXiv.1609.02907]
73. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, et al. Clinical concept embeddings learned from massive sources of multimodal medical data. Biocomputing 2019:295-306 [FREE Full text] [doi: 10.1142/9789811215636_0027]
74. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
75. D'Souza J, Ng V. Sieve-based entity linking for the biomedical domain. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015 Presented at: ACL 2015; July 26-31, 2015; Beijing, China. [doi: 10.3115/v1/p15-2049]
76. Barrault L, Biesialska M, Bojar O, Costa-jussà MR, Federmann C, Graham Y, et al. Findings of the 2020 conference on machine translation (WMT20). In: Proceedings of the Fifth Conference on Machine Translation. 2020 Presented at: WMT 2020; November 19-20, 2020; Online. [doi: 10.18653/v1/w19-5301]
77. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002 Presented at: ACL '02; July 7-12, 2002; Philadelphia, PA. [doi: 10.3115/1073083.1073135]
78. Lin CY, Och FJ. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004 Presented at: ACL '04; July 21-26, 2004; Barcelona, Spain. [doi: 10.3115/1218955.1219032]

79. Björk BC, Solomon D. The publishing delay in scholarly peer-reviewed journals. J Informetr 2013 Oct;7(4):914-923. [doi: 10.1016/j.joi.2013.09.001]
80. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
81. Jonnagaddala J, Liaw ST, Ray P, Kumar M, Chang NW, Dai HJ. Coronary artery disease risk assessment from unstructured electronic health records using text mining. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S203-S210 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.003] [Medline: 26319542]
82. Sezgin E, Hussain SA, Rust S, Huang Y. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. JMIR Form Res 2023 Mar 07;7:e43014 [FREE Full text] [doi: 10.2196/43014] [Medline: 36881467]
83. Humm BG, Archer P, Bense H, Bernier C, Goetz C, Hoppe T, et al. New directions for applied knowledge-based AI and machine learning. Informatik Spektrum 2022 Dec 30;46(2):65-78. [doi: 10.1007/S00287-022-01513-9]
84. Yang L, Chen H, Li Z, Ding X, Wu X. Give us the facts: enhancing large language models with knowledge graphs for fact-aware language modeling. IEEE Trans Knowl Data Eng 2024 Jul;36(7):3091-3110. [doi: 10.1109/tkde.2024.3360454]
85. Confalonieri R, Del Prado FM, Agramunt S, Malagarriga D, Faggion D, Weyde T, et al. An ontology-based approach to explaining artificial neural networks. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2019 Presented at: ECML PKDD 2019; September 16-20, 2019; Würzburg, Germany.
86. Farfán Sedano FJ, Terrón Cuadrado M, García Rebolledo EM, Castellanos Clemente Y, Serrano Balazote P, Gómez Delgado A. Implementation of SNOMED CT to the medicines database of a general hospital. Stud Health Technol Inform 2009;148:123-130. [Medline: 19745242]
87. sopankhosla / MedFilter. GitHub. URL: https://github.com/sopankhosla/MedFilter [accessed 2024-06-04]
88. Sterne JA, Hernán MA, McAleenan A, Reeves BC, Higgins JP. Chapter 25: assessing risk of bias in a non-randomized study. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5. London, UK: The Cochrane Collaboration; 2024.

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
BLEU: Bilingual Evaluation Understudy
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resource
KG: knowledge graph
LLM: large language model
LM: language model
MCN: Medical Concept Normalization
NL: natural language
NLG: natural language generation
NLP: natural language processing
NLU: natural language understanding
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 04.06.24; peer-reviewed by S Zhu, T Karen, HJT van Mens, C Gaudet-Blavignac; comments to author 03.07.24; revised version received 22.07.24; accepted 15.09.24; published 07.10.24.

Please cite as:
Chang E, Sung S
Use of SNOMED CT in Large Language Models: Scoping Review
JMIR Med Inform 2024;12:e62924
URL: https://medinform.jmir.org/2024/1/e62924
doi:10.2196/62924
PMID:39374057

Review

Application of Spatial Analysis on Electronic Health Records to Characterize Patient Phenotypes: Systematic Review

Abolfazl Mollalo[1], PhD; Bashir Hamidi[1], MPH; Leslie A Lenert[1], MD; Alexander V Alekseyenko[1], PhD

Biomedical Informatics Center, Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Abolfazl Mollalo, PhD
Biomedical Informatics Center
Department of Public Health Sciences
Medical University of South Carolina
22 Westedge Street
Suite 200
Charleston, SC, 29403
United States
Phone: 1 8437922970
Email: mollalo@musc.edu

Abstract

Background: Electronic health records (EHRs) commonly contain patient addresses that provide valuable data for geocoding and spatial analysis, enabling more comprehensive descriptions of individual patients for clinical purposes. Despite the widespread use of EHRs in clinical decision support and interventions, no systematic review has examined the extent to which spatial analysis is used to characterize patient phenotypes.

Objective: This study reviews advanced spatial analyses that used individual-level health data from EHRs within the United States to characterize patient phenotypes.

Methods: We systematically evaluated English-language, peer-reviewed studies from the PubMed/MEDLINE, Scopus, Web of Science, and Google Scholar databases from inception to August 20, 2023, without imposing constraints on study design or specific health domains.

Results: A substantial proportion of studies (>85%) were limited to geocoding or basic mapping without implementing advanced spatial statistical analysis, leaving only 49 studies that met the eligibility criteria. These studies used diverse spatial methods, with a predominant focus on clustering techniques, while spatiotemporal analysis (frequentist and Bayesian) and modeling were less common. A noteworthy surge (n=42, 86%) in publications was observed after 2017. The publications investigated a variety of adult and pediatric clinical areas, including infectious disease, endocrinology, and cardiology, using phenotypes defined over a range of data domains such as demographics, diagnoses, and visits. The primary health outcomes investigated were asthma, hypertension, and diabetes. Notably, patient phenotypes involving genomics, imaging, and notes were limited.

Conclusions: This review underscores the growing interest in spatial analysis of EHR-derived data and highlights knowledge gaps in clinical health, phenotype domains, and spatial methodologies. We suggest that future research should focus on addressing these gaps and harnessing spatial analysis to enhance individual patient contexts and clinical decision support.

(JMIR Med Inform 2024;12:e56343) doi:10.2196/56343

KEYWORDS

clinical phenotypes; electronic health records; geocoding; geographic information systems; patient phenotypes; spatial analysis

Introduction

Electronic health records (EHRs) have significantly enriched clinical decision support by providing relatively cost-effective, time-efficient, and convenient sources of a large population of patient records [1,2]. Because EHRs often contain patient addresses, spatial analysis can enable value addition via high-resolution geocoding. The simplest of such analyses may be mapping, which can promote a better understanding of health disparities. Further, patient geocoding can link external data such as environmental, demographic, and socioeconomic factors for more refined patient phenotyping and a more profound

understanding of patient exposures for targeted interventions [3].

The possibilities for applying spatial analysis on individual-level, EHR-derived data are beyond geocoding, basic mapping, or external data linkage. For instance, spatial network analysis examines proximity to the sources of pollution [4], measures accessibility to health care facilities [5], and optimizes resource allocations to mitigate health disparities [6]. Spatial clustering pinpoints statistically significant spatial and spatiotemporal hotspots and cold spots [7], especially when considering longitudinal EHRs data. Moreover, spatial and spatiotemporal modeling can identify localized patterns, trends, and relationships within a specific region [8,9]. Identifying underserved communities through spatial analysis can enhance clinical decision support to implement targeted interventions such as screening, vaccination, or health education campaigns.

Despite the availability of advanced spatial analysis methods, most studies primarily focus on basic mapping or geocoding. Moreover, while these methodologies have the potential to better describe the context of individual patients in biomedical studies, there is a need for their improved application to derive more meaningful insights. To accurately address medical conditions, identify a disease in a patient, and scale that to cohorts of patients, phenotyping is required [10]. Phenotypes are a combination of observable traits, symptoms, and characteristics. They can contain inclusion and exclusion criteria (eg, diagnoses, procedures, laboratory reports, and medications) and can be used to recruit patients who fit the necessary criteria for clinical trials.

A prior systematic review used spatially linked EHRs data to investigate the effects of social, physical, and built environments on health outcomes [11]. Another study highlighted the need to integrate spatial data related to individual patients into health care decision-making and practice [12]. Nonetheless, this is the first comprehensive study that systematically reviews US-based studies that used spatial analysis for analyzing EHR-derived data in characterizing patient phenotypes for clinical decision support and interventions. This review collates and synthesizes existing literature that used individual-level health data from EHRs in conjunction with advanced spatial analyses and patient phenotyping. Thus, the main objectives of this review are (1) to evaluate the degree to which advanced spatial methods are currently being used with individual-level data sourced from EHRs in the United States, (2) to identify areas of spatial analyses most applicable to biomedical studies, (3) to categorize publications concerning their biomedical and clinical areas and the specific patient phenotypes they target, and (4) to highlight knowledge gaps and propose future research directions for harnessing the potential of spatial analysis to enhance the context of individual-level data sourced from EHRs for biomedical studies.

Methods

Overview

This systematic review was performed using the protocols outlined by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) to identify the studies that satisfy the eligibility criteria for subsequent data extraction and synthesis (Multimedia Appendix 1).

Data Source

A comprehensive search for peer-reviewed studies was carried out using abstracts and titles screening within the PubMed/MEDLINE, Scopus, and Web of Science databases using the search terms in Table 1. The search was conducted on August 29, 2023, without limitations on study design or specific health domains.

Table 1. The search strategy key terms.

Theme[a]	Key terms
Spatial analysis	(“Geospatial*” OR “Geo-spatial*” OR “Spatio-Temporal” OR “Spatial Temporal” OR “Space-Time” OR “Space Time” OR “Spatiotemporal” OR “Geocod*” OR “ Spatial Autocorrelation” OR “Spatial Interpolation” OR “Spatial Epidemiology” OR “Spatial Data” OR “Spatial Modeling” OR “Spatial Modelling” OR “Spatial Mapping” OR “Geographic Mapping” OR “Georeferenc*” OR “Spatial Analys*” OR “Spatial Inequalit*” OR “Spatial Disparit*” OR “Spatial Dependenc*” OR “Spatial Access*” OR “Geographical Mapping” OR “Geographical Visualization” OR “Geographic Visualization” OR “Geovisualization” OR “Geographical Information System*” OR “Geographic Information System*” OR “Geofencing” OR “Geographical Distribution*” OR “Geographic Distribution*” OR “Spatial Statistic*” OR “Spatial Bayesian” OR “Spatial Hotspot*” OR “Spatial Cluster*” OR “Geographic Cluster*” OR “Geographic Hotspot*” OR “Remote Sensing” OR “Global Positioning System” OR “Spatial Pattern*” OR “Spatial Data Mining” OR “Spatial Variabilit*” OR “Spatial Heterogeneit*” OR “Geostatistic*” OR “Spatial Covariance” OR “Spatial Regression” OR “Spatial Uncertaint*” OR “Spatial Point Pattern*” OR “Kriging” OR “Cartography” OR “Spatial Decision Support System*” OR “OpenStreetMap” OR “Location-Based Services” OR “Spatial Quer*” OR “GIS” OR “Web GIS” OR “Satellite Imager*” OR “ArcGIS” OR “QGIS” OR “Risk Mapping”) AND
EHR[b]	(“EHR” OR “EMR” OR “EPR” OR “Electronic Health Record*” OR “Electronic Medical Record*” OR “Electronic Patient Record*” OR “EDW” OR “Enterprise Data Warehouse” OR “RDW” OR “Research Data Warehouse”)

[a]The selected studies that used spatial analysis of EHR data were manually excluded if they lacked patient phenotype characteristics or were not conducted based on the US data.

[b]EHR: electronic health record.

Search Strategy

The initial search comprised 2 main categories. The first category included a broad set of key terms related to spatial analysis. The second category used the key terms associated with EHR. Henceforth, our reference to EHRs will also encompass electronic medical records (EMRs), electronic patient records (EPRs), enterprise data warehouses (EDWs), and research data warehouses (RDWs). The Boolean operator AND was applied to synthesize the 2 categories.

For PubMed/MEDLINE, Scopus, and Web of Science, we used a consistent search strategy tailored to the specific features and functionalities of each platform. We used the advanced search options available on these databases to input the key terms from Table 1. The search was conducted across titles and abstracts. For Google Scholar, due to its distinct search engine and more limited filtering options compared to the other databases, we conducted broad search queries with the same key terms. We then manually reviewed the results to identify and include relevant studies that met our criteria.

Study Selection

The retrieved abstracts and titles were imported into Covidence systematic review software (Veritas Health Innovation), where duplicate records between original databases are automatically eliminated. Two reviewers (AM and BH) independently assessed the eligibility of the studies based on the following inclusion and exclusion criteria.

The studies were eligible for primary inclusion if they (1) were composed in English; (2) were original peer-reviewed studies; (3) used individual-level patient data derived from EHRs, EMRs, EPRs, EDWs, or RDW; and (4) incorporated at least 1 form of spatial methods. Conversely, the studies were excluded if they (1) were not peer-reviewed (eg, letters, editorials, reviews, case reports, abstracts, and grey literature), (2) solely geocoded addresses or generated basic visualizations (eg, dot map and choropleth map) without any spatial analysis, and (3) not based on the US data.

The reviewers (AM and BH) independently reviewed the full texts of all remaining studies. The studies also were excluded if they lacked phenotype characteristics. Further, we manually checked the references for all the selected studies for possible inclusion. A third reviewer (AVA) was consulted to break ties.

Data Extraction

Upon identifying studies that satisfied all inclusion criteria, two reviewers (AM and BH) extracted the following items for each study: title, publication year, country and region, sample size, study period, spatial methodologies, and key findings from the spatial methods. Moreover, studies were assessed to identify clinical domains (including primary and secondary when applicable), health conditions or problems, and themes (including social determinants of health [SDOH], environmental factors, ecological aspects, climate, microbiome, genomics, and clinical phenotypic characteristics). Previous publications have emphasized the importance of data domain sources in phenotyping, underscoring the need for validating the created phenotype [13] and using multiple data sources. Thus, in cases where the included publications did not provide details of data sources but instead referenced previously published works, referenced publications were reviewed. Additionally, we cataloged the types of EHRs that served as the sources.

Narrative Synthesis

There is no universally accepted classification for spatial analysis methods. In this review, we have adopted and refined a classification framework based on the study of Nazia et al [14], which initially categorized methods into frequentist and Bayesian approaches and spatial and spatiotemporal methods. This classification was further broken down into descriptive, clustering, and modeling techniques [15]. Therefore, following data extraction, the studies were categorized into the following spatial methodology classifications: descriptive, clustering, modeling (frequentist), spatiotemporal (frequentist), and Bayesian. The phenotype characteristics were extracted and recorded as free text. It should be noted that the categories were not mutually exclusive.

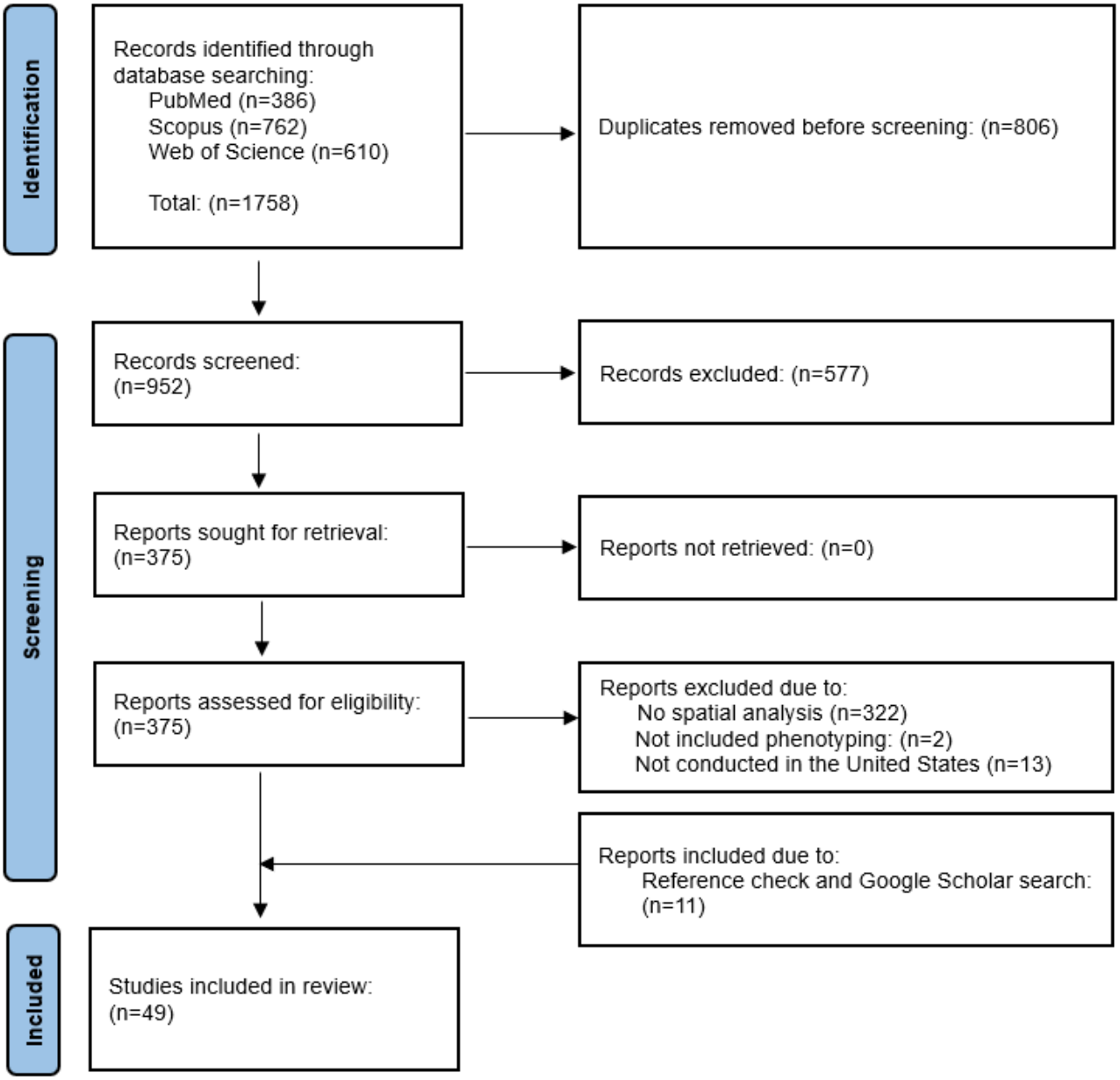
The quality appraisal of the studies was not feasible due to the substantial heterogeneity in spatial methodologies and health domains. The geospatial distribution of the included studies was visualized using ArcGIS Pro software (version 3.0; ESRI).

Results

Study Selection

The initial search yielded 1758 references. After removing duplicate records, we identified 952 studies for abstract and title screening, from which 375 were selected for full-text review. Of these, 322 studies were excluded as they only contained geocoding or basic mapping without any spatial analysis. Additionally, 15 studies were omitted due to the absence of patient phenotype characteristics (n=2) or were not based on US data (n=13). We further manually searched references and Google Scholar and found 11 new studies that met the eligibility criteria. Therefore, 49 studies that fulfilled the inclusion criteria were retained for data extraction and synthesis. Figure 1 depicts the PRISMA flowchart for the study selection process.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) study selection flowchart.



Temporal and Geographic Distribution of Studies

Of the 49 included studies, a limited number (n=7, 14%) were published prior to 2017. The earliest study included in this study was published in 2011, and the publication frequency has experienced a significant upsurge since 2017 (n=42, 86%), likely due to increased adoption of EHR systems and growing familiarity with spatial analysis techniques among researchers. There was only one study [16] at the national level. General characteristics of the included studies are presented in Table 2. Most studies were concentrated in North Carolina (n=8, 16%), Pennsylvania (n=6, 12%), California (n=6, 12%), and Illinois (n=4, 8%). Figure 2 illustrates the geospatial distribution of studies at the state level in the United States.

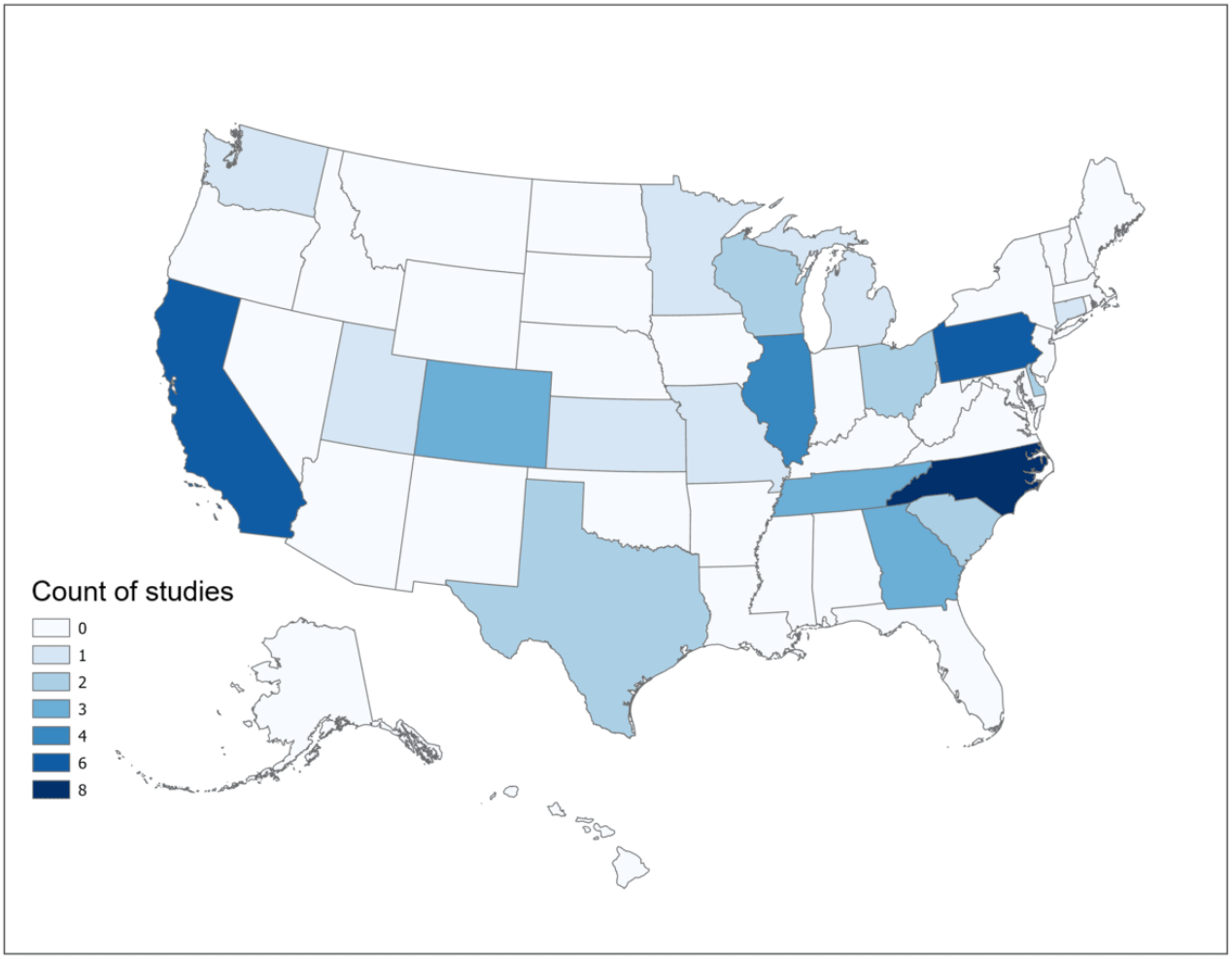
Table 2. General characteristics of the included studies.

No.	Author	Year	Region	Sample size, n	Study period
1	Ali et al [7]	2019	Atlanta	4613	2002-2010
2	Beck et al [17]	2018	Cincinnati	24,428	2011-2016
3	Bravo et al [18]	2018	Durham	147,000	2007-2011
4	Bravo et al [19]	2019	Durham	147,351	2007-2011
5	Bravo et al [20]	2019	Durham	41,203	2007-2011
6	Brooks et al [21]	2020	Delaware	5421	2020
7	Carey et al [22]	2021	Utah	366	2006-2015
8	Casey et al [23]	2016	Pennsylvania	20,569	2006-2013
9	Chang et al [8]	2015	Wisconsin	103,690	2007-2009
10	Cobert et al [24]	2020	Durham	10,352	2013-2018
11	Davidson et al [25]	2018	Denver	21,578	2011-2012
12	DeMass et al [26]	2023	South Carolina	2195	2019-2020
13	Epstein et al [27]	2014	Los Angeles	5390	2007-2011
14	Gaudio et al [28]	2023	Tennessee	2240	2015-2021
15	Georgantopoulos et al [29]	2020	South Carolina	3736	1999-2015
16	Ghazi et al [30]	2022	Twin Cities, Minnesota	20,289	2012-2019
17	Grag et al [31]	2023	Chicago	777,994	2007-2012
18	Grunwell et al [32]	2022	Georgia	1403	2015-2020
19	Hanna-Attisha et al [33]	2016	Flint, Michigan	1473	2013-2015
20	Immergluck et al [34]	2019	Atlanta	13,938	2002-2010
21	Jilcott et al [35]	2011	Eastern North Carolina	744	2007-2008
22	Kane et al [36]	2023	Kansas and Missouri	2427	2011-2020
23	Kersten et al [37]	2018	San Francisco	47,175	2007-2011
24	Lantos et al [38]	2018	North Carolina	3527	N/A[a]
25	Lantos et al [39]	2017	Durham	3527	≤2015
26	Lê-Scherban et al [40]	2019	Philadelphia	3778	2016
27	Lieu et al [41]	2015	Northern California	154,424	2000-2011
28	Lipner et al [42]	2017	Colorado	479	2008-2015
29	Liu et al [43]	2021	Cincinnati and Houston	88,013	2011-2016
30	Mayne et al [44]	2019	Chicago	14,309	2015-2017
31	Mayne et al [45]	2018	Chicago	4748	2009-2013
32	Oyana et al [46]	2017	Memphis	28,793	2005-2015
33	Patterson and Grossman [16]	2017	Nationwide	~100 million	2003-2010
34	Pearson and Werth [47]	2019	Philadelphia	642	2000-2017
35	Samuels et al [48]	2022	New Haven	6366	2013-2017
36	Schwartz et al [49]	2011	Pennsylvania	47,769	2009-2010
37	Sharif-Askary et al [50]	2018	North Carolina	558	1998-2013
38	Sidell et al [51]	2022	Southern California	446,440	2020-2021
39	Siegel et al [52]	2022	Delaware	3449	2012-2020
40	Soares et al [6]	2017	Pennsylvania	2049	2011-2012
41	Sun et al [53]	2022	Southern California	395,927	2008-2018
42	Tabano et al [54]	2017	Denver	31,275	2009-2011

No.	Author	Year	Region	Sample size, n	Study period
43	Wakefield et al [55]	2020	Memphis	3754	2015-2017
44	Wilson et al [56]	2022	Chicago	39,211	2014-2016
45	Winckler et al [57]	2023	Southern California	7896	2017-2019
46	Xie et al [3]	2017	Philadelphia	27,604	2011-2014
47	Xie et al [58]	2023	Washington	242,637	2015-2019
48	Zhan et al [59]	2021	Central Texas	21,923	2019
49	Zhao et al [60]	2021	Wisconsin	43,752	2007-2012

[a]Not applicable.

Figure 2. Geospatial distribution of the included studies at the state level in the United States.



Spatial Methodologies

Overview

Most studies focused on frequentist methods compared to the Bayesian methods. Among frequentist methods, the most prevalent category was clustering (n=29), followed by descriptive (n=12), modeling (n=6), and spatiotemporal analyses (n=2). More detailed explanations of the spatial methods used in this study are provided in Multimedia Appendix 2.

Descriptive Analyses

Descriptive analyses were categorized into four groups: spatial sampling (n=2), spatial overlay (n=2), proximity analysis (n=4), and spatial interpolation (n=4).

Spatial Sampling

A 2 SD ellipse method is used to optimize spatial sampling density. This ellipse contains almost 95% of the locations of patients and is used to ensure that the collected samples reflect the underlying spatial pattern in data, particularly when resources are limited [61]. Lantos et al [38] and Lantos et al [39] adopted this approach when sampling women who

underwent cytomegalovirus antibody testing during pregnancy, especially in peripheral areas with limited subject representation.

Spatial Overlay

Spatial overlay integrates various spatial data sources, often maps, to represent their shared features. Wakefield et al [55] overlaid the map of major radiation treatment interruptions based on race onto the map of median household income. Their analysis implied that regions with higher income levels experienced lower rates of radiation treatment interruption. Samuels et al [48] spatially joined patient addresses to the nearest city parcels and computed an estimate of the incidence of emergency department visits for asthma for each parcel [48].

Proximity Analysis

Proximity analysis includes measuring distances between geographic features to identify nearby features within a defined distance or buffer zone to uncover proximity patterns [62]. Wilson et al [56] created temporal and spatial buffers to assess the correlation between individual exposure to violent crime and blood pressure. Schwartz et al [49] evaluated the associations between environmental factors and BMI within a 0.5-mile network buffer from the place of residence. Casey et al [23] investigated the associations between prenatal residential greenness and birth outcomes within 250-m and 1250-m buffers. Using a geographic information system service area network analysis, Jilcott et al [35] examined BMI percentile and proximity to fast-food and pizza establishments among adolescents within 0.25-mile Euclidean and network buffer zones.

Spatial Interpolation

Ordinary Kriging is one of the most widely used spatial interpolation techniques that leverages the spatial autocorrelation structure of observed locations to estimate values at unmeasured locations [63]. Hanna-Attisha et al [33] applied ordinary Kriging with a spherical semivariogram model based on observations of the children's elevated blood lead level geocoded to the home address to visualize blood lead level variations before and after water source changes. Mayne et al [44] interpolated the levels of neighborhood physical disorder based on an exponential variogram. Patterson and Grossman [16] demonstrated spatial variations for the incidence rates of each *International Classification of Diseases, Ninth Revision* diagnostic code based on an exponential variogram. Sun et al [53] estimated monthly average concentrations of fine particulate matter to investigate the associations between air pollution exposure during pregnancy and gestational diabetes mellitus.

Spatial Clustering

Overview

Spatial clustering techniques assess whether health outcomes are random, uniform, or clustered and pinpoint the locations of clusters [64]. Spatial clustering was the most widely used category (n=29) among all studied categories. Moran *I* clustering and cluster detection were the most frequent techniques (n=10), followed by kernel/point density estimation (n=5), spatial scan statistics (n=4), and Getis-Ord Gi* statistics (n=4).

Kernel/Point Density Estimation

Kernel density estimation generates a smooth surface to visualize areas of the most significant spatial intensity by calculating a distance-weighted count of events within a specified radius per unit area [65]. Several studies adopted kernel density estimation to analyze patterns, including cholera hospitalization [58], comparison of the spatial intensity of chronic kidney disease with nonchronic kidney disease patients [30], and comparison of the spatial intensity of breast cancer and nonbreast cancer [52]. Using the point density function, Beck et al [17] pinpointed hotspots of inpatient bed-day rates within a 2-mile radius of a medical center, and Kane et al [36] estimated the number of participants per square mile.

Global and Local Moran I

Global Moran *I* (GMI) evaluates the overall pattern for spatial autocorrelation [66] by inferring if a variable is spatially clustered or overdispersed versus being randomly distributed under the null hypothesis [66]. Local Moran *I* (LISA) is used to locate statistically significant clusters including hotspots, cold spots, and outliers [67]. GMI has been adopted to analyze spatial clustering of health outcomes including gestational diabetes mellitus [53], day-of-surgery cancellation [43], obesity [54], and COVID-19 [51]. All exhibited clustered patterns. Xie et al [58] analyzed 3 groups: depression, obesity, and comorbid cases, confirmed clustering for all outcomes, and identified spatial clusters and outliers. Pearson and Werth [47] found random distributions for dermatomyositis (DM) and subtypes, classic DM, and clinically amyopathic DM. Meanwhile, Davidson et al [25] pinpointed clusters with higher or lower depression prevalence, and Winckler et al [57] identified a cluster of low use of acute pediatric mental health interventions in less-densely populated rural border areas.

GMI and semivariograms or variograms can also identify spatial autocorrelation in model residuals. If detected, the models are adjusted accordingly to avoid biased estimates. For example, Lipner et al [42] modeled nontuberculous mycobacteria disease, shifting the use from a nonspatial Bayesian model to a spatial model when spatial autocorrelation was found in residuals. Similarly, Georgantopoulos et al [29] incorporated spatial random effects into a prostate cancer model due to significant autocorrelation in the residuals. Sharif-Askary et al [50] used variograms to assess spatial dependency in cleft lip or palate, leading to a geostatistical model over standard logistic regression. Conversely, Casey et al [23] found no spatial autocorrelation in nonspatial model residuals.

The bivariate GMI quantifies the overall spatial dependence between two distinct variables (positive value indicates high values of one variable are surrounded by high values of the other or low values are surrounded by low values, while negative value implies high values of one variable are surrounded by low values of the other) [68]. Bivariate LISA assesses the relationship between the two variables at the local level. Pearson and Werth [47] used bivariate GMI for the prevalence of DM, classic DM, and clinically amyopathic DM with airborne toxics but found no overall spatial dependencies. However, bivariate LISA identified local dependencies at the zip code level. Garg et al [31] applied bivariate GMI and found significant overall

associations between longer (average) distances to the nearest supermarket and higher incidence of diabetes, and bivariate LISA identified significant "high-high" relationships at the zip code level. Gaudio et al [28] used bivariate LISA and found no local association between radiation therapy interruption and social vulnerability index at the zip code level.

Getis-Ord Gi*

The Getis-Ord Gi* statistic identifies high- or low-value clusters (hotspots and cold spots) by assessing deviations of health outcomes at locations from the average within a defined neighborhood [69]. Lê-Scherban et al [40] measured racial residential segregation by examining the deviations in the African American residents in each census tract from the mean of neighboring tracts. Similarly, Mayne et al [45] measured racial residential segregation for the percentage of non-Hispanic Black residents. Ali et al [7] identified significant community-onset methicillin-resistant Staphylococcus aureus (CO-MRSA) hotspots with distinct patterns between cases and controls. Kersten et al [37] detected the high- and low-value clusters for the child opportunity index and median household income.

Spatial Scan Statistics

The spatial scan statistics technique identifies high- and low-risk clusters and estimates their relative risks [70]. It also can incorporate covariates to characterize underlying patterns [71]. Lipner et al [42] found that people living in zip codes within the primary cluster had an almost 2.5 times greater risk of nontuberculous mycobacteria disease. Lieu et al [41] identified clusters of underimmunization and vaccine refusal among children, with rates ranging from 18% to 23% inside the clusters compared to 11% outside.

The technique can also pinpoint cold spots. Brooks et al [21] identified areas with significantly lower COVID-19 testing than expected, indicating a need for interventions. Zhan et al [59] observed significantly low rates of up-to-date colorectal cancer screening.

Spatial Modeling (Frequentist)

Among the included studies, the generalized additive models (GAMs) emerged as the most frequently used spatial models. GAMs can account for spatial autocorrelation by incorporating smooth functions (such as thin-plate regression) of spatial coordinates [72], allowing the estimate of geographic variation with or without covariate adjustments. GAMs were used to identify the spatial variabilities in asthma prevalence [3,8] and cytomegalovirus [38,39], although such variations often diminished when adjusted for demographic factors such as race and age. Less commonly used geospatial models were generalized linear mixed effects [51] and spatial error [43] models.

Spatiotemporal Analysis

Only 2 studies explored spatiotemporal patterns, and no spatiotemporal modeling was conducted. Oyana et al [46] used space-time scan statistics to study the spatiotemporal patterns of childhood asthma and found a significant frequency increase (2009-2013) and a rising trend from 4 to 16 per 1000 children (2005-2015). Ali et al [7] used the space-time cube tool and emerging hotspot analysis to analyze the spatial-temporal trends and evolving patterns of CO-MRSA from 2002 to 2010. They identified several types of space-time hotspots of CO-MRSA including new, consecutive, intensifying, sporadic, and oscillating hotspots.

Bayesian Analysis

The studies using Bayesian methods were categorized into empirical Bayes smoothing (n=5) and Bayesian modeling (n=6).]

Empirical Bayes Smoothing

The empirical Bayes smoothing was used by Lê-Scherban et al [40], Liu et al [43], Tabano et al [54], and Xie et al [58] to stabilize estimated rates in areas with limited data points by borrowing information from the overall population [73]. Zhao et al [60] used nonparametric kernel smoothing to estimate the prevalence of childhood obesity in areas with sparse observations (n<20 individuals) [60].

Bayesian Modeling

Bayesian modeling can account for spatial and temporal dependencies and quantify uncertainty by specifying prior distributions [74]. Among the studies, the conditional autoregressive (CAR) prior emerged as the most used, with 2 variants: intrinsic and multivariate CAR. Intrinsic CAR was used to assess the spatial variations in diabetes in relationship with racial isolation [18], hypertension related to racial isolation [19], and type 2 diabetes mellitus with the built environment [20]. Multivariate CAR was used to identify areas with higher or lower-than-expected prostate cancer while controlling for risk factors [29]. Moreover, hierarchical Bayesian that can incorporate hierarchical structures for modeling [75] was used to investigate spatial distributions of patients admitted for drug-related reasons concerning the area deprivation index [24]. Bayesian negative binomial hurdle models that can account for excessive zeros and overdispersion were used to examine spatial variation between patient responses to the questions concerning unhealthy home environments and the mean number of emergency department visits after screening [26].

Phenotyping

Clinical Domain Characteristics and Themes

The largest category of studies was classified under the infectious disease (n=7), endocrinology (n=7), and oncology (n=6) domains. Additionally, 19 studies had a pediatric domain or focus, as noted with an additional column in Table 3. Maternal and newborn care was classified as its own domain (n=8), but it overlapped with other domains such as nephrology, endocrinology, and infectious disease.

Table 3. Clinical domains and condition or problem of focus for each publication.

Condition by clinical domain[a]	Secondary clinical domain[b]	Pediatric population involved
Pediatric		
DoSC[c] [43]	—[d]	✓
EBLL[e] [33]	—	✓
Disparities in inpatient bed-day rates [17]	—	✓
Maternal and newborn care		
Under immunization; vaccine refusal [41]	—	✓
Preterm birth; small for gestational age; hypertensive disorder of pregnancy [44]	—	
Preterm birth; small for gestational age; low birth weight; low Apgar score [23]	—	
Hypertension [56]	—	
Hypertension [19]	—	
Hypertension; diabetes [40]	Endocrinology	
Hypertension; diabetes; CKD[f] [31]	Endocrine; nephrology	
Hypertension, disorder of pregnancy [45]	Maternal and newborn care	
Endocrinology		
GDM[g] [53]	Maternal and newborn care	
T2DM[h] [18]	—	
T2DM [20]	—	
Obesity [54]	—	
Obesity [49]	—	✓
Obesity [35]	—	✓
Obesity [60]	—	✓
Obesity; depression [58]	Psychiatry	
Psychiatry		
Acute pediatric mental health interventions or services [57]	—	✓
Depression [25]	—	
Telemedicine use in developmental-behavioral pediatrics [6]	—	✓
Drug overdoses [24]	Emergency medicine	
Emergency medicine		
Disparities in pediatric acute care visit frequency and diagnoses [37]	—	✓
Disparities in use of PICU[i] [27]	—	✓
Emergency department use [26]	—	
Pulmonary		
Asthma, emergency department asthma visits [48]	Emergency medicine	
Asthma [32]	—	✓
Asthma [46]	—	✓
Asthma [3]	—	
Asthma [8]	—	
Infectious disease		
Coccidioidomycosis [22]	Pulmonary	

Condition by clinical domain[a]	Secondary clinical domain[b]	Pediatric population involved
Community-associated MRSA[j] [34]	—	✓
Community-onset-MRSA [7]	—	✓
COVID-19 [21]	—	
COVID-19 [51]	—	
CMV[k] [39]	Maternal and newborn care	✓
CMV [38]	—	✓
Nontuberculous mycobacterial infection [42]	—	
Oncology		
RTI[l] [55]	—	
RTI [28]	—	
Colorectal cancer screening [59]	—	
Prostate cancer [29]	—	
TNBC[m] [52]	—	
Disparities in genomic answers for kids (GA4K) [36]	—	✓
Maxillofacial		
Cleft lip or palate [50]	—	✓
Nephrology		
CKD [30]	—	
Rheumatology		
Dermatomyositis [47]	Neurology; dermatology	
All domains		
Geospatial variation of disease incidence [16]	—	

[a]Condition or problem of focus column displays the general condition of the study and may not directly correspond to the phenotype.
[b]Publications with more than 1 clinical domain and those with a pediatric component are noted as such.
[c]DoSC: day-of-surgery cancellation.
[d]Not applicable.
[e]EBLL: elevated blood lead levels.
[f]CKD: chronic kidney disease.
[g]GDM: gestational diabetes mellitus.
[h]T2DM: diabetes mellitus, type 2.
[i]PICU: pediatric intensive care unit.
[j]MRSA: methicillin-resistant Staphylococcus aureus.
[k]CMV: cytomegalovirus.
[l]RTI: radiation treatment interruption.
[m]TNBC: triple-negative breast cancer.

The relationship between the clinical domains and the "conditions or problems of focus" in each study was examined (Table 3). In some cases, direct correspondence was observed, while in other instances, the "condition or problems of focus" differed from the phenotype of the patient cohort. In many studies, one or more overlapping domains were observed (eg, rheumatology, neurology, and dermatology for the study of DM). Asthma (n=5), hypertension (n=5), and diabetes (n=4) were studied most frequently. Three studies did not focus on any health condition but rather on examining disparities in either a data source or a specific domain or cohort (eg, disparities in the use of pediatric intensive care units).

Every study was attributed to at least one prominent theme, with the possibility of multiple themes. SDOH themes were prevalent in many studies. To organize and present this information, we used the domains defined by the Healthy People 2030 framework [76]. There are 5 domains in the SDOH framework (Table 4), with the corresponding counts of these domains being seen as themes of the studies. Most studies had 1 or more SDOH themes (n=42). Many studies focused either on all the domains or SDOH holistically without particular focus on any specific domain (n=32). However, some studies contained prominent themes that were not directly related to SDOH, which were phenotypic features (n=4), followed by

environmental (n=3), and ecological (n=2), with climate, genomics, and microbiome, each contributing one study.

Table 4. SDOH[a] themes examined within the framework of Healthy People 2030 SDOH domains [76].

Labels and SDOH domains	Counts, n
SDOH 1	
Economic stability (employment, food insecurity, housing instability, poverty)	2
SDOH 2	
Education access and quality (early childhood development and education, enrollment in higher education, high school graduation, language, and literacy)	N/A[b]
SDOH 3	
Health access and quality (access to health services, access to primary care, health literacy)	5
SDOH 4	
Neighborhood and built environment (access to foods that support healthy dietary patterns, crime and violence, environmental conditions, quality of housing)	14
SDOH 5	
Social and community context (civic participation, discrimination, incarceration, social cohesion)	5
All 5 SDOH domains or SDOH as a whole	36
Non-SDOH focus	8

[a]SDOH: social determinants of health.

[b]Not applicable.

Clinical Phenotype Features

For each publication, clinical phenotype definitions were extracted (Multimedia Appendix 3). In almost all studies, phenotype definitions included demographic details such as patient age, race, and gender, along with some diagnostic characteristics (eg, asthma diagnosis). Only a limited number of phenotypes were observed to be validated (n=8). The most frequently observed method for phenotype validation was a manual chart review of all matches or a sample of matched charts. None of the studies with chart review as a validation method shared information on the match rate. Additionally, only two studies [20,58] were observed to use validated eMERGE Network computable phenotypes from the Phenotype Knowledgebase [77-79].

Discussion

Principal Findings

This systematic review is the first comprehensive investigation of spatial methodologies within EHR-derived data in the United States. The findings reveal that a considerable portion of studies predominantly focus on basic mapping or geocoding, with a limited use of advanced spatial analysis methods. Spatial clustering and descriptive analysis were the most used methods, while space-time modeling, either frequentist or Bayesian, was not widely applied. The diverse use of spatial analysis for EHR-derived data in different health domains highlights the potential to incorporate spatial methods to enhance the context of individual patients for future biomedical research. We found limited use of EHR-derived data for spatial analysis, probably due to the challenge of safeguarding patient privacy. Address data, crucial for spatial analysis, is highly confidential and often restricted from sharing. Researchers and institutions often use geographic masking techniques [6,80] to balance data use and privacy protection by altering the precise geographic coordinates while preserving the overall spatial characteristics of data. Encouraging the adoption of spatial analysis could promote biomedical knowledge sharing and collaboration.

The use of EHRs data for spatial analysis can present several challenges, particularly in accurately geocoding patient addresses. Issues, such as address formatting errors, incomplete or outdated addresses, and potential inaccuracies in geocoding services, can influence the outcome of spatial analysis [81]. Advanced geocoding algorithms and manual verification processes can mitigate these issues. For instance, Goldberg et al [82] developed a web-based system for rapid manual intervention of previously geocoded data, significantly improving the match rate and quality of individual geocodes with minimal time and effort. Additionally, when addresses are only available at the zip code level, additional nuances arise as zip code boundaries are often not well-defined and can change over time [83]. Spatial smoothing techniques and zip code centroids can mitigate some of these challenges. We recommend standardizing address formats before geocoding (using tools like the US Postal Service address verification), using advanced geocoding services, leveraging higher-resolution geographical data when possible, and integrating multiple spatial scales to enhance the accuracy and reliability of spatial analysis using EHRs data.

We acknowledge that not all patient phenotypes are inherently suited for spatial analysis, and integrating genomics, imaging, and clinical notes phenotypes can be particularly challenging. However, evidence suggests that spatial techniques can provide valuable insights even in these areas where their application may initially appear challenging. For instance, Baker et al [84] demonstrated the effectiveness of spatial analysis in genomics

by combining single-nucleotide polymorphism genotyping with geospatial K-function analysis. Their study of typhoid in Nepal found significant geographic clustering of cases. Canino [85] developed a robust framework that integrated biological data with geographic information from EMRs. Their system identified correlations between patient profiles and geographic factors such as environmental exposures related to pollution. Future interdisciplinary studies can explore developing frameworks that integrate genomics or notes with geospatial datasets to reveal complex relationships and patterns.

The application of spatiotemporal analysis of EHR-derived data was mainly limited to exploring spatiotemporal clusters with no spatiotemporal modeling. This might be due to the technical expertise required for analysis, data complexity, availability of longitudinal data, and computational challenges. The Bayesian framework offers a more adaptable framework to handle complex spatial and temporal dependencies, control confounding variables [86], and incorporate prior information, such as existing medical literature and expert opinions, resulting in more interpretable results [87,88]. Moreover, spatiotemporal Bayesian modeling can aid in understanding disease trends and progressions, seasonality, and long-term shifts at the local levels [89]. Bayesian modeling can also account for uncertainty in parameter estimates and predictions to assess the reliability of findings before implementing interventions [90]. Thus, future research should delve into spatial and spatiotemporal modeling, focusing on Bayesian approaches. Moreover, ignoring spatial dependence in modeling can bias parameter estimates [9,91,92]. Additional state-of-the-art methods, such as space-time autoregressive models and generalized additive models for location scale and shape, also provide flexibility in modeling complex relationships. Spatiotemporal point process models also contribute by analyzing the distribution of health events and underlying states over space and time.

Among the health conditions studied, chronic and infectious diseases emerged as the most frequently investigated domains compared to others. This disparity may be attributed to the pressing public health concerns posed by diseases with immediate impacts that often attract more funding and resources for research initiatives [93,94]. The historically high mortality rates of these conditions likely led to continuous research. Furthermore, the nature of spatial contamination and the spread of infectious diseases has historically driven the development of spatial analysis for clinical purposes, exemplified by John Snow's seminal cholera investigation. Surprisingly, despite the plethora of funding in cancer research, we only found a small number of studies within the cancer domain, which may likewise be attributed to and indicative of the pressing needs of other domains such as infectious disease.

We observed recurring and prominent themes related to the SDOH. This emphasis may result from the growing maturity and increased awareness within the biomedical informatics community regarding the significant influence of social, economic, and environmental factors on health outcomes. Understanding the roles of SDOH in health disparities will likely lead to the implementation of integrative health interventions that address the needs of individuals affected by these health disparities. These interventions can likewise be enhanced by incorporating spatial perspectives.

Another missed opportunity is the limited use of computable phenotypes—automated algorithms designed for characterizing diseases and enrolling patients in studies. Most studies primarily depended on the manual application of inclusion and exclusion criteria to define phenotypes. While this method may be suitable in certain scenarios, it often necessitates greater depth and granularity to consistently and accurately capture the intended patient cohorts. The accuracy and precision of the manual approach can vary depending on the data sources and clinical domains. Notably, only 2 of the studies in this review used computable phenotypes, indicating a limited adoption of this essential and potentially transformative approach, highlighting a noteworthy area for growth. Furthermore, only 5 studies carried out any form of chart review validation. Validation methods, including chart reviews, genetic markers, and clinical variables, are indispensable in phenotyping to guarantee the accurate characterization of the desired cohorts. This applies even to computable phenotypes within specific medical domains [95].

Limitations

This study has several main limitations. First, we only considered English-language studies, possibly introducing language bias. Additionally, selection bias is possible due to database availability. However, we mitigated these limitations by searching Google Scholar and conducting backward reference checking to identify relevant studies that might yet be identified through our initial search strategy. Finally, we used a query search strategy with limited keywords, which inherently restricted the scope of studies we could retrieve, potentially omitting studies that did not use these specific terms in their abstract or title.

Our rationale to focus exclusively on US data was driven by our familiarity with the reliability and availability of EHR-based systems within the country. Moreover, we recognize that spatial analyses of health data in regions, such as Europe, Asia, Australia, and Canada, use different terminologies and labels for their systems, which might not align with our search terms for EHRs or EMRs. For instance, Canada's national administrative databases and electronic discharge records could encompass significant work not captured by our key terms, a situation that can be generalized to other countries. To avoid inconsistencies arising from varying data labeling and storage systems across different regions, we opted to concentrate on the United States. Nevertheless, future research should endeavor to include and explore contributions from these regions to provide a more comprehensive understanding of emerging trends in spatial analysis in characterizing patient phenotypes.

Conclusions

This systematic review provided a comprehensive overview of the current use of spatial analysis in EHR-based research in the United States and underscored the pivotal role that spatial analysis can play in clinical decision support and interventions. The use of EHR-derived spatial analysis is on an upward trajectory, parallel with the widespread adoption of EHR

systems. The volume of studies on this topic is anticipated to continue to grow. The primary health outcomes investigated were asthma, hypertension, and diabetes. Notably, patient phenotypes involving genomics, imaging, and notes that are notoriously high-dimensional and add to the computational intensity of spatial methods were limited. This review also highlighted the need for additional exploration of spatial analysis techniques, including but not limited to spatiotemporal Bayesian analysis and modeling, particularly in computable phenotypes or patient phenotypes involving genomics, imaging, and notes.

Acknowledgments

We would like to express our gratitude to Professor Gregory Glass from the University of Florida for his constructive review of the earlier version of the manuscript. We would also like to thank Clemson University librarian Karen Burton and Medical University of South Carolina librarian Ayaba Logan, MPH, MLIS, whose expertise in library and information sciences facilitated our systematic review. AM, BH, and AVA are supported by the South Carolina SmartState Endowed Center for Environmental and Biomedical Panomics (CEABP). AVA is supported by South Carolina Cancer Disparities Research Center (SC CADRE) from NIH/NCI U54 CA210962. BH is a trainee supported by the SC Biomedical Informatics and Data Science for Health Equity Research Training (SC BIDS4HEALTH) from NIH/NLM T15 LM013977.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 31 KB - medinform_v12i1e56343_app1.docx]

Multimedia Appendix 2
Detailed descriptions of the spatial methods used.
[DOCX File , 30 KB - medinform_v12i1e56343_app2.docx]

Multimedia Appendix 3
Clinical phenotype definitions and spatial method used for each publication.
[DOCX File , 95 KB - medinform_v12i1e56343_app3.docx]

References

1. Kuo A, Dang S. Secure messaging in electronic health records and its impact on diabetes clinical outcomes: a systematic review. Telemed e-Health 2016;22(9):769-777. [doi: 10.1089/tmj.2015.0207] [Medline: 27027337]
2. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019;6(1):1-25. [doi: 10.1186/s40537-019-0217-0]
3. Xie S, Greenblatt R, Levy MZ, Himes BE. Enhancing electronic health record data with geospatial information. AMIA Jt Summits Transl Sci Proc 2017;2017:123-132 [FREE Full text] [Medline: 28815121]
4. He J, Ghorveh MG, Hurst JH, Tang M, Alhanti B, Lang JE, et al. Evaluation of associations between asthma exacerbations and distance to roadways using geocoded electronic health records data. BMC Public Health 2020;20(1):1626 [FREE Full text] [doi: 10.1186/s12889-020-09731-0] [Medline: 33121457]
5. Schooley BL, Horan TA, Lee PW, West PA. Rural veteran access to healthcare services: investigating the role of information and communication technologies in overcoming spatial barriers. Perspect Health Inf Manag 2010;7(Spring):1f [FREE Full text] [Medline: 20697468]
6. Soares N, Dewalle J, Marsh B. Utilizing patient geographic information system data to plan telemedicine service locations. J Am Med Inform Assoc 2017;24(5):891-896 [FREE Full text] [doi: 10.1093/jamia/ocx011] [Medline: 28339932]
7. Ali F, Immergluck LC, Leong T, Waller L, Malhotra K, Jerris RC, et al. A spatial analysis of health disparities associated with antibiotic resistant infections in children living in Atlanta (2002-2010). EGEMS (Wash DC) 2019;7(1):50 [FREE Full text] [doi: 10.5334/egems.308] [Medline: 31565665]
8. Chang TS, Gangnon RE, Page CD, Buckingham WR, Tandias A, Cowan KJ, et al. Sparse modeling of spatial environmental variables associated with asthma. J Biomed Inform 2015;53:320-329 [FREE Full text] [doi: 10.1016/j.jbi.2014.12.005] [Medline: 25533437]
9. Mollalo A, Mohammadi A, Mavaddati S, Kiani B. Spatial analysis of COVID-19 vaccination: a scoping review. Int J Environ Res Public Health 2021;18(22):12024 [FREE Full text] [doi: 10.3390/ijerph182212024] [Medline: 34831801]
10. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]

11. Schinasi LH, Auchincloss AH, Forrest CB, Roux AVD. Using electronic health record data for environmental and place based population health research: a systematic review. Ann Epidemiol 2018;28(7):493-502. [doi: 10.1016/j.annepidem.2018.03.008] [Medline: 29628285]
12. Simpson CL, Novak LL. Place matters: the problems and possibilities of spatial data in electronic health records. : American Medical Informatics Association; 2013 Presented at: AMIA Annual Symposium Proceedings; October 03, 2021; California, USA URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900146/
13. Hamidi B, Flume PA, Simpson KN, Alekseyenko AV. Not all phenotypes are created equal: covariates of success in e-phenotype specification. J Am Med Inform Assoc 2023;30(2):213-221 [FREE Full text] [doi: 10.1093/jamia/ocac157] [Medline: 36069977]
14. Nazia N, Butt ZA, Bedard ML, Tang W, Sehar H, Law J. Methods used in the spatial and spatiotemporal analysis of COVID-19 epidemiology: a systematic review. Int J Environ Res Public Health 2022;19(14):8267 [FREE Full text] [doi: 10.3390/ijerph19148267] [Medline: 35886114]
15. Moore DA, Carpenter TE. Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiol Rev 1999;21(2):143-161. [doi: 10.1093/oxfordjournals.epirev.a017993] [Medline: 10682254]
16. Patterson MT, Grossman RL. Detecting spatial patterns of disease in large collections of electronic medical records using neighbor-based bootstrapping. Big Data 2017;5(3):213-224 [FREE Full text] [doi: 10.1089/big.2017.0028] [Medline: 28933946]
17. Beck AF, Riley CL, Taylor SC, Brokamp C, Kahn RS. Pervasive income-based disparities in inpatient bed-day rates across conditions and subspecialties. Health Aff (Millwood) 2018;37(4):551-559 [FREE Full text] [doi: 10.1377/hlthaff.2017.1280] [Medline: 29608357]
18. Bravo MA, Anthopolos R, Kimbro RT, Miranda ML. Residential racial isolation and spatial patterning of type 2 diabetes mellitus in Durham, North Carolina. Am J Epidemiol 2018;187(7):1467-1476. [doi: 10.1093/aje/kwy026] [Medline: 29762649]
19. Bravo MA, Batch BC, Miranda ML. Residential racial isolation and spatial patterning of hypertension in Durham, North Carolina. Prev Chronic Dis 2019;16:E36 [FREE Full text] [doi: 10.5888/pcd16.180445] [Medline: 30925142]
20. Bravo MA, Anthopolos R, Miranda ML. Characteristics of the built environment and spatial patterning of type 2 diabetes in the urban core of Durham, North Carolina. J Epidemiol Community Health 2019;73(4):303-310. [doi: 10.1136/jech-2018-211064] [Medline: 30661032]
21. Brooks M, Brown C, Liu W, Siegel SD. Mapping the ChristianaCare response to COVID-19: clinical insights from the value institute's geospatial analytics core. Dela J Public Health 2020;6(2):66-70 [FREE Full text] [doi: 10.32481/djph.2020.07.018] [Medline: 34467114]
22. Carey A, Gorris ME, Chiller T, Jackson B, Beadles W, Webb BJ. Epidemiology, clinical features, and outcomes of coccidioidomycosis, Utah, 2006-2015. Emerg Infect Dis 2021;27(9):2269-2277 [FREE Full text] [doi: 10.3201/eid2709.210751] [Medline: 34423764]
23. Casey JA, James P, Rudolph KE, Wu CD, Schwartz BS. Greenness and birth outcomes in a range of Pennsylvania communities. Int J Environ Res Public Health 2016;13(3):311 [FREE Full text] [doi: 10.3390/ijerph13030311] [Medline: 26978381]
24. Cobert J, Lantos PM, Janko MM, Williams DGA, Raghunathan K, Krishnamoorthy V, et al. Geospatial variations and neighborhood deprivation in drug-related admissions and overdoses. J Urban Health 2020;97(6):814-822 [FREE Full text] [doi: 10.1007/s11524-020-00436-8] [Medline: 32367203]
25. Davidson AJ, Xu S, Oronce CIA, Durfee MJ, McCormick EV, Steiner JF, et al. Monitoring depression rates in an urban community: use of electronic health records. J Public Health Manag Pract 2018;24(6):E6-E14 [FREE Full text] [doi: 10.1097/PHH.0000000000000751] [Medline: 29334514]
26. DeMass R, Gupta D, Self S, Thomas D, Rudisill C. Emergency department use and geospatial variation in social determinants of health: a pilot study from South Carolina. BMC Public Health 2023;23(1):1527 [FREE Full text] [doi: 10.1186/s12889-023-16136-2] [Medline: 37563566]
27. Epstein D, Reibel M, Unger JB, Cockburn M, Escobedo LA, Kale DC, et al. The effect of neighborhood and individual characteristics on pediatric critical illness. J Community Health 2014;39(4):753-759 [FREE Full text] [doi: 10.1007/s10900-014-9823-0] [Medline: 24488647]
28. Gaudio E, Ammar N, Gunturkun F, Akkus C, Brakefield W, Wakefield DV, et al. Defining radiation treatment interruption rates during the COVID-19 pandemic: findings from an academic center in an underserved urban setting. Int J Radiat Oncol Biol Phys 2023;116(2):379-393 [FREE Full text] [doi: 10.1016/j.ijrobp.2022.09.073] [Medline: 36183931]
29. Georgantopoulos P, Eberth JM, Cai B, Emrich C, Rao G, Bennett CL, et al. Patient- and area-level predictors of prostate cancer among South Carolina veterans: a spatial analysis. Cancer Causes Control 2020;31(3):209-220. [doi: 10.1007/s10552-019-01263-2] [Medline: 31975155]
30. Ghazi L, Drawz PE, Berman JD. The association between fine particulate matter (PM) and chronic kidney disease using electronic health record data in urban Minnesota. J Expo Sci Environ Epidemiol 2022;32(4):583-589 [FREE Full text] [doi: 10.1038/s41370-021-00351-3] [Medline: 34127789]

31. Garg G, Tedla YG, Ghosh AS, Mohottige D, Kolak M, Wolf M, et al. Supermarket proximity and risk of hypertension, diabetes, and CKD: a retrospective cohort study. Am J Kidney Dis 2023;81(2):168-178 [FREE Full text] [doi: 10.1053/j.ajkd.2022.07.008] [Medline: 36058428]
32. Grunwell JR, Opolka C, Mason C, Fitzpatrick AM. Geospatial analysis of social determinants of health identifies neighborhood hot spots associated with pediatric intensive care use for life-threatening asthma. J Allergy Clin Immunol Pract 2022;10(4):981-991.e1 [FREE Full text] [doi: 10.1016/j.jaip.2021.10.065] [Medline: 34775118]
33. Hanna-Attisha M, LaChance J, Sadler RC, Champney Schnepp A. Elevated blood lead levels in children associated with the flint drinking water crisis: a spatial analysis of risk and public health response. Am J Public Health 2016;106(2):283-290. [doi: 10.2105/AJPH.2015.303003] [Medline: 26691115]
34. Immergluck LC, Leong T, Malhotra K, Parker TC, Ali F, Jerris RC, et al. Geographic surveillance of community associated MRSA infections in children using electronic health record data. BMC Infect Dis 2019;19(1):170 [FREE Full text] [doi: 10.1186/s12879-019-3682-3] [Medline: 30777016]
35. Jilcott SB, Wade S, McGuirt JT, Wu Q, Lazorick S, Moore JB. The association between the food environment and weight status among eastern North Carolina youth. Public Health Nutr 2011;14(9):1610-1617. [doi: 10.1017/S1368980011000668] [Medline: 21486525]
36. Kane NJ, Cohen AS, Berrios C, Jones B, Pastinen T, Hoffman MA. Committing to genomic answers for all kids: evaluating inequity in genomic research enrollment. Genet Med 2023;25(9):100895 [FREE Full text] [doi: 10.1016/j.gim.2023.100895] [Medline: 37194653]
37. Kersten EE, Adler NE, Gottlieb L, Jutte DP, Robinson S, Roundfield K, et al. Neighborhood child opportunity and individual-level pediatric acute care use and diagnoses. Pediatrics 2018;141(5):e20172309 [FREE Full text] [doi: 10.1542/peds.2017-2309] [Medline: 29626164]
38. Lantos PM, Hoffman K, Permar SR, Jackson P, Hughes BL, Kind A, et al. Neighborhood disadvantage is associated with high cytomegalovirus seroprevalence in pregnancy. J Racial Ethn Health Disparities 2018;5(4):782-786 [FREE Full text] [doi: 10.1007/s40615-017-0423-4] [Medline: 28840519]
39. Lantos PM, Hoffman K, Permar SR, Jackson P, Hughes BL, Swamy GK. Geographic disparities in cytomegalovirus infection during pregnancy. J Pediatric Infect Dis Soc 2017;6(3):e55-e61 [FREE Full text] [doi: 10.1093/jpids/piw088] [Medline: 28201739]
40. Lê-Scherban F, Ballester L, Castro JC, Cohen S, Melly S, Moore K, et al. Identifying neighborhood characteristics associated with diabetes and hypertension control in an urban African-American population using geo-linked electronic health records. Prev Med Rep 2019;15:100953 [FREE Full text] [doi: 10.1016/j.pmedr.2019.100953] [Medline: 31367515]
41. Lieu TA, Ray GT, Klein NP, Chung C, Kulldorff M. Geographic clusters in underimmunization and vaccine refusal. Pediatrics 2015;135(2):280-289. [doi: 10.1542/peds.2014-2715] [Medline: 25601971]
42. Lipner EM, Knox D, French J, Rudman J, Strong M, Crooks JL. A geospatial epidemiologic analysis of nontuberculous mycobacterial infection: an ecological study in Colorado. Ann Am Thorac Soc 2017;14(10):1523-1532 [FREE Full text] [doi: 10.1513/AnnalsATS.201701-081OC] [Medline: 28594574]
43. Liu L, Ni Y, Beck AF, Brokamp C, Ramphul RC, Highfield LD, et al. Understanding pediatric surgery cancellation: geospatial analysis. J Med Internet Res 2021 Sep 10;23(9):e26231. [doi: 10.2196/26231] [Medline: 34505837]
44. Mayne SL, Pellissier BF, Kershaw KN. Neighborhood physical disorder and adverse pregnancy outcomes among women in Chicago: a cross-sectional analysis of electronic health record data. J Urban Health 2019;96(6):823-834 [FREE Full text] [doi: 10.1007/s11524-019-00401-0] [Medline: 31728900]
45. Mayne SL, Yellayi D, Pool LR, Grobman WA, Kershaw KN. Racial residential segregation and hypertensive disorder of pregnancy among women in Chicago: analysis of electronic health record data. Am J Hypertens 2018;31(11):1221-1227 [FREE Full text] [doi: 10.1093/ajh/hpy112] [Medline: 30010764]
46. Oyana TJ, Podila P, Wesley JM, Lomnicki S, Cormier S. Spatiotemporal patterns of childhood asthma hospitalization and utilization in Memphis Metropolitan area from 2005 to 2015. J Asthma 2017;54(8):842-855 [FREE Full text] [doi: 10.1080/02770903.2016.1277537] [Medline: 28055280]
47. Pearson DR, Werth VP. Geospatial correlation of amyopathic dermatomyositis with fixed sources of airborne pollution: a retrospective cohort study. Front Med (Lausanne) 2019;6:85 [FREE Full text] [doi: 10.3389/fmed.2019.00085] [Medline: 31069228]
48. Samuels EA, Taylor RA, Pendyal A, Shojaee A, Mainardi AS, Lemire ER, et al. Mapping emergency department asthma visits to identify poor-quality housing in New Haven, CT, USA: a retrospective cohort study. Lancet Public Health 2022;7(8):e694-e704 [FREE Full text] [doi: 10.1016/S2468-2667(22)00143-8] [Medline: 35907420]
49. Schwartz BS, Stewart WF, Godby S, Pollak J, Dewalle J, Larson S, et al. Body mass index and the built and social environments in children and adolescents using electronic health records. Am J Prev Med 2011;41(4):e17-e28. [doi: 10.1016/j.amepre.2011.06.038] [Medline: 21961475]
50. Sharif-Askary B, Bittar PG, Farjat AE, Liu B, Vissoci JRN, Allori AC. Geospatial analysis of risk factors contributing to loss to follow-up in cleft lip/palate care. Plast Reconstr Surg Glob Open 2018;6(9):e1910. [doi: 10.1097/GOX.0000000000001910] [Medline: 30349785]

51. Sidell MA, Chen Z, Huang BZ, Chow T, Eckel SP, Martinez MP, et al. Ambient air pollution and COVID-19 incidence during four 2020-2021 case surges. Environ Res 2022;208:112758 [FREE Full text] [doi: 10.1016/j.envres.2022.112758] [Medline: 35063430]
52. Siegel SD, Brooks MM, Sims-Mourtada J, Schug ZT, Leonard DJ, Petrelli N, et al. A population health assessment in a community cancer center catchment area: triple-negative breast cancer, alcohol use, and obesity in New Castle County, Delaware. Cancer Epidemiol Biomarkers Prev 2022;31(1):108-116 [FREE Full text] [doi: 10.1158/1055-9965.EPI-21-1031] [Medline: 34737210]
53. Sun Y, Li X, Benmarhnia T, Chen JC, Avila C, Sacks DA, et al. Exposure to air pollutant mixture and gestational diabetes mellitus in Southern California: results from electronic health record data of a large pregnancy cohort. Environ Int 2022;158:106888 [FREE Full text] [doi: 10.1016/j.envint.2021.106888] [Medline: 34563749]
54. Tabano DC, Bol K, Newcomer SR, Barrow JC, Daley MF. The spatial distribution of adult obesity prevalence in Denver County, Colorado: an empirical bayes approach to adjust EHR-derived small area estimates. EGEMS (Wash DC) 2017;5(1):24 [FREE Full text] [doi: 10.5334/egems.245] [Medline: 29881741]
55. Wakefield DV, Carnell M, Dove AP, Edmonston DY, Garner WB, Hubler A, et al. Location as destiny: identifying geospatial disparities in radiation treatment interruption by neighborhood, race, and insurance. Int J Radiat Oncol Biol Phys 2020;107(4):815-826. [doi: 10.1016/j.ijrobp.2020.03.016] [Medline: 32234552]
56. Wilson WW, Chua RFM, Wei P, Besser SA, Tung EL, Kolak M, et al. Association between acute exposure to crime and individual systolic blood pressure. Am J Prev Med 2022;62(1):87-94 [FREE Full text] [doi: 10.1016/j.amepre.2021.06.017] [Medline: 34538556]
57. Winckler B, Nguyen M, Khare M, Patel A, Crandal B, Jenkins W, et al. Geographic variation in acute pediatric mental health utilization. Acad Pediatr 2023;23(2):448-456. [doi: 10.1016/j.acap.2022.07.026] [Medline: 35940570]
58. Xie SJ, Kapos FP, Mooney SJ, Mooney S, Stephens KA, Chen C, et al. Geospatial divide in real-world EHR data: analytical workflow to assess regional biases and potential impact on health equity. AMIA Jt Summits Transl Sci Proc 2023;2023:572-581 [FREE Full text] [Medline: 37350875]
59. Zhan FB, Morshed N, Kluz N, Candelaria B, Baykal-Caglar E, Khurshid A, et al. Spatial insights for understanding colorectal cancer screening in disproportionately affected populations, Central Texas, 2019. Prev Chronic Dis 2021;18:E20 [FREE Full text] [doi: 10.5888/pcd18.200362] [Medline: 33661726]
60. Zhao YQ, Norton D, Hanrahan L. Small area estimation and childhood obesity surveillance using electronic health records. PLoS One 2021;16(2):e0247476 [FREE Full text] [doi: 10.1371/journal.pone.0247476] [Medline: 33606784]
61. Zhao P, Kwan MP, Zhou S. The uncertain geographic context problem in the analysis of the relationships between obesity and the built environment in Guangzhou. Int J Environ Res Public Health 2018;15(2):308 [FREE Full text] [doi: 10.3390/ijerph15020308] [Medline: 29439392]
62. Yu W. Spatial co-location pattern mining for location-based services in road networks. Expert Syst Appl 2016;46:324-335. [doi: 10.1016/j.eswa.2015.10.010]
63. Moazeni M, Maracy MR, Dehdashti B, Ebrahimi A. Spatiotemporal analysis of COVID-19, air pollution, climate, and meteorological conditions in a metropolitan region of Iran. Environ Sci Pollut Res Int 2022;29(17):24911-24924 [FREE Full text] [doi: 10.1007/s11356-021-17535-x] [Medline: 34826084]
64. Diggle PJ. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. Boca Raton, FL: CRC Press; 2013:300.
65. Okabe A, Satoh T, Sugihara K. A kernel density estimation method for networks, its computational method and a GIS‐based tool. Int J Geogr Inf Sci 2009;23(1):7-32. [doi: 10.1080/13658810802475491]
66. Fu WJ, Jiang PK, Zhou GM, Zhao KL. Using Moran's I and GIS to study the spatial pattern of forest litter carbon density in a subtropical region of southeastern China. Biogeosciences 2014;11(8):2401-2409. [doi: 10.5194/bg-11-2401-2014]
67. Anselin L. Local indicators of spatial association—LISA. Geogr Anal 2010;27(2):93-115. [doi: 10.1111/j.1538-4632.1995.tb00338.x]
68. Lee SI. Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. J Geogr Syst 2001;3:369-385. [doi: 10.1007/s101090100064]
69. Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 1995;27(4):286-306. [doi: 10.1111/j.1538-4632.1995.tb00912.x]
70. Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods 1997;26(6):1481-1496. [doi: 10.1080/03610929708831995]
71. Sheehan TJ, DeChello LM, Kulldorff M, Gregorio DI, Gershman S, Mroszczyk M. The geographic distribution of breast cancer incidence in Massachusetts 1988 to 1997, adjusted for covariates. Int J Health Geogr 2004;3(1):17 [FREE Full text] [doi: 10.1186/1476-072X-3-17] [Medline: 15291960]
72. Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 2007;30(5):609-628. [doi: 10.1111/j.2007.0906-7590.05171.x]
73. Kumar VS, Devika S, George S, Jeyaseelan L. Spatial mapping of acute diarrheal disease using GIS and estimation of relative risk using empirical Bayes approach. Clin Epidemiol Global Health 2017;5(2):87-96. [doi: 10.1016/j.cegh.2016.07.004]

74. Wah W, Ahern S, Earnest A. A systematic review of Bayesian spatial-temporal models on cancer incidence and mortality. Int J Public Health 2020;65(5):673-682. [doi: 10.1007/s00038-020-01384-5] [Medline: 32449006]
75. Shiffrin RM, Lee MD, Kim W, Wagenmakers EJ. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. Cogn Sci 2008;32(8):1248-1284 [FREE Full text] [doi: 10.1080/03640210802414826] [Medline: 21585453]
76. Social determinants of health. U.S. Department of Health and Human Services. URL: https://health.gov/healthypeople/priority-areas/social-determinants-health [accessed 2024-01-01]
77. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011;4:13 [FREE Full text] [doi: 10.1186/1755-8794-4-13] [Medline: 21269473]
78. Overweight & obesity statistics. National Institute of Diabetes and Digestive and Kidney Diseases. 2021. URL: https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity [accessed 2023-09-18]
79. Depression. PheKB. 2018. URL: https://phekb.org/phenotype/depression [accessed 2023-09-18]
80. Zandbergen PA. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. Adv Med 2014;2014:567049 [FREE Full text] [doi: 10.1155/2014/567049] [Medline: 26556417]
81. Roongpiboonsopit D, Karimi HA. Quality assessment of online street and rooftop geocoding services. Cartogr Geogr Inf Sci 2010;37(4):301-318. [doi: 10.1559/152304010793454318]
82. Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG. An effective and efficient approach for manually improving geocoded data. Int J Health Geogr 2008;7:60 [FREE Full text] [doi: 10.1186/1476-072X-7-60] [Medline: 19032791]
83. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, et al. Geocoding in cancer research: a review. Am J Prev Med 2006;30(2 Suppl):S16-S24. [doi: 10.1016/j.amepre.2005.09.011] [Medline: 16458786]
84. Baker S, Holt KE, Clements ACA, Karkey A, Arjyal A, Boni MF, et al. Combined high-resolution genotyping and geospatial analysis reveals modes of endemic urban typhoid fever transmission. Open Biol 2011;1(2):110008 [FREE Full text] [doi: 10.1098/rsob.110008] [Medline: 22645647]
85. Canino G. A system for geoanalysis of clinical and geographical data. 2014 Presented at: SIGSPATIAL '14: 22nd SIGSPATIAL International Conference on Advances in Geographic Information Systems; November 4, 2014; Dallas, TX p. 57-62. [doi: 10.1145/2676629.2676635]
86. Aswi A, Cramb SM, Moraga P, Mengersen K. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. Epidemiol Infect 2018;147:e33 [FREE Full text] [doi: 10.1017/S0950268818002807] [Medline: 30369335]
87. Bharadiya JP. A review of Bayesian machine learning principles, methods, and applications. Int J Innovative Sci Res Technol 2023;8(5):2033-2038. [doi: 10.5281/zenodo.8020825]
88. Walsh AS, Louis TA, Glass GE. Detecting multiple levels of effect during survey sampling using a Bayesian approach: point prevalence estimates of a hantavirus in hispid cotton rats (Sigmodon hispidus). Ecol Modell 2007;205(1-2):29-38. [doi: 10.1016/j.ecolmodel.2007.01.016]
89. Hanzlicek GA, Raghavan RK, Ganta RR, Anderson GA. Bayesian space-time patterns and climatic determinants of bovine anaplasmosis. PLoS One 2016;11(3):e0151924 [FREE Full text] [doi: 10.1371/journal.pone.0151924] [Medline: 27003596]
90. Wintle BA, McCarthy MA, Volinsky CT, Kavanagh RP. The use of Bayesian model averaging to better represent uncertainty in ecological models. Conserv Biol 2003;17(6):1579-1590. [doi: 10.1111/j.1523-1739.2003.00614.x]
91. Anselin L, Varga A, Acs Z. Geographical spillovers and university research: a spatial econometric perspective. Growth Change 2002;31(4):501-515. [doi: 10.1111/0017-4815.00142]
92. Mollalo A, Tatar M. Spatial modeling of COVID-19 vaccine hesitancy in the United States. Int J Environ Res Public Health 2021;18(18):9488 [FREE Full text] [doi: 10.3390/ijerph18189488] [Medline: 34574416]
93. Carter AJ, Nguyen CN. A comparison of cancer burden and research spending reveals discrepancies in the distribution of research funding. BMC Public Health 2012;12:526 [FREE Full text] [doi: 10.1186/1471-2458-12-526] [Medline: 22800364]
94. Varnousfaderani SD, Musazadeh V, Ghalichi F, Kavyani Z, Razmjouei S, Faghfouri AH, et al. Alleviating effects of coenzyme Q10 supplements on biomarkers of inflammation and oxidative stress: results from an umbrella meta-analysis. Front Pharmacol 2023;14:1191290 [FREE Full text] [doi: 10.3389/fphar.2023.1191290] [Medline: 37614320]
95. Brown JS, Maro JC, Nguyen M, Ball R. Using and improving distributed data networks to generate actionable evidence: the case of real-world outcomes in the food and drug administration's sentinel system. J Am Med Inform Assoc 2020;27(5):793-797 [FREE Full text] [doi: 10.1093/jamia/ocaa028] [Medline: 32279080]

Abbreviations

CAR: conditional autoregressive
CO-MRSA: community-onset methicillin-resistant Staphylococcus aureus
DM: dermatomyositis
EDW: enterprise data warehouse
EHR: electronic health record

EMR: electronic medical record
EPR: electronic patient record
GAM: generalized additive model
GMI: global Moran I
LISA: local Moran I
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SDOH: social determinants of health

Edited by Q Chen; submitted 13.01.24; peer-reviewed by C Zhao, S Tian, CY Hsu, A Zgodic; comments to author 21.06.24; revised version received 30.07.24; accepted 11.09.24; published 15.10.24.

Please cite as:
Mollalo A, Hamidi B, Lenert LA, Alekseyenko AV
Application of Spatial Analysis on Electronic Health Records to Characterize Patient Phenotypes: Systematic Review
JMIR Med Inform 2024;12:e56343
URL: https://medinform.jmir.org/2024/1/e56343
doi:10.2196/56343
PMID:

Targeted Development and Validation of Clinical Prediction Models in Secondary Care Settings: Opportunities and Challenges for Electronic Health Record Data

I S van Maurik[1], PhD; H J Doodeman[1], MSc; B W Veeger-Nuijens[1], MSc; R P M Möhringer[2], MSc; D R Sudiono[2,3], MD; W Jongbloed[4], PhD; E van Soelen[1], MD

1

2

3

4

Corresponding Author:
I S van Maurik, PhD

Abstract

Before deploying a clinical prediction model (CPM) in clinical practice, its performance needs to be demonstrated in the population of intended use. This is also called "targeted validation." Many CPMs developed in tertiary settings may be most useful in secondary care, where the patient case mix is broad and practitioners need to triage patients efficiently. However, since structured or rich datasets of sufficient quality from secondary to assess the performance of a CPM are scarce, a validation gap exists that hampers the implementation of CPMs in secondary care settings. In this viewpoint, we highlight the importance of targeted validation and the use of CPMs in secondary care settings and discuss the potential and challenges of using electronic health record (EHR) data to overcome the existing validation gap. The introduction of software applications for text mining of EHRs allows the generation of structured "big" datasets, but the imperfection of EHRs as a research database requires careful validation of data quality. When using EHR data for the development and validation of CPMs, in addition to widely accepted checklists, we propose considering three additional practical steps: (1) involve a local EHR expert (clinician or nurse) in the data extraction process, (2) perform validity checks on the generated datasets, and (3) provide metadata on how variables were constructed from EHRs. These steps help to generate EHR datasets that are statistically powerful, of sufficient quality and replicable, and enable targeted development and validation of CPMs in secondary care settings. This approach can fill a major gap in prediction modeling research and appropriately advance CPMs into clinical practice.

(JMIR Med Inform 2024;12:e57035) doi:10.2196/57035

KEYWORDS

clinical prediction model; electronic health record; targeted validation; EHR; EMR; prediction models; validation; CPM; secondary care; machine learning; artificial intelligence; AI

Background

In health care, distinct tiers of care, namely primary, secondary, and tertiary care, play vital roles in addressing patients' diverse medical needs. Patients requiring specialized medical attention or hospital care are generally treated in secondary care settings. Approximately one-third of primary care patients are referred to secondary care, and the majority of these patients are treated and monitored in this setting [1]. Tertiary care consists of highly specialized services for highly complex diseases. Less than 5% of patients require care in a tertiary setting. The distribution of patients across primary and secondary care settings may differ between countries and health care systems; some countries require a referral from primary care to enter secondary care, while in other countries patients have direct access to medical specialists without a referral. While there can be significant variability in primary and secondary care structures, variablity in tertiary care structures are generally less pronounced. This is because tertiary care focuses on highly specialized and complex conditions that are often standardized based on international research and protocols. Academic hospitals and specialized centers provide similar highly specialized care worldwide.

Due to the complexity of care, strong research facilities, and involvement in clinical trials, most clinical understanding and knowledge of medical conditions come from patients treated in tertiary settings [2,3]. Similarly, in this setting many clinical prediction models (CPMs) are developed. A CPM is a statistical or artificial intelligence-based tool used in health care to predict future health events in individual patients using a set of predictors or risk factors. CPMs have the potential to combine and weigh large amounts of patient information, enabling the

stratification of patients based on their risk of future health events. This informs decision-making processes and may guide the allocation of resources and interventions. While such models are also developed in primary and secondary settings, CPMs developed in tertiary settings may have great potential to be useful in secondary care, where the patient case mix is broad and practitioners need to triage patients efficiently.

However, the usefulness of such CPMs depends significantly on their quality in the population of intended use. Recent discussions emphasize the importance of targeted validation, which is the assessment of a CPM's quality in the specific population for which it is intended. Yet, this specification of the population of intended use is often lacking in publications [4]. Secondary health care settings, where large numbers of patients with specialized medical needs are treated, accumulate vast amounts of data in electronic health records (EHR) on a daily basis. Despite this potential, CPMs are often not developed or validated on data from secondary care populations due to the scarcity of appropriate datasets. This is known as the "validation gap." In this viewpoint, we discuss the opportunities and challenges faced when considering EHR data from secondary health care settings for the development or validation of CPMs.

Importance of Targeted Validation of CPMs

The performance of CPMs is significantly influenced by the case mix of patients (ie, baseline characteristics of the patients) and the prevalence of the outcome [5-7]. The case mix of a secondary care population is essentially different from a tertiary care population. Due to these case mix differences, a CPM developed in tertiary care often performs poorly in secondary care populations [8].

For instance, in cardiovascular risk prediction models, such disparities in patient characteristics and outcomes between tertiary and secondary care settings substantially impact model performance. Research by Wynants et al [3] highlights the challenges of model transportability and generalizability. In a review, they showed that 23 out of 50 studies did not describe the population of intended use. In those studies that reported health care setting, all participating centers were in tertiary or academic settings. One of the studies applied a tertiary CPM in a secondary care setting. In this secondary care setting, patients were older, the outcome was less prevalent, and patients more often had (multiple) risk factors such as diabetes and hypertension. Under these circumstances, the CPM severely overestimated event probabilities when applied to secondary care. Similarly, in chronic obstructive pulmonary disease management, the use of CPMs is complicated by variations in patient profiles across health care settings. While primary and secondary care cohorts exhibit marked heterogeneity in health status, tertiary care cohorts tend to comprise more homogeneous samples [2].

These examples, along with many others in medical literature [9,10], demonstrate poor model performance, specifically poor calibration, of tertiary CPMs in the population of intended use. Arguably, these prediction models are most useful at lower levels of care, where the patient case mix is broad and practitioners need to triage patients efficiently [11]. More concretely, an overestimation of event probabilities means that patients could be incorrectly categorized as high-risk based on a CPM that is poorly calibrated to the target population. Such inaccurate risk prediction can be misleading and may negatively influence clinical practice; it may lead to false expectations from the patientor professional, or patients may make personal decisions in anticipation (or absence) of an event [12]. CPM specialists argue that poor calibration may render an algorithm less clinically useful than a competing model with lower discriminative ability but is well calibrated [9].

While checklists exist to improve reporting quality [13,14] and assess the risk of bias [15] in CPM development and validation, targeted validation remains an uncommon practice. Sperrin et al [4] rightly argue that we should report the intended population of use more explicitly. This means that if, for example, a CPM is intended to aid decision-making in a secondary care setting in the Netherlands, then it should be developed and validated in a secondary care setting in the Netherlands. Such targeted validation requires data from the population of intended use. The difficulty lies in the scarcity of structured or rich datasets from secondary settings available to assess the quality of a CPM. Addressing this validation gap remains a challenge in CPM literature and hampers the implementation of CPMs in clinical practice, emphasizing the importance of leveraging EHR data from secondary health care settings for CPM development and validation.

EHR Datasets and Text Mining Tools

Every day, hospitals collect an enormous amount of health information in EHRs. Data in these EHRs have structured and unstructured formats. Structured EHR data comprise data in fixed numerical or categorical areas, such as diagnoses, prescriptions, and laboratory values, while unstructured data includes clinical documentation such as notes, referral letters, or discharge summaries produced by health care personnel [16]. These documents are inputted as free text into EHRs and offer a complete picture of a patient's condition. It is estimated that more than 70% of EHR data is stored as free text. Even information that seems structured, such as a total score from a questionnaire, is often stored as free text in the EHR in letters or notes. To conduct good research in general, this data should be converted into structured formats and datasets. Specifically, to validate a CPM, it is required that a certain predictor, which is part of the CPM, is collected and recorded in a consistent manner. Leveraging the value of unstructured data is key to generating meaningful insights from clinical data [16-18].

Text mining tools and natural language processing (NLP) techniques allow us to transform unstructured documents into a structured format to enable analysis and the generation of high-quality CPMs [19]. Text mining applications are increasingly used in research and computational settings, but are now also commercialized in software applications (for example, CTcue from IQVIA or Amazon Comprehend Medical from Amazon Web Services) that allow hospitals to generate structured data and subsequently cohorts of patients with a

specific disease more efficiently from their electronic medical files.

With respect to targeted development and validation of CPMs, specific predictors can be found more easily, especially when the CPM is based on commonly used clinical measures and data. These datasets are often very large and are therefore statistically powerful [16].

EHR Data Quality

Overview

Leveraging EHR data for research brings challenges with regard to data quality. These records are prone to ascertainment bias and missingness [18,20], especially concerning free text data, where semantic and context understanding are required to correctly classify types of information. Furthermore, data quality depends on how and if a clinician records information in the EHR [19]. This may be even more problematic in secondary teaching hospitals, which have a higher turnover of personnel. Another challenge is information overload, which poses a substantial problem in accessing a particular, significant piece of information from vast datasets. A recent systematic review shows additional technical challenges such as lack of labeled data, spelling correction, medical abbreviation, negation detection, and clinical entity recognition [21].

Data quality is a key contributor to the quality and success of developed CPMs: “rubbish in, rubbish out.” While NLP software is developing rapidly and their quality improves, their output needs to be checked and validated carefully. When using EHR data for the development and validation of CPMs, alongside the widely accepted checklists, we propose additionally considering the following steps in the data extraction process:

Step 1: Include a Clinician, Nurse, or Health Care Professional as the Local EHR Expert and in the EHR Data Extraction Process

This is not always the case, as data extraction may be conducted by supporting staff, business intelligence specialists, or students and interns. However, clinicians have firsthand knowledge of their patients’ conditions, treatments, and histories. Clinicians and health care professionals may be aware of certain patient details that are not well-documented in the EHR, such as informal diagnoses or symptoms not coded in the system. Including this information helps create a more comprehensive and accurate dataset, informing the EHR data capturing process. With regard to unstructured data; discuss the clinical workflow and how and when specific clinical notes are made. As a simple example: when extracting data from the “medical history” part of medical notes, you might find “Hypertension: -.” Does this mean that information on hypertension for this patient is missing, is not applicable, or is absent?

With regard to structured data, check (if applicable) whether protocol changes occurred in the period of interest. Unlike research databases, major protocol changes are not documented in EHRs. In a hospital setting, system updates are regularly performed, new equipment is purchased, or measurement methods are changed. This is not documented in the EHR of individual patients. When using EHR data for research purposes, such as developing or validating a CPM, these organizational factors should be considered. For example, if the clinical chemistry laboratory first measured thyroid hormone FT4 with a Beckman Coulter analyzer with normal values between 7 - 16 pmol/L and later switched to Siemens with normal values between 11 - 21 pmol/L, this significantly influences the outcome of CPMs including FT4. Another example is the measurement of the tumor marker carcinoembryonic antigen, where levels of carcinoembryonic antigen measured with Siemens are approximately 25% lower compared to those measured with Beckman Coulter. Harmonization of such laboratory results within a hospital, but also between hospitals, is therefore important and requires knowledge of protocol changes over time.

Step 2: Perform Validity Checks on the Generated Dataset

Data validation and verification are broadly accepted exercises in research settings. It is the process of checking whether entered data is accurate and consistent. This may encompass the crosschecking of data in a random set of cases, which may be even more relevant in research where data is derived from EHRs. EHR data are complex and heterogeneous, originating from different systems, formats, and medical practices. This variability can introduce inconsistencies and errors. Validation and verification processes are essential to standardize the data, correct inconsistencies, and ensure uniformity in the data used for research.

In addition to checking the data quality of specific variables extracted from EHRs, we advise also executing a crosscheck on the included cases. Specifically, if software is used to compose a cohort, let a clinician provide a list of patients that they believe should be included in the generated dataset, and check whether that is indeed the case (ie, do I find the cases that I should find?). This is important for a number of reasons. First, clinicians can identify patients who meet specific criteria based on nuances that may not be captured in the EHR data alone. This clinical insight is invaluable for ensuring that the correct patients are included in the research cohort. Second, automated systems rely on predefined algorithms to identify patients, but these algorithms can sometimes miss relevant cases or include irrelevant ones. Lastly, clinicians can provide supplementary information to fill gaps in the EHR data, enhancing the completeness and richness of the dataset. This additional information can improve the robustness of the research outcomes.

Step 3: Deliver Information or Metadata on How Certain Variables Are Constructed

Information should be provided and made publicly available on whether a variable is composed from structured codes or from a search in unstructured free-text (for example, reports) and include a list of search terms used (or excluded). Delivering detailed information or metadata on how certain variables are constructed, and understanding whether these variables come from structured or unstructured electronic patient record data, enhances data quality and integrity during the data extraction process and is crucial for transparency and reproducibility [22].

Knowing whether a variable comes from structured (eg, coded fields or predefined formats) or unstructured (eg, free-text notes or narratives) data is essential as they have different characteristics. Structured data is generally more reliable, easier to analyze, and in some cases similar across hospitals (eg, Anatomical Therapeutic Chemical Classification System codes). It follows a predefined format, making it straightforward to extract and use in statistical analyses. Unstructured data, on the other hand, is rich in detailed information but more challenging to analyze due to variability and complexity. NLP and other sophisticated methods are often required to extract meaningful information from unstructured data. Clear documentation of variable construction enhances the impact and credibility of research findings, making it easier for clinicians and policy makers to apply the results in real-world settings.

Conclusion

CPMs may be particularly valuable in secondary care settings, and the introduction of software applications for text mining of EHRs allows the generation of structured "big" datasets. However, the imperfection of EHRs as a research database requires careful validation of data quality. On using EHR data for the development and validation of CPMs, alongside the widely accepted checklists, we propose to additionally consider three practical steps: (1) let a local EHR expert (clinician or nurse) be involved in the data extraction process, (2) perform validity checks on the generated datasets, and (3) provide metadata on how variables were constructed from EHRs. If successful, such datasets are statistically powerful and enable targeted development and validation of CPMs in secondary care settings, filling a major gap in prediction modeling research.

Authors' Contributions

ISvM interpreted the literature and wrote this paper. All other authors revised this paper. All authors read and approved the final paper.

Conflicts of Interest

None declared.

References

1. Heins M, et al. Zorg door de huisarts: Nivel Zorgregistraties Eerste Lijn. Jaarcijfers 2021 en trendcijfers 2017-2021. Nivel. 2022. URL: https://www.nivel.nl/sites/default/files/bestanden/1004273.pdf [accessed 2024-10-23]
2. de Klein MM, Peters JB, van 't Hul AJ, et al. Comparing health status between patients with COPD in primary, secondary and tertiary care. NPJ Prim Care Respir Med 2020 Sep 8;30(1):39. [doi: 10.1038/s41533-020-00196-7] [Medline: 32901030]
3. Wynants L, Kent DM, Timmerman D, Lundquist CM, Van Calster B. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. Diagn Progn Res 2019;3:6. [doi: 10.1186/s41512-019-0046-9] [Medline: 31093576]
4. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagn Progn Res 2022 Dec 22;6(1):24. [doi: 10.1186/s41512-022-00136-8] [Medline: 36550534]
5. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023 Feb 24;21(1):70. [doi: 10.1186/s12916-023-02779-w] [Medline: 36829188]
6. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol 2013 Mar 6;13:33. [doi: 10.1186/1471-2288-13-33] [Medline: 23496923]
7. Nieboer D, van der Ploeg T, Steyerberg EW. Assessing discriminative performance at external validation of clinical prediction models. PLoS ONE 2016;11(2):e0148820. [doi: 10.1371/journal.pone.0148820] [Medline: 26881753]
8. Smid DE, Spruit MA, Houben-Wilke S, et al. Burden of COPD in patients treated in different care settings in the Netherlands. Respir Med 2016 Sep;118:76-83. [doi: 10.1016/j.rmed.2016.07.015] [Medline: 27578474]
9. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med 2019 Dec 16;17(1):230. [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
10. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 2015 Feb;35(2):162-169. [doi: 10.1177/0272989X14547233] [Medline: 25155798]
11. Weimar C, Diener HC, Alberts MJ, et al. The Essen stroke risk score predicts recurrent cardiovascular events: a validation within the REduction of Atherothrombosis for Continued Health (REACH) registry. Stroke 2009 Feb;40(2):350-354. [doi: 10.1161/STROKEAHA.108.521419] [Medline: 19023098]
12. Lipkus IM. Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. Med Decis Making 2007;27(5):696-713. [doi: 10.1177/0272989X07307271] [Medline: 17873259]
13. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 6;162(1):W1-73. [doi: 10.7326/M14-0698] [Medline: 25560730]
14. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 7;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]

15. Fernandez-Felix BM, López-Alcalde J, Roqué M, Muriel A, Zamora J. CHARMS and PROBAST at your fingertips: a template for data extraction and risk of bias assessment in systematic reviews of predictive models. BMC Med Res Methodol 2023 Feb 17;23(1):44. [doi: 10.1186/s12874-023-01849-0] [Medline: 36800933]
16. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;Suppl 1(Suppl 1):S48-S61. [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
17. Ehrenstein V, Kharrazi H, Lehmann H, Taylor T. Chapter 4. Obtaining data from electronic health records. In: Gliklich RE LM, Dreyer NA, editors. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide: Agency for Healthcare Research and Quality; 2019.
18. Hek K, Rolfes L, van Puijenbroek EP, et al. Electronic health record-triggered research infrastructure combining real-world electronic health record data and patient-reported outcomes to detect benefits, risks, and impact of medication: development study. JMIR Med Inform 2022 Mar 16;10(3):e33250. [doi: 10.2196/33250] [Medline: 35293877]
19. Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med 2023 Mar;155:106649. [doi: 10.1016/j.compbiomed.2023.106649] [Medline: 36805219]
20. Khurshid S, Reeder C, Harrington LX, et al. Cohort design and natural language processing to reduce bias in electronic health records research. NPJ Digit Med 2022 Apr 8;5(1):47. [doi: 10.1038/s41746-022-00590-0] [Medline: 35396454]
21. Tornero-Costa R, Martinez-Millana A, Azzopardi-Muscat N, Lazeri L, Traver V, Novillo-Ortiz D. Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review. JMIR Ment Health 2023 Feb 2;10:e42045. [doi: 10.2196/42045] [Medline: 36729567]
22. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]

Abbreviations

CPM: clinical prediction model
EHR: electronic health record
NLP: natural language processing

Edited by C Lovis; submitted 02.02.24; peer-reviewed by K Rahmani, M Sperrin; revised version received 11.07.24; accepted 21.07.24; published 24.10.24.

Please cite as:
van Maurik IS, Doodeman HJ, Veeger-Nuijens BW, Möhringer RPM, Sudiono DR, Jongbloed W, van Soelen E
Targeted Development and Validation of Clinical Prediction Models in Secondary Care Settings: Opportunities and Challenges for Electronic Health Record Data
JMIR Med Inform 2024;12:e57035
URL: https://medinform.jmir.org/2024/1/e57035
doi:10.2196/57035

Unintended Consequences of Data Sharing Under the Meaningful Use Program

Irmgard Ursula Willcockson, PhD; Ignacio Herman Valdes, MS, MD

Astronaut, LLC, 7505 Fannin Street, Suite 170, Houston, TX, United States

Corresponding Author:
Ignacio Herman Valdes, MS, MD

Abstract

Interoperability has been designed to improve the quality and efficiency of health care. It allows the Centers for Medicare and Medicaid Services to collect data on quality measures as a part of the Meaningful Use program. Covered providers who fail to provide data have lower rates of reimbursement. Unintended consequences also arise at each step of the data collection process: (1) providers are not reimbursed for the extra time required to generate data; (2) patients do not have control over when and how their data are provided to or used by the government; and (3) large datasets increase the chances of an accidental data breach or intentional hacker attack. After detailing the issues, we describe several solutions, including an appropriate data use review board, which is designed to oversee certain aspects of the process and ensure accountability and transparency.

(JMIR Med Inform 2024;12:e52675) doi:10.2196/52675

KEYWORDS

electronic health records; EHR; medical record; interoperability; health information interoperability; clinical burden; Medicare; Medicaid; reimbursement; data science; data governance; data breach; cybersecurity; privacy

Introduction

Background

Interoperability has been an overarching goal of the American health care industry since the American Recovery and Reinvestment Act in 2009. Provider-to-provider sharing of patient information was designed to improve the safety, quality, and efficiency of patient care. The next target for interoperability was provider-to-patient sharing of information. The 21st Century Cures Act called for patients to have electronic access to their health care record [1]. Health Level 7 Fast Healthcare Interoperability Resources was developed for secure data exchange between computer systems using different information storage methods [2].

Interoperability is a prerequisite for meaningful use (MU). The Health Information Technology for Economic and Clinical Health Act, a component of the American Recovery and Reinvestment Act, included incentives for providers to adopt electronic health records (EHRs) as well as penalties for failing MU [3]. The Meaningful Use program requires providers to share data with the Centers for Medicare and Medicaid Services (CMS) to ensure satisfactory health care quality. Using the Quality Reporting Document Architecture [4], providers are required to submit data in support of the quality measures applicable to their practice. If covered providers do not submit data, their rate of reimbursement is reduced.

Although interoperability has been designed to improve the quality and efficiency of patient care, there are unintended consequences of data sharing. To date, few articles have examined the negative consequences of data sharing by providers to governmental entities [5,6]. Data must be generated before it can be shared, so this article begins with clinical data generation.

Ethical Considerations

This study does not include human subject research (no human subject experimentation or intervention was conducted) and so does not require institutional review board (IRB) approval.

Data Generation

Data are generated by both providers and patients during a visit. While most data are generated to serve clinical needs, some are generated solely to satisfy later MU reporting.

Uniform Data Generation

The criteria for satisfying a particular MU requirement do not consider differences in specialty. For example, annual screening for depression is one MU measure (Centers for Medicare and Medicaid Services 2, version 12, 2023) [7]. Patients with depression are frequently undiagnosed in general medicine settings and therefore remain untreated. Explicit screening for depression in these settings is important to identify patients and offer appropriate treatment [8,9]. Administering a validated screening instrument is one way to satisfy the screening portion of the MU requirement.

In contrast, psychiatrists screen patients for depression using implicit methods such as interviewing the patient [10,11]. Requiring them to report an instrument score for each patient

does not further the goal of identifying and treating patients with depression. It does, however, increase the burden of data collection and documentation on both the provider and patient.

Proposed Solution: Broader Criteria for Meeting MU Requirements

Instead of forcing all providers to perform the same task to meet a particular MU requirement, providers should be allowed limited flexibility to decide how best to meet the intent of the requirement.

Free-Rider Problem

In the broadest definition, a free rider receives a benefit without contributing to the cost of that benefit's production [12]. Billing codes determine the provider's compensation for a given patient visit. These codes in part represent the amount of time spent on a given activity. However, there is no billing code for collecting data only required for CMS reporting. Providers are penalized for not providing the data but are not compensated for their time spent; thus, the CMS is acting as a free rider.

Proposed Solution: Current Procedural Terminology Code Modifiers

Current Procedural Terminology code modifiers further describe a procedure code without changing its definition [13]. Creating a modifier specifically for MU data collection would allow providers to bill for the time spent on clinical data collection. Additional payment for data collection should be made by the CMS, since it receives the reports.

Data Ownership

Access to Data

Ownership of data has philosophical implications that differ from ownership of real property [14,15]. A more useful framework for understanding the consequences of MU data collection may be access to data. Patients have a compelling interest in managing their own data. Patients who do not trust their provider with safeguarding their data may withhold information, leading to adverse outcomes [6]. This is especially true for marginalized groups [16].

Despite advances in interoperability and data-sharing mandates [1], neither providers nor patients can usually access all data pertaining to their role. Health care entities create data silos and deny access to patients and providers who are not, or are no longer, part of the entity. Data do not commonly follow patients who are seen by multiple providers. Outside of closed systems such as the Veterans Affairs, in which patients receive all care within the system, data do not follow providers who work in multiple health care settings, or who change jobs.

Proposed Solution: Improved Interoperability Processes

While interoperability is constantly improving, the process of data sharing continues to be cumbersome. We propose the following process: When a patient begins receiving care at a particular entity, the consent form should include data sharing with providers outside of the entity. If the patient does not opt out, any provider who has the patient's demographics and certifies that they need access to protected health information (PHI) should receive it. Allowing patients to consent ahead of time lowers the burden on both the patients and the providers.

Data Sharing

Overview

Data can be shared with the CMS either by a provider via EHRs or by an insurance company. The CMS receives complete charts, including all PHI. Neither providers nor patients are directly involved in this process, and they may not know when or how often data sharing occurs.

Calculating Compliance

Many MU measures include a numerical criterion. To calculate the percentage of patients who have had a particular test or screening requires collecting not only the charts of patients satisfying the measure (the numerator) but also the charts of patients not satisfying the measure (the denominator). In practice, this means that any chart may be included in any dataset, and that many charts are included in multiple datasets. Patients and providers have no control over this process and no way of knowing which charts are included in which dataset.

Lack of Patient Consent

Patient consent forms include consent for sharing data with other members of their health care team, designated family members, and insurance companies. Currently, consent forms do not include sharing data with the CMS as part of MU requirements. Patients are unaware that their data are shared with the CMS, via their providers and their insurance companies, and have no mechanism to give or deny consent.

Proposed Solution: Updated Consent Forms and Opt Outs

Including a section on data sharing with the CMS in patient consent forms is straightforward. Both providers and insurance companies can include a section explaining that PHI may be shared with the CMS and allow patients to opt out.

However, opting out could have unintended consequences. Some patients may hesitate to share any health data with the CMS. More patients may hesitate to share sensitive data, such as mental health or substance misuse data, with the CMS. Therefore, the CMS would need to decide how to handle compliance calculations. Giving patients the opportunity to opt out of data sharing also requires changes to EHR programming.

Data Use and Potential Misuse

The intent of MU data collection is to improve the efficiency and effectiveness of health care. However, once collected, data can be used for any number of other purposes. At present, it is unclear what safeguards are in place to prevent other branches of government from accessing the data for their own purposes. For example, California allows a person to obtain a driver's license without proof of legal immigration status. The personal data collected can then be accessed by the Department of Homeland Security to perform civil immigration enforcement [17].

Data Breaches

The US government has experienced multiple data breaches over the last 15 years, both accidental and through hacking [18-21]. Improvements in processes at least partially address accidental breaches. Hacking by sophisticated foreign entities is more difficult to prevent. Hackers may be attracted by the combination of the volume of data and its potential sensitivity [22,23].

Proposed Solution: An Appropriate Data Use Review Board

We propose the creation of an independent board, modeled after an IRB, to address many of the concerns related to data collection and use. Research on human subjects requires the approval of an IRB [24,25]. This requirement was put in place in the United States after a series of egregiously unethical experiments was conducted. Each IRB is required to have at least 5 members, with at least one whose main concern is scientific, one whose main concern is nonscientific, and one who is not affiliated with the academic institution in which the proposed research would take place.

As the data recipient is the CMS, a part of the federal government, we propose the following composition for an appropriate data use review board, made up of a minimum of 6 members to allow for group decision-making:

- At least one member who is employed at the CMS
- At least one member who is a clinician providing data
- At least one member who represents patients
- At least one member with communication experience
- At least one member who is a biomedical informatician with big data expertise
- At least one member who owns and operates a small market-share EHR software company

Each of the members of the appropriate data use review board has a critical role to play. As the data recipient, the CMS needs to lay out both the data required and the rationale behind collecting the data. The clinician and EHR owner provide important insight into the impact of data collection on workflows, as well as the feasibility of modifying software to streamline data collection and reporting. The patient representative provides the patients' perspective as well as communication guidance. As any decision reached by the review board needs to be communicated clearly and effectively to patients, the communication specialist and patient representative would work together to craft and disseminate necessary information. The biomedical informatician can assist the CMS with deciding on data needs as well as suggest the most current data analysis methods. They can also help the patient representative and communication specialist explain different ways to protect patient data from unauthorized disclosure.

After an open application and vetting process, members to the board should be appointed by a bipartisan committee of the US House. Their term of service should be 4-5 years to allow members to become proficient in their roles.

Proposed uses of data should be approved by the committee and communicated to the public. Public input should be sought through a variety of means and become an important aspect of decision-making. Disclosure of data use and data breaches should be prompt and effective, without further compromising data security.

Effect of MU on the EHR Ecosystem

MU requirements affect not only patients and providers but also EHR companies. MU certification and recertification is time consuming and costly. While EHR companies with a large market share can justify the expense and pass the cost on to their customers, companies with a smaller market share cannot. Lack of certification leads to decreased market share, thereby encouraging consolidation across the industry.

Consolidation may be advantageous to the government because it is easier to negotiate with fewer companies. However, consolidation leads to increased costs for providers and less competition. It makes independent, autonomous practice, away from corporate monocultures, very difficult. Customer service also suffers because customers have fewer choices.

Customer service is not the only casualty; innovation is also affected. Smaller companies are more likely to produce innovative products. However, given the high bar of MU certification, bringing these innovations to the market often proves to be cost prohibitive. Similarly, though open-source software has also driven down costs and spurred innovation [26], given the expense associated with MU certification, many companies committed to open-source software have stopped providing their code freely.

Conclusion

MU was made possible by progressive advances in interoperability. While CMS data collection has the potential to advance health care, it leads to the aggregation of large datasets that are vulnerable to unintentional data breaches and data misuse. Since the data collection is largely invisible to both providers and patients, an appropriate data use review board is needed to protect all participants.

Acknowledgments

We want to thank Ross Koppel for insightful comments on an early draft of this paper.

Conflicts of Interest

IHV is the founder and chief executive officer of a small market-share electronic health record company.

References

1. H.R.6 - 21st Century Cures Act. Congress.gov. 2015. URL: https://www.congress.gov/bill/114th-congress/house-bill/6?s=1&r=6 [accessed 2024-11-04]
2. What is HL7 FHIR? Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/sites/default/files/page/2021-04/What%20Is%20FHIR%20Fact%20Sheet.pdf [accessed 2024-11-04]
3. H.R.1 - American Recovery and Reinvestment Act of 2009. Congress.gov. 2009. URL: https://www.congress.gov/bill/111th-congress/house-bill/1?s=10&r=1 [accessed 2024-11-04]
4. Sethi K. Introduction to QRDA. eCQI Resource Center. 2015 Aug 12. URL: https://ecqi.healthit.gov/system/files/qrda_basics_08_12_2015_a_508.pdf [accessed 2024-11-04]
5. Spithoff S, Stockdale J, Rowe R, McPhail B, Persaud N. The commercialization of patient data in Canada: ethics, privacy and policy. CMAJ 2022 Jan 24;194(3):E95-E97. [doi: 10.1503/cmaj.210455] [Medline: 35074837]
6. Turner GM, Monaco C. Doctor-patient relationship compromised by 'oppressive' quality reporting requirements. Forbes. 2018. URL: https://www.forbes.com/sites/gracemarieturner/2018/04/05/doctor-patient-relationship-compromised-by-oppressive-quality-reporting-requirements/ [accessed 2024-10-11]
7. Quality ID #134: preventive care and screening: screening for depression and follow-up plan. Quality Payment Program. URL: https://qpp.cms.gov/docs/QPP_quality_measure_specifications/CQM-Measures/2023_Measure_134_MIPSCQM.pdf [accessed 2024-11-05]
8. Bailey RK, Mokonogho J, Kumar A. Racial and ethnic differences in depression: current perspectives. Neuropsychiatr Dis Treat 2019 Feb 22;15:603-609. [doi: 10.2147/NDT.S128584] [Medline: 30863081]
9. Goodwin RD, Dierker LC, Wu M, Galea S, Hoven CW, Weinberger AH. Trends in U.S. depression prevalence from 2015 to 2020: the widening treatment gap. Am J Prev Med 2022 Nov;63(5):726-733. [doi: 10.1016/j.amepre.2022.05.014] [Medline: 36272761]
10. Nordgaard J, Sass LA, Parnas J. The psychiatric interview: validity, structure, and subjectivity. Eur Arch Psychiatry Clin Neurosci 2013 Jun;263(4):353-364. [doi: 10.1007/s00406-012-0366-z] [Medline: 23001456]
11. Nordgaard J, Parnas J. A semi structured, phenomenologically-oriented psychiatric interview: descriptive congruence in assessing anomalous subjective experience and mental status. Clin Neuropsychiatry 2012 Jun;9(3):1-6 [FREE Full text]
12. Hardin R. The free rider problem. Stanford Encyclopedia of Philosophy. URL: https://plato.stanford.edu/entries/free-rider/ [accessed 2024-10-11]
13. What are CPT code modifiers? how are they used? Medical Billing Analysts. 2022 Jan 10. URL: https://www.medicalbillinganalysts.com/what-are-cpt-code-modifiers-how-are-they-used [accessed 2024-10-11]
14. Hummel P, Braun M, Dabrock P. Own data? ethical reflections on data ownership. Philos Technol 2021 Sep;34(3):545-572. [doi: 10.1007/s13347-020-00404-9]
15. Evans BJ. Much ado about data ownership. Harv J Law Technol 2011;25(1):70-113 [FREE Full text]
16. Nong P, Williamson A, Anthony D, Platt J, Kardia S. Discrimination, trust, and withholding information from providers: implications for missing data and inequity. SSM Popul Health 2022 Apr 7;18:101092. [doi: 10.1016/j.ssmph.2022.101092] [Medline: 35479582]
17. How California driver's license records are shared with the Department of Homeland Security. National Immigration Law Center. 2018 Dec 16. URL: https://www.nilc.org/issues/immigration-enforcement/how-calif-dl-records-shared-with-dhs/ [accessed 2024-10-11]
18. Lord N. Top 10 biggest government data breaches of all time in the U.S. Digital Guardian. 2024 Aug 22. URL: https://www.digitalguardian.com/blog/top-10-biggest-us-government-data-breaches-all-time [accessed 2024-10-11]
19. Thrush G, Cameron C. Hackers breach U.S. marshall system with sensitive personal data. The New York Times. 2023 Feb 27. URL: https://www.nytimes.com/2023/02/27/us/politics/us-marshals-ransomware-hack.html [accessed 2024-10-11]
20. Fung B. Why the US government hack is literally keeping security experts awake at night. CNN. 2020 Dec 16. URL: https://www.cnn.com/2020/12/16/tech/solarwinds-orion-hack-explained/index.html [accessed 2024-10-11]
21. Cybersecurity resource center. US Office of Personnel Management. URL: https://www.opm.gov/cybersecurity-resource-center/ [accessed 2024-11-04]
22. Basil NN, Ambe S, Ekhator C, Fonkem E. Health records database and inherent security concerns: a review of the literature. Cureus 2022 Oct 11;14(10):e30168. [doi: 10.7759/cureus.30168] [Medline: 36397924]
23. Brown ML, Brown JF, Nikakhtar N. Personal health data at risk of foreign exploitation. Wiley. 2022 Feb 2. URL: https://www.wiley.law/alert-Personal-Health-Data-at-Risk-of-Foreign-Exploitation [accessed 2024-10-11]
24. Institutional review boards (IRBs) and protection of human subjects in clinical trials. US Food and Drug Administration. 2019 Sep 11. URL: https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/institutional-review-boards-irbs-and-protection-human-subjects-clinical-trials [accessed 2024-10-11]
25. Moon MR. The history and role of institutional review boards: a useful tension. Virtual Mentor 2009 Apr 1;11(4):311-321 [FREE Full text] [doi: 10.1001/virtualmentor.2009.11.4.pfor1-0904] [Medline: 23195065]
26. Brock A. What is open source, and why does it matter today? Open Access Government. 2022 Feb 8. URL: https://www.openaccessgovernment.org/open-source-technology/129261/ [accessed 2024-10-11]

Abbreviations

CMS: Centers for Medicare and Medicaid Services
EHR: electronic health record
IRB: institutional review board
MU: meaningful use
PHI: protected health information

Edited by C Lovis; submitted 12.09.23; peer-reviewed by J Ehrsam, L Heryawan, M Mun, S Gordon; revised version received 31.07.24; accepted 17.08.24; published 14.11.24.

Please cite as:
Willcockson IU, Valdes IH
Unintended Consequences of Data Sharing Under the Meaningful Use Program
JMIR Med Inform 2024;12:e52675
URL: https://medinform.jmir.org/2024/1/e52675
doi:10.2196/52675

Practical Applications of Large Language Models for Health Care Professionals and Scientists

Florian Reis[1], MD; Christian Lenz[1], MSc, MD, PhD; Manfred Gossen[2,3], PhD; Hans-Dieter Volk[4,5], MD, PhD; Norman Michael Drzeniek[4,5], MD, PhD

1

2

3

4

5

Corresponding Author:
Norman Michael Drzeniek, MD, PhD

Abstract

With the popularization of large language models (LLMs), strategies for their effective and safe usage in health care and research have become increasingly pertinent. Despite the growing interest and eagerness among health care professionals and scientists to exploit the potential of LLMs, initial attempts may yield suboptimal results due to a lack of user experience, thus complicating the integration of artificial intelligence (AI) tools into workplace routine. Focusing on scientists and health care professionals with limited LLM experience, this viewpoint article highlights and discusses 6 easy-to-implement use cases of practical relevance. These encompass customizing translations, refining text and extracting information, generating comprehensive overviews and specialized insights, compiling ideas into cohesive narratives, crafting personalized educational materials, and facilitating intellectual sparring. Additionally, we discuss general prompting strategies and precautions for the implementation of AI tools in biomedicine. Despite various hurdles and challenges, the integration of LLMs into daily routines of physicians and researchers promises heightened workplace productivity and efficiency.

(JMIR Med Inform 2024;12:e58478) doi:10.2196/58478

KEYWORDS

artificial intelligence; healthcare; chatGPT; large language model; prompting; LLM; applications; AI; scientists; physicians; health care

Introduction

Large language models (LLMs), exemplified by OpenAI's ChatGPT, have garnered significant attention as text-based artificial intelligence (AI) tools in record time [1]. While they excel in mimicking natural language patterns, LLMs lack inherent factual accuracy and struggle with tasks requiring, for instance, mathematical reasoning at the graduate level [2]. Current research explores the association between LLMs and external databases [3], as well as the combination of LLMs with preexisting data and software tools as seen in ChatGPT's incorporation within Microsoft Copilot. This facilitates the step toward retrieving and articulating coherent responses using factual information from web-based sources, which is particularly relevant in fields like health care and the natural sciences where supplementation with external and up-to-date data is an essential prerequisite for meeting professional standards. ChatGPT, for instance, already demonstrates versatile applications in the medical domain, spanning from the identification of research topics, aiding in medical education, and supporting clinical and laboratory diagnosis to enhancing knowledge dissemination among health care professionals, extracting medical knowledge, and engaging in medical consultation [4,5]. However, despite LLMs' seemingly ideal suitability for enhancing personal productivity and keeping up to date with latest research, many professionals in our personal circle, including one of the coauthors, report discouragement after a few initial experiences with widely available LLMs. From our experience, criticism that dampens the enthusiasm of inexperienced users for adopting LLMs in work routine includes (1) superficial responses from chatbots, which read well but fail to capture the level of factual detail often expected in professional context; (2) confabulated information that is factually wrong or has no factual basis (also called "hallucination"); (3) lack of citation and source disclosure; (4) limited control over output format; and (5) failure to implement adaptation requests or criticism. Fortunately, for many use cases, these limitations can be effectively overcome by precisely instructing the LLM. These input instructions, known as "prompts," can be formulated in natural language rather than code, thus enabling intuitive usage. Nevertheless, there are several recommendations on how to effectively interact with

an LLM to obtain the desired output. This viewpoint article provides practical guidance along with an overview of potential LLM applications in the everyday work of health care professionals and scientists who have limited experience with LLMs so far. Selected examples will illustrate why LLMs offer added value for these applications and how they can be used effectively. Finally, we will discuss overarching aspects to consider when using LLMs.

General Recommendations for Using LLMs

Before delving into specific use cases within the medical and scientific context, let us first introduce some universally applicable suggestions for optimizing prompts. Drawing from our experience, prevalent recommendations can be broadly categorized into specifying the precise task for the LLM, elucidating relevant contextual factors, and delineating the desired output (Multimedia Appendix 1).

To align the outcome closely with personal expectations, it can be advantageous to construct one or more exemplary outcomes and append them as explicitly classified examples to the prompt, separated by quotation marks or parentheses. Based on our usage experience, the choice of language itself also influences output quality; hence, formulating the prompt in a widely used language, ideally English, is recommended, as it enables the LLM to draw from a wider database. To refine the desired output, unambiguous language should be used, preferably using positive rather than negative formulations. Particularly with complex tasks, the "Chain of Thought Prompting" technique can yield improved results [6]. This technique enhances the reasoning capabilities of LLMs by breaking down multistep problems into a series of intermediate reasoning steps. For generating suitable prompts, it is also feasible to integrate the LLM by providing the program with relevant instructions and subsequently request feedback to iteratively refine the desired prompt. Further guidance on optimizing prompts can be found within the realm of "Prompt Engineering" and on websites of various popular LLM providers.

Use Cases

There are numerous opportunities for the use of LLMs in the professional environment of scientists and health care professionals [eg, 4,5]. The use cases presented in this viewpoint, therefore, only provide a spotlight on this broad field and by no means an all-encompassing overview. Based on personal experience, our reasons for selecting the applications presented in the following sections are that they are easy to implement also for individuals with limited prior LLM experience, quickly yield satisfactory results, and are transferable to similar application scenarios (Table 1).

Table . Overview of presented use cases in this article and the complexity of their corresponding prompts. For each category of prompt complexity, at least 1 illustrative example with the original prompt and output is given in the course of this article.

	Use case	Description	Prompt complexity
1	Translation	Executing customizable translations	Low
2	Text editing	Refining text and extracting information	Low
3	Information compilation	Constructing comprehensive overviews and specialized insights	Medium
4	Idea elaboration	Integrating ideas into cohesive narratives	Medium
5	Training and education	Developing personalized learning schedules and crafting educational materials	Medium
6	Personalized sparring partner	Facilitating intellectual sparring and creative brainstorming	High

Customizing Translations

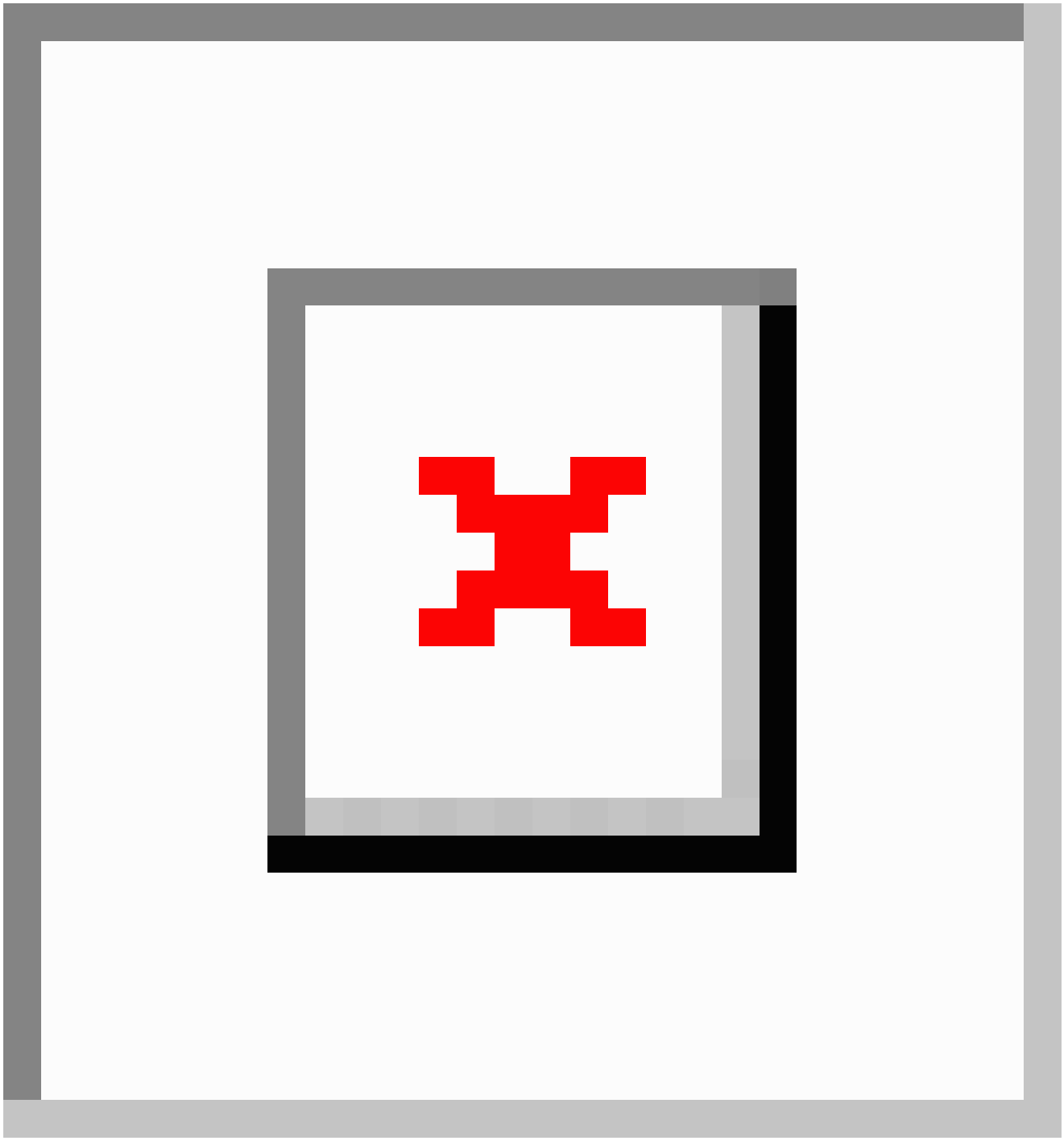
For widely spoken languages, the milestone of ensuring a translation quality comparable to that of translation software has been achieved, for instance, by ChatGPT [7]. In contrast with dedicated translation tools, leveraging an LLM provides diverse customization opportunities for vocabulary, tone, and style, combining general benefits of a proficient translation software with specific strengths inherent in an LLM. A translation prompt should encompass details on the target language, desired degree of alignment with the original text (literal vs freely translated), intended target audience (eg, PhD vs elementary school students), and writing style (formal vs informal). A noteworthy application involves the analysis of scientific information, wherein an LLM not only translates but also succinctly condenses and linguistically adjusts the content to the recipient while retaining its meaning. This facilitates communication and understanding not only across language barriers but also across diverse backgrounds and levels of education, which is particularly crucial in medical and scientific communication with laypeople.

An example for a translation prompt could look like the following: "Translate the following text given in quotation marks into Spanish. Stick closely to the original text. Adopt a formal writing style. Make sure the text is understandable for an adult layperson without medical training."

Editing Text and Extracting Information

One of the core strengths of LLMs lies in handling text, such as summarizing and efficiently organizing written information. For this use case, the text to be summarized or edited can be either typed or uploaded to the LLM as a document, with the latter feature, for instance, available from ChatGPT version 4.0 onwards. The output length, format, and style can be tailored by a precise prompt, as shown in Figure 1. By specifying the audience or assigning a writer persona, the output style and depth of factual detail can be adjusted.

Figure 1. Tailored summaries of an uploaded document depending on the intended target audience. Two different prompts asking for a summary of the same uploaded document (blue). A full-length PDF of one of the authors' open access scientific articles was uploaded as the input [8]. The prompt on the left specifies word count, output format, and audience for a summary to be used in a professional context. The prompt on the right exemplifies how large language models can be used to extract key information and make the content widely accessible for any audience. Answers given by OpenAI's ChatGPT-4 are shown in gray.



Information Compilation: Creating Comprehensive Overviews and Specialized Insights

As the body of available knowledge grows both in depth and breadth, LLMs are a helpful tool to gain specialized insights or general overviews over topics of interest. However, when LLMs are confronted with factual questions, hallucination can occur, meaning that the returned information is syntactically and grammatically accurate but factually incorrect. Special attention should be paid to this matter by meticulously verifying all facts,

references, sources, and links provided by the LLM. Furthermore, while LLMs without internet access often explicitly mention that their "knowledge" is limited to a specific time period, from our practical usage experience, there has rarely been an indication that the prompt's answer might be beyond the capabilities of the program. For the current version (4.0) of OpenAI's ChatGPT, which has internet access, allowing it to execute a web search and mine information, it has been demonstrated that the incidence of hallucinations is significantly lower than that of the previous version (ChatGPT-3.5) [9]. By summarizing and comparing findings from various sources, LLMs are especially useful for exploring a topic in which the user may not be an expert yet. They can also aid in highlighting trends across different sources or extracting divergent positions on a debated topic. However, performing a thorough web search with the help of LLMs often requires specific instructions.

For instance, while the prompt "What are the latest developments in quantum computing?" triggered a web search and quoted 3 sources, the more detailed prompt, "Search the web for research on quantum computing from 2023. What are the latest developments in the field?" yielded results based on 6 sources and a link leading to the web search. Adding "Include your level of confidence, sources and date of answer, and whether your answer is speculative. Make it a markdown table at the end of your answer" returned a long answer based on 5 references and included a statement on research confidence and speculation, as well as a tabular summary of the results and sources. Asking the AI chatbot to "Consider at least 10 sources" returned 8 brief paragraphs based on 12 different sources (Multimedia Appendix 2).

Elaborating Ideas: Integrating Bullet Points Into Flowing Text

Standardized documentation tasks are a frequent requirement in medical-scientific fields. A substantial amount of time is dedicated to formulating individual pieces of information into full-text narratives. By acquainting an LLM with one's preferred or desired writing style, it becomes possible to transform bullet-pointed information into full text. Initially, this involves providing the LLM with examples of target frameworks in terms of structure, content, and style. Once an example is established, additional information can be incorporated and processed with reference to this example. This approach thus represents the opposing process to condensing an extensive text to fewer key points. A specific application in the medical field might be the generation of discharge summaries and medical reports [10]. Initially, structural layout and linguistic style need to be defined. Subsequently, information needed for generating a medical summary report can be listed in bullet points as a prompt and transformed into full text by the LLM.

An exemplary prompt for this use case could appear as follows:

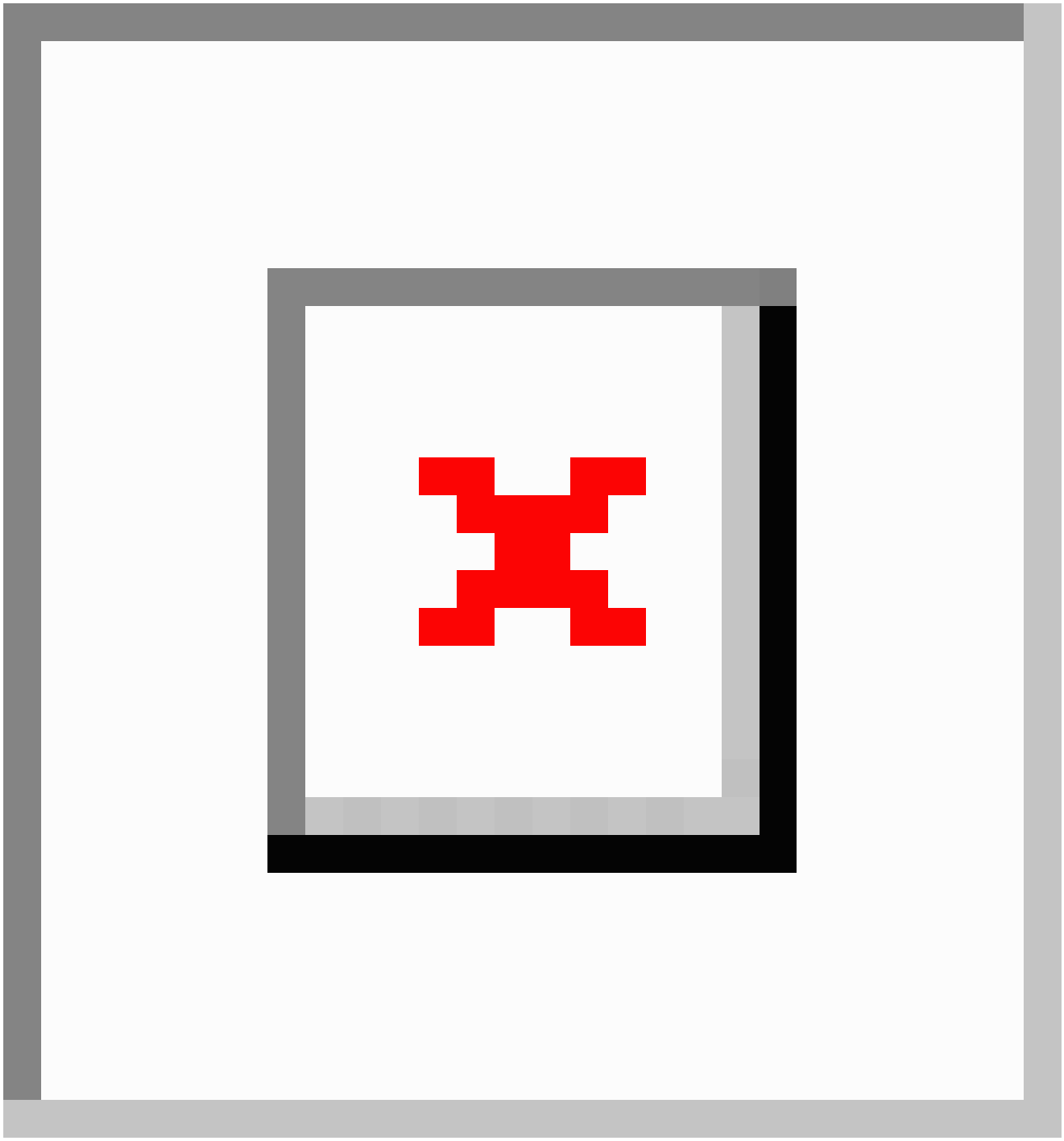
> *Here are two examples (named Example 1 and 2, each in quotation marks) demonstrating the structural and content-based layout of an epicrisis authored by me: ("Example 1," "Example 2"). Now, utilize the following pieces of information to compose a coherent text and align them linguistically and structurally to the two examples mentioned above: (pieces of information in bullet-point style)*

For such applications, special precautions, such as anonymizing all data that could be used to identify a person, must be strictly observed.

Training and Education: Developing Customized Learning Schedules and Teaching Material for Students

Leveraging LLMs in the medical domain also offers applications to training and education. Their adaptive learning capabilities hold promise, for instance, in crafting personalized learning plans for medical professionals. LLMs can also harness vast repositories of medical information and scientific data to curate tailored educational materials. By analyzing a learner's proficiency level, preferred learning modality, areas of interest along with requirements and demands, LLMs can generate personalized curricula and suggest resources, textbooks, or web-based courses. In the context of student education, LLMs present opportunities for the development and structuring of seminars that optimize learning outcomes. By aggregating and analyzing diverse sources of medical and scientific information, LLMs can assist educators in designing seminars that align with specific learning objectives and curriculum requirements [11]. As shown in Figure 2, LLMs can aid in crafting and structuring seminar topics, designing interactive presentations, and suggesting clinical scenarios for case studies, following a single prompt.

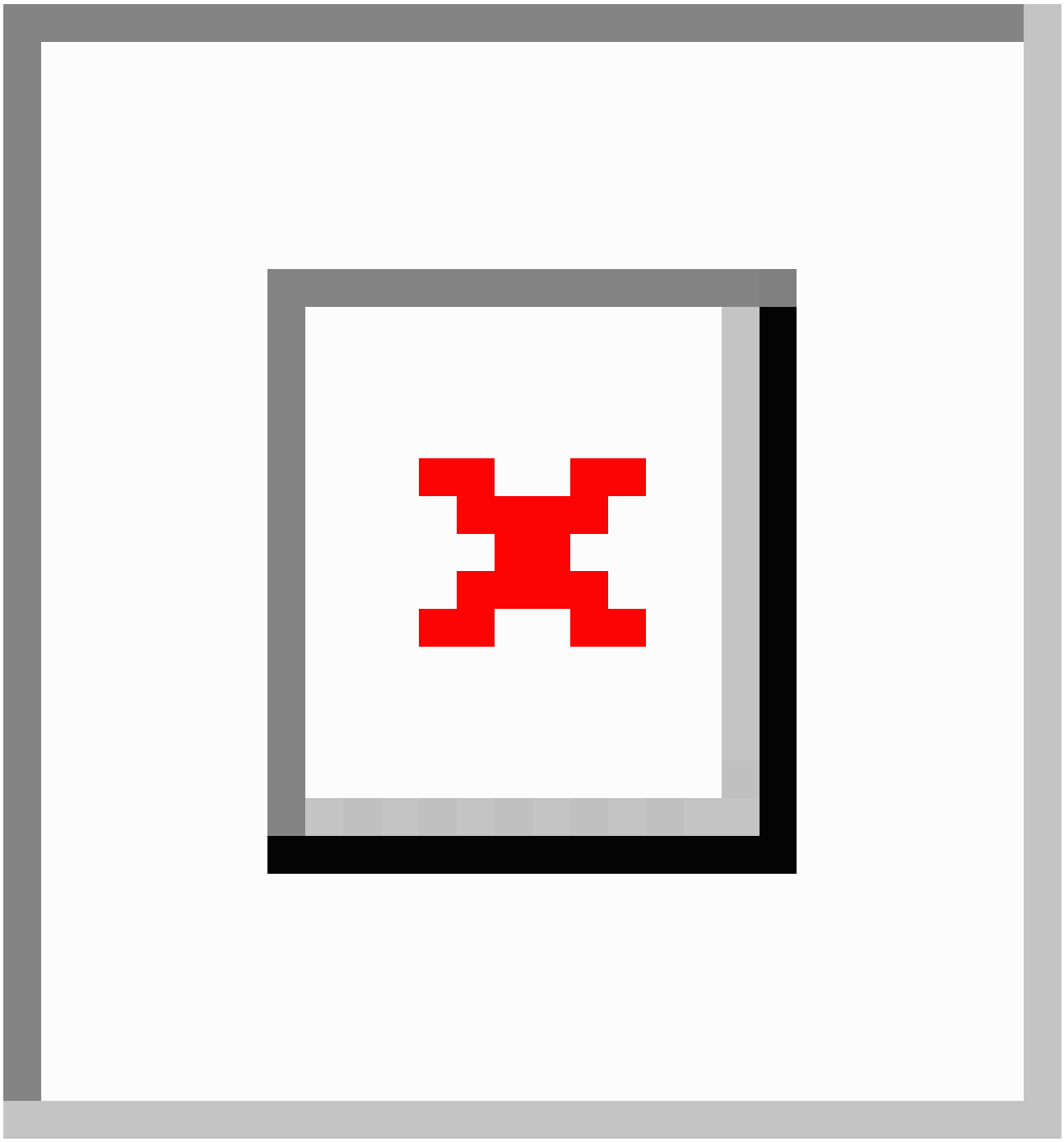
Figure 2. Designing customized teaching material to facilitate medical education. An example prompt asking for a course schedule on a given topic for a specific audience (blue). The response by ChatGPT-4 is shown in gray. The output is presented in tabular format.



Personalized Sparring Partner: Challenging Ideas, Supporting Brainstorming, and Being Queried

This chapter delineates the transformative potential of LLMs as cognitive partners, elucidating their ability to challenge preconceived notions, fuel imaginative thinking, foster structured brainstorming, and provide feedback on the user's theories or hypotheses. Thus, using LLMs as personalized assistants might help researchers overcome writer's block and break out of their creative rut by providing diverse perspectives. The iterative dialogue with an LLM facilitates a reflective process for individuals, compelling them to articulate their ideas cogently. This fosters a deeper understanding of the proposed concepts and enables refinement through robust debate and structured discourse. This feature can be used, for instance, by prompting the LLM to be queried on a specific topic, as shown in Figure 3. It is helpful to first introduce the setting and specify the desired output. To help the LLM process the task, the respective prompt can be subdivided into several shorter inputs.

Figure 3. Using ChatGPT as a personal sparring partner to simulate interactions and discussions. Exemplary prompts asking for a study plan on rheumatology (blue) and then to be tested on the contents of the proposed plan. The task is broken down into 2 prompts, resulting in a conversational interaction. First, ChatGPT-4 replies with a detailed schedule and specific advice for 21 study sessions over the course of 1 week (gray). The first answer is truncated and shown in smaller font due to limited space. Next, ChatGPT-4 is asked to test the user on the proposed topics. Upon receiving an answer, the large language model first corrects the user's answer and then proceeds to pose the next question, thus correctly referring to the previously defined task. EULAR: European Alliance of Associations for Rheumatology.



Technical Hurdles and Challenges

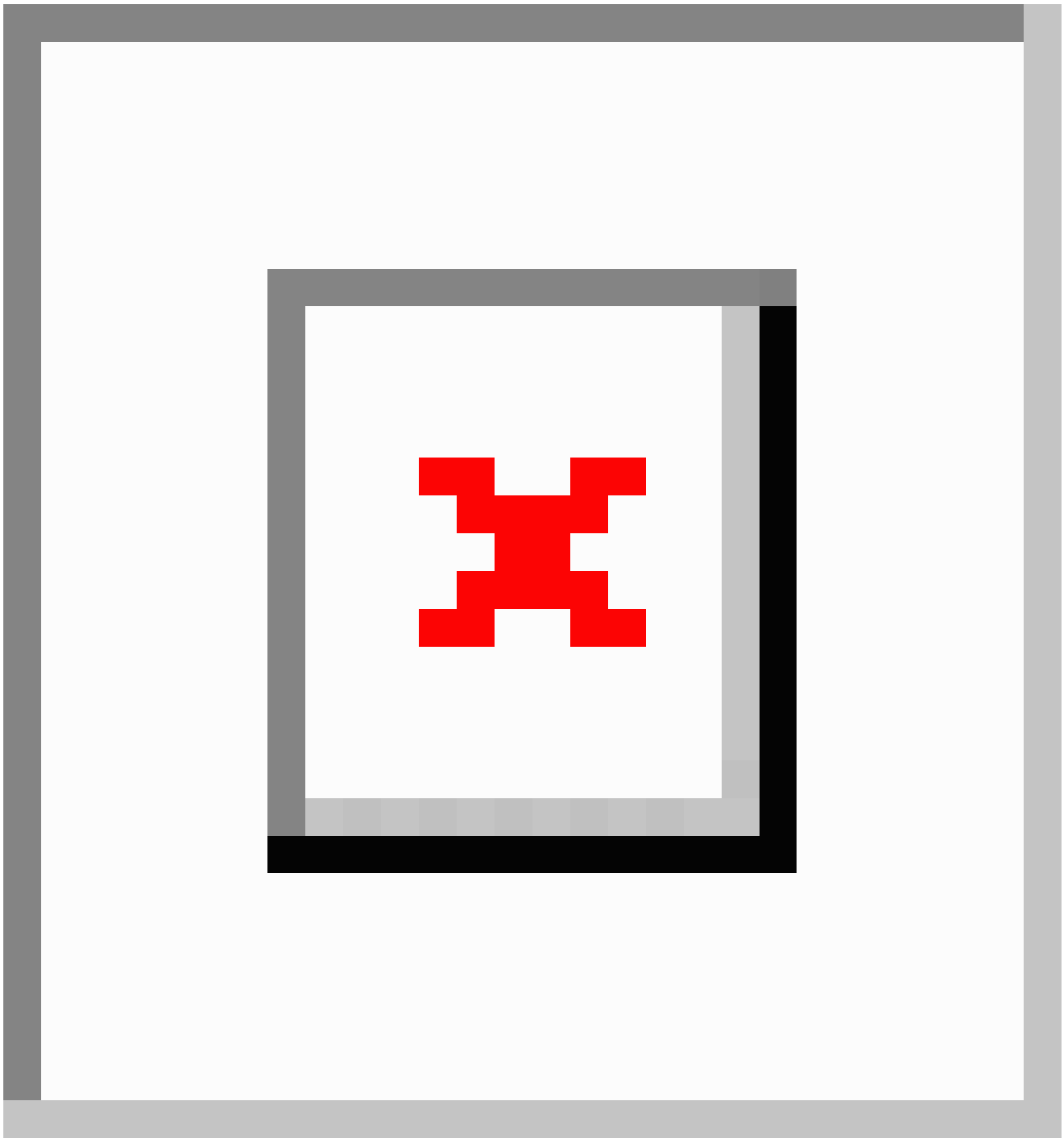
Regarding LLMs, the generation of factually incorrect responses or an output that may be well worded but disconnected from the input is known as a hallucination and may discourage further use of LLMs, particularly among professional users who rely on a high level of output accuracy. Consequently, using LLMs as mere fact-finding tools not only carries substantial risks of misinformation but also fails to fully exploit the actual strength of LLMs in handling extensive volumes of text. Therefore, we do not recommend using LLMs for looking up hard facts or for conducting rigorous literature searches. While completely preventing hallucinations is likely unattainable, various prompting strategies, such as the previously mentioned chain-of-thought prompting, specifying concrete example outputs (few-shot prompting), or uploading reference documents (retrieval-augmented generation) offer empirical evidence for reducing hallucinations [12] and can also be applied by novice users. On a broader level, another technical hurdle to achieving a meaningful impact in health care is the seamless integration

of LLMs into existing systems and workflows [13]. Resource constraints, a lack of clinical validation for these tools, and opaqueness in decision-making are exemplary barriers in this regard, which could be overcome in the future when more practical application experience has been gained and clear regulatory guidelines and frameworks have been established.

Random Versus Deterministic Responses

A limitation of the ChatGPT web interface is the randomness of the responses, meaning that the model may provide a slightly different reply for each new prompt. This is particularly evident with long and complex outputs, which can vary greatly in length, tone, and overall quality, even when the same prompt has been used. The use of an application programming interface (API) offers the same GPT models and functionalities as regular web chats but comes with expanded features for users who wish to purchase a customized model and integrate it into their business software. The API not only enables users to integrate the various available GPT models into their own applications and products, but also allows for more precise customization of the LLM, including a temperature setting. This parameter controls the randomness of the LLM's responses, with a low temperature setting resulting in a more consistent and deterministic output, while a higher temperature generates more varied and creative responses. To demonstrate this, we asked GPT-4 to tell a joke, as shown in Figure 4. At a temperature of 0, the model repeated the same joke on scientists and atoms 3 times, varying its response by only 1 word. With increasing temperature, the program responded less often with its usual atom joke. A low temperature setting may be useful to produce replicable outputs of similar length, style, or structure. A high temperature setting would potentially be useful, when similar prompts are used for repetitive tasks, but a varying output is required, such as when composing invitation emails for recurring professional events without reusing identical content every time.

Figure 4. Setting the output temperature via the application programming interface (API) to generate deterministic results. An identical prompt was entered into the OpenAI API 3 times using the GPT-4 model (chat was cleared between prompts). When the temperature (Temp) was set to 0, the output showed only minimal variation as the model behaved deterministically. At higher values of Temp=1 or Temp=2, the output variety also increased.



Responsible Use of LLMs

The transformative potential of using and integrating LLMs into clinical workflows and health care systems must be approached with caution and guided by rigorous ethical principles and regulatory oversight. One of the foremost challenges lies in patient privacy and data security. As LLMs rely on vast amounts of patient data for training, there are major concerns regarding privacy breaches and unauthorized access [13]. Robust data anonymization, secure storage, and adherence to privacy regulations are essential. Consent for data use must be transparently obtained. From an ethical perspective, LLMs may inadvertently perpetuate biases present in their training data, leading to unequal treatment recommendations [14]. Rigorous bias detection and model fine-tuning are necessary, while legal frameworks should ensure algorithmic fairness and accountability. Lastly, transparency of decision-making processes is crucial, as clinicians need insights into how LLMs arrive at their recommendations [15]. All these limitations should be considered when LLMs are used by health care professionals or scientists. Striking a balance between technological advancement and human safety, multiple governments, companies, and institutions have established guidelines for responsible LLM use. The World Health Organization, for instance, emphasizes the significance of

adhering to ethical principles when using LLMs in health by outlining 6 core principles: protecting autonomy, promoting human well-being, ensuring transparency, fostering responsibility, ensuring inclusiveness, and promoting sustainable AI [16].

Conclusion

LLMs are potent tools, providing immediate and user-friendly access to substantial computational capabilities for a broad range of individuals, while bypassing the necessity for programming skills. Despite certain technical, legal, and ethical challenges, the integration of LLMs into the daily practices of medical professionals and researchers holds potential for significant enhancements in productivity and efficiency. However, as for any tool, the impact of LLMs will be shaped by their intended application. In summary, to address the various challenges arising from the use of LLMs, we believe that a multistage evaluation process is indicated: first, it should be determined which tasks can be appropriately outsourced to such a tool and weighed up against potential implementation hurdles [17,18]. Issues with regulatory, ethical, and legal requirements also need to be examined. Subsequently, the generated material must be thoroughly checked for content accuracy and appropriate references to mitigate the risk of hallucinations or plagiarism [18,19]. Finally, any use of these tools must be transparently disclosed, with accountability always resting with the human user [20]. Overall, the orientation toward a bioethical framework composed of the principles of beneficence, nonmaleficence, autonomy, and justice provides valuable guidance for the deployment of LLMs in a medical context [21]. We advocate for responsible LLM usage, necessitating a comprehensive understanding of their limitations, an awareness of potential data biases, and a consistent adherence to ethical standards.

Acknowledgments

No external funding in the preparation of this paper. When preparing this manuscript, the authors used OpenAI's ChatGPT-4 to translate expressions and suggest vocabulary for improved readability (search prompts shown in Multimedia Appendix 2). After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Authors' Contributions

FR conceptualized the study; designed the methodology; was in charge of the project's administration, software, validation, and visualization; and drafted the manuscript. CL conceptualized the study, designed the methodology, supervised the study, and reviewed and edited the manuscript. MG designed the methodology, was responsible for validation and supervision, and reviewed and edited the manuscript. HDV acquired the funding, designed the methodology, was responsible for validation and supervision, and reviewed and edited the manuscript. NMD conceptualized the study; acquired the funding; designed the methodology; was responsible for the software, validation, visualization; drafted the manuscript; and reviewed and edited the manuscript.

Conflicts of Interest

FR and CL are current employees of Pfizer Pharma GmbH, Berlin, Germany. The opinions expressed in this article are those of the authors and not necessarily those of Pfizer. Pfizer was neither financially involved in the creation nor in the publication of this article. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Different input styles for prompt optimization using a text-based or structured tabular approach. This self-constructed example illustrates 2 possible approaches to formulating prompts: input as continuous text (column 1), akin to directives provided to a human assistant, or, alternatively, structured as bullet points (column 2 and 3). The latter option also allows for saving the prompt structure, facilitating easier modification, and reuse for similar tasks.
[DOCX File, 15 KB - medinform_v12i1e58478_app1.docx]

Multimedia Appendix 2
Web search prompts using ChatGPT-4.
[DOCX File, 30 KB - medinform_v12i1e58478_app2.docx]

References

1. Rate of adoption for major internet and technology services in 2022. Statista. URL: https://www.statista.com/statistics/1360613/adoption-rate-of-major-iot-tech/ [accessed 2024-01-15]
2. Frieder S, Pinchetti L, Chevalier A, et al. Mathematical capabilities of ChatGPT. arXiv. Preprint posted online on Jan 31, 2023. [doi: 10.48550/arXiv.2301.13867]
3. Zakka C, Chaurasia A, Shad R, et al. Almanac: retrieval-augmented language models for clinical medicine. arXiv. Preprint posted online on Mar 1, 2023. [doi: 10.48550/arXiv.2303.01229]

4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
6. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 28, 2022. [doi: 10.48550/arXiv.2201.11903]
7. Jiao W, Wang W, Huang JT, Wang X, Shi S, Tu Z. Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv. Preprint posted online on Jan 20, 2023. [doi: 10.48550/arXiv.2301.08745]
8. Drzeniek NM, Kahwaji N, Schlickeiser S, et al. Immuno-engineered mRNA combined with cell adhesive niche for synergistic modulation of the MSC secretome. Biomaterials 2023 Mar;294:121971. [doi: 10.1016/j.biomaterials.2022.121971] [Medline: 36634491]
9. Research GPT-4. OpenAI. URL: https://openai.com/research/gpt-4 [accessed 2024-02-03]
10. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107-e108. [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
11. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023 Apr;103:102274. [doi: 10.1016/j.lindif.2023.102274]
12. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. arXiv. Preprint posted online on Jan 22, 2024. [doi: 10.48550/arXiv.2401.11817]
13. Denecke K, May R, Rivera Romero O, LLMHealthGroup. Potential of large language models in health care: Delphi study. J Med Internet Res 2024 May 13;26:e52399. [doi: 10.2196/52399] [Medline: 38739445]
14. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023 Jun;5(6):e333-e335. [doi: 10.1016/S2589-7500(23)00083-3] [Medline: 37120418]
15. Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N. Black box warning: large language models and the future of infectious diseases consultation. Clin Infect Dis 2024 Apr 10;78(4):860-866. [doi: 10.1093/cid/ciad633] [Medline: 37971399]
16. WHO calls for safe and ethical AI for health. World Health Organization. URL: https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health [accessed 2024-06-24]
17. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nat New Biol 2023 Feb;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
18. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. N Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
19. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
20. Thorp HH. ChatGPT is fun, but not an author. Science 2023 Jan 27;379(6630):313. [doi: 10.1126/science.adg7879]
21. Ong JCL, Chang SYH, William W, et al. Medical ethics of large language models in medicine. NEJM AI 2024 Jun 27;1(7). [doi: 10.1056/AIra2400038]

Abbreviations

AI: artificial intelligence
API: application programming interface
LLM: large language model

Edited by A Castonguay; submitted 16.03.24; peer-reviewed by L Zhu, Q Jin; revised version received 09.07.24; accepted 11.07.24; published 05.09.24.

Please cite as:
Reis F, Lenz C, Gossen M, Volk HD, Drzeniek NM
Practical Applications of Large Language Models for Health Care Professionals and Scientists
JMIR Med Inform 2024;12:e58478
URL: https://medinform.jmir.org/2024/1/e58478
doi:10.2196/58478

Data Lake, Data Warehouse, Datamart, and Feature Store: Their Contributions to the Complete Data Reuse Pipeline

Antoine Lamer[1,2,3], PhD; Chloé Saint-Dizier[1,2], MSc; Nicolas Paris[3], MSc; Emmanuel Chazard[1], MD, PhD

1

2

3

Corresponding Author:
Antoine Lamer, PhD

Abstract

The growing adoption and use of health information technology has generated a wealth of clinical data in electronic format, offering opportunities for data reuse beyond direct patient care. However, as data are distributed across multiple software, it becomes challenging to cross-reference information between sources due to differences in formats, vocabularies, and technologies and the absence of common identifiers among software. To address these challenges, hospitals have adopted data warehouses to consolidate and standardize these data for research. Additionally, as a complement or alternative, data lakes store both source data and metadata in a detailed and unprocessed format, empowering exploration, manipulation, and adaptation of the data to meet specific analytical needs. Subsequently, datamarts are used to further refine data into usable information tailored to specific research questions. However, for efficient analysis, a feature store is essential to pivot and denormalize the data, simplifying queries. In conclusion, while data warehouses are crucial, data lakes, datamarts, and feature stores play essential and complementary roles in facilitating data reuse for research and analysis in health care.

(JMIR Med Inform 2024;12:e54590) doi:10.2196/54590

KEYWORDS

data reuse; data lake; data warehouse; feature extraction; datamart; feature store

Introduction

Over the last few decades, the widespread adoption and use of health information systems (HISs) have transitioned a substantial amount of clinical data from manual to electronic format [1]. HISs collect and deliver data for care, administrative, or billing purposes. In addition to these initial uses, HISs also offer opportunities for data reuse, defined as “non-direct care use of personal health information” [2], such as research, quality of care, activity management, or public health [3]. Hospitals have gradually adopted data warehouses to facilitate data reuse [4,5]. Even if the data warehouse is a popular concept, data reuse is not limited to feeding and querying a data warehouse. In this viewpoint, our objective is to outline the different components of the data reuse pipeline and how they complement and interconnect with each other. This definition is derived from our personal experiences and insights gained through collaboration with colleagues at various institutions [5-8]. Additionally, we draw on the collective experiences shared by professionals in the field, contributing to a comprehensive understanding of data reuse practices in diverse health care settings. The pipeline is illustrated in Figure 1 and detailed below. Table 1 compares characteristics of each component. Last, Multimedia Appendix 1 provides examples of data, structures, and architectures for each component of the data reuse pipeline.

Figure 1. Components of the complete pipeline for data reuse. EHR: electronic health records; ETL: extract-transform-load.

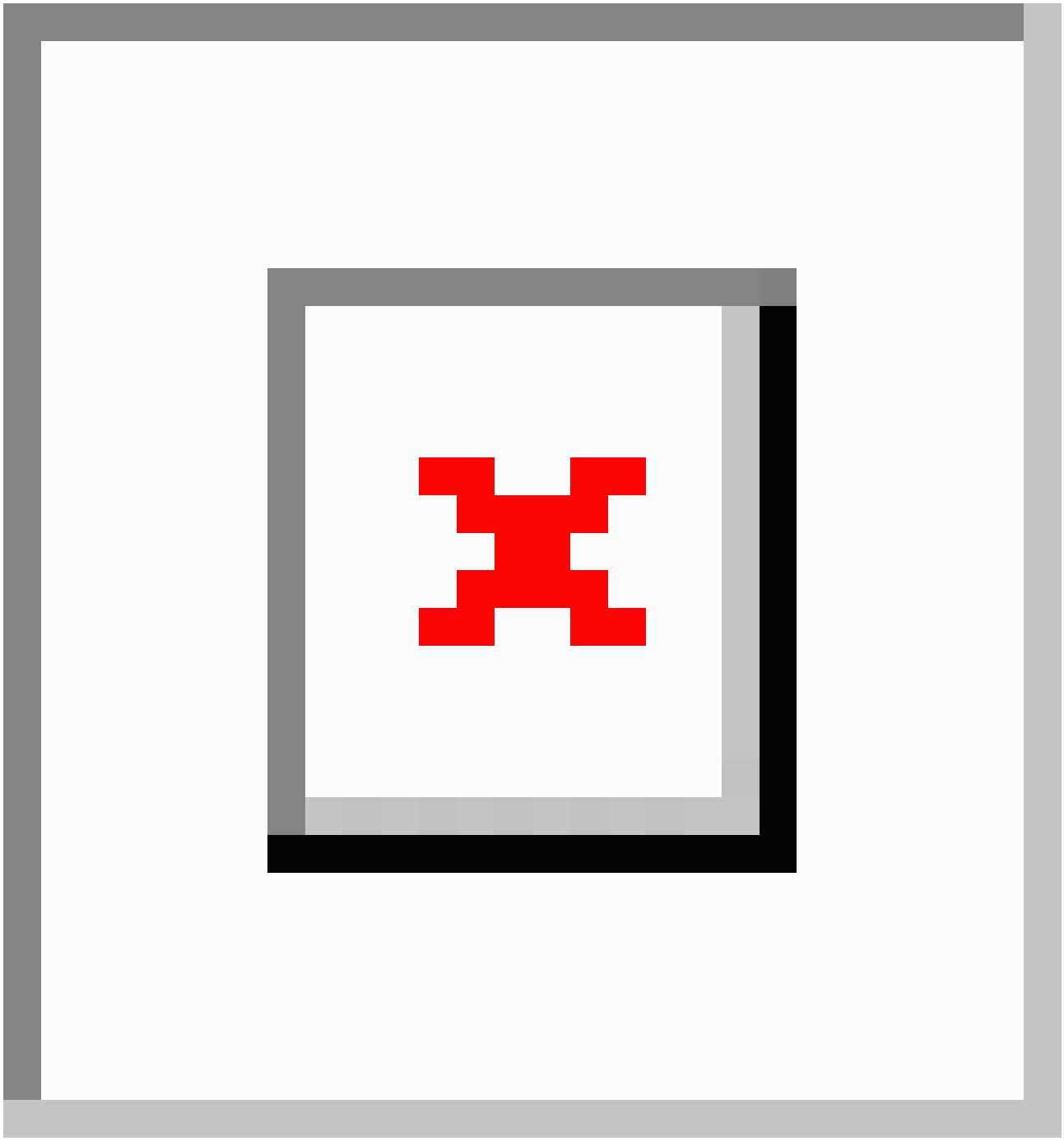


Table . Characteristics of each component of the data reuse pipeline.

Characteristics	Software	Data lake	Data warehouse	Datamarts	Feature store
Content	Data and metadata	Data and metadata	Data	Features	Features and metadata about feature
Architecture	Distributed	Centralized	Centralized	Centralized	Centralized
Detail level	Fine-grained	Fine-grained	Fine-grained	Aggregated	Aggregated
Data	Raw	Raw	Cleaned	Cleaned	Cleaned
Nomenclature	Heterogeneous	Heterogeneous	Standardized	Standardized	Standardized
Data model	Normalized	Normalized	Normalized	Normalized	Denormalized
Data structure	Row-oriented	Row-oriented	Row-oriented	Row-oriented	Column-oriented
Purpose	Transactional software purpose	Ad hoc exploratory queries	All purposes	Prespecified purpose	Prespecified purpose

Ethical Considerations

This study does not include human participants research (no human participants experimentation or intervention was conducted) and so does not require institutional review board approval.

Health Information System

The raw data stored in the HIS are distributed across multiple software, making it impossible to cross-reference information between sources due to variations in data formats, ranging from tabular to hierarchical structures and free text [9]. Different technologies and distinct identification schemes for patients, admissions, or any other records compound the complexity. Additionally, direct write access to the software databases is typically unavailable, as software editors rarely grant such privileges to prevent any potential disruption to routine software operation. In transactional software databases, data consist of meticulously organized and highly accurate records presented in rows. These records are collected with great precision to fulfill the specific functions of the software. Alongside the data, a wealth of metadata is also present, including information regarding data collection (eg, information on the individuals inputting data, record timestamps, and biomedical equipment identifiers), as well as software configurations. Notably, a significant portion of these metadata may not be directly relevant to our research purposes, as they primarily support the routine functioning of the software.

The Data Lake

An optional first component of a comprehensive data reuse pipeline is the data lake [10-14]. A data lake is a centralized, flexible, and scalable data storage system that ingests and stores raw data from multiple heterogeneous sources in its original format [12,15]. Data are stored in a fine-grained, row-oriented, and raw format, in a secure and cost-effective environment. These raw data still encompass diverse formats, from structured data to unstructured text documents, images, songs, videos, and sensor data, ensuring that a wide spectrum of information is readily available for various data analytics endeavors [12].

The technologies implemented for the data lake can include the usual relational databases, such as PostgreSQL or Oracle, but also NoSQL databases and big data technologies, such as the Hadoop Distributed File System or Apache Hudi for the storage and Apache Spark, Hadoop MapReduce, or Apache Kudu for the data processing.

Unlike structured data typically integrated into data warehouses, the data lake refrains from immediate structuring or transformation, allowing for a more agile and adaptable approach. This flexibility enables exploration, manipulation, and, if necessary, transformation of the data to fulfill specific research or analytical requirements. By delaying the application of predefined data models, the data lake cultivates an environment where information can be uncovered without predetermined hypotheses. This includes insights that may not have been evident during the initial phases of data collection and storage. The system further facilitates on-the-fly query processing and data analysis [12,15].

In a data pipeline without a data lake, it is essential to finalize the extract-transform-load (ETL) process before leveraging the data. This introduces a time delay, as it necessitates identifying relevant data in the HIS, updating the data warehouse data model for their accommodation, and subsequently designing and implementing the ETL.

In addition, when interpreting the results, if it becomes apparent that relevant data are missing for the analysis, it requires updating both the ETL process and the data model to incorporate the missing data. This iterative cycle of identifying, modifying, and reimplementing can lead to prolonged timelines and may hinder the agility of the data analysis process. Therefore, a data lake approach proves advantageous in providing a more flexible and dynamic environment for data exploration and analysis, potentially avoiding some of these challenges encountered in a traditional pipeline.

The Data Warehouse

The data warehouse stands as the most prevalent component of the pipeline and acts as a centralized repository of integrated data from 1 or more disparate sources [5,8,16-19]. It stores historical and current fine-grained data in a format optimized for further use. This involves a single storage technology, a

consistent naming convention for tables and fields, and coherent identifiers across data sources. This is a departure from the data lake where all these elements varied between sources.

The data warehouse is supplied through an ETL process [9,18]. The primary objective of this process is to select and extract relevant data from the HIS or other external resources [19]. During this initial phase, the majority of metadata linked to software operations (such as usage logs or interface settings), monitors, and individuals inputting data are usually excluded. Indeed, these types of metadata do not relate to patient care information and would introduce an unnecessary volume of data. Subsequently, the ETL process enhances the raw data by identifying and correcting any abnormal or erroneous information. Following this refinement, data are integrated into a unified data model independent of the source software [9,19]. Notably, there is a strong focus on harmonizing identifiers from diverse data sources to ensure data integrity and streamline queries involving information from multiple origins. The ETL process is also responsible for regularly updating the data warehouse with new information recorded in the original data sources.

The data warehouse, as a relational database, is typically implemented using systems like PostgreSQL, Oracle, SQL Server, Apache Impala, or Netezza. However, for a data warehouse, exploring NoSQL technologies such as MongoDB, Cassandra, or Couchbase can also be interesting, offering advantages in handling unstructured or semistructured data, and providing scalability for large-scale data storage and retrieval [20]. The ETL process can be developed using 2 types of technologies. The first one, with programming languages such as R (R Core Team), Python (Python Software Foundation), or Java (Oracle Corporation), can be used, coupled with a scheduler like Apache Airflow (Apache Software Foundation), to organize the execution of scripts and retrieval of logs and error messages. The second kind of application is graphical user interface software, such as Talend (Talend) or Pentaho (Hitachi Vantara). They do not require programming capacities, because graphical components, corresponding to data management operations, are organized through a drag-and-drop interface.

To foster collaboration among institutions and facilitate the sharing of tools, methods, and results, several initiatives have emerged to offer common data models (CDM). As a result, table and field names are standardized following a common nomenclature, and local vocabularies and terminologies are mapped to a shared vocabulary. Among these CDMs, the Observational Medical Outcomes Partnership CDM was developed by the Observational Health Data Sciences and Informatics community, which brings together multiple countries and thousands of users [21] and led to methodological and practical advancements [22,23].

As a result, the data warehouse functions as a unified, centralized, and normalized repository, for both fined-grained historical data and metadata, and continues to present information in a row-oriented format. The modeling approach presented by Inmon [24] and described as a “subject-oriented, nonvolatile, integrated, time-variant collection of data” implies that data are stored persistently without any assumptions as to their future use, thus remaining open-ended in their usage.

The Datamarts

While the data warehouse serves as a unique standardized repository, primarily dedicated to data storage, querying these data can be time-consuming due to the volume and distribution of data in the relational model. Furthermore, raw data integrated into the data warehouse may not be readily aligned with specific research or analytical questions, as these data lack the necessary aggregated features. For instance, the data warehouse retains all biological measurements (eg, potassium and sodium), while what will be stored in the datamart are the features related to the biology values, such as the occurrence of hypokalemia, hyperkalemia, hyponatremia, or hypernatremia. Thus, the datamart acts as a dedicated resource for transforming the data into usable and meaningful information [19,25,26]. This transformation process involves feature extraction, achieved through the application of algorithms and domain-specific rules [6,7]. The outcome is data that are tailored to address specific research questions or analytical needs. For instance, within a clinical setting, the datamart can convert raw mean arterial pressure values into a format suitable for detecting perioperative hypotension [5].

Moreover, datamarts can be organized in the form of online analytical processing (OLAP) cubes, offering a multidimensional view of the data [27]. This cubical structure allows for in-depth analysis, enabling users to efficiently explore and navigate across various dimensions such as time, geography, or specific categories, gaining profound and contextualized insights. These datamarts are often modeled in either a snowflake or star schema, optimizing their structure for the creation of OLAP cubes. The star schema, with its central fact table surrounded by dimension tables, or the snowflake schema, which further normalizes dimension tables to minimize data redundancy, both serve to facilitate the creation of these OLAP cubes. Such schemas play a pivotal role in enhancing the efficiency of multidimensional data analysis within the OLAP environment, providing a structured framework for faster and more comprehensive insights.

In the context of health care, an example of an OLAP cube could encompass dimensions such as patient (eg, age and gender); time (eg, admission and discharge dates); medical conditions (eg, primary and secondary diagnoses and medical procedures); hospital unit (eg, information on services, departments, and bed types); health care provider (eg, physicians); and outcome (eg, length of stay, treatment outcomes, and medical costs). The cube would include various facts, such as the number of patients, average length of stay, and average treatment costs. This multidimensional structure allows health care professionals to conduct in-depth analyses, explore trends over time, compare costs across different hospital units, and assess the impact of medical interventions on patient outcomes [19,28].

Datamarts, owing to the structured nature of their data, are typically stored on relational databases (eg, PostgreSQL, Oracle, and SQL Server) [25]. In the case of OLAP cubes, this may

include Apache Kylin or other proprietary OLAP tools built on relational databases [28,29].

In contrast to Inmon's [24] approach, the Kimball [9] bottom-up approach places datamarts at the core, with their design driven primarily by business requirements. However, by directly developing datamarts, the Kimball approach may overlook some crucial data that were not initially identified as relevant during the business requirements phase.

As a result, the datamart stands as a centralized component for cleaned and aggregated features for dedicated purposes, still stored in row-oriented structure.

The Feature Store

The feature store addresses the limitations of the traditional row-oriented, relational database structure typically used in datamarts. This architecture, which relies on multiple tables, may not fully meet various analytical requirements. For instance, effective statistical analysis often necessitates a single, flat file with column-oriented variables, mandating the transformation of data from a row-based to a column-based format within the feature store. This process streamlines data access, simplifying complex queries into straightforward selections from a single table. Consequently, the feature store emerges as a centralized repository housing well-documented, curated, and access-controlled features. In addition to features extracted from datamarts, which are often calculated by algorithms derived from business rules, the feature store can also receive features generated by machine learning algorithms [30].

The design of the feature store aims to provide data scientists with direct access to these features, eliminating the need for additional data cleaning, aggregation, or pivoting [31]. This specialized role enhances efficiency and promotes the use of high-quality, analysis-ready data, significantly contributing to the effectiveness of data-driven research in the health care organization. Notably, the feature store not only stores the features themselves but also their associated metadata, documenting how they were calculated and used [31]. It ensures the preservation of all feature versions, guaranteeing the reproducibility of analyses.

When derived from business rules, features are stored in relational databases (eg, PostgreSQL, Oracle, and SQL Server) or in a NoSQL data store such as MongoDB to also store metadata. When features originate from machine learning models, they are stored and shared from big data platforms such as Databricks or Hopsworks [30,32].

As the final component of the data reuse pipeline, the feature store plays a pivotal role in various analytical applications within the health care organization. It significantly contributes to the creation of insightful dashboards and automated reports, delivering real-time and historical information. In research, its most crucial contribution lies in generating denormalized flat tables, similar to questionnaire data tailored for statistical analyses.

Conclusions

In this opinion paper, we propose standardized nomenclature and definitions for the components of a data reuse pipeline. Table 2 summarizes the advantages and limitations of each component in this pipeline.

While the data warehouse serves as a necessary initial stage, the integration of datamarts and a feature store enhances its effectiveness. Datamarts compute pertinent information from raw data, while the feature store organizes it into columns, streamlining data set construction. Additionally, the data lake emerges as a valuable resource for storing raw data in a single location, allowing for exploitation without having to wait for the entire pipeline to be developed.

Notably, in a data pipeline without a data lake, the requirement to complete the ETL process before analysis introduces delays. This involves identifying relevant data in the HIS, adapting the data warehouse data model, and implementing the ETL. Additionally, discovering missing data during result interpretation prompts iterative updates to both the ETL process and the data model, potentially prolonging timelines and hindering data analysis agility.

It is important to emphasize that the specific components and their characteristics described here are not rigidly fixed and can vary based on the unique organizational needs and configurations. For instance, the inclusion of a data lake and feature store is often discretionary, influenced by factors such as the scale and intricacy of source data, the quantity of features, the scope of research projects, the team's size, and the imperative for study reproducibility over time.

Table . Advantages and disadvantages of the components of the data reuse pipeline.

Component	Advantages	Disadvantages
Data lake	• All data sources on the same server • Independence from source software • On-the-fly query processing and data analysis without the need for the complete development of an extract-transform-load (ETL) process	• Inconsistencies in data formats and structures • Lack of standard schema can make querying complex • Analyses reproducibility
Data warehouse	• Querying data from both administrative and biology systems is facilitated by the unified data model (ie, data from both systems are linked, and the model conventions are consistent) • Relevant data are retained at the finest level of detail (eg, dates, diagnoses, and all biology values), enabling the answering of numerous questions without necessarily identifying them beforehand	• ETL process must be implemented to standardize the data • Multidimensional data model with several statistical units • Fine-grained data is not directly usable and adapted for statistical analysis or decision-making
Datamarts	• Features are ready to be used directly	• Features are still organized with a row-format (ie, 1 feature per row) in several datamarts
Feature store	• Using features directly, without the need for data management tasks such as joining datamarts or pivoting to reorganize features into columns	• Having developed the entire pipeline beforehand

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of data, structures, and architectures of components of the data reuse pipeline.
[DOCX File, 68 KB - medinform_v12i1e54590_app1.docx]

References

1. Adler-Milstein J, DesRoches CM, Kralovec P, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist. Health Aff 2015 Dec;34(12):2174-2180. [doi: 10.1377/hlthaff.2015.0992]
2. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007 Jan;14(1):1-9. [doi: 10.1197/jamia.M2273] [Medline: 17077452]
3. Safran C. Reuse of clinical data. Yearb Med Inform 2014 Aug 15;9(1):52-54. [doi: 10.15265/IY-2014-0013] [Medline: 25123722]
4. Wisniewski MF, Kieszkowski P, Zagorski BM, Trick WE, Sommers M, Weinstein RA. Development of a clinical data warehouse for hospital infection control. J Am Med Inform Assoc 2003 Sep;10(5):454-462. [doi: 10.1197/jamia.M1299] [Medline: 12807807]
5. Lamer A, Moussa MD, Marcilly R, Logier R, Vallet B, Tavernier B. Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project. J Clin Monit Comput 2023 Apr;37(2):461-472. [doi: 10.1007/s10877-022-00898-y] [Medline: 35933465]
6. Chazard E, Ficheur G, Caron A, et al. Secondary use of healthcare structured data: the challenge of domain-knowledge based extraction of features. Stud Health Technol Inform 2018;255:15-19. [Medline: 30306898]
7. Lamer A, Fruchart M, Paris N, et al. Standardized description of the feature extraction process to transform raw data into meaningful information for enhancing data reuse: consensus study. JMIR Med Inform 2022 Oct 17;10(10):e38936. [doi: 10.2196/38936] [Medline: 36251369]
8. Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: a case study in France. PLOS Digit Health 2023 Jul;2(7):e0000298. [doi: 10.1371/journal.pdig.0000298] [Medline: 37410797]

9. Kimball R. The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses: John Wiley & Sons; 1998.
10. Wieder P, Nolte H. Toward data lakes as central building blocks for data management and analysis. Front Big Data 2022 Aug;5:945720. [doi: 10.3389/fdata.2022.945720] [Medline: 36072823]
11. Madera C, Laurent A. The next information architecture evolution: the data lake wave. Presented at: MEDES'16: The 8th International Conference on ManagEment of Digital EcoSystems; Nov 1 to 4, 2016; Biarritz, France p. 174-180. [doi: 10.1145/3012071.3012077]
12. Sarramia D, Claude A, Ogereau F, Mezhoud J, Mailhot G. CEBA: a data lake for data sharing and environmental monitoring. Sensors (Basel) 2022 Apr 2;22(7):2733. [doi: 10.3390/s22072733] [Medline: 35408347]
13. Che H, Duan Y. On the logical design of a prototypical data lake system for biological resources. Front Bioeng Biotechnol 2020 Sep;8:553904. [doi: 10.3389/fbioe.2020.553904] [Medline: 33117777]
14. HV S, Rao BD, J MK, Rao BD. Design an efficient data driven decision support system to predict flooding by analysing heterogeneous and multiple data sources using data lake. MethodsX 2023 Dec;11:102262. [doi: 10.1016/j.mex.2023.102262] [Medline: 37448950]
15. Hai R, Koutras C, Quix C, Jarke M. Data lakes: a survey of functions and systems. IEEE Trans Knowl Data Eng 2023 Dec 1;35(12):12571-12590. [doi: 10.1109/TKDE.2023.3270101]
16. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University hospital clinical data warehouse: a 8-years follow-up experience. Int J Med Inform 2017 Jun;102:21-28. [doi: 10.1016/j.ijmedinf.2017.02.006] [Medline: 28495345]
17. Chen W, Xie F, Mccarthy DP, et al. Research data warehouse: using electronic health records to conduct population-based observational studies. JAMIA Open 2023 Jul;6(2):ad039. [doi: 10.1093/jamiaopen/ooad039] [Medline: 37359950]
18. Fleuren LM, Dam TA, Tonutti M, et al. The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients. Crit Care 2021 Aug 23;25(1):304. [doi: 10.1186/s13054-021-03733-z] [Medline: 34425864]
19. Agapito G, Zucco C, Cannataro M. COVID-WAREHOUSE: a data warehouse of Italian COVID-19, pollution, and climate data. Int J Environ Res Public Health 2020 Aug 3;17(15):5596. [doi: 10.3390/ijerph17155596] [Medline: 32756428]
20. McClay W. A Magnetoencephalographic/encephalographic (MEG/EEG) brain-computer interface driver for interactive iOS mobile videogame applications utilizing the Hadoop Ecosystem, MongoDB, and Cassandra NoSQL databases. Diseases 2018 Sep 28;6(4):89. [doi: 10.3390/diseases6040089] [Medline: 30274210]
21. Blacketer C. The Book of OHDSI: Observational Health Data Sciences and Informatics; 2021. URL: https://ohdsi.github.io/TheBookOfOhdsi/ [accessed 2024-11-09]
22. Schuemie MJ, Gini R, Coloma PM, et al. Replication of the OMOP experiment in Europe: evaluating methods for risk identification in electronic health record databases. Drug Saf 2013 Oct;36(S1):S159-S169. [doi: 10.1007/s40264-013-0109-8] [Medline: 24166232]
23. Lane JCE, Weaver J, Kostka K, et al. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. Lancet Rheumatol 2020 Nov;2(11):e698-e711. [doi: 10.1016/S2665-9913(20)30276-9] [Medline: 32864627]
24. Inmon WH. Building the Data Warehouse: Wiley; 1992.
25. Hinchcliff M, Just E, Podlusky S, Varga J, Chang RW, Kibbe WA. Text data extraction for a prospective, research-focused data mart: implementation and validation. BMC Med Inform Decis Mak 2012 Sep 13;12:106. [doi: 10.1186/1472-6947-12-106] [Medline: 22970696]
26. Kim HS, Kim H, Jeong YJ, et al. Development of clinical data mart of HMG-CoA reductase inhibitor for varied clinical research. Endocrinol Metab (Seoul) 2017 Mar;32(1):90-98. [doi: 10.3803/EnM.2017.32.1.90] [Medline: 28256114]
27. Hristovski D, Rogac M, Markota M. Using data warehousing and OLAP in public health care. Proc AMIA Symp 2000:369-373. [Medline: 11079907]
28. Vik S, Seidel J, Smith C, Marshall DA. Breaking the 80:20 rule in health research using large administrative data sets. Health Informatics J 2023;29(2):146045822311805. [doi: 10.1177/14604582231180581] [Medline: 37269132]
29. Ranawade SV, Navale S, Dhamal A, Deshpande K, Ghuge C. Online analytical processing on Hadoop using Apache Kylin. Int J Appl Inf Syst 2017 May 5;12(2):1-5. [doi: 10.5120/ijais2017451682]
30. Armgarth A, Pantzare S, Arven P, et al. A digital nervous system aiming toward personalized IoT healthcare. Sci Rep 2021 Apr 8;11(1):7757. [doi: 10.1038/s41598-021-87177-z] [Medline: 33833303]
31. Sen S, Woodhouse MR, Portwood JL, Andorf CM. Maize Feature Store: a centralized resource to manage and analyze curated maize multi-omics features for machine learning applications. Database (Oxford) 2023 Nov 6;2023:baad078. [doi: 10.1093/database/baad078] [Medline: 37935586]
32. Rajendran S, Obeid JS, Binol H, et al. Cloud-based federated learning implementation across medical centers. JCO Clin Cancer Inform 2021 Jan;5:1-11. [doi: 10.1200/CCI.20.00060] [Medline: 33411624]

Abbreviations

CDM: common data model
ETL: extract-transform-load
HIS: health information system
OLAP: online analytical processing

Edited by C Lovis; submitted 27.11.23; peer-reviewed by R Anand, I Reinecke, N Ahmadi, O Steichen; revised version received 11.03.24; accepted 05.04.24; published 17.07.24.

Please cite as:
Lamer A, Saint-Dizier C, Paris N, Chazard E
Data Lake, Data Warehouse, Datamart, and Feature Store: Their Contributions to the Complete Data Reuse Pipeline
JMIR Med Inform 2024;12:e54590
URL: https://medinform.jmir.org/2024/1/e54590
doi:10.2196/54590

Impact of Large Language Models on Medical Education and Teaching Adaptations

Li Zhui[1], PhD; Nina Yhap[2], PhD; Liu Liping[3], PhD; Wang Zhengjie[4], PhD; Xiong Zhonghao[5], MM; Yuan Xiaoshu[6], BSc; Cui Hong[6], MM; Liu Xuexiu[7], MM; Ren Wei[1], PhD

1
2
3
4
5
6
7

Corresponding Author:
Ren Wei, PhD

Abstract

This viewpoint article explores the transformative role of large language models (LLMs) in the field of medical education, highlighting their potential to enhance teaching quality, promote personalized learning paths, strengthen clinical skills training, optimize teaching assessment processes, boost the efficiency of medical research, and support continuing medical education. However, the use of LLMs entails certain challenges, such as questions regarding the accuracy of information, the risk of overreliance on technology, a lack of emotional recognition capabilities, and concerns related to ethics, privacy, and data security. This article emphasizes that to maximize the potential of LLMs and overcome these challenges, educators must exhibit leadership in medical education, adjust their teaching strategies flexibly, cultivate students' critical thinking, and emphasize the importance of practical experience, thus ensuring that students can use LLMs correctly and effectively. By adopting such a comprehensive and balanced approach, educators can train health care professionals who are proficient in the use of advanced technologies and who exhibit solid professional ethics and practical skills, thus laying a strong foundation for these professionals to overcome future challenges in the health care sector.

(JMIR Med Inform 2024;12:e55933) doi:10.2196/55933

KEYWORDS

large language models; medical education; opportunities; challenges; critical thinking; educator

Introduction

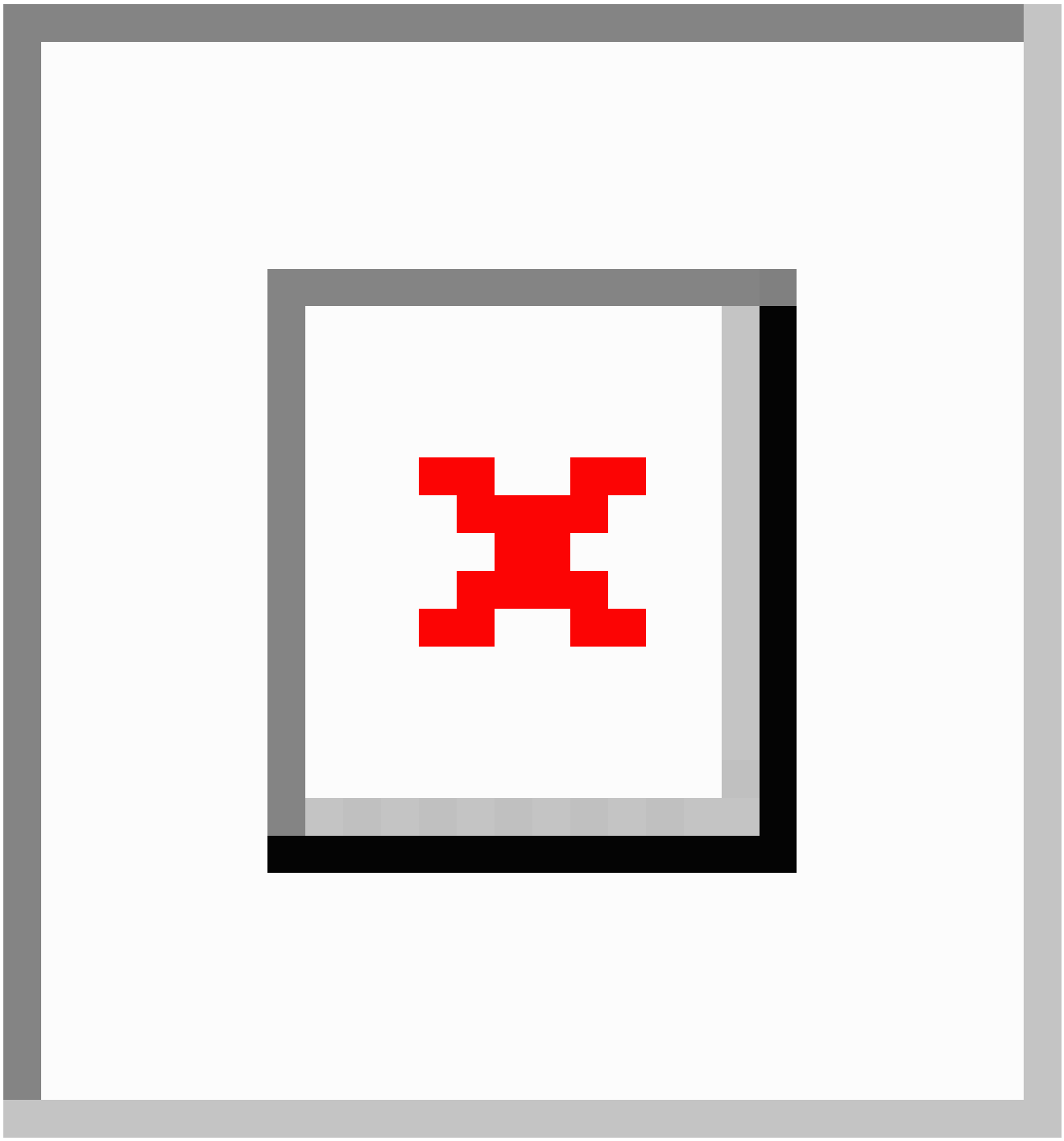
Technological advancements have significantly shaped medical education, leading to transformative shifts in approaches to teaching and learning. The recent rapid evolution of artificial intelligence (AI) technology has entailed unprecedented changes and challenges in this field, particularly due to the emergence of large language models (LLMs). LLMs, which are extensively trained on vast text data sets, are advanced AI systems best exemplified by OpenAI's GPT series, Google's Gemini, and Twitter's Grok [1]. By using complex neural network architectures, especially those based on transformer design, LLMs can identify and replicate subtle linguistic nuances. Their capabilities extend beyond the generation of text; they can also produce corresponding images or even videos based on text inputs, thereby addressing a wide variety of linguistic tasks. Notably, the use of LLMs in medical education is increasingly becoming a focal point of global research and discussion [2,3]. In the context of medical education, LLMs can answer questions regarding medical concepts, generate simulated cases, assist researchers, and support clinical decision-making [4,5]. These applications not only optimize the efficiency of teaching but also offer students a more personalized, digital, and adaptable learning experience [6]. However, as this tool is integrated into medical education, challenges pertaining to the accuracy of information, privacy, potential overreliance on such technology, and academic integrity have emerged [7].

This viewpoint article, which is based on current international research, thoroughly explores the impacts associated with the integration of LLMs into medical education. The article also analyzes in depth the adaptive changes that teaching methods should make in response to these opportunities and challenges. The primary goal of this article is to provide a nuanced perspective on the use of technological innovation in medical education and to serve as a reference for future teaching practices and strategies.

Opportunities in Medical Education

The various uses of LLMs are gradually changing traditional medical education. The opportunities presented by LLMs in the context of medical education are illustrated in Figure 1.

Figure 1. Opportunities associated with the use of large language models in medical education. CME: continuing medical education.



Enhancing the Efficiency and Quality of Teaching

The integration of LLMs into medical education heralds a transformative advance in the efficiency and quality of instruction. A study that focused on the ChatGPT 3.5 platform in the context of a problem-based learning teaching model revealed that the application of ChatGPT significantly enhanced collaboration and participation among members of learning groups. This approach also increased students' motivation, encouraging them to pose questions more actively [8]. A randomized controlled study that compared ChatGPT-assisted learning with traditional literature-based learning reported that students' examination scores significantly improved as a result of ChatGPT-assisted learning [9]. By leveraging these advanced tools, educators can swiftly craft comprehensive syllabi, develop engaging lecture materials, and create detailed textbooks. This approach is particularly effective within the frameworks of case-based learning and problem-based learning, which emphasize contextual and practical learning [10,11]. LLMs offer 2 advantages in educational settings. First, they can

generate visual content, such as illustrative diagrams to depict disease progression or 3D models of anatomical structures, and they can even source existing educational imagery on the internet to meet specific learning objectives. This capability not only enriches teaching materials but also caters to diverse learning styles, thus facilitating students' retention and understanding of complex concepts. Second, LLMs excel in the tasks of analyzing and interpreting visual data. They can help educators and students decipher complex medical images such as computed tomography (CT) or magnetic resonance imaging (MRI) images. Through such digital and visually driven learning experiences, LLMs significantly enhance students' comprehension and retention of information regarding challenging medical subjects, such as pathology or clinical diagnosis. The use of LLMs in medical education substantially improves the efficiency of preparing instructional material as well as the quality of the material itself. By enhancing students' collaboration skills and learning motivation, this approach effectively improves learning outcomes.

Facilitating Personalized Learning

Personalized learning involves tailoring instructional content and strategies to the specific needs, abilities, interests, and learning styles of students. Compared to traditional educational models, LLMs offer a more flexible, digital, and immediate approach to learning [12]. Traditional education often relies on standardized curriculum structures and educator guidance, while LLMs can adapt to each student's unique learning needs and style, thereby providing students with customized guidance and feedback. Such assistance not only enhances student engagement but also helps students grasp knowledge at their own pace [13,14].

Xu et al [15], who conducted interviews with 6 higher education professors and 3experts in information and communication technology, revealed that ChatGPT significantly improves students' cognitive skills, such as their ability to process information, solve problems, and understand concepts. It also enhances students' noncognitive skills, such as motivation and self-efficacy, and facilitates their development of metacognitive skills, thus helping students plan and adjust their learning. These improvements are crucial for enhancing personalized and interdisciplinary education in higher education [15]. This support is particularly crucial for students in dynamic fields such as medicine, in which context triggering intrinsic motivation for self-learning and offering a timely understanding of the latest research findings in depth are essential.

Moreover, ChatGPT facilitates a discussion-based learning approach that is closely connected to real-world scenarios. This type of interaction not only facilitates knowledge acquisition but also serves as a platform through which students can engage LLMs as debate partners, thereby fostering the development of critical thinking skills by enabling students to present and evaluate different viewpoints [16]. Accordingly, students can use LLMs to create a tailored learning journey based on their own interests, knowledge gaps, and learning pace, thereby crafting a personalized educational path that aligns with their individual needs. Additionally, a study that evaluated students' reasoning abilities by referencing published case reports and simulated clinical cases revealed that ChatGPT significantly enhances the clinical reasoning abilities of medical professionals [17]. This kind of reasoning ability is developed through personalized feedback based on the individual learning progress and needs of each person. Through these digital and customized teaching solutions, LLMs effectively meet the personalized learning needs of students. This approach can enhance students' engagement and promote the development of crucial skills such as critical thinking and logical reasoning, thereby helping students overcome the challenges entailed by an ever-evolving knowledge landscape effectively.

Reinforcing Clinical Skills

Traditional clinical skills training relies heavily on real clinical cases or standardized patients (SPs), which can allow students to interact with patients and address authentic clinical challenges directly. However, this approach has significant drawbacks, including high costs, temporal and spatial constraints, the inclusion of only a limited variety of medical conditions, issues related to patient privacy, and the inability to provide timely feedback regarding learning outcomes.

The integration of LLMs into clinical skills training can effectively address these limitations, as demonstrated by a pioneering study conducted in China. The use of ChatGPT to simulate SPs not only addresses the scarcity of SPs but also obviates the need for supplementary training, thereby conserving substantial human and material resources. Crucially, LLMs such as ChatGPT can emulate a diverse array of SPs, thereby delivering intelligent, articulate, and vivid responses that are tailored to specific scenarios [18]. This innovation enables students to engage in repetitive practice in a low-risk environment that can obviate concerns regarding patient welfare and privacy, thereby enhancing students' diagnostic reasoning and communication skills [19].

The use of LLMs to simulate clinical patients offers other significant advantages. First, given that the accessibility of LLMs—for instance, ChatGPT 3.5 and Microsoft Gemini—do not require registration and can be seamlessly integrated into browsers—clinical simulation exercises can become ubiquitous, thus transcending the traditional confines of clinical settings. This flexibility significantly enhances the adaptability of learning [12]. Second, the ability of LLMs to replicate a spectrum of diseases, ranging from common ailments to rare conditions, broadens the scope of educational content and experiences, thereby addressing students' lack of exposure to rare clinical cases. According to Scherr et al [5], LLM simulations facilitate the early development of independent diagnostic and therapeutic reasoning among medical students. Furthermore, LLMs can adaptively refine their responses based on student interactions, thereby offering a more authentic reflection of real-world patient-provider communication than is possible using conventional, static simulation models. Moreover, LLMs provide an inexhaustible supply of free simulation opportunities, thus democratizing access to medical education, particularly for students from economically

disadvantaged backgrounds or for institutions with limited resources [5].

Although LLMs currently do not extend to procedural skill training in medical education, such as training in surgical tasks, including debridement and suturing, the prospective integration of LLMs with virtual reality and augmented reality technologies could represent a transformative advance in this regard. This combination could facilitate the establishment of a multidimensional, digital, and highly authentic simulation learning environment, thereby leading to the advent of a new era of innovative teaching methodologies and content in the context of medical education. Such advancements can address students' diverse educational needs, enrich the learning landscape, and imbue the domain with fresh momentum.

Improving Medical Teaching Assessments and Examinations

Traditional medical education assessments, which include standardized examinations, interviews, and clinical skills evaluations, require considerable preparation, detailed grading, and extensive time to compile students' scores. In contrast, LLMs such as ChatGPT represent a dynamic shift in this context by streamlining the creation of diverse, integrated assessment questions; they are, thus, closely aligned with instructors' needs. A study conducted in Hong Kong, Singapore, and the United Kingdom highlighted the ability of the ChatGPT to generate 50 multiple-choice questions rapidly, with the questions thus generated exhibiting a quality comparable to that of questions developed by university professors while offering a significant reduction in preparation time [20]. This efficiency extends to case-based questions, in which context the outputs of ChatGPT have been reported to be equally robust [21].

The evolving research on LLMs has focused on their precision regarding addressing a spectrum of medical questions, thus highlighting their ability to enhance educational assessments. The success of the ChatGPT in passing rigorous examinations regarding medical theory, such as the United States Medical Licensing Examination and the Objective Structured Clinical Examination in Singapore, highlights the proficiency of this LLM in domain-specific knowledge and its potential applicability across a variety of medical assessments [22,23]. In addition to generating questions, LLMs excel in the tasks of delivering immediate, detailed feedback, identifying student errors, and highlighting areas for improvement [24]. Furthermore, LLMs provide deep insights into examination results, thereby offering educators real-time insights into the learning challenges faced by students. This analytical capability fosters a more nuanced understanding of student performance, thus enabling educators to tailor their teaching strategies effectively [25].

The integration of LLMs into medical education heralds a more efficient, personalized assessment process, thereby decreasing educators' workloads while providing students with a clearer comprehension of their academic progress. This innovative approach enriches the educational experience, shifting from a conventional focus on scores to a more holistic perspective on the efficacy of learning and teaching.

Enhancing the Efficiency of Medical Research

Information retrieval in the context of medical research has long been associated with challenges such as information overload, time-consuming classification processes, and a lack of personalized suggestions. Due to their intelligent information retrieval capabilities, LLMs can offer more precise and personalized literature search suggestions based on researchers' queries. This technology can help researchers swiftly locate relevant information and effectively mitigate issues pertaining to information overload [26]. Moreover, LLMs can summarize lengthy papers and offer reasonable advice regarding research methods, experimental design, and statistical analysis, thereby accelerating the research process and enhancing the efficiency of research [27]. Additionally, LLMs can facilitate data analysis and visualization, enabling researchers to interpret and summarize research results efficiently and streamlining the research workflow [2].

LLMs also provide cross-lingual assistance to researchers from non-English-speaking countries, thereby improving the quality of research [28-30]. Numerous studies have indicated that ChatGPT can help researchers refine papers and that the ability of LLMs to write abstracts is comparable to that of humans, thus highlighting the potential of this technology regarding providing support for academic writing [31-33]. Indeed, scholar King [34] successfully published a paper produced with ChatGPT that required no further language editing. Additionally, ChatGPT can help researchers draft cover letters, thus enabling them to convey the significance and relevance of their research effectively when submitting papers to journals [35,36]. The combined impact of these capabilities can significantly boost the ability of researchers from non-English-speaking backgrounds to engage in international academic communication and support the results of their research.

Support for Continuing Medical Education and Professional Development

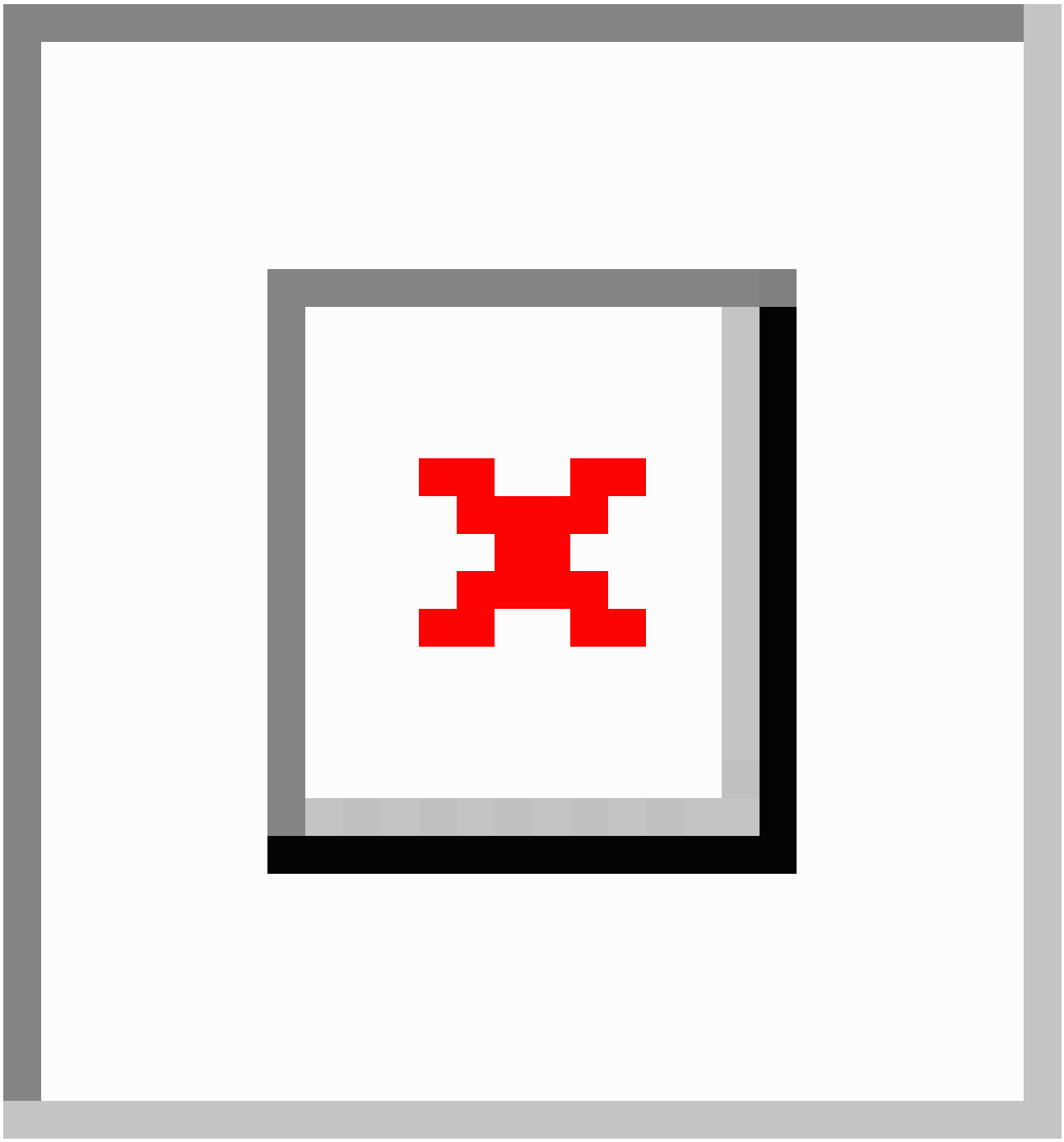
Rapid advancements in the field of medicine require health care professionals to engage in lifelong learning to enhance their professional competence continuously. In this context, LLMs represent an efficient, convenient, and flexible approach to continuing medical education. For instance, LLMs can systematically and specifically help medical professionals remain up-to-date regarding the most recent developments in medical research, treatment guidelines, and clinical practices [24]. Several comparative studies focusing on various LLMs and search engines have consistently revealed that LLMs can generate evidence-based medical recommendations tailored to specific topics [37,38]. The experience of clinical doctors is often limited due to their frequent encounters with common and prevalent cases, which often causes them to lack sufficient expertise regarding rare diseases or emerging epidemics, as was observed at the onset of the COVID-19 pandemic. In this regard,

LLMs can address the knowledge gaps of health care professionals by collecting global medical information or providing clinical experiences with respect to the diagnosis and treatment of specific diseases. By using LLMs, health care professionals can pursue continuing education in a manner that requires minimal economic and time investments, thereby enhancing their clinical decision-making capabilities. This approach ensures that professionals in the field of health care can effortlessly remain up-to-date in terms of their knowledge and skills despite their demanding clinical schedules [39].

Challenges Associated With Medical Education

Despite the numerous opportunities in the field of medical education associated with the use of LLMs, this approach also entails a series of challenges and limitations. The challenges posed by LLMs in the context of medical education are summarized in Figure 2.

Figure 2. Challenges associated with the use of large language models in medical education. AI: artificial intelligence.



Concerns Pertaining to the Accuracy of Information

In medical education, the accuracy of knowledge is of paramount importance, as even minor errors can threaten patient safety. For students who have only limited background knowledge, identifying inaccuracies or misleading information can be particularly challenging. Alkaissi and McFarlane [40] reported that ChatGPT can generate fabricated data, cite nonexistent literature, and provide incorrect citations, a phenomenon known as AI hallucination. Additionally, LLMs may sometimes provide varying responses to identical queries [40,41]. Therefore, the current LLMs cannot yet be fully relied upon in education or research fields. A detailed evaluation of 180 questions answered by the ChatGPT, which was conducted by 33 doctors from 17 medical specialties, revealed that while some responses were highly accurate, significant errors occurred in response to more complex questions [42].Other studies in the medical field have also confirmed this, showing that the ambiguities or errors in LLMs' responses to complex questions stem from their limited depth of understanding and reasoning capabilities [43,44]. Furthermore, the levels of accuracy exhibited by different language models varied. A study of the accuracy of multiple LLMs in the context of generating answers to dental questions revealed that ChatGPT-4 was significantly more accurate than ChatGPT 3.5, Google Bard, and Microsoft Bing Chat. However, all these LLMs provided irrelevant information, vague answers, or even entirely incorrect responses [38]. Similar conclusions have been reported in several other studies [45-47]. This limitation is primarily due to the training data used for LLMs, which are largely drawn from the internet and often lack rigorous screening and quality control. Consequently, the inclusion of errors, biases, or outdated information directly impacts the accuracy of the outputs of LLMs. Additionally, while current LLMs such as Google Gemini can directly incorporate references from open-access journals into their responses, thus significantly enhancing confidence in the accuracy of the information, access to the subscription sections of certain authoritative databases, such as PubMed, remains limited. This constraint continues to impede the accuracy and comprehensiveness of their responses [48,49].

Concerns Pertaining to Overreliance and Academic Integrity

Students may depend excessively on LLMs to perform learning tasks, such as writing papers or preparing for examinations. While the content generated by LLMs is original, students who submit text generated by LLMs directly as their own work might engage in plagiarism and academic dishonesty [34,50]. In particular, the integration of LLMs with search engines, which allow direct access to original texts and multimedia, significantly increases the risk of plagiarism. One study indicated that when university professors are presented with abstracts generated by LLMs and those written by humans, they struggle to determine whether these abstracts were authored by a machine or by a human [31]. In a study conducted in 2023, reviewers were able to identify only 63% of fabricated abstracts generated by ChatGPT [27]. This finding suggests that the quality of the text produced by LLMs has equaled or exceeded the quality of the text produced by professionals, thereby providing opportunities for students to engage in academic misconduct by using LLMs to generate summaries, data, or even complete drafts of papers. Furthermore, the ability of LLMs to achieve high scores on exams such as the United States Medical Licensing Exam, as mentioned previously, inevitably raises the possibility of students using them to cheat on exams. Overreliance on LLMs not only entails the risk of academic misconduct but also may impair students' development of independent research and critical thinking abilities, causing them to have only a shallow grasp of topics and affecting their ability to accurately diagnose complex medical conditions. According to a survey of 370 undergraduate medical students in India, more than 53% of these medical students expressed concerns about the possibility that the widespread application of LLMs might lead to such overreliance, which could hinder their development of clinical reasoning skills [51]. Mechanically extracting answers from LLMs inevitably deprives students of invaluable face-to-face engagement with teachers and peers, thus impeding the development of critical thinking skills, which must be refined through dynamic and diverse intellectual exchanges. Therefore, while LLMs offer significant reference value in medical education, it is crucial to establish appropriate guidelines and balanced strategies to ensure that students can develop independent thinking skills and clinical judgment.

Bias and Transparency

In medical education, when LLMs are employed, it is essential to scrutinize potential biases in the training data and algorithms underlying these LLMs as well as to address the opacity of the models' decision-making processes [52]. If an LLM overemphasizes a specific treatment while neglecting others, it might bias students' perceptions, potentially instilling in them a narrow view of patient care. In January 2024, Zack et al [53] used the Azure OpenAI application interface to evaluate ChatGPT-4 and discovered that the model inadequately represented demographic diversity in terms of medical conditions. It consistently generated clinical vignettes that reinforced stereotypes associated with specific demographic groups. Furthermore, the differential diagnoses produced by the ChatGPT-4 regarding standardized clinical vignettes tended to include stereotypical biases pertaining to certain races, ethnicities, and genders. The assessments and plans generated by the model were significantly linked to demographic attributes; the model thus often recommended more costly procedures and exhibited disparities in terms of patients' perceptions. This widespread unfair response primarily arises from training data and algorithmic biases. Even if the training data are unbiased, bias in the algorithmic design of LLMs can still lead to biased outputs [54]. The harm of LLMs that output-biased information lies not only in the decline of teaching quality but also in their potential to distort medical students' ethical principles, leading to unfair treatment of clinical patients and even endangering their health and lives.

When faced with responses from LLMs, users often struggle to understand how these responses are generated or the underlying

logic, a phenomenon known as the "black box" effect, indicating a lack of transparency in the model [55]. In medical education, the lack of transparency in LLMs may prevent teachers and students from accurately assessing the accuracy of information, posing risks to patient safety in clinical practice [56]. Additionally, failure to comprehend the logic behind LLMs' responses may lead to rigid thinking among students, overlooking the importance of logical thinking in medical curriculum learning and thereby weakening the cultivation of critical thinking.

Risks Pertaining to Privacy and Data Security

AI has long been used to assist in facilitating the diagnosis and treatment of diseases by examining patients' medical histories and examination images [36]. In medical education, when sensitive patient information such as names, ages, diagnoses, and other details is input for case-based teaching simulations or even as raw data for training LLMs, as well as CT or MR images containing patient information for image interpretation or diagnostic purposes, LLMs may inadvertently disclose patient privacy during the reasoning process or through their "memory" effects [52,57]. Even more concerning, current research has shown that even anonymized information could lead to privacy breaches through cross-referencing with other available web-based data, resulting in the reidentification of personal information [57]. Studies have demonstrated that with only 15 demographic characteristics, 99.98% of personal information can be reidentified [58]. As a cloud-based service, LLMs face various cybersecurity risks regarding data storage and processing, including hacking attacks or the infiltration of malicious software. Any such security vulnerabilities could lead to the leakage of sensitive data, thereby leading to significant risks and legal liabilities [59]. Companies that develop chatbots, such as OpenAI, may employ users' personal information for various purposes, including service analysis, improvement, and research. Such companies also reserve the right to share users' personal information with third parties without prior notification or explicit consent from users [60]. All these factors introduce risks pertaining to privacy and data security to the process of medical education.

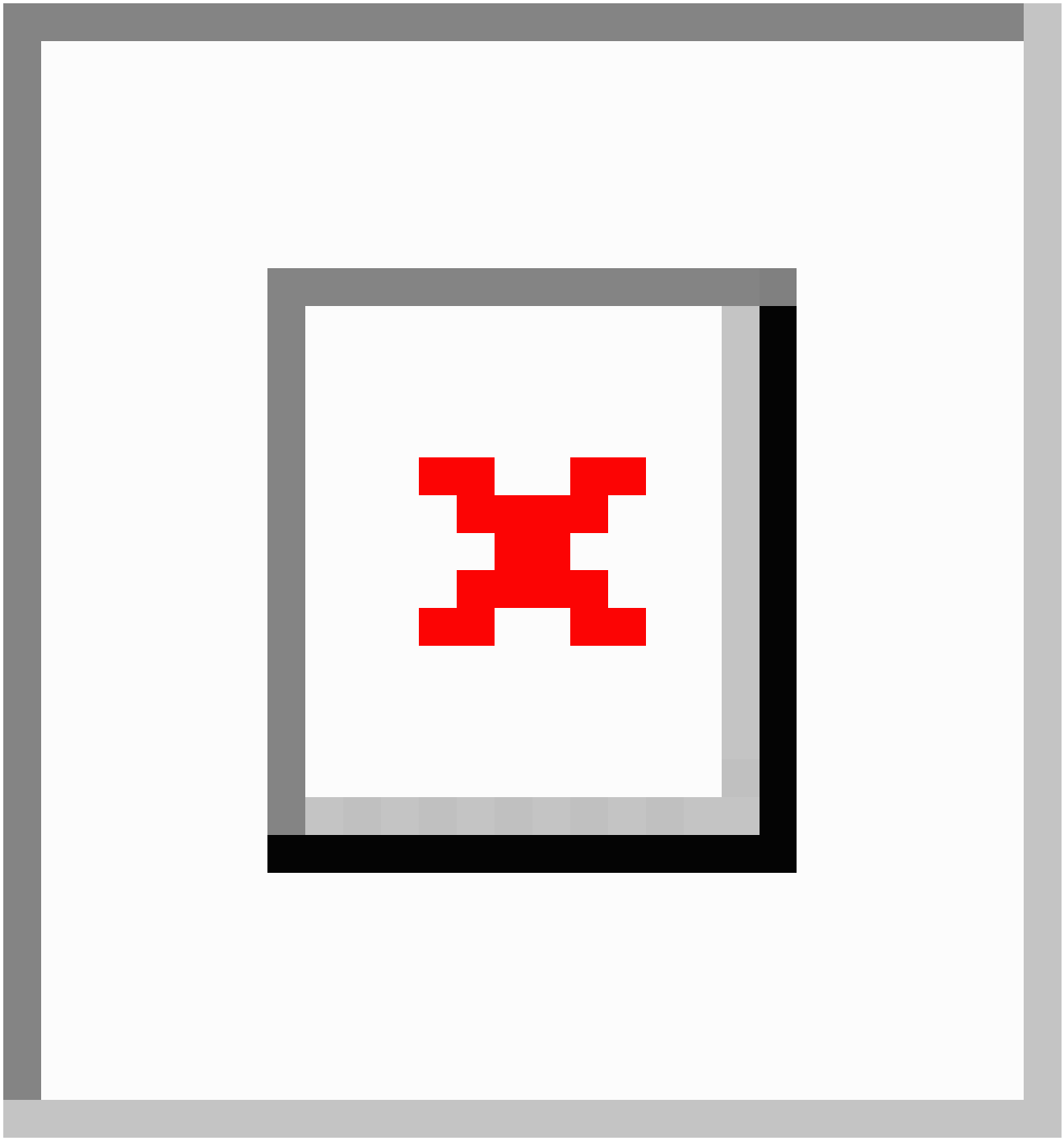
Lack of Emotional Intelligence

Cultivating empathy is a cornerstone of medical education. Although language models such as the ChatGPT can be trained to mimic empathetic language and used to facilitate patient communication [61], artificial empathy cannot replicate genuine, subtle empathy on the part of health care providers, which is crucial for patient care and can easily be identified by patients [62]. Prof Marc Succi from Harvard University believes that excessive reliance on artificial emotional communication in medical settings can exacerbate societal loneliness [63]. The subtlety, profundity, and diversity of empathy that is expressed in the context of human-to-human interactions are currently irreplaceable in LLMs. When LLMs simulate patients, all the emotional responses they exhibit are expressed as textual descriptions, thus preventing the student from interpreting or responding to nonverbal cues such as facial expressions, tone of voice, and gestures. These cues are even more crucial for genuine human empathy than verbal expressions [64]. The empathy conveyed through the programmed responses of LLMs poses a challenge for medical students' attempts to acquire the basic experience necessary to engage in interpersonal relationships by interacting with these models. The establishment and development of empathy also involve addressing the inherent complexity and uncertainty associated with real interactions, in which context the diverse backgrounds, beliefs, and values of real patients lead to a wide range of emotional expressions in interpersonal communication. LLMs struggle to simulate such personalized and diverse emotional responses. Furthermore, cultivating empathy requires ongoing practice over time. LLMs are static models that cannot engage in long-term interaction and provide personalized feedback; accordingly, they are unable to facilitate the deep practices necessary for students to develop empathy.

Changes in Teaching Methods

LLMs are currently revolutionizing numerous aspects of medical education. As we exploit the power of these advanced technologies to promote innovation in medical education, we must also address the various challenges they entail. In the context of teaching activities, educators must continue to play a leadership role in the classroom, emphasize the value of traditional classroom settings, enhance students' critical thinking abilities, and highlight the crucial role of practical experience. In this manner, we can not only benefit from the advantages of LLMs but also ensure the comprehensiveness and effectiveness of the educational process. The specific content of the adaptive changes made by educators in response to the challenges entailed by LLMs is summarized in Figure 3.

Figure 3. Main content of adaptive changes in educators' teaching methods. LLM: large language model.



To ensure the reliability of the information imparted during medical education, educators must exercise critical scrutiny, especially in light of students' tendency to trust their teachers to a considerable extent. Compared to one-on-one human-machine interactions, one-to-many scenarios during teaching may exacerbate the adverse effects of misinformation transmission. When using chatbots such as Microsoft's Copilot or Google's Gemini, which can access and cite real-time data directly from the internet, it is crucial to recognize that medical students may lack the necessary skills to evaluate the authenticity of such information. Therefore, employing multiple methods to verify information and teaching students how to discern the truthfulness of information has become an urgent and important educational need. Educators should corroborate AI-generated content by providing references to trusted sources such as PubMed, the Cochrane Library, or UpToDate. Additionally, seeking insights from subject matter experts can provide valuable perspectives on complex or nuanced medical topics. Cross-verification, which involves comparing the responses provided by various language models, serves as another means of such validation, which can lead to further investigation of discrepancies. This not only involves effectively transmitting precise knowledge but also helps in developing students' critical thinking skills.

Simply prohibiting the use of language models to address challenges pertaining to academic integrity is not a practical solution. Educational institutions must establish explicit policies regarding academic misconduct in the context of these

technologies. Educators should clearly define how assignments should be completed and inform students of the consequences of academic dishonesty. Just as the response to King's [34] inquiry to ChatGPT about the strategies college professors should use to prevent students from cheating with ChatGPT suggested, designing assignments that incorporate a variety of assessment methods is an effective approach. This approach includes changing traditional homework methods and supervising students' independent completion of assignments related to specific topics in class. The introduction of classroom presentations, group discussions, practical activities, and video productions, which are tasks that LLMs cannot complete independently, can highlight students' knowledge and skills. Additionally, when students inevitably use LLMs to complete assignments, they should be required to clarify the role of LLMs in their work and provide examples of such human-computer interactions for reference. Educators can also use plagiarism detection software to identify copied or unoriginal content in submitted papers [34]. These changes in teaching methods can not only foster critical thinking but also prevent overreliance on AI as well as academic dishonesty.

The development of critical thinking skills is also essential in light of the biases and unfairness entailed by LLMs. Although educators may not have the capacity to alter the biases inherent in training data or LLM algorithms, it is imperative for us to scrutinize the outputs of LLMs critically and to cultivate a similar critical perspective on the part of our students. When case-based instruction is employed, inclusivity should be ensured in terms of diverse populations and medical scenarios to avoid reinforcing stereotypical associations with specific demographic groups or medical conditions. LLM-generated clinical vignettes should be used as pedagogical tools by prompting students to engage in structured activities such as group discussions, role-playing, or debates. These activities should focus on dissecting the reasoning process, diagnostic conclusions, and therapeutic recommendations presented in the cases in question, thus emphasizing the process of identifying and critiquing potential biases and logical inconsistencies. Additionally, the broader ethical, societal, and cultural ramifications of the deployment of LLM-generated insights in medical decision-making should be explored. Students should be encouraged to engage in reflective practices after interacting with LLMs, thus prompting them to assess their decision-making trajectory and outcomes introspectively and to recognize and address their inherent biases and cognitive patterns.

Educators should prioritize hands-on experience and advocate for experiential learning anchored in authentic clinical settings. Regarding data privacy and security, it is crucial to adhere strictly to legal standards regarding patient confidentiality. Educators should aim to minimize the use of actual patient data in LLMs and prefer to use synthetic or hypothetical scenarios to avert any potential privacy breaches. In cases in which it is necessary to input patient-specific medical history, examination images, and other sensitive information into LLMs, it is imperative for students to obtain informed consent from the patient, obtain the necessary approval from an ethics committee, and implement appropriate data anonymization measures. Additionally, it is vital to provide medical students with comprehensive education and training regarding patient privacy protection, including lawful and ethical procedures for the collection, use, and storage of patient information. By actively engaging in clinical practices that emphasize the importance of patient privacy and preventing data leakages, medical students can cultivate a profound understanding of the legal and ethical responsibilities associated with patient privacy protection.

An emphasis on clinical practice is essential for the development of fundamental clinical skills and the promotion of profound empathic connections, both of which are rooted in medical practice. To cultivate empathy, practical teaching methods such as role-playing and simulated patient interactions can be used to enhance students' effective doctor-patient communication skills. These skills include listening techniques, nonverbal communication, and appropriate expressions of empathy and care. Engaging with real patients allows medical students to observe, learn from, and emulate the empathetic behaviors of experienced instructors and to apply these insights to patient care. Accumulating experiences in empathy by engaging in diverse patient interactions and a process of immersive reflection can enable students to understand and empathize with patients deeply, thereby equipping them to address the complexities of patient care effectively. Moreover, the incorporation of disciplines such as psychology and other disciplines in the humanities into medical curricula can provide students with a theoretical foundation for understanding patients' emotional and psychological states, thereby enriching the practical application of their skills in clinical settings.

We must acknowledge the positive role played by LLMs in education, and educators must ensure that this technology can complement traditional teaching methods rather than replace them. Teachers should respond to the opportunities and challenges associated with the use of LLMs by adapting their teaching methods and content, promoting critical thinking among students, and emphasizing the importance of practical experience to ensure that teachers continue to play a leading role in medical education. In the future, establishing unified ethical principles or guidelines for the application of LLMs in medical education could better guide educators and students on how to effectively and safely use LLMs. Such a balanced approach can ensure that future medical professionals are better equipped by combining technological advancements with the enduring values of medical education.

Conclusion

We explored the diverse opportunities and challenges associated with the use of LLMs in medical education. While LLMs offer new avenues for exploration and innovation in the context of medical education, educators must recognize that these technologies serve as supplementary tools to traditional teaching methodologies. The expertise and foresight of educators remain paramount regarding safeguarding the quality of teaching. Educators must possess the ability to exploit these emerging technologies while continuing to play a guiding role in student learning, nurture critical thinking skills, and consistently emphasize the significance of clinical practice in medical education, thus ensuring that students use LLMs judiciously

and efficaciously. By obtaining a comprehensive understanding of the strengths and limitations of LLMs and making corresponding adjustments to their teaching approaches, educators can steer medical education toward a more innovative, intelligent, and practice-oriented future.

Acknowledgments

We appreciate Ms Ding Ding for her assistance in the literature search.

Authors' Contributions

LZ and RW contributed to the conceptual design, data collection, drafting, final revision, and final submission of this article; NY, LL, and WZ contributed to the data collection, language correction, and revision; and XZ and YX contributed to the data collection and drafting. CH and LX contributed to the language correction and final revision. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Menz BD, Kuderer NM, Bacchi S, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ 2024 Mar 20;384:e078538. [doi: 10.1136/bmj-2023-078538] [Medline: 38508682]
2. Thorp HH. ChatGPT is fun, but not an author. Science 2023 Jan 27;379(6630):313-313. [doi: 10.1126/science.adg7879]
3. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell 2023 May 4;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
4. Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024 Mar 29;7(1):82. [doi: 10.1038/s41746-024-01074-z] [Medline: 38553625]
5. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. JMIR Med Educ 2023 Nov 10;9:e49877. [doi: 10.2196/49877] [Medline: 37948112]
6. Qiu J, Li L, Sun J, et al. Large AI models in health informatics: applications, challenges, and the future. IEEE J Biomed Health Inform 2023 Dec;27(12):6074-6087. [doi: 10.1109/JBHI.2023.3316750] [Medline: 37738186]
7. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 2023 Jun 1;9:e48291. [doi: 10.2196/48291] [Medline: 37261894]
8. Hamid H, Zulkifli K, Naimat F, Che Yaacob NL, Ng KW. Exploratory study on student perception on the use of chat AI in process-driven problem-based learning. Curr Pharm Teach Learn 2023 Dec;15(12):1017-1025. [doi: 10.1016/j.cptl.2023.10.001] [Medline: 37923639]
9. Kavadella A, Dias da Silva MA, Kaklamanos EG, Stamatopoulos V, Giannakopoulos K. A mixed-methods evaluation of ChatGPT's real-life implementation in undergraduate dental education. JMIR Med Educ 2024 Jan 31;10:e51344. [doi: 10.2196/51344] [Medline: 38111256]
10. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. Med Teach 2024 May;46(5):657-664. [doi: 10.1080/0142159X.2023.2271159] [Medline: 37862566]
11. Liu J, Liu F, Fang J, Liu S. The application of Chat Generative Pre-trained Transformer in nursing education. Nurs Outlook 2023;71(6):102064. [doi: 10.1016/j.outlook.2023.102064] [Medline: 37879261]
12. Lee H. The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ 2023 Mar 14. [doi: 10.1002/ase.2270] [Medline: 36916887]
13. Carr SE, Canny BJ, Wearn A, et al. Twelve tips for medical students experiencing an interruption in their academic progress. Med Teach 2022 Oct;44(10):1081-1086. [doi: 10.1080/0142159X.2021.1921134] [Medline: 33969788]
14. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? Nurse Educ Pract 2023 Jan;66:103537. [doi: 10.1016/j.nepr.2022.103537] [Medline: 36549229]
15. Xu X, Wang X, Zhang Y, Zheng R. Applying ChatGPT to tackle the side effects of personal learning environments from learner and learning perspective: an interview of experts in higher education. PLoS ONE 2024 Jan 3;19(1):e0295646. [doi: 10.1371/journal.pone.0295646] [Medline: 38170691]
16. Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine. Clin Transl Med 2023 Mar;13(3):e1206. [doi: 10.1002/ctm2.1206] [Medline: 36854881]
17. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 2023 Dec 13;13(1):22129. [doi: 10.1038/s41598-023-49483-6] [Medline: 38092821]

18. Liu X, Wu C, Lai R, et al. ChatGPT: when the artificial intelligence meets standardized patients in clinical training. J Transl Med 2023 Jul 6;21(1):447. [doi: 10.1186/s12967-023-04314-0] [Medline: 37415217]
19. Heng JJY, Teo DB, Tan LF. The impact of chat generative pre-trained transformer (ChatGPT) on medical education. Postgrad Med J 2023 Sep 21;99(1176):1125-1127. [doi: 10.1093/postmj/qgad058] [Medline: 37466157]
20. Cheung BHH, Lau GKK, Wong GTC, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-a multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One 2023 Aug;18(8):e0290691. [doi: 10.1371/journal.pone.0290691] [Medline: 37643186]
21. Coşkun Ö, Kıyak YS, Budakoğlu I. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2024 Mar 13:1-7. [doi: 10.1080/0142159X.2024.2327477] [Medline: 38478902]
22. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312. [doi: 10.2196/45312] [Medline: 36753318]
23. Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023 Aug;229:172. [doi: 10.1016/j.ajog.2023.04.020]
24. Seetharaman R. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? J Med Syst 2023 May 9;47(1):61. [doi: 10.1007/s10916-023-01957-w] [Medline: 37160568]
25. Meyer JG, Urbanowicz RJ, Martin PCN, et al. ChatGPT and large language models in academia: opportunities and challenges. BioData Min 2023 Jul 13;16(1):20. [doi: 10.1186/s13040-023-00339-9] [Medline: 37443040]
26. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Feb;614(7947):224-226. [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
27. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023 Apr;307(2):e230163. [doi: 10.1148/radiol.230163] [Medline: 36700838]
28. Biswas S. ChatGPT and the future of medical writing. Radiology 2023 Apr;307(2):e223312. [doi: 10.1148/radiol.223312] [Medline: 36728748]
29. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. Radiology 2023 Apr;307(2):e230171. [doi: 10.1148/radiol.230171] [Medline: 36728749]
30. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 2023 Apr 26;6(1):75. [doi: 10.1038/s41746-023-00819-6] [Medline: 37100871]
31. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023 Jan;613(7944):423. [doi: 10.1038/d41586-023-00056-7] [Medline: 36635510]
32. Hwang T, Aggarwal N, Khan PZ, et al. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. PLoS ONE 2024 Feb 14;19(2):e0297701. [doi: 10.1371/journal.pone.0297701] [Medline: 38354135]
33. Cheng SL, Tsai SJ, Bai YM, et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res 2023 Dec 25;25:e51229. [doi: 10.2196/51229] [Medline: 38145486]
34. King MR, chatGPT. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. Cell Mol Bioeng 2023 Feb;16(1):1-2. [doi: 10.1007/s12195-022-00754-8] [Medline: 36660590]
35. Deveci CD, Baker JJ, Sikander B, Rosenberg J. A comparison of cover letters written by ChatGPT-4 or humans. Dan Med J 2023 Nov 23;70(12):A06230412. [Medline: 38018708]
36. King MR. The future of AI in medicine: a perspective from a chatbot. Ann Biomed Eng 2023 Feb;51(2):291-295. [doi: 10.1007/s10439-022-03121-w] [Medline: 36572824]
37. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus Google search for medical knowledge acquisition. Otolaryngol Head Neck Surg 2024 Jun;170(6):1484-1491. [doi: 10.1002/ohn.465] [Medline: 37529853]
38. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580. [doi: 10.2196/51580] [Medline: 38009003]
39. Mesko B. The ChatGPT (Generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. J Med Internet Res 2023 Jun 22;25:e48392. [doi: 10.2196/48392] [Medline: 37347508]
40. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb 19;15(2):e35179. [doi: 10.7759/cureus.35179] [Medline: 36811129]
41. Wong RSY, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023 Nov 21;9:e47274. [doi: 10.2196/47274] [Medline: 37988149]
42. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq 2023 Feb 28:rs.3.rs-2566942. [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]

43. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 2023 Jun;15(6):e40977. [doi: 10.7759/cureus.40977] [Medline: 37519497]
44. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol 2024 Jan;90(1):232-238. [doi: 10.1111/bcp.15896] [Medline: 37626010]
45. Zúñiga Salazar G, Zúñiga D, Vindel CL, et al. Efficacy of AI chats to determine an emergency: a comparison between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI chat. Cureus 2023 Sep;15(9):e45473. [doi: 10.7759/cureus.45473] [Medline: 37727841]
46. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023 Jun;307(5):e230922. [doi: 10.1148/radiol.230922] [Medline: 37310252]
47. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023 Sep;95:104770. [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
48. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: is ChatGPT a blessing or blight in disguise? Med Educ Online 2023 Dec;28(1):2181052. [doi: 10.1080/10872981.2023.2181052] [Medline: 36809073]
49. Haman M, Školník M. Using ChatGPT to conduct a literature review. Account Res 2023 Mar 6:1-3. [doi: 10.1080/08989621.2023.2185514] [Medline: 36879536]
50. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport 2023 Apr;40(2):615-622. [doi: 10.5114/biolsport.2023.125623] [Medline: 37077800]
51. Biri SK, Kumar S, Panigrahi M, Mondal S, Behera JK, Mondal H. Assessing the utilization of large language models in medical education: insights from undergraduate medical students. Cureus 2023 Oct;15(10):e47468. [doi: 10.7759/cureus.47468] [Medline: 38021810]
52. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023 Aug 11;25:e48009. [doi: 10.2196/48009] [Medline: 37566454]
53. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024 Jan;6(1):e12-e22. [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
54. Khera R, Butte AJ, Berkwits M, et al. AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. JAMA 2023 Sep 5;330(9):818-820. [doi: 10.1001/jama.2023.15481] [Medline: 37566406]
55. Stokel-Walker C, Van Noorden R. What ChatGPT and Generative AI mean for science. Nature 2023 Feb;614(7947):214-216. [doi: 10.1038/d41586-023-00340-6] [Medline: 36747115]
56. Patino GA, Amiel JM, Brown M, Lypson ML, Chan TM. The promise and perils of artificial intelligence in health professions education practice and scholarship. Acad Med 2024 May 1;99(5):477-481. [doi: 10.1097/ACM.0000000000005636] [Medline: 38266214]
57. Jegorova M, Kaul C, Mayor C, et al. Survey: leakage and privacy at inference time. IEEE Trans Pattern Anal Mach Intell 2023 Jul;45(7):9090-9108. [doi: 10.1109/TPAMI.2022.3229593] [Medline: 37015684]
58. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019 Jul 23;10(1):3069. [doi: 10.1038/s41467-019-10933-3] [Medline: 31337762]
59. Li J. Security implications of AI chatbots in health care. J Med Internet Res 2023 Nov 28;25:e47551. [doi: 10.2196/47551] [Medline: 38015597]
60. Open AI. Data usage for consumer services FAQ. OpenAI Help Center. URL: https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq [accessed 2023-12-29]
61. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
62. Guidi C, Traversa C. Empathy in patient care: from 'clinical empathy' to 'empathic concern'. Med Health Care Philos 2021 Dec;24(4):573-585. [doi: 10.1007/s11019-021-10033-4] [Medline: 34196934]
63. Koranteng E, Rao A, Flores E, et al. Empathy and equity: key considerations for large language model adoption in health care. JMIR Med Educ 2023 Dec 28;9:e51199. [doi: 10.2196/51199] [Medline: 38153778]
64. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. JMIR Med Educ 2023 Aug 14;9:e50945. [doi: 10.2196/50945] [Medline: 37578830]

Abbreviations

AI: artificial intelligence
CT: computed tomography
LLMs: large language models

MRI: magnetic resonance imaging
SPs: standardized patients

Edited by C Lovis; submitted 29.12.23; peer-reviewed by A Kavadella, H Alkaissi, M King, M Alshiekh, Y Yu; revised version received 25.04.24; accepted 08.06.24; published 25.07.24.

Please cite as:
Zhui L, Yhap N, Liping L, Zhengjie W, Zhonghao X, Xiaoshu Y, Hong C, Xuexiu L, Wei R
Impact of Large Language Models on Medical Education and Teaching Adaptations
JMIR Med Inform 2024;12:e55933
URL: https://medinform.jmir.org/2024/1/e55933
doi:10.2196/55933

How to Elucidate Consent-Free Research Use of Medical Data: A Case for "Health Data Literacy"

Gesine Richter[1,2], PhD; Michael Krawczak[3], MSc, PhD

1

2

3

Corresponding Author:
Gesine Richter, PhD

Abstract

The extensive utilization of personal health data is one of the key success factors of modern medical research. Obtaining consent to the use of such data during clinical care, however, bears the risk of low and unequal approval rates and risk of consequent methodological problems in the scientific use of the data. In view of these shortcomings, and of the proven willingness of people to contribute to medical research by sharing personal health data, the paradigm of informed consent needs to be reconsidered. The European General Data Protection Regulation gives the European member states considerable leeway with regard to permitting the research use of health data without consent. Following this approach would however require alternative offers of information that compensate for the lack of direct communication with experts during medical care. We therefore introduce the concept of "health data literacy," defined as the capacity to find, understand, and evaluate information about the risks and benefits of the research use of personal health data and to act accordingly. Specifically, health data literacy includes basic knowledge about the goals and methods of data-rich medical research and about the possibilities and limits of data protection. Although the responsibility for developing the necessary resources lies primarily with those directly involved in data-rich medical research, improving health data literacy should ultimately be of concern to everyone interested in the success of this type of research.

(JMIR Med Inform 2024;12:e51350) doi:10.2196/51350

KEYWORDS

health data literacy; informed consent; broad consent; data sharing; data collection; data donation; data linkage; personal health data

Data-Rich Research, Broad Consent, and Informedness

Various initiatives around the world are currently working on the technical and organizational requirements to make data from different sources and contexts usable for medical research (eg, MyHealthRecord in Australia, FINDATA in Finland, and the Medical Informatics Initiative in Germany). The starting points of these endeavors often are local, regional, or national health care data repositories that must nevertheless be highly linkable to allow full exploitation of their scientific value. This connectivity requirement implies that the data cannot be fully anonymized before being moved into the research domain.

One of the ethical prerequisites for research on humans—and thus for research using identifiable personal health data—is the informed consent of the data subjects. However, being properly informed requires that those affected (1) are capable of making self-determined decisions in the first place; (2) were informed about the nature, benefits, and risks of the research in question; (3) have understood the importance of this information; and (4) are able to decide voluntarily and without coercion for or against participation.

Not least because of the increasing relevance of hypotheses-free research approaches (keyword: big data), the storage and use of data for future, currently undeterminable purposes also play an increasingly important role in medical research. Recent studies have shown that patients and members of the general public are very willing to share personal health data for research (eg, [1]), even if no information about the purposes and aims of the research can be provided at the time consent is given. Notably, this attitude turned out to be mainly motivated by altruism, solidarity, and the idea of reciprocity. Since the paradigm of project-related informed consent is difficult to transfer to such unspecific practice, the World Medical Association changed its regulations on research with identifiable data when revising the Declaration of Helsinki in 2013 [2]. There was no longer a requirement for specific information about the subjects of future research, thereby paving the way for a new form of "broad consent."

In essence, "broad consent" means the one-off, unspecific agreement to the use of one's personal data for medical research without knowing who will access the data when and to what end. However, since the data in question are usually collected in a clinical care context, the suitability and practicality of broad

consent as a legitimation for their research use is limited. First, the temporal and spatial linking of the consent process to care measures can lead to incorrect therapeutic [3] and diagnostic [4] assumptions on the side of the patient. Second, in the time available, it is hardly possible to create sufficient understanding of the benefits and risks of the envisaged research, despite great efforts to ensure that the corresponding information and consent documents are legible. Finally, asking for consent during clinical care bears a substantial risk of low and unequal approval rates, which can lead to methodological problems in the scientific use of the data.

In view of these shortcomings, and of the proven willingness of people to contribute to medical research by sharing personal health data, the means to achieve practically feasible and truly informed consent needs to be reconsidered. In particular, is consent-free data use for medical research, combined with the possibility of straightforward opt-out by the data subjects after thorough consideration, a better option for legitimizing the secondary use of health data? This question is all the more justified as numerous studies in the United Kingdom, Iceland, Norway, Sweden, and Germany, among others, have shown a generally positive attitude of people toward such a regulation (eg, United Kingdom [5]; United Kingdom, Iceland, Norway, and Sweden [1]; Norway [6]; and Germany [7,8]).

In the following, we will first introduce “data donation” as an opt-out approach to legitimizing the secondary research use of personal medical data. Since opt-out would imply that patients are no longer informed directly about the research-associated risks and benefits, alternative ways of information provision must be explored in the context of data donation if the paradigm of informedness was to be maintained. We therefore also introduce the concept of “health data literacy,” defined as the capacity to find, understand, and evaluate information about data-rich medical research. Although a case for general health data literacy can be made independently of the issue of patient consent, its consideration becomes particularly urgent for the latter if the framework of consenting was to change from opt-in to opt-out.

Data Donation: Consent-Free Research Use of Medical Data Plus Opt-Out

The European General Data Protection Regulation (EU-GDPR) gives European member states considerable leeway with regard to permitting the research use of health data without consent. While Article 9 Paragraph 1 of the EU-GDPR clearly prohibits the processing of personal genetic, biometric, or health data, Article 9 Paragraph 2(j) explicitly exempts processing for scientific research purposes [9]. In addition, Article 89 allows national legislation to provide for this exception, subject to appropriate safeguards for the rights and freedom of the data subjects.

In Germany, the ethical, legal, technological, and organizational framework of the consent-free use of health data was examined in 2020 in a detailed report to the Federal Ministry of Health [10]. In addition to its legal admissibility, the report addressed the scientific benefits of such an approach, its impact upon the right of informational self-determination, and the necessity and possibilities for fair involvement of the data subjects. The authors concluded that it would be possible in Germany to replace the requirement for explicit consent for research with personal medical data by an equivalent legal permission, combined with an easy-to-exercise opt-out. Under certain conditions, such “data donation” (as it was termed in the report) would be both legally possible and ethically reasonable.

The above notwithstanding, the authors were also unequivocal that the actual process of data access by potential users should be independent of whether access is legitimized by opt-in or opt-out. The involvement of an ethics board or a use-and-access committee that reviews and decides data applications remains essential in both cases. Notably, such institutions also play an important role in weighing the potential risks and benefits of individual research projects, a legitimation mechanism that was deliberately placed on the same level as consent by the EU-GDPR.

Importantly with a view to the following considerations, the report clarified that, in addition to technical and organizational protective measures, one prerequisite for the acceptability of data donation would be that patients and citizens were sufficiently well informed about it. This proviso inevitably leads to the question of how sufficient knowledgeability can be achieved if the decision about sharing one’s data for research purposes is no longer made actively, following thorough verbal explanation, but passively by exercising or not exercising a right of objection.

Limits of Top-Down “Informability”: the COVID-19 Infodemic as an Example

Since data donation, in the above sense, would be temporally and spatially decoupled from medical care and instead be anchored in everyday life, alternative offers of information would have to compensate for the lack of direct communication with medical or scientific experts [11]. Yet, the COVID-19 pandemic recently highlighted that the expansion of top-down media campaigns alone is not sufficient to adequately convey the complex aspects of medical research to the general public. Instead, it turned out that, despite the general increase in information provided, many people who opposed vaccination in the first place still were not sufficiently receptive to scientific facts [12]. Moreover, even some kind of social grouping occurred along people’s vaccination status, and the COSMO study carried out in Germany and Austria revealed that the stronger the identification with being unvaccinated, the lower the inclination to change this status, and the greater the feeling of discrimination [13]. Obviously, the ability to become informed (“informability”) had reached its limits in view of the amount of information available, a paradox that lamentably also had a negative impact upon the effectiveness of public health measures taken.

In connection with the COVID-19 pandemic, the World Health Organization (WHO) coined the term “infodemic” for the increasingly observed susceptibility of people to fake news as a result of reduced informability. According to the WHO, the

infodemic caused a high degree of uncertainty in the population, a greater willingness to engage in health-damaging and risk-taking behavior, and an increased distrust of the health authorities [14]. The "Infodemic Management" called for by the WHO aimed to enable the population to better understand information from health experts and to become more resistant to misinformation [15].

Ways to Better Informability?

In view of its complexity, it seems unrealistic to convey all relevant information about the research use of personal health data at once. We therefore propose "health data literacy" as a basis for better informability of the general population and, hence, as a means to uphold the paradigm of informed consent even in the context of data donation in the above sense. For a well-informed general public, data donation would indeed mean nothing more than a change in decision format—from opt-in to opt-out.

In a narrower sense, the word "literacy" stands for the ability to read and, thereby, to acquire education and knowledge. According to the Organisation for Economic Co-operation and Development (OECD), understanding and interpreting written material should enable citizens to develop their own potential and to fully participate in societal affairs [16]. The starting point of our considerations on health data literacy therefore will be a class of communication models that focus upon the possible causes of limited informability.

One decisive factor for the success of communication is the thought system of the recipient. Since we often have little time to consider large amounts of everyday information, we believe statements that we have heard very often to be more credible than others [17]. This effect is reinforced by the phenomenon of group polarization: those who share a widespread opinion on complex issues are more likely to be reserved about new information and tend to believe whatever confirms their own viewpoint rather than information that does not fit. This selective form of information intake can, for example, increase polarization in social disputes even in the presence of reliable evidence and information [18]. The concept of health data literacy picks up on the basic idea of these communication models and aims to create anchor points in the knowledge base of people, where information on the benefits and risks of data-rich medical research can be stored and evaluated.

Value congruence approaches aim in a similar direction, in that they try to increase trust in certain institutions [19,20]. Such trust will be greater when more individuals perceive that their interests and values are shared by the institution in question, because trust is also largely based upon the perception of common values. This applies all the more to institutions that use health data for research, and it is therefore in the best interest of such institutions to develop and represent values that are highly rated by the public [19,20]. In this context, widespread health data literacy could form the breeding ground for the perception of a congruence of values and, thus, for greater trust in the recipients and beneficiaries of data donation.

The Concept of "Health Data Literacy"

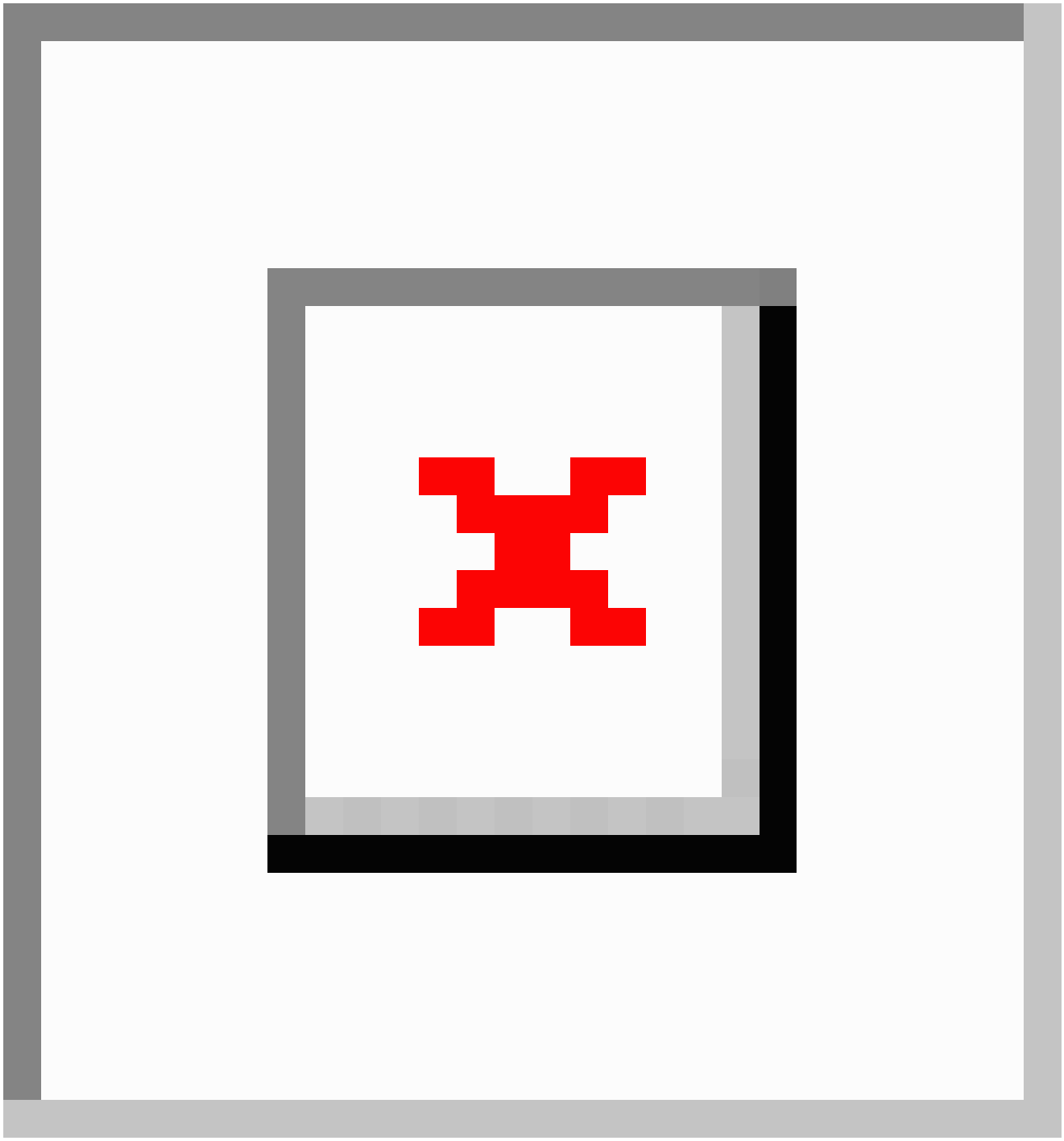
An individual's health data literacy is positioned between their health literacy and their data literacy, where the latter in particular has been promoted politically, for example, by the data strategy of the German federal government [21].

- In view of the increasingly specific treatment options promised by so-called "precision medicine," citizens would be well advised to take an interest in issues related to disease prevention and medical care [22]. The associated term "health literacy" summarizes both the motivation and the ability to find, understand, evaluate, and apply the information underlying personal health–related decisions [23]. Numerous international studies have measured and compared the level of health literacy in different populations (eg, [24]), as well as spurring considerations as to how health literacy can be increased (eg, [25]).
- The term "data literacy" refers to knowledge about data and their use in general, including legal, ethical, and social aspects. Data literacy thus forms the basis of personal self-determination in an increasingly digitalized society [26]. The aim of data literacy is an ability to weigh one's own personal rights against the potential benefits of making personal data available to others [27].

In combining both abovementioned terms, "health data literacy" stands for the capacity to find, understand, and evaluate information about the risks and benefits of medical research with personal health data; to compare this information with one's own values; and to act accordingly. Health data literacy is thus a transformer of information into informed action, aimed at a level of thematic familiarity that enables self-determined decision-making about the sharing of one's own health data with the research community. Specifically, health data literacy should at the very least include basic knowledge about the goals and methods of data-rich medical research and about the possibilities and limits of data protection.

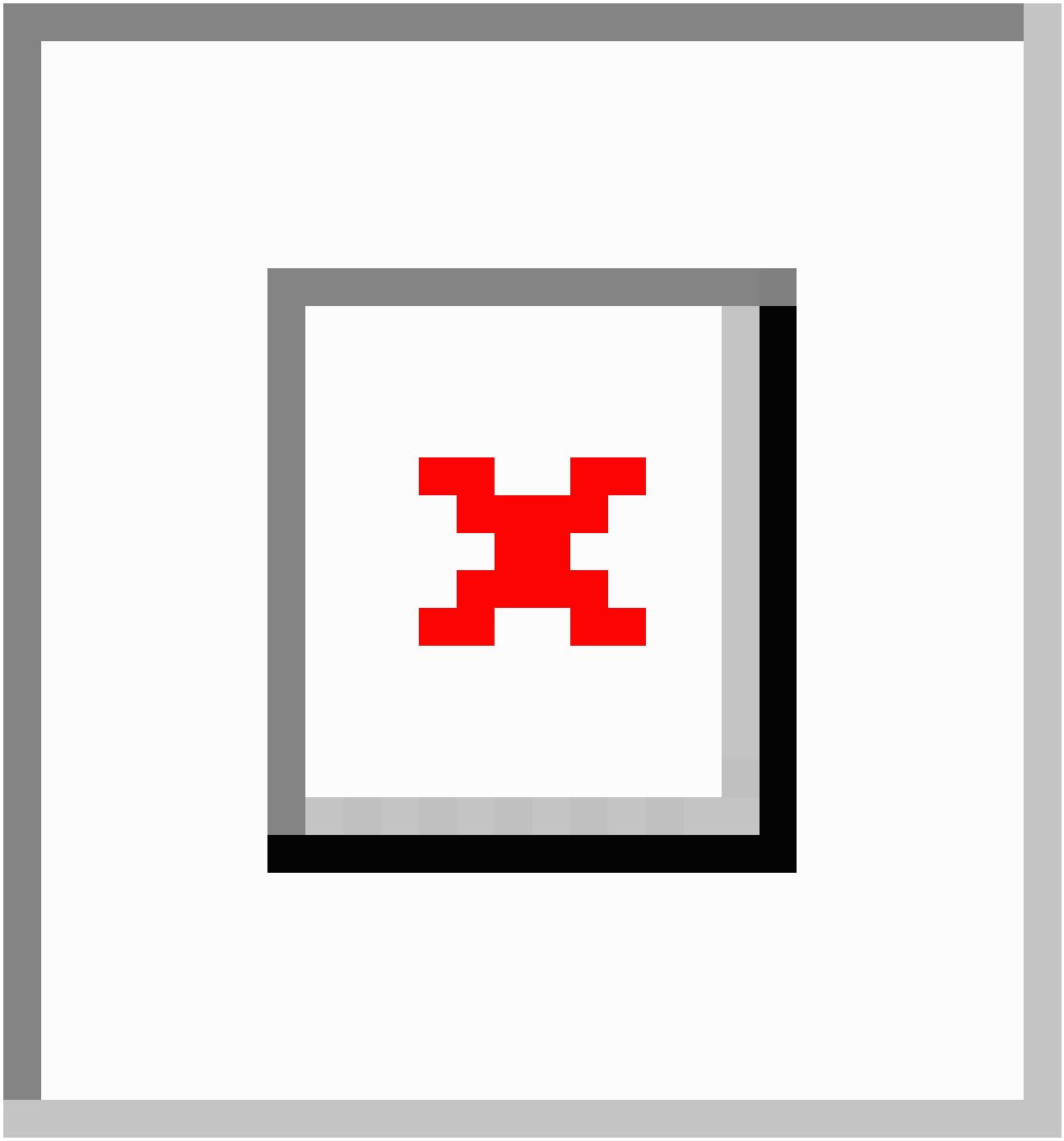
The increasing relevance of personal health data for medical research has led to a large number of measures to increase the societal acceptance of the use of such data. However, legislative regulations on data governance and data protection, as well as efforts to increase patient involvement and public information, are likely to have greater impact when they are met with more adequate prior knowledge in the sense of health data literacy (Figure 1).

Figure 1. Health data literacy as a breeding ground for the societal acceptance of data donation.



When the mental anchor points set by health data literacy receive information on scientific successes, new technical and organizational developments, as well as possible setbacks of data-rich medical research (keyword: transparency), this information can be evaluated competently by the recipient and compared to their own expectations. In the aftermath of such reflections, informed self-determination and sufficient trust in regulations and institutions can develop (Figure 2).

Figure 2. Transparency nourishes confidence and trust in data-rich medical research.



Outlook: Feasibility and Implementation of Health Data Literacy

Numerous international studies among patients and in the general population have revealed a broad positive attitude toward the provision of personal health data to medical research (eg, [28]). This approval was consistently found to be driven by a sense of reciprocity, that is, a wish to give something back after benefiting from research (eg, [29,30]). Evidence also emerged for the widespread belief in a social duty of citizens to contribute their own data to research, independent of their personal benefit [31,32]. At the same time, however, a craving for more detailed information was observed, up to and including the view that every individual is responsible themselves to find out about the nature and benefits of research with personal health data (eg, [33]).

In summary, we are thus in a situation where (1) there is little doubt about the need to utilize personal health data from different contexts to achieve the goals of modern medical research, (2) the consent-free use of such data meets broad approval by the general public, and (3) there is a widespread willingness of people to acquire the knowledge necessary to make a self-determined decision about data donation. The most compelling argument for general health data literacy is therefore self-evident: widespread background knowledge of the risks and benefits of data-rich medical research would allow the paradigm of informedness to be maintained even if consent to

participation in research is implement by opt-out, rather than opt-in.

However, the appeal of general health data literacy undoubtedly goes beyond the issue of data donation. Its necessity arises from the increasing complexity of data-rich medical research, which can no longer be explained adequately via waiting room leaflets or doctor consultations. We are also aware that improved health data literacy could, in principle, help to reduce some of the misunderstandings of patients that we somehow held against broad consent when advocating data donation. However, in view of the many advantages of data donation summarized above, we think that only little importance should be attached to this possibility.

Attempts to establish general health data literacy should strive for a certain level of competence across as broad a proportion of the population as possible. This goal not only expresses fairness and ensures equal representation of different societal groups in medical research but can also help to reduce the vulnerability to fake information as a potential threat to public health, as observed during the COVID-19 pandemic. Achieving equity in practice will require the development and provision of target group–specific offers of information and education. One particularly efficient way to increase health data literacy across the board would be to start this process in school, as suggested previously to strengthen health literacy [25]. This approach is not only easy to implement in practice; it would also offer the opportunity to use children as multipliers among friends and family.

Further research is needed to determine exactly what kind of information should be communicated, in what form, and to whom to improve health data literacy in a given population. These questions are ideally answered through cocreation research involving representatives of different target groups to enhance the credibility of the education curriculum and content among end users. However, although the responsibility for developing the necessary resources lies primarily with those directly involved in data-rich medical research, improving health data literacy should ultimately be of concern to everyone interested in the success of this type of research.

Acknowledgments

We acknowledge financial support by Deutsche Forschungsgemeinschaft (DFG) within the funding program Open Access-Publikationskosten. We are most grateful to Claudia Bozzaro, Kiel University, for discussing health data literacy with us.

Authors' Contributions

The idea of health data literacy was first conceived by GR; GR and MK jointly developed the concept further and authored the manuscript.

Conflicts of Interest

None declared.

References

1. Viberg Johansson J, Bentzen HB, Shah N, et al. Preferences of the public for sharing health data: discrete choice experiment. JMIR Med Inform 2021 Jul 5;9(7):e29614. [doi: 10.2196/29614] [Medline: 36260402]
2. WMA Declaration of Helsinki – ethical principles for medical research involving human subjects. World Medical Association. 2022 Sep 6. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [accessed 2024-01-18]
3. Appelbaum PS, Roth LH, Lidz CW, Benson P, Winslade W. False hopes and best data: consent to research and the therapeutic misconception. Hastings Cent Rep 1987 Apr;17(2):20-24. [Medline: 3294743]
4. Nobile H, Vermeulen E, Thys K, Bergmann MM, Borry P. Why do participants enroll in population biobank studies? a systematic literature review. Expert Rev Mol Diagn 2013 Jan;13(1):35-47. [doi: 10.1586/erm.12.116] [Medline: 23256702]
5. Jones LA, Nelder JR, Fryer JM, et al. Public opinion on sharing data from health services for clinical and research purposes without explicit consent: an anonymous online survey in the UK. BMJ Open 2022 Apr 27;12(4):e057579. [doi: 10.1136/bmjopen-2021-057579] [Medline: 35477868]
6. Eikemo H, Roten LT, Vaaler AE. Research based on existing clinical data and biospecimens: a systematic study of patients' opinions. BMC Med Ethics 2022 Jun 16;23(1):60. [doi: 10.1186/s12910-022-00799-4] [Medline: 35710552]
7. Richter G, Trigui N, Caliebe A, Krawczak M. Attitude towards consent-free research use of personal medical data in the general German population. Heliyon 2024 Mar 11;10(6):e27933. [doi: 10.1016/j.heliyon.2024.e27933] [Medline: 38509969]
8. Köngeter A, Schickhardt C, Jungkunz M, Bergbold S, Mehlis K, Winkler EC. Patients' willingness to provide their clinical data for research purposes and acceptance of different consent models: findings from a representative survey of patients with cancer. J Med Internet Res 2022 Aug 25;24(8):e37665. [doi: 10.2196/37665] [Medline: 36006690]
9. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive

95/46/EC (General Data Protection Regulation). EUR-Lex. URL: http://data.europa.eu/eli/reg/2016/679/2016-05-04 [accessed 2024-01-18]
10. Strech D, von Kielmansegg S, Zenker S, Krawczak M, Semler SC. Wissenschaftliches Gutachten „Datenspende“ – Bedarf für die Forschung, ethische Bewertung, rechtliche, informationstechnologische und organisatorische Rahmenbedingungen [Article in German]. Bundesministerium für Gesundheit. 2020 Mar 30. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Ministerium/Berichte/Gutachten_Datenspende.pdf [accessed 2024-01-18]
11. Jungkunz M, Köngeter A, Spitz M, et al. Stellungnahme zur Etablierung der sekundären Forschungsnutzung von Behandlungsdaten in Deutschland: Ergebnisse des Verbundprojekts LinCDat: „Learning from Clinical Data. Ethical, Social and Legal Aspects" [Article in German]. Forum Marsilius-Kolleg 2022 Nov 24;21. [doi: 10.11588/fmk.2022.1.91697]
12. Rathore FA, Farooq F. Information overload and infodemic in the COVID-19 pandemic. J Pak Med Assoc 2020 May;70(Suppl 3)(5):S162-S165. [doi: 10.5455/JPMA.38] [Medline: 32515403]
13. COSMO PANEL—Langzeitstudie zum Erleben und Verhalten von Geimpften und Ungeimpften in Deutschland und Österreich [Article in German]. COSMO. 2022 Mar 18. URL: https://projekte.uni-erfurt.de/cosmo2020/web/summary/panel2/ [accessed 2022-12-13]
14. Infodemic. World Health Organization. URL: https://www.who.int/health-topics/infodemic#tab=tab_1 [accessed 2023-07-21]
15. 1st WHO infodemic manager training. World Health Organization. 2020 Nov. URL: https://www.who.int/teams/epi-win/infodemic-management/1st-who-training-in-infodemic-management [accessed 2023-07-21]
16. Adult literacy. Organisation for Economic Co-operation and Development (OECD). URL: https://www.oecd.org/education/innovation-education/adultliteracy.htm [accessed 2023-07-21]
17. Kahneman D. Schnelles Denken, Langsames Denken: Siedler Verlag, München; 2021.
18. Lord CG, Ross L, Lepper MR. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. J Pers Soc Psychol 1979 Nov;37(11):2098-2109. [doi: 10.1037//0022-3514.37.11.2098]
19. Sheehan M, Friesen P, Balmer A, et al. Trust, trustworthiness and sharing patient data for research. J Med Ethics 2020 May 18:medethics-2019-106048. [doi: 10.1136/medethics-2019-106048] [Medline: 32424061]
20. Holland S, Cawthra J, Schloemer T, Schröder-Bäck P. Trust and the acquisition and use of public health information. Health Care Anal 2022 Mar;30(1):1-17. [doi: 10.1007/s10728-021-00436-y] [Medline: 34751865]
21. Datenstrategie der Bundesregierung: Eine Innovationsstrategie für gesellschaftlichen Fortschritt und nachhaltiges Wachstum - Kabinettfassung [Article in German]. Bundesregierung. 2021 Jan 27. URL: https://www.publikationen-bundesregierung.de/pp-de/publikationssuche/datenstrategie-der-bundesregierung-1845632 [accessed 2023-07-21]
22. Budin-Ljøsne I, Harris JR. Ask not what personalized medicine can do for you--ask what you can do for personalized medicine. Public Health Genomics 2015 Mar 6;18(3):131-138. [doi: 10.1159/000373919] [Medline: 25766382]
23. Sørensen K, van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
24. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058. [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
25. Schaeffer D, Hurrelmann K, Bauer U. Nationaler Aktionsplan Gesundheitskompetenz Die Gesundheitskompetenz in Deutschland Stärken: KomPart Verlagsgesellschaft mbH; 2018.
26. Renz A, Etsiwah B, Burgueno Hopf AT. Datenkompetenz. Whitepaper: Weizenbaum-Institut für die vernetzte Gesellschaft; 2021.
27. Hummel P, Braun M, Augsberg S, von Ulmenstein U, Dabrock P. Datensouveränität Governance-Ansätze Für Den Gesundheitsbereich: Springer; 2021:11. [doi: 10.1007/978-3-658-33755-1]
28. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016 Nov 10;17(1):73. [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]
29. Richter G, Krawczak M, Lieb W, Wolff L, Schreiber S, Buyx A. Broad consent for health care-embedded biobanking: understanding and reasons to donate in a large patient sample. Genet Med 2018 Jan;20(1):76-82. [doi: 10.1038/gim.2017.82] [Medline: 28640237]
30. A public dialogue on genomic medicine: time for a new social contract? Ipsos MORI. 2019. URL: https://www.ipsos.com/sites/default/files/ct/publication/documents/2019-04/public-dialogue-on-genomic-medicine-full-report.pdf [accessed 2024-01-18]
31. Skatova A, Goulding J. Psychology of personal data donation. PLoS One 2019 Nov 20;14(11):e0224240. [doi: 10.1371/journal.pone.0224240] [Medline: 31747408]
32. Richter G, Borzikowsky C, Hoyer BF, Laudes M, Krawczak M. Secondary research use of personal medical data: patient attitudes towards data donation. BMC Med Ethics 2021 Dec 15;22(1):164. [doi: 10.1186/s12910-021-00728-x] [Medline: 34911502]
33. Platt J, Raj M, Büyüktür AG, et al. Willingness to participate in health information networks with diverse data use: evaluating public perspectives. EGEMS (Wash DC) 2019 Jul 25;7(1):33. [doi: 10.5334/egems.288] [Medline: 31367650]

Abbreviations

EU-GDPR: European General Data Protection Regulation
OECD: Organisation for Economic Co-operation and Development
WHO: World Health Organization

Edited by C Lovis; submitted 28.07.23; peer-reviewed by G Arnolda, LD C, S McLennan, S Wiertz; revised version received 19.01.24; accepted 21.04.24; published 18.06.24.

Please cite as:
Richter G, Krawczak M
How to Elucidate Consent-Free Research Use of Medical Data: A Case for "Health Data Literacy"
JMIR Med Inform 2024;12:e51350
URL: https://medinform.jmir.org/2024/1/e51350
doi:10.2196/51350

Viewpoint

Using a Natural Language Processing Approach to Support Rapid Knowledge Acquisition

Taneya Y Koonce[1*], MSLS, MPH; Dario A Giuse[2*], MS, Dr Ing; Annette M Williams[1], MLS; Mallory N Blasingame[1], MA, MSIS; Poppy A Krump[1], MSIS; Jing Su[1], MS, MSIS, MD; Nunzia B Giuse[1,2], MLS, MD

[1]Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN, United States

[2]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Taneya Y Koonce, MSLS, MPH
Center for Knowledge Management
Vanderbilt University Medical Center
3401 West End
Suite 304
Nashville, TN, 37203
United States
Phone: 1 6159365790
Email: taneya.koonce@vumc.org

Abstract

Implementing artificial intelligence to extract insights from large, real-world clinical data sets can supplement and enhance knowledge management efforts for health sciences research and clinical care. At Vanderbilt University Medical Center (VUMC), the in-house developed Word Cloud natural language processing system extracts coded concepts from patient records in VUMC's electronic health record repository using the Unified Medical Language System terminology. Through this process, the Word Cloud extracts the most prominent concepts found in the clinical documentation of a specific patient or population. The Word Cloud provides added value for clinical care decision-making and research. This viewpoint paper describes a use case for how the VUMC Center for Knowledge Management leverages the condition-disease associations represented by the Word Cloud to aid in the knowledge generation needed to inform the interpretation of phenome-wide association studies.

(JMIR Med Inform 2024;12:e53516) doi:10.2196/53516

KEYWORDS

natural language processing; electronic health records; machine learning; data mining; knowledge management; NLP

Introduction

The rapid advancement and availability of artificial intelligence (AI) approaches provide biomedical informatics groups with opportunities for exploring and generating insights from internal and external data at scale to enhance health sciences research and clinical care [1,2]. One such opportunity is using natural language processing (NLP) to extract usable knowledge from the vast amounts of structured and unstructured clinical data captured daily via the electronic health record (EHR). Insights from this process can be used to inform patient care, target information provision, and generate research hypotheses. This paper presents some of the activities that such usable knowledge makes possible.

Vanderbilt University Medical Center (VUMC) maintains an electronic health repository containing data for over 4.6 million individuals, going back to 1995, which includes structured data (eg, laboratory results and vital signs), textual data (eg, provider notes and radiology interpretations), reports (eg, electrocardiograms and pulmonary function test results), and image data. Included in this vendor-agnostic repository are all VUMC patient data captured from the in-house developed StarPanel EHR (VUMC) dating back to 2001 [3] and VUMC's current vendor-based EHR (Epic; Epic Systems Corporation), which was implemented in 2017 [4]. Roughly 850,000 new documents are added daily.

To identify and quickly represent the most critical information about a particular patient or population from this large data set, VUMC established the Word Cloud, a real-time and at-scale concept extraction tool that uses NLP to create a visual, time-oriented representation of clinical data [5-7]. The Word Cloud NLP uses a rules-based, finite-state machine approach

to process all nonimage incoming documents in real time and extract coded concepts using the Unified Medical Language System (UMLS) terminology [8]. With a processing speed of more than 50,000 documents per minute, the Word Cloud NLP is faster than currently available concept extraction NLP tools such as Apache cTakes (50,000 documents per hour; Apache Software Foundation, Mayo Clinic) [9] and MetaMap (22 citations per minute; National Library of Medicine) [10]. The rapid speed allows for better integration into the clinical workflow as real time–generated Word Cloud concepts are immediately presented to health care providers as they access the feature in the medical chart. The system handles all linguistic phenomena in clinical text, including acronyms, abbreviations, misspellings, negation, family history, uncertainty, and differential diagnosis. Excluding image data, the entire EHR repository is included in the Word Cloud NLP database, which uses close to 14,000 UMLS concepts to index 1.7 billion documents. In addition to the individual patient concepts, which include pointers to the original documents, the database also includes population-level associations of any pair of concepts.

The original purpose of the Word Cloud data was to provide a user interface that displays all concepts extracted from a patient's clinical documents in a word cloud display, with the size of each concept indicating how often the concept was documented for the patient. This interface is available to all users of the EHR. The Word Cloud data have been used since 2019 to generate clinical alerts for a variety of situations, such as flagging patients with implanted cardiac devices and a positive blood culture, patients with signs of serious inflammation due to immune checkpoint inhibitors, or patients with Andersen-Tawil syndrome who might be candidates for enrollment into a research study. The Word Cloud data drive real-time decision support by injecting detected concepts back into the VUMC EHR [11]. Because all the concepts extracted by the Word Cloud NLP are stored in the enterprise data lake, these data are also available for retrospective research and can be easily combined with other data such as the International Statistical Classification of Diseases codes or coded medications data [11].

The Center for Knowledge Management (CKM) has explored how the Word Cloud can be leveraged by information scientists engaged in EHR projects. The CKM facilitates the discovery and integration of external knowledge into medical practice and promotes curation, archiving, and reuse of internal knowledge across VUMC [12-15]. This viewpoint paper details how the CKM's innovative application of the Word Cloud enhances knowledge generation processes and describes future directions for NLP in knowledge management.

Case Description

Collaborations with medical center researchers comprise the majority of the CKM's partnership activities. A recent project to inform the interpretation of phenome-wide association studies (PheWASs) using evidence-linked knowledge bases illustrates these types of partnerships [16]. PheWASs examine relationships between markers (genetic or nongenetic) and phenotypes, producing extensive lists of possibly relevant marker-phenotype associations [17,18]. A methodological approach to compare known associations with PheWAS results can make it easier to identify potentially novel PheWAS outcomes [16]. Knowledge bases—created in part from synthesized evidence sources and primary literature documenting disease causes, risks, and complications—can be used for these comparisons.

For this research collaboration, the CKM created a "condition flowchart" with the causes, risk factors, and complications of a given medical condition. The sources consulted to create the flowchart include evidence synthesis resources (eg, UpToDate; UpToDate, Inc), medical textbooks (eg, Goldman-Cecil Medicine), and consumer health websites (eg, MedlinePlus; National Library of Medicine). From each source, the CKM team identified all causes, risk factors, and complications for the condition of interest and added them to the flowchart. Our collaborators then used the flowchart to create a knowledge base of phecodes for the PheWAS analysis. During flowchart creation, the CKM leveraged the Word Cloud to identify meaningful disease-condition associations—based on real-world population-level data—and target appropriate primary literature to substantiate the observed linkages.

Identifying Meaningful Condition Associations From the EHR

Each flowchart focuses on a specific clinical condition (eg, hypertension and hypotension), which is searched against the Word Cloud. Using a population-level analysis feature, the Word Cloud returns a list of all UMLS concepts represented in the EHR records of patients with the specified condition. The expected value is calculated for each UMLS concept [19] and the ratio of actual-to-expected patient cases (ie, strength of association) is then used to rank the list of causes, risk factors, and complications on the flowchart. This ranking thus provides our team with rapid knowledge acquisition of what is associated with the condition of interest. The actual-to-expected ratio for concept 1 and concept 2 is computed as follows:

where T=total population size, $a_{1,2}$=number of patients with both concept 1 and concept 2, n_1=number of patients with concept 1, and n_2=number of patients with concept 2.

A strength of association ratio of 15 or higher indicates that the concept occurs more often than expected by chance and signifies a meaningful relationship between the UMLS concept and the condition. Figure 1 provides an example of the UMLS concepts most associated with orthostatic hypotension in 71,996 patients.

Figure 1. Snapshot of the Word Cloud population-level list of UMLS concepts associated with orthostatic hypotension. Concepts are listed in descending order by the strength-of-association ratio, that is, the ratio of actual to expected number of cases in the VUMC EHR with the pairwise association of UMLS concepts. The population frequency of each term is also displayed. The ratio is used to rank the condition flowchart. CUI: concept unique identifier; EHR: electronic health record; Misc.: miscellaneous; MsgBskts: message baskets; NewRes: new results; Pop. Freq.: population frequency; Pt.Chart: patient chart; Pt.Lists: patient lists; StNotes: Star Notes; UMLS: Unified Medical Language System; VUMC: Vanderbilt University Medical Center; WhBoards: white boards.

Pt.Chart | StNotes | Forms | Panels | Pt.Lists | TaskList | MsgBskts | WhBoards | NewRes | SignDrafts | Misc.

= | 1 | 2 | 3

Enter one or more UMLS concepts to see population-level associations for all patients with those concepts documented. Ctrl-click o

Term: Show Use a panel

orthostatic hypotension ×

C0020651 71996 orthostatic hypotension

Total: 71996 patients

Filter by type: - All - system: - All -

CUI	Name	Expect	Actual	Ratio	Pop. Freq.
C4076686	supine hypertension	49.565	2721	54.898	2916
C0393911	autonomic failure	50.006	2698	53.953	2942
C1868528	autonomic hypotension	2.363	125	52.907	139
C2315246	pandysautonomia	0.714	36	50.428	42
C0877382	delayed renal graft function	0.544	26	47.801	32
C3662086	multisystem degeneration of autonomic nervous system	0.629	30	47.702	37
C0037019	shy-drager syndrome	8.686	410	47.204	511
C1959798	baroreflex failure syndrome	5.337	237	44.405	314
C0393571	multisystem atrophy	14.482	631	43.572	852
C0877157	pancreas transplant rejection	1.377	56	40.674	81
C5395071	autoimmune ganglionopathy	0.153	6	39.222	9
C1963757	dopamine dysregulation syndrome	0.561	21	37.439	33
C0920233	pisa syndrome	1.122	40	35.656	66

The Word Cloud also aids in identifying concepts most applicable to guide the ranking by displaying each term's UMLS semantic type. In the UMLS Metathesaurus, each concept term is assigned to 1 or more of 127 types in the vocabulary's hierarchical semantic network [20]. Semantic types most relevant for comparison with the condition flowchart include disease or syndrome, injury or poisoning, mental or behavioral dysfunction, sign or symptom, finding, and congenital abnormality. The Word Cloud provides a filter to exclude concepts with semantic types nonrelevant to this task (eg, procedures).

The Word Cloud often lists multiple UMLS concepts that can be grouped to correspond with a single term on the condition flowchart. For example, the Word Cloud concepts associated with orthostatic hypotension include Shy-Drager syndrome, multisystem degeneration of autonomic nervous system, and multisystem atrophy (Figure 1). In 1998, Shy-Drager syndrome was newly categorized as a multisystem atrophy and is no longer the preferred term [21]; the UMLS also lists it as a narrower concept of the term "multiple system atrophy" [8]. In the UMLS, the relationship between "multiple system atrophy" and "multisystem degeneration of autonomic nervous system" is vaguely and imprecisely defined as an "RO" relationship type. RO relationships are described as "other than synonymous, narrower, or broader," however, in this case, the RO determination in the UMLS lacks the relationship attribute that is normally included [8]. The phecodeX map, the term mapping table used for the CKM collaborator's PheWAS, matches "multisystem atrophy" to the phecode "multi-system degeneration of the autonomic nervous system" [22]. Given the evolution of the Shy-Drager syndrome terminology, the UMLS, and the phecodeX mapping, we subsequently considered all 3 of the Word Cloud concepts as a group of related terms; the highest ratio within the group was then used to rank order the condition flowchart.

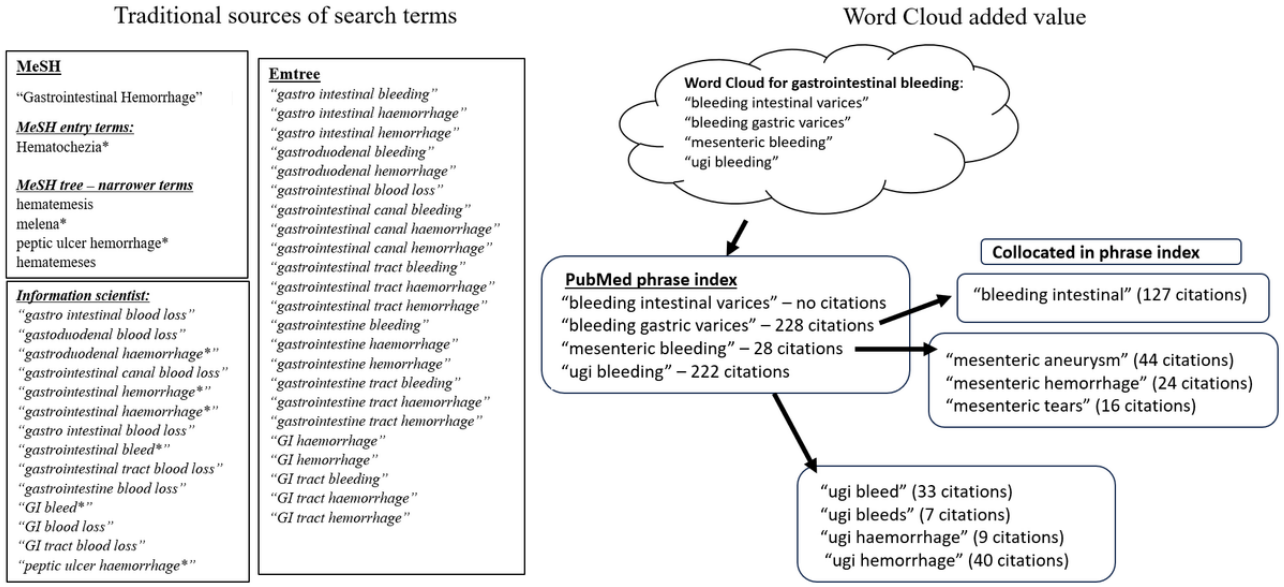
Through the combined processes of documenting actual-to-expected case ratios of the Word Cloud's relevant UMLS concepts, excluding nonrelevant semantic types, and grouping related concepts, our team creates rank-ordered lists of disease causes, risk factors, and complications reflecting our medical center's real-world clinical data.

Substantiating Disease-Condition Associations With Evidence

Providing primary literature to substantiate the associations on the condition flowchart is a key component of our research collaboration. CKM information scientists derive synonyms from the Word Cloud to strengthen the search strategy. When conducting a search, they first compile controlled vocabulary and synonyms from Medical Subject Headings and Emtree [23,24]. Next, they brainstorm additional permutations and extract terms from a scan of the literature; these keywords are subsequently checked for inclusion in the PubMed phrase index.

Figure 2 shows an example of this process for the UMLS concept "gastrointestinal bleeding." Consulting the Word Cloud identified 4 phrases that were not in the initial list of search strategy terms; 3 were found in the PubMed phrase index. Additional terminology was found by scanning collocated terms in the phrase index.

Figure 2. Word Cloud concepts leading to supplemental terminology for a search strategy on gastrointestinal hemorrhage. An asterisk denotes the truncation of a term or phrase to capture permutations. GI: gastrointestinal; MeSH: Medical Subject Headings; UGI: upper gastrointestinal.



In addition to aiding with identifying terminology and concepts to build upon our search strategies, we increasingly realize the importance of the Word Cloud's actual-to-expected patient case ratio for locating appropriate evidence. When creating the condition flowchart, our team may encounter associations for which it is difficult to locate substantiating evidence. In these cases, a Word Cloud ratio that is nonexistent, or lower than 15, can aid in validating literature scarcity. For example, snake bites can lead to nonseptic distributive shock [25]. In the Word Cloud, the association between snake bite and distributive shock has a low ratio of 5.38. Substantiating evidence for the association was found only in case reports and case series (ie, studies with few patients). Similarly, searching for literature to support hypertrophic cardiomyopathy as a cause of obstructive shock yielded only case reports as the best available evidence. A review of the Word Cloud UMLS concepts revealed a ratio of 8.7. In these instances, the evidence may still be used, but the low ranking, due to the low ratio, aids in understanding the strength of association when compared with other causes, risk factors, and complications listed on the condition flowchart.

Conclusions

This viewpoint paper describes a novel use of an institution's AI-driven, large-scale aggregation of condition-specific patient data extracted from free-text clinical documents. The Word Cloud NLP system can inform and guide knowledge generation processes by enhancing our ability to represent, substantiate, and prioritize condition associations for use in PheWAS interpretation.

The VUMC Word Cloud NLP is a valuable resource that provides real-time concept extraction from all clinical documentation and makes the resulting data viewable interactively, available for real-time decision support and alerting, and available as a rich source of coded data for research. An important limitation, however, is that this type of resource would be expensive and difficult to port directly to other institutions, thus limiting its generalizability. The emergence of generative AI, and in particular large language models, makes it conceivable that some of these limitations might be reduced in the near future; for example, large language models might be used to perform a significant portion of the concept extraction task, turning clinical free text into sets of terms which might then be mapped to coded terminologies (such as the UMLS). This possibility is still largely hypothetical and will need to be investigated to evaluate whether it is feasible, performant, and economically viable.

It is also worth noting that in addition to the population-level analysis features offered by the Word Cloud as described in our research collaboration for PheWAS analysis, the CKM also uses its capability of providing summary views of individual patient charts for other projects, such as our synthesized evidence provision services [14,26]. In response to providers' complex clinical questions, information scientists consult the visual display of the Word Cloud to gain a holistic understanding of each patient's comorbidities, medications, and other prominent clinical history. This greatly facilitates our ability to generate tailored syntheses of the published evidence that are personalized to each specific patient case [26]. Additional applications of the Word Cloud and other AI tools are also under exploration at our center, including the use of AI for scaling the maintenance of evidence syntheses over time [27-29]. Through both of these approaches—leveraging the Word Cloud NLP for population-level concept analysis and individual patient-level assessment—the CKM achieves the rapid knowledge acquisition strategy critical for informing clinical health care and research at our institution.

Acknowledgments

This project did not receive any specific external funding.

Authors' Contributions

DAG and NBG wholly developed the work's concept and design. TYK, DAG, AMW, and PAK contributed to the conduct of the case study methods. TYK, DAG, AMW, MNB, PAK, JS, and NBG participated in the writing, editing, and critical review of this paper. AMW and MNB helped visualize the case study details. NBG provided oversight of case study activities.

Conflicts of Interest

None declared.

References

1. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med 2023;155:106649. [doi: 10.1016/j.compbiomed.2023.106649] [Medline: 36805219]
2. Robinson PN, Haendel MA. Ontologies, knowledge representation, and machine learning for translational research: recent contributions. Yearb Med Inform 2020;29(1):159-162 [FREE Full text] [doi: 10.1055/s-0040-1701991] [Medline: 32823310]
3. Giuse DA. Supporting communication in an integrated patient record system. AMIA Annu Symp Proc 2003;2003:1065 [FREE Full text] [Medline: 14728568]
4. Johnson KB, Ehrenfeld JM. An EPIC switch: preparing for an electronic health record transition at Vanderbilt University Medical Center. J Med Syst 2017;42(1):6 [FREE Full text] [doi: 10.1007/s10916-017-0865-6] [Medline: 29164347]
5. Madani S, Giuse D, McLemore M, Weitkamp A. Augmenting NLP results by leveraging SNOMED CT relationships for identification of implantable cardiac devices from patient notes. : SNOMED International; 2019 Presented at: SNOMED CT Expo; Oct 31-Nov 1, 2019; Kuala Lumpur, Malaysia URL: http://tinyurl.com/5awcneyr
6. Tan H, Giuse D, Kumah-Crystal Y. Usability of a word cloud visualization of the problem list. Washington, DC: American Medical Informatics Association; 2020 Presented at: AMIA Clinical Informatics Conference; May 19-21, 2020; Virtual URL: https://brand.amia.org/m/1b1f63ea67b0c099/original/CIC2020-Visual-Abstract-Collection-FINAL.pdf
7. Krause KJ, Shelley J, Becker A, Walsh C. Exploring risk factors in suicidal ideation and attempt concept cooccurrence networks. AMIA Annu Symp Proc 2023;2022:644-652 [FREE Full text] [Medline: 37128429]
8. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
9. Apache cTAKES. Apache Software Foundation. URL: https://ctakes.apache.org/ [accessed 2023-12-13]
10. MetaMap speed. National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/Speed.pdf [accessed 2023-12-13]
11. Albert D, Weitkamp AO, Giuse D, Wright A. Building a pipeline for clinical alerts generated via natural language processing. 2022 Presented at: American Medical Informatics Association Annual Symposium; 2022 Nov; Washington, DC URL: https://knowledge.amia.org/event-data/esearch?q=Building%20a%20pipeline%20for%20clinical%20alerts%20generated%20via%20natural%20language%20processing&size=n_20_n
12. Giuse NB, Kusnoor SV, Koonce TY, Ryland CR, Walden RR, Naylor HM, et al. Strategically aligning a mandala of competencies to advance a transformative vision. J Med Libr Assoc 2013;101(4):261-267 [FREE Full text] [doi: 10.3163/1536-5050.101.4.007] [Medline: 24163597]
13. Giuse NB, Koonce TY, Jerome RN, Cahall M, Sathe NA, Williams A. Evolution of a mature clinical informationist model. J Am Med Inform Assoc 2005;12(3):249-255 [FREE Full text] [doi: 10.1197/jamia.M1726] [Medline: 15684125]
14. Blasingame MN, Williams AM, Su J, Naylor HM, Koonce TY, Epelbaum MI, et al. Bench to bedside: detailing the catalytic roles of fully integrated information scientists. Mount Laurel, NJ: Special Libraries Association; 2019 Presented at: Special Libraries Association Annual Meeting; June 18, 2019; Cleveland, OH URL: https://www.sla.org/learn-2/research/sla-contributed-papers/2019-contributed-papers/
15. Giuse NB, Williams AM, Giuse DA. Integrating best evidence into patient care: a process facilitated by a seamless integration with informatics tools. J Med Libr Assoc 2010;98(3):220-222 [FREE Full text] [doi: 10.3163/1536-5050.98.3.009] [Medline: 20648255]
16. Stead WW, Lewis A, Giuse NB, Koonce TY, Bastarache L. Knowledgebase strategies to aid interpretation of clinical correlation research. J Am Med Inform Assoc 2023;30(7):1257-1265. [doi: 10.1093/jamia/ocad078] [Medline: 37164621]
17. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26(9):1205-1210 [FREE Full text] [doi: 10.1093/bioinformatics/btq126] [Medline: 20335276]

18. Bastarache L, Denny JC, Roden DM. Phenome-wide association studies. JAMA 2022;327(1):75-76 [FREE Full text] [doi: 10.1001/jama.2021.20356] [Medline: 34982132]
19. Bland M. An Introduction to Medical Statistics, 4th Edition. Oxford, UK: Oxford University Press; 2015.
20. Semantic network. UMLS® Reference Manual. Bethesda, MD: National Library of Medicine; 2009. URL: https://www.ncbi.nlm.nih.gov/books/NBK9679/ [accessed 2023-12-13]
21. Fecek C, Nagalli S. Shy-Drager Syndrome. Treasure Island, FL: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK560502/ [accessed 2023-12-13]
22. PhecodeX (Extended). PheWAS Resources. URL: https://phewascatalog.org/phecode_x [accessed 2023-12-13]
23. Medical Subject Headings (MeSH). National Library of Medicine. 2023. URL: https://www.nlm.nih.gov/mesh/meshhome.html [accessed 2023-12-13]
24. Emtree. Elsevier. URL: https://www.elsevier.com/products/embase/emtree [accessed 2024-01-10]
25. Gaieski DF, Mikkelsen ME. Definition, classification, etiology, and pathophysiology of shock in adults. UpToDate. 2023. URL: https://www.uptodate.com/contents/definition-classification-etiology-and-pathophysiology-of-shock-in-adults [accessed 2023-12-13]
26. Koonce TY, Giuse DA, Blasingame MN, Su J, Williams AM, Biggerstaff PL, et al. The personalization of evidence: using intelligent datasets to inform the process. Washington, DC: American Medical Informatics Association; 2020 Presented at: AMIA Fall Symposium; November 2020; Virtual.
27. Koonce TY, Blasingame MN, Williams AW, Clark JD, DesAutels SJ, Giuse DA, et al. Building a scalable knowledge management approach to support evidence provision for precision medicine. Washington, DC: American Medical Informatics Association; 2022 Presented at: AMIA Informatics Summit; March 2022; Chicago, IL.
28. Blasingame MN, Su J, Zhao J, Clark JD, Koonce TY, Giuse NB. Using a semi-automated approach to update clinical genomics evidence summaries. Chicago, IL: Medical Library Association; 2023 Presented at: Medical Library Association and Special Libraries Association Annual Meeting; May 2023; Detroit, MI.
29. Su J, Blasingame MN, Zhao J, Clark JD, Koonce TY, Giuse NB. Using a performance comparison to evaluate four distinct AI-assisted citation screening tools. Chicago, IL: Medical Library Association; 2024 Presented at: Medical Library Association Annual Meeting; May 2024; Portland, OR.

Abbreviations

AI: artificial intelligence
CKM: Center for Knowledge Management
EHR: electronic health record
NLP: natural language processing
PheWAS: phenome-wide association study
UMLS: Unified Medical Language System
VUMC: Vanderbilt University Medical Center

Edited by G Eysenbach, C Lovis; submitted 10.10.23; peer-reviewed by P Han, D Chrimes; comments to author 08.12.23; revised version received 15.12.23; accepted 04.01.24; published 30.01.24.

Please cite as:
Koonce TY, Giuse DA, Williams AM, Blasingame MN, Krump PA, Su J, Giuse NB
Using a Natural Language Processing Approach to Support Rapid Knowledge Acquisition
JMIR Med Inform 2024;12:e53516
URL: https://medinform.jmir.org/2024/1/e53516
doi:10.2196/53516
PMID:38289670

Viewpoint

The Current Status and Promotional Strategies for Cloud Migration of Hospital Information Systems in China: Strengths, Weaknesses, Opportunities, and Threats Analysis

Jian Xu[1], MSc, MPH

Department of Health Policy, Beijing Municipal Health Big Data and Policy Research Center, Beijing, China

Corresponding Author:
Jian Xu, MSc, MPH
Department of Health Policy
Beijing Municipal Health Big Data and Policy Research Center
Building 1, Number 6 Daji Street
Tongzhou District
Beijing, 101160
China
Phone: 86 01055532146
Email: xujian@163.com

Abstract

Background: In the 21st century, Chinese hospitals have witnessed innovative medical business models, such as online diagnosis and treatment, cross-regional multidepartment consultation, and real-time sharing of medical test results, that surpass traditional hospital information systems (HISs). The introduction of cloud computing provides an excellent opportunity for hospitals to address these challenges. However, there is currently no comprehensive research assessing the cloud migration of HISs in China. This lack may hinder the widespread adoption and secure implementation of cloud computing in hospitals.

Objective: The objective of this study is to comprehensively assess external and internal factors influencing the cloud migration of HISs in China and propose promotional strategies.

Methods: Academic articles from January 1, 2007, to February 21, 2023, on the topic were searched in PubMed and HuiyiMd databases, and relevant documents such as national policy documents, white papers, and survey reports were collected from authoritative sources for analysis. A systematic assessment of factors influencing cloud migration of HISs in China was conducted by combining a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis and literature review methods. Then, various promotional strategies based on different combinations of external and internal factors were proposed.

Results: After conducting a thorough search and review, this study included 94 academic articles and 37 relevant documents. The analysis of these documents reveals the increasing application of and research on cloud computing in Chinese hospitals, and that it has expanded to 22 disciplinary domains. However, more than half (n=49, 52%) of the documents primarily focused on task-specific cloud-based systems in hospitals, while only 22% (n=21 articles) discussed integrated cloud platforms shared across the entire hospital, medical alliance, or region. The SWOT analysis showed that cloud computing adoption in Chinese hospitals benefits from policy support, capital investment, and social demand for new technology. However, it also faces threats like loss of digital sovereignty, supplier competition, cyber risks, and insufficient supervision. Factors driving cloud migration for HISs include medical big data analytics and use, interdisciplinary collaboration, health-centered medical service provision, and successful cases. Barriers include system complexity, security threats, lack of strategic planning and resource allocation, relevant personnel shortages, and inadequate investment. This study proposes 4 promotional strategies: encouraging more hospitals to migrate, enhancing hospitals' capabilities for migration, establishing a provincial-level unified medical hybrid multi-cloud platform, strengthening legal frameworks, and providing robust technical support.

Conclusions: Cloud computing is an innovative technology that has gained significant attention from both the Chinese government and the global community. In order to effectively support the rapid growth of a novel, health-centered medical industry, it is imperative for Chinese health authorities and hospitals to seize this opportunity by implementing comprehensive strategies aimed at encouraging hospitals to migrate their HISs to the cloud.

(JMIR Med Inform 2024;12:e52080) doi:10.2196/52080

KEYWORDS

hospital information system; HIS; cloud computing; cloud migration; Strengths, Weaknesses, Opportunities, and Threats analysis

Introduction

In the 21st century, innovative business models have emerged in Chinese hospitals, such as online diagnosis and treatment, cross-regional multidepartment consultation, real-time sharing of medical test results, and continuous public health surveillance. However, most hospitals still rely on traditional hospital information systems (HISs) designed for in-hospital management that are inadequate to support the development of these new business models [1]. Cloud computing has emerged as a promising global information technology recognized as a new infrastructure for future economic growth [2,3]. Since 2010, it has also been prioritized by the Chinese government as a "national strategic emerging industry" [4]. The adoption of cloud computing technologies can significantly reduce hospitals' costs associated with system construction and maintenance [5], expand medical services to partner institutions or patients outside the hospital [6], provide more secure network protection than self-built data centers [7], and facilitate large-scale collection and analysis of clinical data essential for scientific clinical decision-making [8]. Based on these advantages, there has been a surge in China's medical cloud service market and application research in recent years [9].

However, despite the increased attention given to cloud computing in various disciplinary domains such as disease monitoring, health surveillance, and clinical diagnosis, there is a lack of research on the cloud migration of HISs. A comprehensive review of the PubMed and HuiyiMd databases only yielded 3 relevant studies: an Iranian study that identified key driving factors for hospitals adopting cloud computing [10], a Greek study that proposed a method for migrating clinical and laboratory data based on local hospital conditions [11], and an American study that focused on essential considerations for chief financial officers before venturing into the cloud [12]. However, none of these studies have adequately addressed the aforementioned issue. Without conducting prior assessments, hospitals may struggle to fully comprehend the external environment, internal conditions, and potential opportunities and risks, thus failing to ensure prudent decision-making. Blindly following trends could pose significant threats to the security, operational efficiency, and maintenance costs of already deployed cloud-based information systems and existing hospital networks [13]. Therefore, this study aims to systematically assess factors influencing the cloud migration of HISs in China, identify associated challenges, and propose corresponding strategies for advancement. It can assist hospitals in gaining a comprehensive understanding of this work while safely implementing their cloud-based medical services. Additionally, it serves as a foundation for formulating policies aligned with Chinese hospital informatization development in the new era by health authorities while being referenced by other countries or regions facing similar challenges.

Methods

Information Sources

The primary data source for this study was obtained from literature databases to understand the practical applications of cloud technology in Chinese hospitals. The articles published between January 1, 2007, and February 21, 2023, were selected from MEDLINE (accessible through PubMed) and HuiyiMd (accessible through the Huiyi Medical Literature Express Service System).

In order to overcome the inherent limitations of academic articles, this study augmented a wealth of pertinent internal and external environmental information by extensively consulting authoritative sources such as government agencies, industry organizations, academic institutions, and market research companies. These sources of information included national policies, action plans, white papers, implementation guidelines, survey reports, and statistical data from the past 10 years.

Search Strategies

The search strategy for PubMed: (((cloud [Title/Abstract]) OR (cloud-based [Title/Abstract])) AND (hospital [Title/Abstract]) AND ("2007/01/01" [Date-Publication]: "2023/02/21" [Date-Publication])). The search strategy for HuiyiMd: ([TI] (cloud AND hospital) OR [AB] (cloud AND hospital) OR [MH] (cloud AND hospital)) AND ([PY]>=2007). The search strategy for authoritative sources involves entering the keywords "hospital" AND "cloud" in the search box on websites.

Inclusion and Exclusion Criteria

Based on specified inclusion and exclusion criteria (Textbox 1), irrelevant articles or those covering the same topic from the same institution were excluded. Subsequently, an Excel (Microsoft) spreadsheet (Multimedia Appendix 1) and a reference list (Multimedia Appendix 2 [1,2,6,8,14-26]) were generated for literature review and Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis.

Textbox 1. Inclusion and exclusion criteria for literature review.

Inclusion criteria

- Article type: fully retrievable
- Language: English, Chinese
- Nationality of the first author: Chinese (including Hong Kong and Taiwan)
- Article topic: the research, development, and application of cloud technology in Chinese hospitals
- Publication date: from January 1, 2007, to February 21, 2023

Exclusion criteria

- Article type: nonretrievable
- Language: other languages
- Nationality of the first author: other countries
- Article topic: other topics
- Publication date: before January 1, 2007; after February 21, 2023

Information Extraction

The accessible articles were assessed based on the following criteria: title, authors, first author, first author affiliation, publication year, journal name, digital object identifier (DOI), PubMed unique identifier (PMID), first author nationality, abstract, and conclusion. Furthermore, the positive and negative impacts, research methods, disciplinary domains, cloud service models, and institutional affiliations were taken into account for further in-depth analysis purposes. The findings were documented and statistically analyzed in Excel.

Analysis Methods

The SWOT analysis is a systematic assessment of strengths (S), weaknesses (W), opportunities (O), threats (T), and other factors that influence a specific topic, objectively describing the current situation of an organization or enterprise and formulating corresponding strategies [14]. It is widely used in strategic decision-making and competitor analysis within organizations or businesses due to its ability to simplify complex problems into essential issues, enabling more focused problem-solving. This study uses the SWOT method to assess the factors impacting China's cloud migration of HISs and proposes promotional strategies.

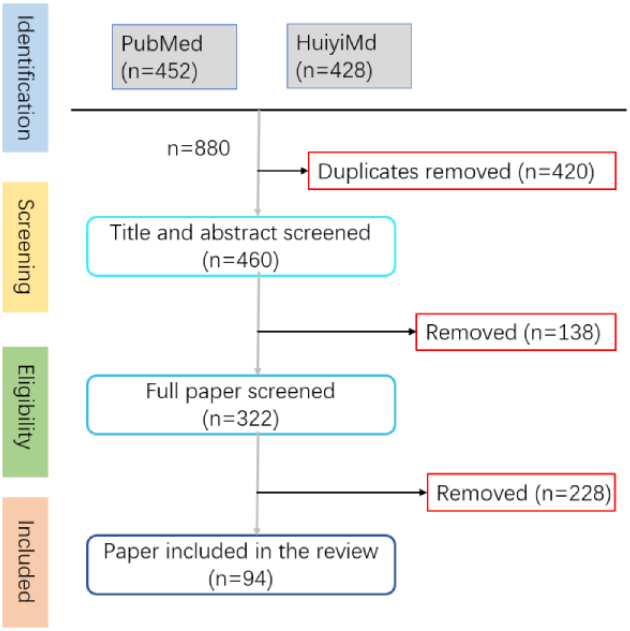
Results

Literature Review

Identification Process

The identification process in this study consists of four steps (Figure 1): (1) A total of 880 articles were retrieved from PubMed and HuiyiMd databases. (2) The search results were amalgamated, resulting in 460 deduplicated articles. (3) Screening the titles and abstracts eliminated 138 irrelevant articles based on the exclusion criteria. (4) The full text of the remaining articles was meticulously examined against predefined inclusion and exclusion criteria, resulting in a final selection of 94 relevant articles.

Figure 1. Flow diagram for the identification process.

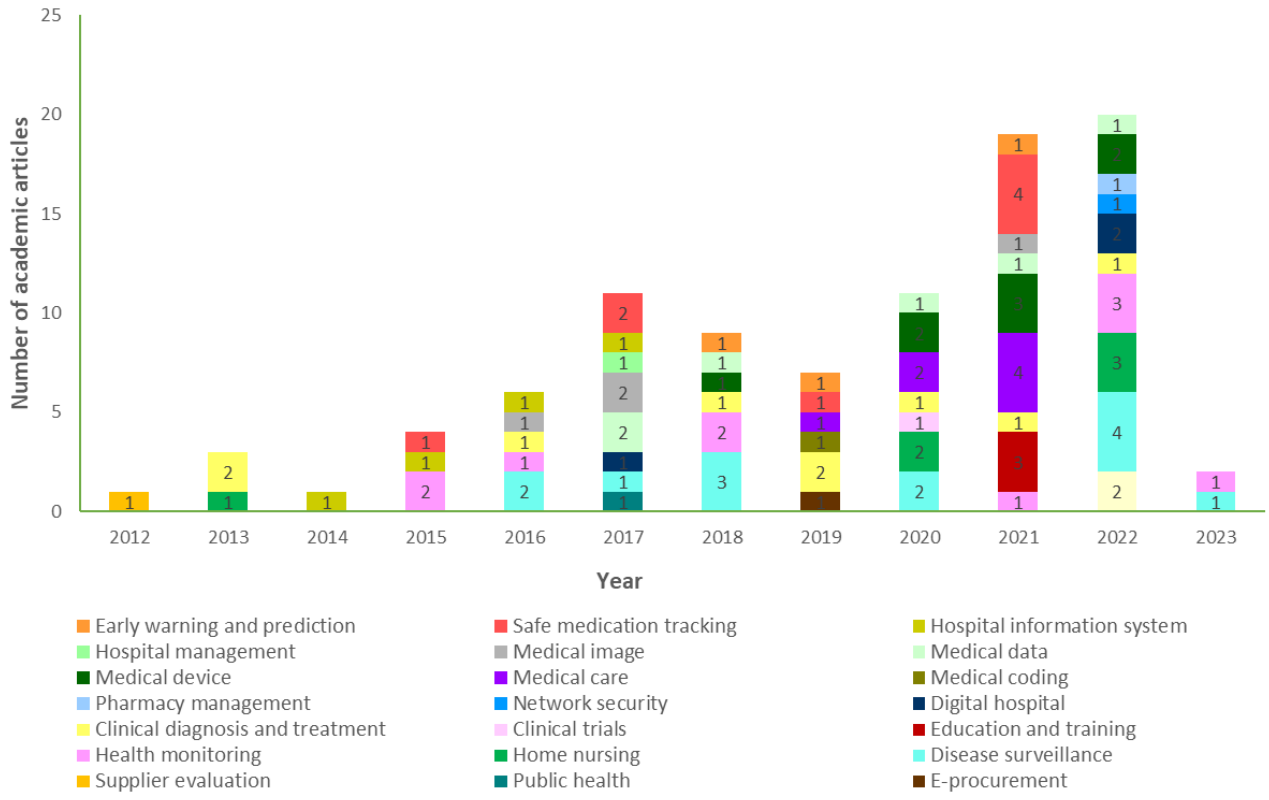


Comprehensive Description of the Literature

Research and Application of Cloud Technology in Hospitals Has Grown Rapidly and Continuously Expanded in Disciplinary Domains

In terms of time line, there was a gradual step-like increase in the number of articles starting in 2012 and reaching its peak at 20 articles in 2022. The compound annual growth rate (CAGR) was approximately 35%, highlighting the escalating quantity of research into and application of cloud technology in hospitals, as shown in Figure 2. The number of disciplines involved has increased from 1 in 2012 to 10 in 2022, encompassing a total of 22 domains. Specifically, research and application are predominantly observed in the domains of disease monitoring, health surveillance, clinical diagnosis and treatment, safe medication tracking, and medical devices, constituting 51% (48/94) of the overall distribution.

Figure 2. Time distribution of disciplinary domains involved in academic articles.



Implementation of Cloud Technology Can Yield Favorable Outcomes for Hospitals to a Certain Extent

The analysis of 94 articles identified 3 categories and 9 research methods (Table 1). The "technology" category was the most prevalent, with 47 (50%) articles focusing on information systems, cloud platforms, and associated technologies. The "experience" category followed closely, with 40 (43%) articles, primarily validating the performance of or applying cloud-based information systems through empirical research, case-control studies, experience sharing, and cohort studies. Finally, the "literature" category consisted of only 7 literature reviews on this subject matter. The consistent findings of these studies demonstrate the implementation of cloud technology in hospitals can yield positive impact to some extent, such as enhancing precision in management practices, expanding disease monitoring capabilities, reducing workload for medical personnel, and providing convenient and cost-effective health care services for patients. However, 5% (n=5) of the articles also acknowledged certain negative impacts, such as underdevelopment of digital methods in hospitals, cybersecurity risks, and low satisfaction rates among physicians and community pharmacists.

Table 1. The correlation between research methods used in academic articles and the institutional affiliations of their first authors.

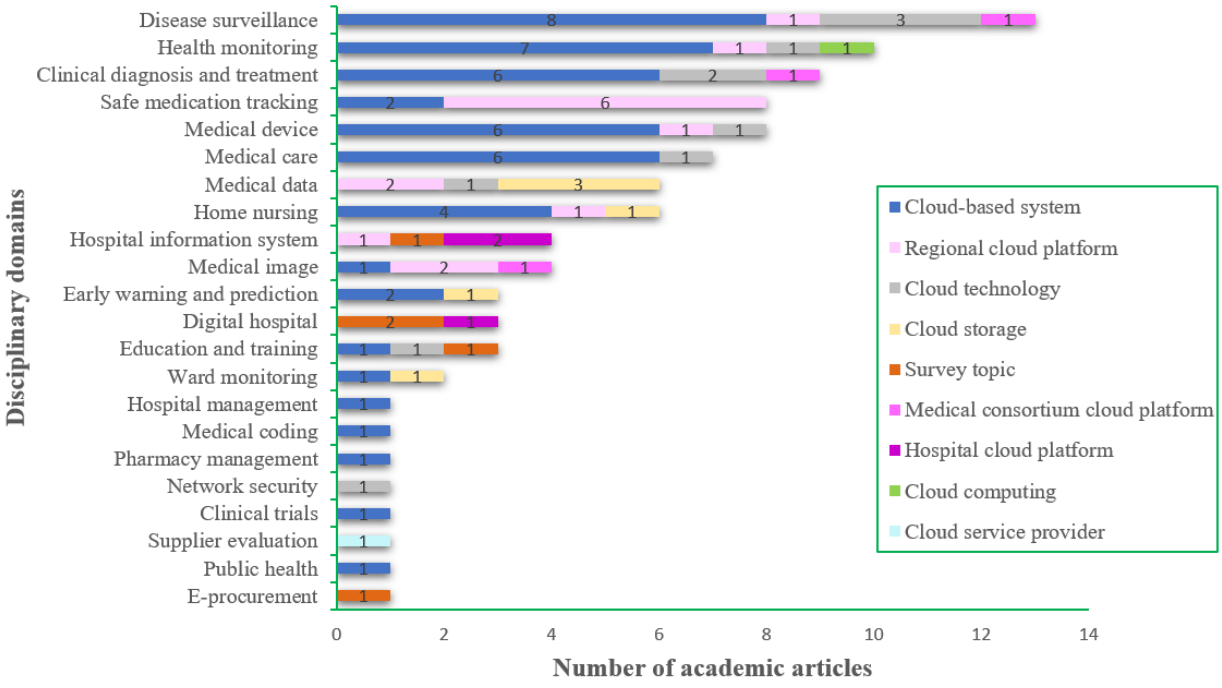
Research methods	Hospitals, n (%)	Universities or colleges, n (%)	Associations or companies, n (%)
Technology			
System research and development	13 (14)	14 (15)	N/A[a]
Cloud platform construction	5 (5)	5 (5)	N/A
Technical research	2 (2)	8 (9)	N/A
Experience			
Empirical research	14 (15)	7 (7)	1 (1)
Case control study	11 (12)	2 (2)	N/A
Summary of experience	3 (3)	N/A	N/A
Cohort study	2 (2)	N/A	N/A
Literature			
Retrospective study	4 (4)	2 (2)	N/A
Standard study	N/A	1 (1)	N/A

[a]N/A: not applicable.

More Than Half of the Studies Focused on Task-Specific Cloud-Based Systems, While Only 1 in 5 Addressed Integrated Cloud Platforms

Out of the 94 articles analyzed, the majority (n=49, 52%) focused on task-specific cloud-based systems in hospitals. In contrast, only 21 (22%) articles discussed or developed integrated cloud platforms for sharing within a region, medical alliance, or hospital. Furthermore, as shown in Figure 3, 67% (n=33) of task-specific cloud-based systems were used in patient-related domains such as disease monitoring, health surveillance, clinical diagnosis and treatment, medical care, and medical devices. A total of 15 regional cloud platforms (16%) were commonly used for safe medication tracking, data storage, and medical imaging. A total of 3 medical alliance cloud platforms (3%) found use in disease treatment-related domains such as disease monitoring, clinical diagnosis, and treatment along with medical imaging. A total of 3 hospital cloud platforms (3%) primarily originated from digital hospitals or HIS upgrades.

Figure 3. The cross-distribution of cloud service models and disciplinary domains covered in academic articles.



More Supplementary Materials Were Collected to Support the SWOT Analysis for This Study

Because the literature review provided insufficient information to support the analysis of internal and external factors for this study, supplementary materials were collected, including national policies, action plans, white papers, implementation guidelines, research reports, and statistical reports from authoritative websites, such as government agencies, industry organizations, academic institutions, and market research companies. In total, 37 supplementary documents were included in the SWOT analysis: 4 policy papers, 11 industry reports, 15 academic articles, 2 dissertations, 2 official bulletins, and 3 news articles. All of them were recorded in an Excel file (Multimedia Appendix 2).

SWOT Analysis

External Opportunities (O)

Politics: Governments Worldwide Prioritize Cloud Technology and Have Implemented Supportive Policies

The United States introduced the Cloud First policy [2] and the CLOUD Act [15]. Similarly, the European Union aims for digital sovereignty through initiatives like the Gaia-X Association and the EU Cloud Computing Strategy [27]. China also prioritized cloud computing as a "national strategic emerging industry" in 2010 and implemented policies to promote its adoption by the government and businesses [28]. Consequently, China has rapidly developed in this field and ranks among global leaders [29].

Economy: Both Nations and Enterprises Have Made Substantial Capital Investments to Foster Its Development

According to Gartner's data from 2011 to 2022, the market scale increased from US $95.24 billion to US $491 billion with a CAGR exceeding 16% [30]. The global medical cloud computing market reached US $39.4 billion, with Asia-Pacific exhibiting the fastest growth rate at 22% per year; China and India are significant contributors to this expansion [9]. In China, the market size has surged from less than US $270 million in 2011 to US $66.91 billion in 2022, with a consistent CAGR surpassing 40%, and will exceed US $150 billion by 2025 [30].

Social: Online Health Care Has Become a Norm in Modern Life

According to the National Telemedicine and Connected Healthcare Center of China, as of June 2021, there were 239 million users accessing health care services online and more than 1600 internet-based hospitals in China [31]. Another survey revealed that approximately 63% (n=465) of the surveyed hospitals (738 hospitals across 30 provinces) used cloud services to some extent in 2022 [32]. Moreover, the COVID-19 pandemic has further fueled the demand for online health care services [6]. Consequently, these services have become as commonplace in our lives as online shopping.

Older Adult Care: The Older Adult Care Industry Urgently Needs Advanced Technological Support

By the end of November 2020, China's population of people aged 60 years and older was 264,018,766, accounting for 19% of the total population, making it the country with both the largest older population and the fastest-aging society worldwide [33]. Consequently, relying solely on their children, nursing homes, or communities to provide older adult care services has become increasingly impractical. Recognizing this challenge, the National Health Commission of China issued a document in October 2019 emphasizing the need to fully harness modern technologies such as cloud computing, artificial intelligence (AI), and the Internet of Things to develop an intelligent service model known as "Internet plus Healthy Aging" [34].

External Threats (T)

Market: Technological Monopolies Pose a Significant Threat to the Digital Autonomy of Nations

A total of 81 (81%) of the Forbes Top 100 cloud computing companies are American [35], and they possess significant technological and capital advantages. They continuously expand their global market share; they captured 60%-70% (JPY ¥1,725-¥2,012 billion [US $11.6-$13.6 million]) of the Japanese cloud market in 2020 [36] and 69% of the European cloud market in 2021 [37]. The passage of the Clarifying Lawful Use of Data Abroad Act (CLOUD Act) by the US Congress in March 2018 caused European countries to feel threatened due to the potential loss of digital sovereignty [15], which prompted them to initiate their "European cloud" project in 2020 [27]. Other nations may face similar challenges.

Competition: Fierce Competition May Lead to Uncertain Levels of Service Quality

In November 2021, 5 bidders for a public cloud service project in Shijiazhuang City, China, submitted bids of CNY ¥0, sparking concerns among stakeholders [38]. The intense competition may lead to unpredictable service performance issues for users, such as limiting hospitals' access to better pricing and a wider range of choices by restricting the interoperability and portability of HISs or causing sudden disruptions in cloud-based medical services after winning the bid, which poses significant risks to patient safety [16].

Security: Hospitals Express Apprehensions Regarding Diverse Cyber-Attacks Targeting Cloud Infrastructure

Security is the primary challenge for cloud-based systems due to various cyberattacks faced by current cloud environments, especially in the health care sector where sensitive data such as personal privacy, health records, diagnostics, and treatments are stored [13]. Even prominent cloud providers like Azure (Microsoft), Docker Hub (Docker), and Everis (NTT DATA) have experienced malicious intrusions [30], while both the United Kingdom's National Health Service (NHS) in 2017 and Ireland's Department of Health information system in 2021 were both targeted and resulted in a complete paralysis [39,40].

Legislation: The Lack of Precise Legislation Hinders the Efficient Implementation and Enforcement of Regulatory Measures

To support the implementation of the "Cloud Normal" and "Internet Plus" strategies, the Chinese government has enacted laws, regulations, and management measures. However, there are limited directly applicable legal provisions for cloud migration of HISs. Imperfect laws and regulations, insufficient safety standards, unclear legal liabilities, and the absence of a damage assessment mechanism hinder the proper development of cloud services. As a result, doctors and patients may encounter challenges in protecting their rights during disputes [17].

Internal Strengths (S)

Data: Hospitals Generate Substantial Volumes of Medical Data on a Daily Basis

Hospitals are natural suppliers of big data. For instance, the Chinese National Cloud-Based Telepathology System (CNCTPS) has collected 23,167 cases and served 9240 users in 4 years (2016-2019), providing comprehensive details from whole-slide images to diagnostic reports [5]. Additionally, medical big data can provide substantial value to both hospitals and patients. For example, the aforementioned CNCTPS application can save patients around US $300,000 per year [5]. Abundant electronic health and care records in the United Kingdom's NHS can reduce hospital operational costs by approximately £5 billion (US $ 3.9 billion) annually and save patient welfare expenses by roughly £4.6 billion (US $3.6 billion) [41]. Furthermore, traditional computing tools are unable to handle the processing and analysis of such massive amounts of data—this is exactly where cloud computing technology excels [18].

Business: The Provision of Comprehensive Medical Services Necessitates Extensive Interdisciplinary Collaboration

Medical services are complex and innovative, requiring synchronization of knowledge, technology, experience, and resources from diverse disciplines. Cloud computing provides extensive connectivity, offering robust support for these tasks, including interdisciplinary expert consultations, collaborative surgeries, and integrating medicine and care [19]. The Huashan Hospital, affiliated with Fudan University, uses a medical consortium cloud platform where experts from higher-level hospitals offer diagnostic advice to lower-level hospitals for subsequent care and daily treatment, ensuring positive outcomes for patients with epilepsy [42].

Application: Multiple Cloud Technology Applications Have Been Effectively Implemented Across Various Medical Domains

As shown in Figure 2, cloud technology is receiving increasingly extensive research and application in the medical field, and even some regional or medical alliances have constructed their own medical cloud platforms to store health data, share medical images, and facilitate collaboration. Furthermore, a national survey conducted in 2022 also confirmed these findings by revealing that out of the 738 surveyed hospitals, 63% (n=465) partially used cloud services across nearly 20 different medical business scenarios [32]. These effective practices can serve as valuable references and support for other hospitals yet to implement such initiatives.

Demand: The Provision of Health-Centered Medical Services Necessitates Advanced Technological Support

The transition from disease-centered to health-centered hospital development in the new era has rendered traditional HISs increasingly inadequate as they were previously designed solely for managing information within hospitals. Cloud computing can significantly expand hospitals' medical services beyond their physical premises, enabling online chronic disease management, individual life-cycle health surveillance, and remote diagnosis for patients in remote areas. This enhancement

empowers hospitals to provide health-centered medical services [20]. The findings of this study also strongly support this notion. As depicted in Figure 3, cloud technology has been extensively used in closely associated domains with patients, encompassing disease monitoring, health surveillance, clinical diagnosis and treatment, and safe medication tracking.

Internal Weaknesses (W)

System: The Complexity of HISs Poses Challenges for Hospitals When Migrating Them to the Cloud

The HISs are the most complex organizational information management systems developed by various contractors in diverse environments, covering a wide range of business functions and user groups [21], as depicted in Figure 3, with only 94 articles included but spanning across 22 distinct disciplinary domains as well. Therefore, the cloud migration of HISs presents significant challenges, particularly for those systems abandoned by development companies due to insolvency or insufficient technical support. Nevertheless, if there existed an all-encompassing and authoritative medical cloud platform enabling hospitals to tailor services based on their specific requirements, it would undoubtedly expedite the overall migration process.

Security: The Security of Existing Hospital Networks Still Faces Numerous Risks

Currently, most HISs still operate in self-constructed networks instead of using cloud-based solutions, which poses information security challenges due to insufficient infrastructure, overreliance on a single protective measure, incomplete regulatory frameworks, and potential vulnerabilities from privilege abuse [22]. For example, the 2019-2020 China Hospital Informatization Survey Report revealed that around 28% (n=282) of surveyed hospitals experienced unplanned core system failures lasting more than 30 minutes [43]. To effectively address these concerns, proficient IT teams like reputable cloud vendors or organizations equipped with advanced technologies such as cloud computing should collaborate rather than solely rely on in-house hospital IT capabilities.

Plan: Strategic Planning and Resource Allocation in Hospitals Exhibit Certain Deficiencies

According to Figure 3, more than half of the research articles focused on hospital-specific systems for various tasks. These systems still adhered to traditional information system designs, had limited scalability and functionality, and operated independently. As a result, there were significant challenges in effectively using cloud computing's computing capabilities, storage capacity, and integrated analysis to generate valuable information supporting government scientific decision-making. The survey results from China's National Health Commission also confirmed this point as many internet hospitals were not fully developed yet and encountered diverse issues [44]. Moreover, only 14% (n=101) of hospitals had migrated their core business operations to the cloud with a mere 13% (n=100) planning to do so in the next 3-5 years [32]. Therefore, it is crucial for hospitals to reorganize and integrate their operations and resources before incorporating their HISs into the cloud in order to meet the demands of cloud capabilities and new health care models.

Personnel: The Allocation of Information Personnel Is Inadequate and Lacks Specialization Levels

With the increasing integration of cloud computing, AI, and robotics, hospitals urgently require highly skilled IT professionals to effectively implement these new technologies [23]. However, a national survey in 2021 revealed that the average number of information department personnel in 9376 secondary and tertiary hospitals was only 6. Most of these personnel held undergraduate or junior college computer degrees and possessed limited interdisciplinary expertise. This falls significantly below national standards [24], particularly for hospitals below grade 2 or in economically underdeveloped areas [45].

Investment: Primary Hospitals Lack Sufficient Investment in Information Technology and Cloud Services

According to a 2020 survey by the National Health Commission of China, most primary hospitals allocated less than 1% of their budgets to HIS development, facing challenges such as unstable funding and support [25]. A nationwide survey conducted in 2022 revealed that only 53% of the surveyed 738 hospitals had expenses related to public cloud services in the previous 2 years, with 54% spending less than US $14,000. Particularly for primary hospitals, establishing a cloud service system is even more financially challenging [32]. Clearly, these primary hospitals require more reliable financial guarantees for the smooth operation of their HISs and cloud services.

Discussion

Principal Findings

Extensive literature review and systematic SWOT analysis indicate that cloud computing is increasingly being applied in nearly 22 discipline domains in Chinese hospitals; it plays a crucial role in monitoring patient-related diseases, health surveillance, clinical diagnosis and treatment, and safe medication tracking. However, more than half of the research and applications are limited to cloud-based systems for specific hospital tasks, which fail to fully leverage the robust integrated analytical capabilities of cloud computing due to limited data scale and functionality that could otherwise generate valuable information supporting hospital or government decision-making processes. Additionally, challenges such as market sovereignty disputes, intense industry competition, network attacks, inadequate regulation, and hospitals' internal weaknesses like complexity of HISs, insufficient resource integration, and limited manpower and investment, hinder widespread adoption of cloud technology among most hospitals that exhibit a relatively weak willingness to migrate their core operations to the cloud within the next 3-5 years. Nevertheless, cloud computing is widely recognized as a novel infrastructure driving global economic growth. Integrating cloud technology in hospitals can enhance medical service quality, foster interdisciplinary collaboration and remote consultations, and promote coordinated development within regional health care economies. Consequently, it is imperative for hospitals and health authorities to pay special

attention to this matter and actively implement diverse strategies to facilitate its advancement. Based on the aforementioned research findings, this study proposes a set of promotional strategies for collective deliberation among peers. The overall framework depicting these proposed strategies is illustrated in Figure 4, which will be further elucidated in subsequent sections.

Figure 4. SWOT analysis and response strategies diagram for cloud migration of HISs. HIS: hospital information system; SWOT: Strengths, Weaknesses, Opportunities, and Threats.

Internal factors / External factors	Strengths (S)	Weaknesses (W)
	1. Data: Hospitals generate substantial volumes of medical data on a daily basis. 2. Business: The provision of comprehensive medical services necessitates extensive interdisciplinary collaboration. 3. Application: Multiple cloud technology applications have been effectively implemented across various medical domains. 4. Demand: The provision of health-centered medical services necessitates advanced technological support.	1. System: The complexity of HISs poses challenges for hospitals when migrating them to the cloud. 2. Security: The security of existing hospital networks still faces numerous risks. 3. Plan: The strategic planning and resource allocation in hospitals exhibit certain deficiencies. 4. Personnel: The allocation of information personnel is inadequate and lacks specialization levels. 5. Investment: Primary hospitals lack sufficient investment in information technology and cloud services.
Opportunities (O)	**OS strategy**	**OW strategy**
1. Politics: Governments worldwide prioritize cloud technology and have implemented supportive policies. 2. Economy: Both nations and enterprises have made substantial capital investments to foster its development. 3. Social: Online health care has become a norm in modern life. 4. Older adult care: The older adult care industry urgently needs advanced technological support.	**Implementing multiple initiatives to encourage more hospitals to migrate their HISs to the cloud.** 1. Taking multiple initiatives to encourage more hospitals to migrate their HISs to the cloud. 2. Guiding hospitals in safely and effectively migrating their systems to the cloud.	**Enhancing hospitals' overall capability for cloud migration of HISs.** 1. Improving the medical information literacy and capabilities of all staff in hospitals. 2. Reengineering hospital's business processes and integrating them into one organic HIS. 3. Increasing investment in and application of cloud technology in primary hospitals.
Threats (T)	**TS strategy**	**TW strategy**
1. Market: Technological monopolies pose a significant threat to the digital autonomy of nations. 2. Competition: Fierce competition may lead to uncertain levels of service quality. 3. Security: Hospitals express apprehensions regarding diverse cyber-attacks targeting cloud infrastructure. 4. Legislation: The lack of precise legislation hinders the efficient implementation and enforcement of regulatory measures.	**Establishing a provincial-level unified medical hybrid multicloud platform.** 1. Selecting the most suitable cloud deployment model for hospitals. 2. Establishing a unified medical cloud platform at the provincial or municipal level.	**Enhancing legal framework and technical support for cloud-based HISs.** 1. Establishing a solid and reliable legal foundation. 2. Providing comprehensive and efficient technical support.

Implementing Multiple Initiatives to Encourage More Hospitals to Migrate Their HISs to the Cloud

The primary objective of this strategy is to address the issue of "whether or not to adopt cloud technology." Based on the outcomes of the SWOT analysis, despite the pressing need for cloud technology to enhance health-centered medical service delivery today, hospitals remain cautious about its implementation due to external threats and internal weaknesses. Furthermore, providing cloud-based medical services has brought about a significant transformation within the medical industry that exceeds traditional independent operations and self-financing models used by hospitals. Therefore, governments should make greater efforts by implementing more active measures such as policy guidance, training planning, demonstration projects, or provision of cloud vouchers, in order to encourage more hospitals to securely migrate their HISs onto the cloud and meet demands for innovative medical services in this modern era.

Enhancing Hospitals' Overall Capability for Cloud Migration of HISs

The strategy aims to address the issue of "what preparations are necessary." As previously mentioned, cloud migration of HISs is a highly intricate system engineering project that requires hospitals to be fully prepared for its successful implementation. These preparations encompass various aspects including, but not limited to the following. First, human resources: hospitals should enhance employees' medical information literacy and application skills through comprehensive training programs, specialized lectures, or successful case studies. Second, material resources: hospitals should redesign and integrate existing systems and resources based on future development, existing foundations, and expert recommendations in order to optimize the use of cloud resources for acquiring more valuable information. Third, financial resources: hospitals require long-term financial investment planning to support the provision of cloud-based medical services. Moreover, health authorities should acknowledge that primary hospitals serve as both frontline institutions addressing medical needs and significant sources of authentic data. Therefore, moderate increases in investment in HIS construction of primary hospitals are necessary to ensure a continuous input of firsthand authentic data. Fourth, tools: a hospital that has robust IT capabilities can leverage free cloud migration consultation and tools provided by major cloud providers such as Alibaba, Tencent, Google, Microsoft, and Amazon Web Services, which can expedite the process of migrating information systems. However, it should be noted that the cloud services used (eg, computing and storage) may incur charges.

Establishing a Provincial-Level Unified Medical Hybrid Multi-Cloud Platform

The strategy aims to address the issue of "how to implement changes efficiently." In response to numerous complex internal and external situations, this study proposes a coping strategy: establishing a unified medical hybrid multi-cloud platform in each province or municipality.

First, the hybrid multi-cloud platform highly aligns with hospital operations. Hospitals require private clouds for storing sensitive and core data, nonpublic community clouds for internal consultations and other collaborations, public clouds for providing more extensive medical services to the public, and adopting a "multi-cloud" strategy to reduce risks such as vendor lock-in or declining service quality.

Second, a medical cloud platform at the provincial or municipal level can achieve maintainable security and more highly effective cloud migration. In comparison to hospitals, health authorities at this level possess stronger technological expertise, greater manpower resources, maintainable financial support, and relevant assets for constructing comprehensive platforms while effectively mitigating internal and external risks. Moreover, this approach can also foster overall advancements in cloud migration and system function quality of hospitals (particularly primary ones), as well as minimize redundant construction and maintenance expenses.

Last but most importantly, the scale of data possessed by one or several individual hospitals is insufficient to constitute true big data, limiting the opportunities for leveraging cloud computing's powerful intelligent analysis capabilities in uncovering hidden objective laws that can support the government to make scientific decisions. Considering factors such as data scale, cloud computing capabilities, and government economic support capacity, a provincial or municipal regional medical cloud platform is a more suitable choice.

Enhancing Legal Framework and Technical Support for Cloud-Based HISs

The primary goal of this strategy is to address the issue of "how to create a supportive environment." As an ancient Chinese proverb states, "A single log cannot support a crumbling building," indicating that relying solely on a provincial-level medical cloud platform is still insufficient to counter all external threats and internal risks faced by hospitals. Therefore, it requires a more proactive role from the government, which strengthens cooperation with relevant departments and enterprises to build a more robust and secure supporting environment for medical cloud platforms. Specifically, 2 aspects of support are necessary. First, credible legal support: although the Chinese government has been improving laws regarding cybersecurity, personal information protection, and data security, its support for cloud-based medical services remains inadequate. For example, in resolving disputes related to cloud medical services, health authorities still rely on laws such as the Physician Practice Law and Regulations on Prevention and Handling of Medical Disputes in China. However, these regulations have not explicitly defined the status and responsibilities of all parties involved in cloud-based medical services, which poses challenges in terms of judgments and penalties [26]. Therefore, it is essential to further refine relevant legislation and update existing regulations regarding doctors' practice rights, insurance responsibility, and reimbursement for medical insurance, ensuring doctors and patients can confidently participate in cloud-based medical services. Second, holistic technical support: as indicated by SWOT analysis results, hospitals often lack professionals with deep knowledge of cutting-edge technologies like cloud computing, AI, and big data. Therefore, establishing an organization like a medical cloud technology association becomes necessary. This organization should consist of experts from various relevant fields, including IT, communication, engineering, cryptography, medicine, and more. Their responsibilities would include devising unified long-term plans and action plans, standardizing contracts between hospitals and cloud service providers, reviewing hospitals' cloud migration plans and contracts, and conducting regular evaluations of the construction and operation of cloud-based HISs.

Conclusions

In conclusion, cloud computing is prioritized by the Chinese government as a "national strategic emerging industry." Despite encountering numerous challenges, the cloud migration of HISs in China exemplifies a prevailing development trend. Therefore, hospitals should adopt an open mindset and focus on enhancing their capabilities to develop medical services based on the cloud. Health authorities should also use more effective strategies to incentivize hospitals to migrate their HISs safely to the cloud, thereby fostering the flourishing growth of a novel health-centered medical industry.

The main contribution of this study is a comprehensive literature review and systematic SWOT analysis on the current status of cloud migration of HISs in China, and corresponding strategies for different combinations of internal and external influencing factors. It can help hospitals gain a clearer understanding of the overall situation while having more specific goals and methods when planning and implementing related work. Moreover, it can serve as a foundation for health authorities to develop policies aligned with the development of hospital informatization in the new era, as well as provide references for other countries or regions facing similar challenges.

There are 2 limitations in this study: first, not all personnel from hospitals contribute to writing papers, thus limiting the comprehensiveness of literature information; second, the SWOT analysis method assumes a distinction between internal and external factors, as well as a differentiation between strengths and weaknesses, overlooking the interrelated effects among influencing factors. To supplement and improve these aspects, more empirical investigation data are needed, along with a more rigorous analysis of the interactions among influencing factors. This will be the focus of my future research.

Data Availability

The data sets generated and analyzed during this study are not publicly available but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Academic paper database.
[XLSX File (Microsoft Excel File), 50 KB - medinform_v12i1e52080_app1.xlsx]

Multimedia Appendix 2
Supplementary database.
[XLSX File (Microsoft Excel File), 13 KB - medinform_v12i1e52080_app2.xlsx]

References

1. Zhou JY, Jiang Q, Ren J. The role and impact of cloud computing in hospital information management. Mod Hosp 2023 Mar 27;23(03):422-424 [FREE Full text] [doi: 10.3969/j.issn.1671-332X.2023.03.027]
2. Jia YW, Zhao D, Jiang KY, Luan GC. The U.S. federal government's cloud computing strategy. E-Government 2011 Jul 1(7):2-16 [FREE Full text] [doi: 10.16582/j.cnki.dzzw.2011.07.001]
3. Unleashing the potential of cloud computing in Europe. European Commission. 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52012DC0529 [accessed 2023-05-12]
4. Outline of the 14th five-year plan (2021-2025) for national economic and social development and vision 2035 of the People's Republic of China. The Government of China. 2021 Mar 12. URL: http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm [accessed 2023-05-12]
5. He XY, Wang LL, Wang L, Gao JH, Cui FF, Ma QQ, et al. Effectiveness of a cloud-based telepathology system in China: large-sample observational study. J Med Internet Res 2021;23(7):e23799 [FREE Full text] [doi: 10.2196/23799] [Medline: 34326037]
6. Gong MC, Liu L, Sun X, Yang Y, Wang S, Zhu H. Cloud-based system for effective surveillance and control of COVID-19: useful experiences from Hubei, China. J Med Internet Res 2020;22(4):e18948 [FREE Full text] [doi: 10.2196/18948] [Medline: 32287040]
7. Catteddu D, Hogben G. Cloud Computing Risk Assessment. The European Union Agency for Cybersecurity (ENISA). 2009 Nov. URL: https://www.enisa.europa.eu/publications/cloud-computing-risk-assessment [accessed 2023-05-16]
8. Wu J, Wang J, Nicholas S, Maitland E, Fan QY. Application of big data technology for COVID-19 prevention and control in China: lessons and recommendations. J Med Internet Res 2020;22(10):e21980 [FREE Full text] [doi: 10.2196/21980] [Medline: 33001836]
9. Healthcare cloud computing market. MarketsandMarkets. 2022. URL: https://www.marketsandmarkets.com/Market-Reports/cloud-computing-healthcare-market-347.html [accessed 2023-05-18]
10. Alipour J, Mehdipour Y, Karimi A, Sharifian R. Affecting factors of cloud computing adoption in public hospitals affiliated with Zahedan University of Medical Sciences: a cross-sectional study in the Southeast of Iran. Digit Health 2021;7:20552076211033428 [FREE Full text] [doi: 10.1177/20552076211033428] [Medline: 34777850]
11. Nikolopoulos M, Karampela I, Tzortzis E, Dalamaga M. Deploying cloud computing in the Greek healthcare system: a modern development proposal incorporating clinical and laboratory data. Stud Health Technol Inform 2018;251:35-38. [Medline: 29968595]
12. Rajendran J. What CFOs should know before venturing into the cloud. Healthc Financ Manage 2013;67(5):40-43. [Medline: 23678688]
13. Gao SM. Network security problems and countermeasures of hospital information system after going to the cloud. Comput Math Methods Med 2022;2022:9725741 [FREE Full text] [doi: 10.1155/2022/9725741] [Medline: 35898480]
14. Puyt R, Lie FB, De Graaf FJ, Wilderom CPM. Origins of SWOT analysis. Acad Manag Proc 2020;2020(1):17416. [doi: 10.5465/ambpp.2020.132]
15. The American Society of International Law. Congress enacts the Clarifying Lawful Overseas Use of Data (CLOUD) act, reshaping U.S. law governing cross-border access to data. Am J Int law 2018;112(3):487-493 [FREE Full text] [doi: 10.1017/ajil.2018.61]
16. Opara-Martins J, Sahandi R, Tian F. Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. J Cloud Comp 2016;5:4 [FREE Full text] [doi: 10.1186/s13677-016-0054-z]
17. Xiao Q. Research on Internet Medical Legal Regulation [Dissertation]. Shenzhen University. 2019. URL: https://cdmd.cnki.com.cn/Article/CDMD-10590-1019908748.htm [accessed 2023-10-18]
18. Berisha B, Mëziu E, Shabani I. Big data analytics in cloud computing: an overview. J Cloud Comput (Heidelb) 2022;11(1):24 [FREE Full text] [doi: 10.1186/s13677-022-00301-w] [Medline: 35966392]
19. Kong L. Application and exploration of collaborative medicine based on cloud computing. Dissertation. 2017. URL: https://d.wanfangdata.com.cn/thesis/Y3337261 [accessed 2023-08-25]

20. Zhang XG. Situation and strategy for the development of hospital informatization in the new era. China J Health Inform Manag 2018;15(4):367-372 [FREE Full text] [doi: 10.3969/j.issn.1672-5166.2018.04.01]
21. Setyonugroho W, Puspitarini AD, Kirana YC, Ardiansyah M. The complexity of the Hospital Information System (HIS) and obstacles in implementation: a mini-review. Enferm Clin 2020;30:233-235 [FREE Full text] [doi: 10.1016/j.enfcli.2020.06.053]
22. Zhang YX, Hu JP, Zhou GH, Xu XD. The development and prospect of health informatization during the 13th Five-Year Plan period. Chin J Health Inform Manag 2021 Jun 28;18(3):297-302 [FREE Full text] [doi: 10.1007/s35764-016-0106-7]
23. Zhao X, Li XH. Thoughts on the development of hospital information construction during the "14th five-year plan". Chin Hosp 2021;25(1):64-66 [FREE Full text] [doi: 10.19660/j.issn.1671-0592.2021.1.20]
24. Li HX, Xu F, Wang K. Research on the current situation of hospital informatization personnel configuration in China: a cross-sectional study. China Health Qual Manag 2022;01:4-7 [FREE Full text] [doi: 10.13912/j.cnki.chqm.2022.29.01.02]
25. Hao XN, Ma CY, Liu ZY, Zhou YC, Liu QK, Zhang S. The effects and problems on the reform of primary health informatization in China. Health Econ Res 2019;07:3-5 [FREE Full text]
26. Jiao YL. Investigation on legal status of internet hospital: From the perspective of "Ningbo cloud hospital". Chin J Health Pol 2017;10(10):72-75 [FREE Full text] [doi: 10.3969/j.issn.1674-2982.2017.10.012]
27. European digital infrastructure and data sovereignty-a policy perspective. European Institute of Innovation & Technology. 2020. URL: https://www.eitdigital.eu/fileadmin/files/2020/publications/data-sovereignty/EIT-Digital-Data-Sovereignty-Summary-Report.pdf [accessed 2023-05-18]
28. Decision on accelerating the cultivation and development of strategic emerging industries. The Government of China. 2010 Oct 18. URL: https://www.gov.cn/zwgk/2010-10/18/content_1724848.htm [accessed 2023-05-24]
29. Assessment report on the global computing index 2022-2023. Institute for Global Industry of Tsinghua University. 2023. URL: https://www.igi.tsinghua.edu.cn/info/1019/1321.htm [accessed 2023-11-16]
30. Cloud computing white paper (2023). China Academy of Information and Communications Technology (CAICT). 2023. URL: http://www.caict.ac.cn/kxyj/qwfb/bps/202307/P020230725521473129120.pdf [accessed 2023-11-16]
31. Zhang XX. The "2021 Internet Hospital Report" has been issued, incorporating analyses from 1,140 data sources and comprehensive examinations from 109 dimensions, thereby unearthing these core trends. Vbdata. 2021. URL: https://www.vbdata.cn/52404 [accessed 2023-06-01]
32. Hospital cloud service application survey report. China Hospital Information Management Association (CHIMA). 2022. URL: https://www.hit180.com/57127.html [accessed 2023-06-23]
33. Bulletin of the seventh national population census (No. 5) - age composition of the population. National Bureau of Statistics of China (CNBS). 2021. URL: http://www.stats.gov.cn/sj/tjgb/rkpcgb/qgrkpcgb/202302/t20230206_1902005.html [accessed 2023-06-02]
34. Guiding opinions on establishing and improving the healthcare system for the elderly. National Health Commission of China (CNHC). 2019. URL: http://www.nhc.gov.cn/lljks/s7785/201911/cf0ad12cb0ec4c96b87704fbbeb5bbde.shtml [accessed 2023-06-16]
35. The cloud 100. Forbes. 2023. URL: https://www.forbes.com/lists/cloud100/?sh=6d17eead7d9c [accessed 2023-09-16]
36. Report on trade practices in cloud services sector. Japan Fair Trade Commission (JFTC). 2022. URL: https://www.jftc.go.jp/en/pressreleases/yearly-2022/June/220722_2EN.pdf [accessed 2023-11-16]
37. Hardesty L. European cloud providers take hit from AWS, Google, Azure, says Synergy. Silverlinings. 2021 Sep 23. URL: https://www.silverliningsinfo.com/platforms/european-cloud-providers-take-hit-from-aws-google-azure-says-synergy [accessed 2023-12-30]
38. Huawei Cloud won the bid of the Shijiazhuang Beiguo Electronics public cloud project: Ali Cloud, Unicom, Telecom, and Xinhua Net lost the bid. NetEase. 2021. URL: https://www.163.com/dy/article/GPRKUNKM05386WWT.html [accessed 2023-10-13]
39. Li RZZ, Hua J. The global healthcare sector often faced ransomware attacks; health authorities should prioritize cybersecurity preparedness. SFC. URL: https://www.sfccn.com/2022/2-10/2MMDE0MDVfMTY5NjM2Mw.html [accessed 2023-10-03]
40. Nipitpon SA. Why the cloud is a lifeline for the NHS. Open Access Government. 2020. URL: https://www.openaccessgovernment.org/why-the-cloud-is-a-lifeline-for-the-nhs/84112/ [accessed 2023-09-21]
41. Realising the value of health care data: a framework for the future. Ernst & Young Global Limited. 2020. URL: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/life-sciences/life-sciences-pdfs/ey-value-of-health-care-data-v20-final.pdf [accessed 2023-10-12]
42. Xu L, Wu DP, Li Y, Zhang L, Wang Y, Wang QY, et al. Application of portable electroencephalograph in patients with epilepsy and establishment of medical platform. National Medical Journal of China 2020;100(47):3764-3767. [doi: 10.3760/cma.j.cn112137-20200703-02023] [Medline: 33379840]
43. Survey report on the informationization status of Chinese hospitals in 2019-2020 (public version). China Hospital Information Management Association (CHIMA). 2021. URL: https://www.chima.org.cn/Html/News/Articles/8684.html [accessed 2023-10-12]
44. Xie XX, Zhou WM, Lin LY, Fan S, Lin F, Wang L, et al. Internet hospitals in China: cross-sectional survey. J Med Internet Res 2017;19(7):e239 [FREE Full text] [doi: 10.2196/jmir.7854] [Medline: 28676472]

45. Liu D, Li T, Liu X, Wang DF. Survey and analysis on digital construction of primary hospitals in Guizhou Province. Chinese Critical Care Medicine 2022 Aug;34(8):863-870. [doi: 10.3760/cma.j.cn121430-20220511-00476] [Medline: 36177932]

Abbreviations

AI: artificial intelligence
CAGR: compound annual growth rate
CLOUD Act: Clarifying Lawful Use of Data Abroad Act
CNCTPS: Chinese National Cloud-Based Telepathology System
DOI: digital object identifier
HIS: hospital information system
NHS: National Health Service
PMID: PubMed unique identifier
SWOT: Strengths, Weaknesses, Opportunities, and Threats

Edited by C Lovis; submitted 22.08.23; peer-reviewed by T Khodaveisi, M Platt, C Xie; comments to author 07.10.23; revised version received 30.11.23; accepted 02.12.23; published 05.02.24.

Please cite as:
Xu J
The Current Status and Promotional Strategies for Cloud Migration of Hospital Information Systems in China: Strengths, Weaknesses, Opportunities, and Threats Analysis
JMIR Med Inform 2024;12:e52080
URL: https://medinform.jmir.org/2024/1/e52080
doi:10.2196/52080
PMID:38315519

Viewpoint

A Roadmap for Using Causal Inference and Machine Learning to Personalize Asthma Medication Selection

Flory L Nkoy[1*], MS, MPH, MD; Bryan L Stone[1], MS, MD; Yue Zhang[2,3], PhD; Gang Luo[4*], PhD

[1]Department of Pediatrics, University of Utah, Salt Lake City, UT, United States

[2]Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, UT, United States

[3]Division of Biostatistics, Department of Population Health Sciences, University of Utah, Salt Lake City, UT, United States

[4]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

[*]these authors contributed equally

Corresponding Author:
Gang Luo, PhD
Department of Biomedical Informatics and Medical Education
University of Washington
850 Republican Street, Building C
Box 358047
Seattle, WA, 98195
United States
Phone: 1 2062214596
Fax: 1 2062212671
Email: gangluo@cs.wisc.edu

Abstract

Inhaled corticosteroid (ICS) is a mainstay treatment for controlling asthma and preventing exacerbations in patients with persistent asthma. Many types of ICS drugs are used, either alone or in combination with other controller medications. Despite the widespread use of ICSs, asthma control remains suboptimal in many people with asthma. Suboptimal control leads to recurrent exacerbations, causes frequent ER visits and inpatient stays, and is due to multiple factors. One such factor is the inappropriate ICS choice for the patient. While many interventions targeting other factors exist, less attention is given to inappropriate ICS choice. Asthma is a heterogeneous disease with variable underlying inflammations and biomarkers. Up to 50% of people with asthma exhibit some degree of resistance or insensitivity to certain ICSs due to genetic variations in ICS metabolizing enzymes, leading to variable responses to ICSs. Yet, ICS choice, especially in the primary care setting, is often not tailored to the patient's characteristics. Instead, ICS choice is largely by trial and error and often dictated by insurance reimbursement, organizational prescribing policies, or cost, leading to a one-size-fits-all approach with many patients not achieving optimal control. There is a pressing need for a decision support tool that can predict an effective ICS at the point of care and guide providers to select the ICS that will most likely and quickly ease patient symptoms and improve asthma control. To date, no such tool exists. Predicting which patient will respond well to which ICS is the first step toward developing such a tool. However, no study has predicted ICS response, forming a gap. While the biologic heterogeneity of asthma is vast, few, if any, biomarkers and genotypes can be used to systematically profile all patients with asthma and predict ICS response. As endotyping or genotyping all patients is infeasible, readily available electronic health record data collected during clinical care offer a low-cost, reliable, and more holistic way to profile all patients. In this paper, we point out the need for developing a decision support tool to guide ICS selection and the gap in fulfilling the need. Then we outline an approach to close this gap via creating a machine learning model and applying causal inference to predict a patient's ICS response in the next year based on the patient's characteristics. The model uses electronic health record data to characterize all patients and extract patterns that could mirror endotype or genotype. This paper supplies a roadmap for future research, with the eventual goal of shifting asthma care from one-size-fits-all to personalized care, improve outcomes, and save health care resources.

(JMIR Med Inform 2024;12:e56572) doi:10.2196/56572

KEYWORDS

asthma; causal inference; forecasting; machine learning; decision support; drug; drugs; pharmacy; pharmacies; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications; medication selection;

respiratory; pulmonary; forecast; ICS; inhaled corticosteroid; inhaler; inhaled; corticosteroid; corticosteroids; artificial intelligence; personalized; customized

Introduction

Asthma is a chronic disease characterized by inflammation, narrowing, and hyperactivity of the airways causing shortness of breath, chest tightness, coughing, and wheezing [1]. Asthma affects about 25 million people in the United States [2]. In 2021, there were 9.8 million exacerbations of asthma symptoms (or asthma attacks) leading to over 980,000 emergency room (ER) visits and over 94,500 hospitalizations [2]. Asthma costs the US economy over US $80 billion in health care expenses each year, work and school absenteeism, and deaths [3].

Inhaled corticosteroid (ICS) is a mainstay treatment for controlling asthma and preventing exacerbations in patients with persistent asthma [4] accounting for over 60% of people with asthma [5,6]. Many types of ICS drugs are used, either alone like fluticasone (Flovent, Arnuity, and Aller-flo), budesonide (Pulmicort, Entocort, and Rhinocort), mometasone (Asmanex), beclomethasone (Beclovent, Qvar, Vancenase, Beconase, Vanceril, and Qnasl), ciclesonide (Alvesco), and so forth, or in combination with a long-acting beta2 agonist like fluticasone/salmeterol (Advair), budesonide/formoterol (Symbicort), mometasone/formoterol (Dulera), and fluticasone/vilanterol (Breo), and so forth [4]. Regular use of appropriate ICSs improves asthma control and reduces airway inflammation, symptoms, exacerbations, ER visits, and inpatient stays [7-9].

Despite the widespread use of ICSs, asthma control remains suboptimal in many people with asthma [10-13] including 44% of children and 60% of adults based on asthma exacerbations in the past year [14,15], 72% of patients based on asthma control test [10], 53% of children and 44% of adults based on asthma attacks in the past year [16], and 59% of children based on the 2007-2013 Medical Expenditure Panel Survey [17]. Suboptimal control leads to recurrent exacerbations, causes frequent ER visits and inpatient stays, and is projected to have an economic burden of US $963.5 billion over the next 20 years [18]. Suboptimal control is due to multiple factors [19-23] including (1) failure to recognize and act on early signs of declining control [24,25], (2) lack of self-management skills, (3) nonadherence to therapy [26], and (4) inappropriate ICS choice for the patient [27-32]. While interventions targeting other factors exist, less attention has been given to inappropriate ICS choice.

Asthma is heterogeneous with variable profiles in terms of clinical presentations (phenotypes) and underlying mechanisms (endotypes) [33,34]. Molecular techniques have revealed a few phenotype and endotype relationships, allowing the categorization of asthma into two main groups (1) T-helper type 2 (Th2)-high (eg, atopic and late onset) and (2) Th2-low (eg, nonatopic, smoking-related, and obesity-related) [33,34]. It is known that within the 2 groups, there are many subgroups [33,35] with different biomarker expressions (eg, immunoglobulin E [IgE], fractional exhaled nitric oxide [FeNO], interleukin [IL]-4, IL-5, and IL-13) [36]. So far, only a few biomarkers have been characterized for use in clinical practice. Despite a few successes using biomarkers for targeted therapy, ICS choice, especially in the primary care setting, is largely by trial and error and many patients remain uncontrolled [37-42].

Besides patient nonadherence and environmental factors, response to ICS treatment is affected by genetic variations in ICS metabolizing enzymes [43,44], regardless of whether the ICS is used alone or is combined with another asthma medication like a long-acting beta2 agonist. Single nucleotide polymorphisms in cap methyltransferase 1 (CMTR1), tripartite motif containing 24 (TRIM24), and membrane associated guanylate kinase, WW and PDZ domain containing 2 (MAGI2) genes were found to be associated with variability in asthma exacerbations [43]. Additional evidence supports that these genes also cause variability in ICS response [44]. Due to genetic variations in cytochrome P (CYP) 450 enzymes that metabolize over 80% of drugs including ICS, up to 50% of people with asthma have altered metabolism to certain ICSs [45-51] impacting asthma control [52,53]. CYP3A5*3/*3 and CYP3A4*22 genotypes were found to be linked to ICS response [54,55]. These studies provide evidence that genetic variations greatly affect ICS responsiveness, although the exact relationships between genetic variations and ICS response remain largely unknown [36,56,57]. Currently, many candidate genes are being studied, and pharmacogenetics has not yet reached routine clinical practice in asthma care.

ICS choice for patients is often dictated by insurance reimbursement, organizational policies, or cost, leading to a one-size-fits-all approach [37-42]. Some insurers require patients to first fail on a cheaper ICS before authorizing a more expensive ICS [39]. Nonmedical switch due to preferred drug formulary change is common and leads to bad outcomes, with 70% of patients reporting more exacerbations after the switch [39]. Patients also often report that they tried a few different ICSs before ending up with the drug that gave them the most relief, with 60% reporting it was hard for their providers to find the effective drug [37-39]. Cycling through various ICSs delays the start of an effective ICS and is neither efficient nor cost-effective [39]. New strategies are needed to allow a faster and more efficient way to tailor ICS selection to each patient's characteristics [36].

While the biologic heterogeneity of asthma is vast, few, if any, biomarkers or genotypes can currently be used to systematically profile all patients with asthma and predict ICS response [36,58,59]. Readily available electronic health record (EHR) data collected during clinical care offer a low-cost, reliable, and more holistic way to profile all patients [36,60]. With a high accuracy of 87%-95% [36], machine learning models using EHR data have been used to profile patients in various areas, for example, to develop a phenotype for patients with Turner syndrome [61], identify low medication adherence profiles [62], find variable COVID-19 treatment response profiles [63], and predict hypertension treatment response [64]. Yet, while machine learning has helped find various asthma profiles [65-72], no prior study has predicted ICS response. Also, prior

studies are mostly from single centers with small sample sizes and have not moved the needle of precision treatment for asthma [58,60].

A decision support tool is greatly needed, especially in the primary care setting, to guide providers to select at the point of care the ICS that will most likely and quickly ease patient symptoms and improve asthma control. Forecasting which patient will respond well to which ICS is the first step toward creating this tool, but no prior study has predicted ICS response, forming a gap.

To shift asthma care from one-size-fits-all to personalized care, improve outcomes, and save health care resources, we make three contributions in this paper, supplying a roadmap for future research: (1) we point out the above-mentioned need for creating a decision support tool to guide ICS selection; (2) we point out the above-mentioned gap in fulfilling this need; and (3) to close this gap, we outline an approach to create a machine learning model and apply causal inference to predict a patient's ICS response in the next year based on the patient's characteristics. We present the central ideas of this approach in the following sections.

Creating a Machine Learning Model and Applying Causal Inference to Predict ICS Response

Overview of Our Approach

We use EHR data from a large health care system to develop a machine learning model and apply casual inference to predict a patient's ICS response based on the patient's characteristics. As endotyping or genotyping all patients is infeasible, our model uses EHR data to characterize all patients and extract patterns that could mirror endotype or genotype. Our model is trained on historical data, and can then be applied to new patients to guide ICS selection during an initial or early encounter for asthma care. The optimal ICS choice identified by our approach can be either an ICS (generic name and dosage) alone or an ICS combined with another asthma medication like a long-acting beta2 agonist.

Both pediatric and adult patients with asthma are treated by primary care providers (PCPs) who are mostly generalists and asthma specialists including allergists, immunologists, and pulmonologists. Large differences exist between PCPs and specialists in terms of knowledge, care patterns, and asthma outcomes, with asthma specialists adhering more often to guideline recommendations [73-76]. A greater difference exists between PCPs and specialists in controller medication use [76]. Compared to PCPs, asthma specialists tend to achieve better outcomes [77], including higher physical functioning [78], better patient-reported care [78], and fewer ER visits and inpatient stays [78-84]. As over 60% of people with asthma are cared for by PCPs [85], our machine learning model primarily targets PCPs, although asthma specialists could also benefit from this model.

The asthma medication ratio (AMR) is the total number of units of asthma controller medications dispensed divided by the total number of units of asthma medications (controllers + relievers) dispensed [86,87]. Higher AMR (≥0.5) is associated with less oral corticosteroid use (a surrogate measure for asthma exacerbations), fewer ER visits and inpatient stays, and lower costs [87-89]. Lower AMR (<0.5) is associated with more exacerbations, ER visits, and inpatient stays [90,91]. Approved by Healthcare Effectiveness Data and Information Set (HEDIS) as a quality measure, AMR is widely used by health care systems [89]. AMR is a reliable reflection of asthma control and gives an accurate assessment of asthma exacerbation risk [92]. We use change in AMR as the prediction target of our model for predicting ICS response, as AMR can be calculated on all patients. In comparison, neither asthma control nor acute outcomes (eg, ER visits, inpatient stays, or oral corticosteroid use) is used as the prediction target, as the former is often missing in EHRs and the latter does not occur in all patients. An effective ICS will lead to less reliever use and increased AMR. An ineffective ICS will lead to more reliever use and reduced AMR. We formerly used EHR data to build accurate models to predict hospital use (ER visit or inpatient stay) for asthma [93-95]. We expect EHR data to have great predictive power for AMR, which is associated with hospital use for asthma [87-91]. Using the AMR can facilitate the dissemination of our approach across health care systems.

We outline the individual steps of our approach in the following sections.

Step 1: Building a Machine Learning Model to Predict a Patient's ICS Response Defined by Changes in AMR

We focus on patients with persistent asthma for whom ICSs are mainly used. We use the HEDIS case definition of persistent asthma [96,97], the already validated [98] and the most commonly used administrative data marker of persistent asthma [97]. A patient is deemed to have persistent asthma if in each of 2 consecutive years, the patient meets at least one of the following criteria: (1) at least 1 ER visit or inpatient stay with a principal diagnosis code of asthma (*ICD-9* [*International Classification of Diseases, Ninth Revision*] 493.0x, 493.1x, 493.8x, 493.9x; *ICD-10* [*International Classification of Diseases, Tenth Revision*] J45.x), (2) at least 2 asthma medication dispensing and at least 4 outpatient visits, each with a diagnosis code of asthma, and (3) at least 4 asthma medication dispensing. In the rest of this paper, we always use patients with asthma to refer to patients with persistent asthma. The prediction target or outcome is the amount of change in a patient's AMR after 1 year. The AMR is computed over a 1-year period [86,87].

We combine patient, air quality, and weather features computed on the raw variables to build the model to predict ICS response. Existing predictive models for asthma outcomes [93-95,99-110] rarely use air quality and weather variables, but these variables impact asthma outcomes [111-117] (eg, short-term exposure to air pollution, even if measured at the regional level, is associated with asthma exacerbations [113-117]). For each such variable, we examine multiple features (eg, mean, maximum, SD, and slope). We examine over 200 patient features listed in our papers' [93-95] appendices and formerly used to predict hospital use for asthma, which is associated with AMR [87-91]. Several examples of these features are comorbidities, allergies, the

number of the patient's asthma-related ER visits in the prior 12 months, the total number of units of systemic corticosteroids ordered for the patient in the prior 12 months, and the number of primary or principal asthma diagnoses of the patient in the prior 12 months. We also use as features the patient's current AMR computed over the prior 12 months [86,87], the generic name and the dosage of the ICS that the patient currently uses, and those of the long-acting beta2 agonist, leukotriene receptor antagonist, biologic or another asthma medication, if any, that is combined with the ICS.

Step 2: Conducting Causal Machine Learning to Identify Optimal ICS Choice

Our goal is to integrate machine learning and G-computation to develop a method to estimate the causal effects of various ICS choices on AMR for patients with specific characteristics. This causal machine learning method [118] processes large data sets by capturing complex nonlinear relationships between features, thereby revealing the cause-and-effect relationships between ICS choice and change in AMR. We use the machine learning model built in step 1. Using G-computation [119,120], an imputation-based causal inference method, we estimate the potential effects of hypothetical ICS choices with specific dosages on changes in AMR after 1 year. G-computation builds on the machine learning model of the outcome as a function of ICS indicators, ICS dosages, and other features to predict AMR outcomes under different counterfactual ICS choice scenarios. CIs are estimated through 10,000 bootstrap resampling with replacement [121].

We apply causal machine learning to estimate the impact of ICS choices on patients with specific characteristics by averaging predicted AMR after 1 year for a given ICS and these characteristics across all participants. This estimation is contrasted with the averaged predicted outcome in the absence of any ICS choice. The ICS choice with the highest and statistically significant contrast estimation is identified as the optimal choice for patients with these characteristics. All hypotheses can be tested at a significance level of .05.

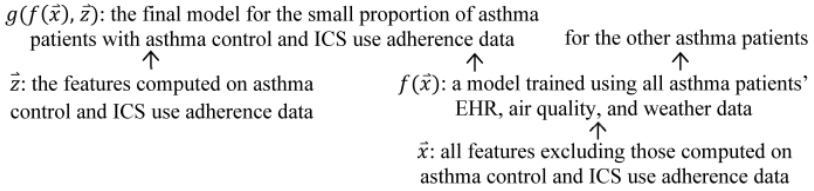
Step 3: Assessing the Impact of Adding External Patient-Reported Asthma Control and ICS Use Adherence Data on the Model's Predictions

EHRs have limitations regarding patient-reported data with extra predictive power such as asthma control and ICS use adherence. For asthma, asthma control and ICS use adherence are critical variables, as (1) a patient's asthma control fluctuates over time and drives the provider's decision to prescribe or adjust ICSs and (2) ICS use adherence impacts the patient's asthma control and helps assess whether the patient is actually responding to an ICS. However, despite their high predictive power for patient outcomes, these variables are not routinely collected or included in EHRs in clinical practice. At Intermountain Healthcare, the largest health care system in Utah, we pioneered the electronic AsthmaTracker, a mobile health (mHealth) app used weekly to assess, collect, and monitor patients' asthma control and actual ICS use adherence [122]. Like most patient-reported data, these patient-reported variables have been collected on only a small proportion of patients with asthma. To date, 1380 patients with asthma have used the app and produced about 45,000 records of weekly asthma control scores and ICS use adherence data (eg, the ICS' name and the number of days an ICS is actually used by the patient in that week). If we train a predictive model using EHR and patient-reported data limited to this small proportion of patients, the model will be inaccurate due to insufficient training data. Yet, for these patients, combining their patient-reported data with the outputs of a model built on all patients' EHR data can help raise the prediction accuracy for them. To realize this, we propose the first method to combine external patient-reported data available on a small proportion of patients with the outputs of a model built on all patients' EHR data to raise prediction accuracy for the small proportion of patients while maintaining prediction accuracy for the other patients.

To illustrate how our method works, we consider the case that the model created in step 1 is built using Intermountain Healthcare EHR data. The weekly asthma control scores and ICS use adherence data collected from the 1380 patients with asthma are unused in step 1. Now we add features (eg, mean, SD, and slope) computed on patient-reported asthma control and ICS use adherence data to raise prediction accuracy for these patients. Among all patients with asthma, only 1% have asthma control and ICS use adherence data. We use the method shown in Figure 1 to combine the asthma control and ICS use adherence data from this small proportion of patients with the outputs of a model trained on EHR, air quality, and weather data of all patients with asthma. We start from the original model built in step 1. This model is reasonably accurate, as it is trained using EHR, air quality, and weather data of all patients with asthma and all features excluding those computed on asthma control and ICS use adherence data. For each patient with asthma control and ICS use adherence data, we apply the model to the patient, obtain a prediction result, and use this result as a feature. We then combine this new feature with the features computed on asthma control and ICS use adherence data to train a second model for these patients using their data. The second model is built upon and thus tends to be more accurate than the original model for these patients. The original model is used for the other patients. Our method is general, works for all kinds of features, and is not limited to any specific disease, prediction target, cohort, or health care system. Whenever a small proportion of patients have extra predictive variables, we could use this method to raise prediction accuracy for these patients while maintaining prediction accuracy for the other patients.

For the patients with asthma control and ICS use adherence data, we compare the mean squared and the mean absolute prediction errors gained by the model built in step 1 and the second model built here. We expect adding asthma control and ICS use adherence data to the model to lower both prediction errors. The error drop rates help reveal the value of routinely collecting asthma control and ICS use adherence data in clinical care to lower prediction errors. Currently, such data are rarely collected.

Figure 1. Our method to raise prediction accuracy for the small proportion of patients with asthma and asthma control and ICS use adherence data while maintaining prediction accuracy for the other patients with asthma. EHR: electronic health record; ICS: inhaled corticosteroid.



Discussion

Principal Findings

Besides the variables mentioned in the "Step 1: Building a machine learning model to predict a patient's ICS response defined by changes in AMR" section, environmental variables beyond air quality and weather and many other factors can impact patient outcomes. Moreover, there are almost infinite possible features. For any first future study that one will do along the direction pointed out in this paper, a realistic goal is to show that using our methods can build decent models and improve asthma care rather than to exhaust all possible useful variables and features and obtain the theoretically highest possible model performance. Not accounting for all possible factors limits the generalizability of these models to medication selection for other diseases.

We use the G-computation method to conduct causal inference. This method relies heavily on correctly specifying the predictive model for ICS response, including accurately identifying all relevant confounders and interactions and incorporating them into the model. Misspecification of the model can lead to biased estimated effects of various ICS choices on AMR. To address this issue, we can adopt several preventive strategies during model development. We engage with subject matter experts to ensure that the model includes all relevant variables and reflects the underlying process. To guide model development and help identify potential sources of bias, we construct a directed acyclic graph that lays out the relationships among the independent and dependent variables. We use machine learning techniques that provide flexible modeling approaches to capture complex relationships among variables. When reporting our findings, we keep transparent about the final model specification and the rationale behind our model building process. We believe using these strategies will mitigate the risk of model misspecification and strengthen the reliability of our estimated effects of various ICS choices on AMR.

AMR is reported to be a reliable reflection of asthma control and of asthma exacerbation risk [92]. In a future study that we plan to do along the direction pointed out in this paper, we can use Intermountain Healthcare data to validate this relationship. Specifically, we use multivariable linear regression to assess the relationship between the AMR computed on EHR data and the patient's asthma control level obtained from the external patient-reported data, while controlling for other factors. We expect to see a strong and positive association between the AMR and the patient's asthma control level.

When creating the model in step 1, we can include medication persistence measures computed on insurance claim data [123], such as the proportion of days covered for ICS, as features. However, this does not obviate the need to examine patient-reported ICS use adherence data in step 3. ICS persistence measures give information on the possession of ICS, but not on actual use of ICS. Each ICS persistence measure is computed at a coarse time granularity as an average value over a long period. In comparison, our patient-reported ICS use adherence data offer information on the actual use of ICS. The data are at a fine time granularity, with 1 set of values per week for a patient. This enables us to compute features on various patterns and trends that can be useful for making predictions.

Conclusions

In asthma care, ICS choice is largely by trial and error and often made by a one-size-fits-all approach with many patients not achieving optimal outcomes. In this paper, we point out the need for creating a decision support tool to guide ICS selection and a gap in fulfilling this need. Then we outline an approach to close this gap via creating a machine learning model and applying causal inference to predict a patient's ICS response in the next year based on the patient's characteristics. This supplies a roadmap for future research.

Authors' Contributions

FLN and GL are co-senior authors mainly responsible for the paper. They conceptualized the presentation approach, performed literature review, and wrote the paper. BLS provided feedback on various medical issues, contributed to conceptualizing the presentation, and revised the paper. YZ wrote the causal inference section. All authors read and approved the final paper.

Conflicts of Interest

GL is an editorial board member of *JMIR AI*. The other authors declare no conflicts of interest.

References

1. Hashmi MF, Tariq M, Cataletto ME. Asthma. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.

2. Most recent national asthma data. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm [accessed 2024-01-22]
3. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018;15(3):348-356 [FREE Full text] [doi: 10.1513/AnnalsATS.201703-259OC] [Medline: 29323930]
4. Inhaled corticosteroids. American Academy of Allergy, Asthma & Immunology. 2023. URL: https://www.aaaai.org/tools-for-the-public/drug-guide/inhaled-corticosteroids [accessed 2024-01-22]
5. Asthma severity among children with current asthma. Centers for Disease Control and Prevention. 2023. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/asthma/asthma_stats/severity_child.htm [accessed 2024-01-22]
6. Asthma severity among adults with current asthma. Centers for Disease Control and Prevention. 2023. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/asthma/asthma_stats/severity_adult.htm [accessed 2024-01-22]
7. Averell CM, Laliberté F, Germain G, Duh MS, Rousculp MD, MacKnight SD, et al. Impact of adherence to treatment with inhaled corticosteroids/long-acting β-agonists on asthma outcomes in the United States. Ther Adv Respir Dis 2022;16:17534666221116997 [FREE Full text] [doi: 10.1177/17534666221116997] [Medline: 36036456]
8. Cardet JC, Papi A, Reddel HK. "As-needed" inhaled corticosteroids for patients with asthma. J Allergy Clin Immunol Pract 2023;11(3):726-734 [FREE Full text] [doi: 10.1016/j.jaip.2023.01.010] [Medline: 36702246]
9. Sadatsafavi M, Lynd LD, De Vera MA, Zafari Z, FitzGerald JM. One-year outcomes of inhaled controller therapies added to systemic corticosteroids after asthma-related hospital discharge. Respir Med 2015;109(3):320-328 [FREE Full text] [doi: 10.1016/j.rmed.2014.12.014] [Medline: 25596136]
10. George M, Balantac Z, Gillette C, Farooqui N, Tervonen T, Thomas C, et al. Suboptimal control of asthma among diverse patients: a US mixed methods focus group study. J Asthma Allergy 2022;15:1511-1526 [FREE Full text] [doi: 10.2147/JAA.S377760] [Medline: 36313858]
11. Sullivan PW, Ghushchyan V, Kavati A, Navaratnam P, Friedman HS, Ortiz B. Trends in asthma control, treatment, health care utilization, and expenditures among children in the United States by place of residence: 2003-2014. J Allergy Clin Immunol Pract 2019;7(6):1835-1842.e2. [doi: 10.1016/j.jaip.2019.01.055] [Medline: 30772478]
12. Zhang S, White J, Hunter AG, Hinds D, Fowler A, Gardiner F, et al. Suboptimally controlled asthma in patients treated with inhaled ICS/LABA: prevalence, risk factors, and outcomes. NPJ Prim Care Respir Med 2023;33(1):19 [FREE Full text] [doi: 10.1038/s41533-023-00336-9] [Medline: 37156824]
13. Nurmagambetov TA, Krishnan JA. What will uncontrolled asthma cost in the United States? Am J Respir Crit Care Med 2019;200(9):1077-1078 [FREE Full text] [doi: 10.1164/rccm.201906-1177ED] [Medline: 31251082]
14. Uncontrolled asthma among children with current asthma, 2018-2020. Centers for Disease Control and Prevention. 2021. URL: https://tinyurl.com/ycdz2mp2 [accessed 2024-01-22]
15. Uncontrolled asthma among adults, 2019. Centers for Disease Control and Prevention. 2020. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/asthma/asthma_stats/uncontrolled-asthma-adults-2019.htm [accessed 2024-01-22]
16. Pate CA, Zahran HS, Qin X, Johnson C, Hummelman E, Malilay J. Asthma surveillance—United States, 2006-2018. MMWR Surveill Summ 2021;70(5):1-32 [FREE Full text] [doi: 10.15585/mmwr.ss7005a1] [Medline: 34529643]
17. Sullivan PW, Ghushchyan V, Navaratnam P, Friedman HS, Kavati A, Ortiz B, et al. National prevalence of poor asthma control and associated outcomes among school-aged children in the United States. J Allergy Clin Immunol Pract 2018;6(2):536-544.e1. [doi: 10.1016/j.jaip.2017.06.039] [Medline: 28847656]
18. Yaghoubi M, Adibi A, Safari A, FitzGerald JM, Sadatsafavi M. The projected economic and health burden of uncontrolled asthma in the United States. Am J Respir Crit Care Med 2019;200(9):1102-1112 [FREE Full text] [doi: 10.1164/rccm.201901-0016OC] [Medline: 31166782]
19. Centers for Disease Control and Prevention (CDC). Asthma hospitalizations and readmissions among children and young adults--Wisconsin, 1991-1995. MMWR Morb Mortal Wkly Rep 1997;46(31):726-729 [FREE Full text] [Medline: 9262074]
20. Li D, German D, Lulla S, Thomas RG, Wilson SR. Prospective study of hospitalization for asthma. A preliminary risk factor model. Am J Respir Crit Care Med 1995;151(3 Pt 1):647-655. [doi: 10.1164/ajrccm.151.3.7881651] [Medline: 7881651]
21. Crane J, Pearce N, Burgess C, Woodman K, Robson B, Beasley R. Markers of risk of asthma death or readmission in the 12 months following a hospital admission for asthma. Int J Epidemiol 1992;21(4):737-744. [doi: 10.1093/ije/21.4.737] [Medline: 1521979]
22. Mitchell EA, Bland JM, Thompson JM. Risk factors for readmission to hospital for asthma in childhood. Thorax 1994;49(1):33-36 [FREE Full text] [doi: 10.1136/thx.49.1.33] [Medline: 8153938]
23. Vargas PA, Perry TT, Robles E, Jo CH, Simpson PM, Magee JM, et al. Relationship of body mass index with asthma indicators in head start children. Ann Allergy Asthma Immunol 2007;99(1):22-28. [doi: 10.1016/S1081-1206(10)60616-3] [Medline: 17650825]
24. Barnes PJ. Achieving asthma control. Curr Med Res Opin 2005;21(Suppl 4):S5-S9. [doi: 10.1185/030079905X61730] [Medline: 16138939]
25. Bloomberg GR, Banister C, Sterkel R, Epstein J, Bruns J, Swerczek L, et al. Socioeconomic, family, and pediatric practice factors that affect level of asthma control. Pediatrics 2009;123(3):829-835 [FREE Full text] [doi: 10.1542/peds.2008-0504] [Medline: 19255010]

26. Bateman ED, Frith LF, Braunstein GL. Achieving guideline-based asthma control: does the patient benefit? Eur Respir J 2002;20(3):588-595 [FREE Full text] [doi: 10.1183/09031936.02.00294702] [Medline: 12358333]
27. Chapman KR, Boulet LP, Rea RM, Franssen E. Suboptimal asthma control: prevalence, detection and consequences in general practice. Eur Respir J 2008;31(2):320-325 [FREE Full text] [doi: 10.1183/09031936.00039707] [Medline: 17959642]
28. Rabe KF, Adachi M, Lai CK, Soriano JB, Vermeire PA, Weiss KB, et al. Worldwide severity and control of asthma in children and adults: the global asthma insights and reality surveys. J Allergy Clin Immunol 2004;114(1):40-47 [FREE Full text] [doi: 10.1016/j.jaci.2004.04.042] [Medline: 15241342]
29. National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): guidelines for the diagnosis and management of asthma-summary report 2007. J Allergy Clin Immunol 2007;120(Suppl 5):S94-S138. [doi: 10.1016/j.jaci.2007.09.043] [Medline: 17983880]
30. Stempel DA, McLaughin TP, Stanford RH, Fuhlbrigge AL. Patterns of asthma control: a 3-year analysis of patient claims. J Allergy Clin Immunol 2005;115(5):935-939 [FREE Full text] [doi: 10.1016/j.jaci.2005.01.054] [Medline: 15867848]
31. Cukovic L, Sutherland E, Sein S, Fuentes D, Fatima H, Oshana A, et al. An evaluation of outpatient pediatric asthma prescribing patterns in the United States. Int J Sci Res Arch 2023;9(1):344-349 [FREE Full text] [doi: 10.30574/ijsra.2023.9.1.0388]
32. Belhassen M, Nibber A, Van Ganse E, Ryan D, Langlois C, Appiagyei F, et al. Inappropriate asthma therapy-a tale of two countries: a parallel population-based cohort study. NPJ Prim Care Respir Med 2016;26:16076 [FREE Full text] [doi: 10.1038/npjpcrm.2016.76] [Medline: 27735927]
33. McIntyre AP, Viswanathan RK. Phenotypes and endotypes in asthma. Adv Exp Med Biol 2023;1426:119-142. [doi: 10.1007/978-3-031-32259-4_6] [Medline: 37464119]
34. Kuruvilla ME, Lee FE, Lee GB. Understanding asthma phenotypes, endotypes, and mechanisms of disease. Clin Rev Allergy Immunol 2019;56(2):219-233 [FREE Full text] [doi: 10.1007/s12016-018-8712-1] [Medline: 30206782]
35. Salter B, Lacy P, Mukherjee M. Biologics in asthma: a molecular perspective to precision medicine. Front Pharmacol 2021;12:793409 [FREE Full text] [doi: 10.3389/fphar.2021.793409] [Medline: 35126131]
36. van der Burg N, Tufvesson E. Is asthma's heterogeneity too vast to use traditional phenotyping for modern biologic therapies? Respir Med 2023;212:107211. [doi: 10.1016/j.rmed.2023.107211] [Medline: 36924848]
37. A study of the qualitative impact of non-medical switching. Alliance for Patient Access. 2019. URL: https://tinyurl.com/2vxwks83 [accessed 2024-01-22]
38. Cost-motivated treatment changes & non-medical switching: commercial health plans analysis. Alliance for Patient Access. 2017. URL: https://tinyurl.com/424dy3xz [accessed 2024-01-22]
39. Collins S. Asthma meds, insurers, and the practice of non-medical drug switching. HealthCentral. 2023. URL: https://www.healthcentral.com/condition/asthma/what-you-need-to-know-about-asthma-meds [accessed 2024-01-22]
40. Landhuis E. OTC budesonide-formoterol for asthma could save lives, money. Medscape Medical News. 2023. URL: https://www.medscape.com/viewarticle/989099 [accessed 2024-01-22]
41. Modglin L. How much do inhalers cost? SingleCare. 2022. URL: https://www.singlecare.com/blog/asthma-inhalers-price-list [accessed 2024-01-22]
42. Gibson PG, McDonald VM, Thomas D. Treatable traits, combination inhaler therapy and the future of asthma management. Respirology 2023;28(9):828-840 [FREE Full text] [doi: 10.1111/resp.14556] [Medline: 37518933]
43. Dahlin A, Denny J, Roden DM, Brilliant MH, Ingram C, Kitchner TE, et al. CMTR1 is associated with increased asthma exacerbations in patients taking inhaled corticosteroids. Immun Inflamm Dis 2015;3(4):350-359 [FREE Full text] [doi: 10.1002/iid3.73] [Medline: 26734457]
44. Keskin O, Farzan N, Birben E, Akel H, Karaaslan C, Maitland-van der Zee AH, et al. Genetic associations of the response to inhaled corticosteroids in asthma: a systematic review. Clin Transl Allergy 2019;9:2 [FREE Full text] [doi: 10.1186/s13601-018-0239-2] [Medline: 30647901]
45. Delgado-Dolset MI, Obeso D, Rodríguez-Coira J, Tarin C, Tan G, Cumplido JA, et al. Understanding uncontrolled severe allergic asthma by integration of omic and clinical data. Allergy 2022;77(6):1772-1785 [FREE Full text] [doi: 10.1111/all.15192] [Medline: 34839541]
46. Liu Q, Hua L, Bao C, Kong L, Hu J, Liu C, et al. Inhibition of spleen tyrosine kinase restores glucocorticoid sensitivity to improve steroid-resistant asthma. Front Pharmacol 2022;13:885053 [FREE Full text] [doi: 10.3389/fphar.2022.885053] [Medline: 35600871]
47. Cardoso-Vigueros C, von Blumenthal T, Rückert B, Rinaldi AO, Tan G, Dreher A, et al. Leukocyte redistribution as immunological biomarker of corticosteroid resistance in severe asthma. Clin Exp Allergy 2022;52(10):1183-1194 [FREE Full text] [doi: 10.1111/cea.14128] [Medline: 35305052]
48. Liang H, Zhang X, Ma Z, Sun Y, Shu C, Zhu Y, et al. Association of CYP3A5 gene polymorphisms and amlodipine-induced peripheral edema in Chinese Han patients with essential hypertension. Pharmgenomics Pers Med 2021;14:189-197 [FREE Full text] [doi: 10.2147/PGPM.S291277] [Medline: 33564260]
49. Wang SB, Huang T. The early detection of asthma based on blood gene expression. Mol Biol Rep 2019;46(1):217-223. [doi: 10.1007/s11033-018-4463-6] [Medline: 30421126]

50. Roberts JK, Moore CD, Romero EG, Ward RM, Yost GS, Reilly CA. Regulation of CYP3A genes by glucocorticoids in human lung cells. F1000Res 2013;2:173 [FREE Full text] [doi: 10.12688/f1000research.2-173.v2] [Medline: 24555085]
51. Moore CD, Roberts JK, Orton CR, Murai T, Fidler TP, Reilly CA, et al. Metabolic pathways of inhaled glucocorticoids by the CYP3A enzymes. Drug Metab Dispos 2013;41(2):379-389 [FREE Full text] [doi: 10.1124/dmd.112.046318] [Medline: 23143891]
52. Roche N, Garcia G, de Larrard A, Cancalon C, Bénard S, Perez V, et al. Real-life impact of uncontrolled severe asthma on mortality and healthcare use in adolescents and adults: findings from the retrospective, observational RESONANCE study in France. BMJ Open 2022;12(8):e060160 [FREE Full text] [doi: 10.1136/bmjopen-2021-060160] [Medline: 36002203]
53. Munoz-Cano R, Torrego A, Bartra J, Sanchez-Lopez J, Palomino R, Picado C, et al. Follow-up of patients with uncontrolled asthma: clinical features of asthma patients according to the level of control achieved (the COAS study). Eur Respir J 2017;49(3):1501885 [FREE Full text] [doi: 10.1183/13993003.01885-2015] [Medline: 28254764]
54. Stockmann C, Reilly CA, Fassl B, Gaedigk R, Nkoy F, Stone B, et al. Effect of CYP3A5*3 on asthma control among children treated with inhaled beclomethasone. J Allergy Clin Immunol 2015;136(2):505-507 [FREE Full text] [doi: 10.1016/j.jaci.2015.02.009] [Medline: 25825214]
55. Stockmann C, Fassl B, Gaedigk R, Nkoy F, Uchida DA, Monson S, et al. Fluticasone propionate pharmacogenetics: CYP3A4*22 polymorphism and pediatric asthma control. J Pediatr 2013;162(6):1222-1227, 1227.e1-2 [FREE Full text] [doi: 10.1016/j.jpeds.2012.11.031] [Medline: 23290512]
56. Smolnikova MV, Kasparov EW, Malinchik MA, Kopylova KV. Genetic markers of children asthma: predisposition to disease course variants. Vavilovskii Zhurnal Genet Selektsii 2023;27(4):393-400 [FREE Full text] [doi: 10.18699/VJGB-23-47] [Medline: 37465198]
57. Kim HK, Kang JO, Lim JE, Ha TW, Jung HU, Lee WJ, et al. Genetic differences according to onset age and lung function in asthma: a cluster analysis. Clin Transl Allergy 2023;13(7):e12282 [FREE Full text] [doi: 10.1002/clt2.12282] [Medline: 37488738]
58. Mohan A, Lugogo NL. Phenotyping, precision medicine, and asthma. Semin Respir Crit Care Med 2022;43(5):739-751. [doi: 10.1055/s-0042-1750130] [Medline: 36220058]
59. Casanova S, Ahmed E, Bourdin A. Definition, phenotyping of severe asthma, including cluster analysis. Adv Exp Med Biol 2023;1426:239-252. [doi: 10.1007/978-3-031-32259-4_11] [Medline: 37464124]
60. Singhal P, Tan ALM, Drivas TG, Johnson KB, Ritchie MD, Beaulieu-Jones BK. Opportunities and challenges for biomarker discovery using electronic health record data. Trends Mol Med 2023;29(9):765-776. [doi: 10.1016/j.molmed.2023.06.006] [Medline: 37474378]
61. Huang SD, Bamba V, Bothwell S, Fechner PY, Furniss A, Ikomi C, et al. Development and validation of a computable phenotype for turner syndrome utilizing electronic health records from a national pediatric network. Am J Med Genet A 2024;194(4):e63495. [doi: 10.1002/ajmg.a.63495] [Medline: 38066696]
62. Blecker S, Schoenthaler A, Martinez TR, Belli HM, Zhao Y, Wong C, et al. Leveraging electronic health record technology and team care to address medication adherence: protocol for a cluster randomized controlled trial. JMIR Res Protoc 2023;12:e47930 [FREE Full text] [doi: 10.2196/47930] [Medline: 37418304]
63. Verhoef PA, Spicer AB, Lopez-Espina C, Bhargava A, Schmalz L, Sims MD, et al. Analysis of protein biomarkers from hospitalized COVID-19 patients reveals severity-specific signatures and two distinct latent profiles with differential responses to corticosteroids. Crit Care Med 2023;51(12):1697-1705. [doi: 10.1097/CCM.0000000000005983] [Medline: 37378460]
64. Hu Y, Huerta J, Cordella N, Mishuris RG, Paschalidis IC. Personalized hypertension treatment recommendations by a data-driven model. BMC Med Inform Decis Mak 2023;23(1):44 [FREE Full text] [doi: 10.1186/s12911-023-02137-z] [Medline: 36859187]
65. Cottrill KA, Rad MG, Ripple MJ, Stephenson ST, Mohammad AF, Tidwell M, et al. Cluster analysis of plasma cytokines identifies two unique endotypes of children with asthma in the pediatric intensive care unit. Sci Rep 2023;13(1):3521 [FREE Full text] [doi: 10.1038/s41598-023-30679-9] [Medline: 36864187]
66. Horne EMF, McLean S, Alsallakh MA, Davies GA, Price DB, Sheikh A, et al. Defining clinical subtypes of adult asthma using electronic health records: analysis of a large UK primary care database with external validation. Int J Med Inform 2023;170:104942 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104942] [Medline: 36529028]
67. Ilmarinen P, Julkunen-Iivari A, Lundberg M, Luukkainen A, Nuutinen M, Karjalainen J, et al. Cluster analysis of Finnish population-based adult-onset asthma patients. J Allergy Clin Immunol Pract 2023;11(10):3086-3096 [FREE Full text] [doi: 10.1016/j.jaip.2023.05.034] [Medline: 37268268]
68. Imoto S, Suzukawa M, Fukutomi Y, Kobayashi N, Taniguchi M, Nagase T, et al. Phenotype characterization and biomarker evaluation in moderate to severe type 2-high asthma. Asian Pac J Allergy Immunol 2023:1-14 [FREE Full text] [doi: 10.12932/AP-021222-1510] [Medline: 37302094]
69. Kim MA, Shin SW, Park JS, Uh ST, Chang HS, Bae DJ, et al. Clinical characteristics of exacerbation-prone adult asthmatics identified by cluster analysis. Allergy Asthma Immunol Res 2017;9(6):483-490 [FREE Full text] [doi: 10.4168/aair.2017.9.6.483] [Medline: 28913987]

70. Matabuena M, Salgado FJ, Nieto-Fontarigo JJ, Álvarez-Puebla MJ, Arismendi E, Barranco P, et al. Identification of asthma phenotypes in the Spanish MEGA cohort study using cluster analysis. Arch Bronconeumol 2023;59(4):223-231 [FREE Full text] [doi: 10.1016/j.arbres.2023.01.007] [Medline: 36732158]
71. Ngo SY, Venter C, Anderson WC3, Picket K, Zhang H, Arshad SH, et al. Clinical features and later prognosis of replicable early-life wheeze clusters from two birth cohorts 12 years apart. Pediatr Allergy Immunol 2023;34(7):e13999 [FREE Full text] [doi: 10.1111/pai.13999] [Medline: 37492911]
72. Zhan W, Wu F, Zhang Y, Lin L, Li W, Luo W, et al. Identification of cough-variant asthma phenotypes based on clinical and pathophysiologic data. J Allergy Clin Immunol 2023;152(3):622-632. [doi: 10.1016/j.jaci.2023.04.017] [Medline: 37178731]
73. Cloutier MM, Akinbami LJ, Salo PM, Schatz M, Simoneau T, Wilkerson JC, et al. Use of national asthma guidelines by allergists and pulmonologists: a national survey. J Allergy Clin Immunol Pract 2020;8(9):3011-3020.e2 [FREE Full text] [doi: 10.1016/j.jaip.2020.04.026] [Medline: 32344187]
74. Vollmer WM, O'Hollaren M, Ettinger KM, Stibolt T, Wilkins J, Buist AS, et al. Specialty differences in the management of asthma. A cross-sectional assessment of allergists' patients and generalists' patients in a large HMO. Arch Intern Med 1997;157(11):1201-1208. [Medline: 9183231]
75. Cloutier MM, Salo PM, Akinbami LJ, Cohn RD, Wilkerson JC, Diette GB, et al. Clinician agreement, self-efficacy, and adherence with the guidelines for the diagnosis and management of asthma. J Allergy Clin Immunol Pract 2018;6(3):886-894.e4 [FREE Full text] [doi: 10.1016/j.jaip.2018.01.018] [Medline: 29408439]
76. Diette GB, Skinner EA, Nguyen TT, Markson L, Clark BD, Wu AW. Comparison of quality of care by specialist and generalist physicians as usual source of asthma care for children. Pediatrics 2001;108(2):432-437. [doi: 10.1542/peds.108.2.432] [Medline: 11483811]
77. Rosman Y, Hornik-Lurie T, Meir-Shafrir K, Lachover-Roth I, Cohen-Engler A, Confino-Cohen R. The effect of asthma specialist intervention on asthma control among adults. World Allergy Organ J 2022;15(11):100712 [FREE Full text] [doi: 10.1016/j.waojou.2022.100712] [Medline: 36440463]
78. Wu AW, Young Y, Skinner EA, Diette GB, Huber M, Peres A, et al. Quality of care and outcomes of adults with asthma treated by specialists and generalists in managed care. Arch Intern Med 2001;161(21):2554-2560 [FREE Full text] [doi: 10.1001/archinte.161.21.2554] [Medline: 11718586]
79. Erickson S, Tolstykh I, Selby JV, Mendoza G, Iribarren C, Eisner MD. The impact of allergy and pulmonary specialist care on emergency asthma utilization in a large managed care organization. Health Serv Res 2005;40(5 Pt 1):1443-1465 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00410.x] [Medline: 16174142]
80. Zeiger RS, Heller S, Mellon MH, Wald J, Falkoff R, Schatz M. Facilitated referral to asthma specialist reduces relapses in asthma emergency room visits. J Allergy Clin Immunol 1991;87(6):1160-1168. [doi: 10.1016/0091-6749(91)92162-t] [Medline: 2045618]
81. Mahr TA, Evans R3. Allergist influence on asthma care. Ann Allergy 1993;71(2):115-120. [Medline: 8346862]
82. Schatz M, Zeiger RS, Mosen D, Apter AJ, Vollmer WM, Stibolt TB, et al. Improved asthma outcomes from allergy specialist care: a population-based cross-sectional analysis. J Allergy Clin Immunol 2005;116(6):1307-1313 [FREE Full text] [doi: 10.1016/j.jaci.2005.09.027] [Medline: 16337464]
83. Wechsler ME. Managing asthma in primary care: putting new guideline recommendations into context. Mayo Clin Proc 2009;84(8):707-717 [FREE Full text] [doi: 10.4065/84.8.707] [Medline: 19648388]
84. Cooper S, Rahme E, Tse SM, Grad R, Dorais M, Li P. Are primary care and continuity of care associated with asthma-related acute outcomes amongst children? A retrospective population-based study. BMC Prim Care 2022;23(1):5 [FREE Full text] [doi: 10.1186/s12875-021-01605-7] [Medline: 35172739]
85. Akinbami LJ, Salo PM, Cloutier MM, Wilkerson JC, Elward KS, Mazurek JM, et al. Primary care clinician adherence with asthma guidelines: the National Asthma Survey of Physicians. J Asthma 2020;57(5):543-555 [FREE Full text] [doi: 10.1080/02770903.2019.1579831] [Medline: 30821526]
86. HEDIS measures and technical resources: asthma medication ratio (AMR). NCQA. 2023. URL: https://www.ncqa.org/hedis/measures/medication-management-for-people-with-asthma-and-asthma-medication-ratio [accessed 2024-01-22]
87. Schatz M, Zeiger RS, Vollmer WM, Mosen D, Mendoza G, Apter AJ, et al. The controller-to-total asthma medication ratio is associated with patient-centered as well as utilization outcomes. Chest 2006;130(1):43-50. [doi: 10.1378/chest.130.1.43] [Medline: 16840381]
88. Kim Y, Parrish KM, Pirritano M, Moonie S. A higher asthma medication ratio (AMR) predicts a decrease in ED visits among African American and Hispanic children. J Asthma 2023;60(7):1428-1437. [doi: 10.1080/02770903.2022.2155183] [Medline: 36461904]
89. Luskin AT, Antonova EN, Broder MS, Chang E, Raimundo K, Solari PG. Patient outcomes, health care resource use, and costs associated with high versus low HEDIS asthma medication ratio. J Manag Care Spec Pharm 2017;23(11):1117-1124 [FREE Full text] [doi: 10.18553/jmcp.2017.23.11.1117] [Medline: 29083971]
90. Andrews AL, Simpson AN, Basco WTJ, Teufel RJ2. Asthma medication ratio predicts emergency department visits and hospitalizations in children with asthma. Medicare Medicaid Res Rev 2013;3(4):mmrr.003.04.a05 [FREE Full text] [doi: 10.5600/mmrr.003.04.a05] [Medline: 24834366]

91. Andrews AL, Brinton DL, Simpson KN, Simpson AN. A longitudinal examination of the asthma medication ratio in children with Medicaid. J Asthma 2020;57(10):1083-1091 [FREE Full text] [doi: 10.1080/02770903.2019.1640727] [Medline: 31313611]
92. Andrews AL, Brinton D, Simpson KN, Simpson AN. A longitudinal examination of the asthma medication ratio in children. Am J Manag Care 2018;24(6):294-300 [FREE Full text] [Medline: 29939504]
93. Tong Y, Messinger AI, Wilcox AB, Mooney SD, Davidson GH, Suri P, et al. Forecasting future asthma hospital encounters of patients with asthma in an academic health care system: predictive model development and secondary analysis study. J Med Internet Res 2021;23(4):e22796 [FREE Full text] [doi: 10.2196/22796] [Medline: 33861206]
94. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020;8(1):e16080 [FREE Full text] [doi: 10.2196/16080] [Medline: 31961332]
95. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Rozema E, et al. Developing a predictive model for asthma-related hospital encounters in patients with asthma in a large, integrated health care system: secondary analysis. JMIR Med Inform 2020;8(11):e22689 [FREE Full text] [doi: 10.2196/22689] [Medline: 33164906]
96. Mosen DM, Macy E, Schatz M, Mendoza G, Stibolt TB, McGaw J, et al. How well do the HEDIS asthma inclusion criteria identify persistent asthma? Am J Manag Care 2005;11(10):650-654 [FREE Full text] [Medline: 16232006]
97. Schatz M, Zeiger RS. Improving asthma outcomes in large populations. J Allergy Clin Immunol 2011;128(2):273-277 [FREE Full text] [doi: 10.1016/j.jaci.2011.03.027] [Medline: 21497885]
98. Schatz M, Zeiger RS, Yang SJ, Chen W, Crawford WW, Sajjan SG, et al. Persistent asthma defined using HEDIS versus survey criteria. Am J Manag Care 2010;16(11):e281-e288 [FREE Full text] [Medline: 21087074]
99. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004;10(1):25-32 [FREE Full text] [Medline: 14738184]
100. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003;9(8):538-547 [FREE Full text] [Medline: 12921231]
101. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998;157(4 Pt 1):1173-1180 [FREE Full text] [doi: 10.1164/ajrccm.157.4.9708124] [Medline: 9563736]
102. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty of half full? J Asthma 1999;36(4):359-370. [doi: 10.3109/02770909909068229] [Medline: 10386500]
103. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, et al. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010;138(5):1156-1165 [FREE Full text] [doi: 10.1378/chest.09-2426] [Medline: 20472862]
104. Loymans RJB, Debray TPA, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TRJ, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-1952.e15 [FREE Full text] [doi: 10.1016/j.jaip.2018.02.004] [Medline: 29454163]
105. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012;141(1):58-65. [doi: 10.1378/chest.11-0020] [Medline: 21885725]
106. Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, et al. The strategy for predicting future exacerbation of asthma using a combination of the asthma control test and lung function test. J Asthma 2009;46(7):677-682. [doi: 10.1080/02770900902972160] [Medline: 19728204]
107. Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, et al. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004;10(5):321-328 [FREE Full text] [Medline: 15152702]
108. Xiang Y, Ji H, Zhou Y, Li F, Du J, Rasmy L, et al. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. J Med Internet Res 2020;22(7):e16981 [FREE Full text] [doi: 10.2196/16981] [Medline: 32735224]
109. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, et al. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006;28(6):1145-1155 [FREE Full text] [doi: 10.1183/09031936.06.00145105] [Medline: 16870656]
110. Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TR, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016;71(9):838-846 [FREE Full text] [doi: 10.1136/thoraxjnl-2015-208138] [Medline: 27044486]
111. Schatz M. Predictors of asthma control: what can we modify? Curr Opin Allergy Clin Immunol 2012;12(3):263-268. [doi: 10.1097/ACI.0b013e32835335ac] [Medline: 22517290]
112. Dick S, Doust E, Cowie H, Ayres JG, Turner S. Associations between environmental exposures and asthma control and exacerbations in young children: a systematic review. BMJ Open 2014;4(2):e003827 [FREE Full text] [doi: 10.1136/bmjopen-2013-003827] [Medline: 24523420]

113. Schwartz J, Slater D, Larson TV, Pierson WE, Koenig JQ. Particulate air pollution and hospital emergency room visits for asthma in Seattle. Am Rev Respir Dis 1993;147(4):826-831. [doi: 10.1164/ajrccm/147.4.826] [Medline: 8466116]
114. Romieu I, Meneses F, Sienra-Monge JJ, Huerta J, Ruiz Velasco S, White MC, et al. Effects of urban air pollutants on emergency visits for childhood asthma in Mexico City. Am J Epidemiol 1995;141(6):546-553. [doi: 10.1093/oxfordjournals.aje.a117470] [Medline: 7900722]
115. Lu P, Zhang Y, Lin J, Xia G, Zhang W, Knibbs LD, et al. Multi-city study on air pollution and hospital outpatient visits for asthma in China. Environ Pollut 2020;257:113638. [doi: 10.1016/j.envpol.2019.113638] [Medline: 31812526]
116. Liu Y, Pan J, Zhang H, Shi C, Li G, Peng Z, et al. Short-term exposure to ambient air pollution and asthma mortality. Am J Respir Crit Care Med 2019;200(1):24-32 [FREE Full text] [doi: 10.1164/rccm.201810-1823OC] [Medline: 30871339]
117. Vagaggini B, Taccola M, Cianchetti S, Carnevali S, Bartoli ML, Bacci E, et al. Ozone exposure increases eosinophilic airway response induced by previous allergen challenge. Am J Respir Crit Care Med 2002;166(8):1073-1077 [FREE Full text] [doi: 10.1164/rccm.2201013] [Medline: 12379550]
118. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. R Soc Open Sci 2022;9(8):220638 [FREE Full text] [doi: 10.1098/rsos.220638] [Medline: 35950198]
119. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model 1986;7(9-12):1393-1512 [FREE Full text] [doi: 10.1016/0270-0255(86)90088-6]
120. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol 2011;173(7):731-738 [FREE Full text] [doi: 10.1093/aje/kwq472] [Medline: 21415029]
121. Efron B. Bootstrap methods: another look at the Jackknife. Ann Stat 1979;7(1):1-26 [FREE Full text] [doi: 10.1214/aos/1176344552]
122. Nkoy FL, Stone BL, Fassl BA, Uchida DA, Koopmeiners K, Halbern S, et al. Longitudinal validation of a tool for asthma self-monitoring. Pediatrics 2013;132(6):e1554-e1561 [FREE Full text] [doi: 10.1542/peds.2013-1389] [Medline: 24218469]
123. Anghel LA, Farcas AM, Oprean RN. An overview of the common methods used to measure treatment adherence. Med Pharm Rep 2019;92(2):117-122 [FREE Full text] [doi: 10.15386/mpr-1201] [Medline: 31086837]

Abbreviations

AMR: asthma medication ratio
CMTR1: cap methyltransferase 1
CYP: cytochrome P
EHR: electronic health record
ER: emergency room
FeNO: fractional exhaled nitric oxide
HEDIS: Healthcare Effectiveness Data and Information Set
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
ICS: inhaled corticosteroid
IgE: immunoglobulin E
IL: interleukin
MAGI2: membrane associated guanylate kinase, WW and PDZ domain containing 2
mHealth: mobile health
PCP: primary care provider
Th2: T-helper type 2
TRIM24: tripartite motif containing 24

Edited by A Benis; submitted 24.01.24; peer-reviewed by H Tibble, A Kaplan; comments to author 01.03.24; revised version received 12.03.24; accepted 25.03.24; published 17.04.24.

Please cite as:
Nkoy FL, Stone BL, Zhang Y, Luo G
A Roadmap for Using Causal Inference and Machine Learning to Personalize Asthma Medication Selection
JMIR Med Inform 2024;12:e56572
URL: https://medinform.jmir.org/2024/1/e56572
doi:10.2196/56572
PMID:38630536

Viewpoint

AI: Bridging Ancient Wisdom and Modern Innovation in Traditional Chinese Medicine

Linken Lu[1]; Tangsheng Lu[2]; Chunyu Tian[1]; Xiujun Zhang[3], PhD

[1]North China University of Science and Technology, Tangshan, China

[2]National Institute on Drug Dependence and Beijing Key Laboratory of Drug Dependence Research, Peking University, Beijing, China

[3]School of Psychology and Mental Health, North China University of Science and Technology, Hebei Province, China

Corresponding Author:
Xiujun Zhang, PhD
School of Psychology and Mental Health
North China University of Science and Technology
21 Bohai Avenue, Caofeidian New Town, Tangshan
Hebei Province, 063210
China
Phone: 86 0315 8805970
Email: zhxj@ncst.edu.cn

Abstract

The pursuit of groundbreaking health care innovations has led to the convergence of artificial intelligence (AI) and traditional Chinese medicine (TCM), thus marking a new frontier that demonstrates the promise of combining the advantages of ancient healing practices with cutting-edge advancements in modern technology. TCM, which is a holistic medical system with >2000 years of empirical support, uses unique diagnostic methods such as inspection, auscultation and olfaction, inquiry, and palpation. AI is the simulation of human intelligence processes by machines, especially via computer systems. TCM is experience oriented, holistic, and subjective, and its combination with AI has beneficial effects, which presumably arises from the perspectives of diagnostic accuracy, treatment efficacy, and prognostic veracity. The role of AI in TCM is highlighted by its use in diagnostics, with machine learning enhancing the precision of treatment through complex pattern recognition. This is exemplified by the greater accuracy of TCM syndrome differentiation via tongue images that are analyzed by AI. However, integrating AI into TCM also presents multifaceted challenges, such as data quality and ethical issues; thus, a unified strategy, such as the use of standardized data sets, is required to improve AI understanding and application of TCM principles. The evolution of TCM through the integration of AI is a key factor for elucidating new horizons in health care. As research continues to evolve, it is imperative that technologists and TCM practitioners collaborate to drive innovative solutions that push the boundaries of medical science and honor the profound legacy of TCM. We can chart a future course wherein AI-augmented TCM practices contribute to more systematic, effective, and accessible health care systems for all individuals.

(JMIR Med Inform 2024;12:e58491) doi:10.2196/58491

KEYWORDS

traditional Chinese medicine; TCM; artificial intelligence; AI; diagnosis

Introduction

Traditional Chinese medicine (TCM) is a vibrant and enduring medical system that has been refined for thousands of years, thus offering a rich tapestry of health and healing practices [1]. With its roots deeply embedded in Chinese philosophy and a profound understanding of the human body's relationship with the natural world, TCM has developed a unique set of diagnostic and therapeutic methodologies. These methodologies include herbal medicine, acupuncture, and the identification of syndrome patterns, which are grounded in the fundamental concepts of yin and yang and the 5-element theory, thus providing a holistic approach to health that addresses the body, mind, and spirit. As a traditional medical system, TCM primarily depends on personal experience and lacks standardized and systematic diagnosis and treatment procedures, which potentially discourages its more widespread adoption. Thus, the rapid expansion of artificial intelligence (AI) can significantly improve the reliability and accuracy of TCM diagnostics, thereby increasing the effective use of therapeutic methods for patients [2]. AI is currently being used in diagnostics to detect abnormalities in medical imaging, such as identifying lung nodules or other suspicious lesions in early cancer screening

[3]. These tools can assist physicians in making an initial diagnosis by analyzing large amounts of imaging data to identify potential signs of cancer and can also help physicians more accurately characterize suspicious lesions, including their shape, volume, histopathological diagnosis, disease stage, and molecular features. AI also has potential therapeutic value for the diagnosis of mental disorders such as major depressive disorder by successfully distinguishing patients from healthy controls through machine learning [4,5]. This characterization is critical for determining treatment options and predicting disease progression. However, applying AI to TCM exhibits a range of both opportunities and challenges. AI can enhance the diagnostic process by enabling physicians to make more accurate assessments of diseases and patient constitutions through the analysis of extensive TCM clinical data. It can also be instrumental in creating personalized treatment plans that are tailored to the unique conditions of each patient. Concurrently, the integration of AI into TCM raises significant concerns about patient privacy and data security. Although the establishment and learning of classification criteria still cannot eliminate subjectivity, efforts must be made to refine these processes. Addressing these issues necessitates the establishment of stringent ethical standards and robust privacy measures.

The successful integration of AI into TCM is contingent upon collaborative efforts across various disciplines, including medicine, computer science, and data science. Building interdisciplinary teams and fostering effective communication are crucial for driving innovation in this field. As technology evolves, it possesses the potential to revolutionize TCM practices provided that it is implemented with careful consideration of the ethical implications and needs of the TCM community. In the realm of TCM, diagnosis is a pivotal element wherein practitioners engage in comprehensive inquiry and conduct head-to-toe examinations of patients. This process is central to collecting health-related data. Such a diagnostic approach facilitates an in-depth evaluation of the patient's overall health status coupled with a nuanced understanding of the fundamental characteristics of the disease. TCM diagnosis is based on a holistic perspective and involves a comprehensive assessment of both physiological and psychological aspects of health rather than relying solely on objective diagnostic criteria such as molecular markers and physiological indicators. There are generally 4 main diagnostic methods in TCM: inspection, auscultation and olfaction, inquiry, and palpation. These facets of TCM have been widely accepted by TCM practitioners worldwide. This aspect of TCM is particularly pronounced for highly skilled physicians who are able to derive diagnostic insights through visual inspection alone, often without an explicit need for detailed procedural explication. Moreover, they believe that nonrational thinking, which encompasses implicit meanings, intuition, inspiration, and imagination, plays a vital role in TCM diagnosis and treatment [6]. In conjunction with the concept of the 4 diagnostic methods, practitioners assess and integrate various clinical data to investigate evidence, identify a disease's root causes, establish treatment strategies, assess treatment efficacy, and anticipate healing progress [7]. The integration of AI into TCM diagnosis respects and uses the intuition and experience of practitioners, thus serving as an auxiliary means of clinical research to evaluate and verify diagnostic results rather than replace human judgment. AI can be designed to analyze complex patterns in patient data, thereby augmenting traditional diagnostic methods. For example, machine learning models can be trained to recognize subtleties in patient inquiry responses, thus enhancing the practitioner's ability to identify the root causes of diseases and tailor treatment strategies. AI systems can be developed to consider the holistic approach that is inherent in TCM. By analyzing a wide range of clinical data, AI can provide a more comprehensive health assessment that aligns with the TCM concept of integrating various aspects of a patient's condition. This not only supports the practitioner's diagnostic process but also aids in anticipating healing progress and the efficacy of treatments. In addition, AI can tailor treatments to patients' individual needs and conditions by considering their unique body states and responses to various therapeutic interventions, thus leveraging the ability of AI to process and learn from large data sets as well as the enormous potential for personalized treatment in TCM. The role of AI in TCM is to augment the expertise of physicians, thus providing insights and analytics that support the holistic and personalized approach to health care that is at the heart of TCM. By doing so, AI can contribute to enhancing patient care and ensuring that the rich heritage of TCM is carried forward and developed in the modern health care era.

AI in TCM Diagnosis

Inspection

In the practice of TCM, the method of inspection primarily focuses on the acquisition of diagnostic information through direct observation. This approach involves assessing the patient's condition by scrutinizing various physical changes across the body. Distinct from the paradigm of Western medicine, which predominantly relies on objective, empirical evidence, TCM tends to base its diagnostic conclusions on subjective interpretations by medical practitioners [8]. The scope of inspection for diagnosis is quite broad. Although the methods comprise craniofacial observations, tongue and face diagnoses are the primary methods used for inspection. An inspection of the tongue's shape, size, color, and texture aids in the assessment of organ function and the development of medical conditions. Facial expression analysis is a diagnostic method that aligns with the theory of 5 zang organs, corresponding to 5 elements and colors. It involves distinguishing different changes in facial color, such as green, red, yellow, white, and black, based on the principles of yin and yang and the 5-element theory [9]. Building on this traditional foundation, technological advancements have introduced new methods to enhance TCM diagnostics. For example, the development by Chen [10] of a neural network–based system marked a significant advancement in this research. This innovative platform automates the diagnostic and treatment process in TCM with a focus on symptom analysis. It empowers physicians to efficiently access crucial medical records, thus providing valuable insights into the effectiveness of TCM treatments for similar conditions. Furthermore, the system streamlines the prescription process, thus enabling precise electronic prescriptions to be quickly dispatched to relevant departments and patients. This integration of AI with TCM improves diagnostic accuracy and facilitates

the sharing of expertise among practitioners, thus ultimately enhancing the standard of care in TCM. AI is currently one of the most discussed topics in medical imaging research. It serves as a significant enabler for handling vast amounts of medical images, thereby deciphering disease features that may be imperceptible to the human eye. Similarly, AI-based facial diagnosis and tongue diagnosis hold promise for further development. Recently, Liu et al [11] reviewed AI methods in the field of tongue diagnosis. They identified two main challenges that hinder development in this field: (1) the authority of data sets and (2) a misconception about a sole reliance on single features for diagnosis in traditional Chinese tongue diagnosis [11]. The combination of AI with this field overcomes the inherent subjectivity of TCM diagnosis and provides a more objective and standardized approach to tongue diagnosis. Technological advances such as multiscale features and the incorporation of previous knowledge have been successfully applied to improve the accuracy and reliability of AI-assisted tongue analysis. In addition, robust data sets and reliable performance evaluations are still needed to address existing problems in the field. The future of intelligent tongue diagnosis is promising, with potential breakthroughs in self-supervised methods, multimodal information fusion, and tongue pathology research that are expected to have a significant impact on research and clinical practice. On the basis of this scenario, we propose potential solutions to address these issues. First, standardizing data sets for tongue diagnosis should be a collaborative effort that involves experts in TCM. Second, leveraging multimodal data in AI is a crucial approach for the AI-driven transformation of TCM.

The future of AI in the inspection component of TCM is poised to transform traditional diagnostic practices through innovative research and practical applications. One of the primary research directions is the development of sophisticated AI algorithms that can analyze and interpret tongue and facial diagnostics at a level of detail that surpasses human perception. By training these algorithms on diverse and high-quality data sets, AI systems can learn to identify subtle patterns and changes that indicate underlying health conditions, thus complementing the expertise of TCM practitioners.

Auscultation and Olfaction

Auscultation and olfaction in TCM involve the use of a physician's hearing to detect changes in a patient's voice and sounds. Olfaction relies on the physician's sense of smell to detect changes in odors. The theoretical basis for these practices in TCM is the belief that a patient's speech sounds and body odors can reflect the physiological and psychological states of their internal organs. Consequently, auscultation and olfaction have long been highly regarded in the field of TCM. However, objective studies and literature on auscultation and olfaction are scarce, which may be attributed to the complex acoustic properties of sounds, including a plethora of natural noises, similar acoustic signals, and diverse chemical compositions of thousands of volatile organic compounds in exhaled gases. These factors have hindered the development of objective research on TCM auscultation and olfaction. Chiu et al [12] introduced quantifiable parameters for TCM auscultation, which allowed for the identification of nonvacuity, qi vacuity, and yin vacuity characteristics in participants. This quantification process enhances the practice of TCM auscultation. There is still a need for more quantitative data on auscultation and further advancements in the application of AI to analyze such quantitative data. The integration of AI into this method is facilitated through the use of advanced sensor technologies and audio analysis algorithms. For example, digital stethoscopes can capture and record bodily sounds with greater clarity. These sounds are then processed by AI algorithms that can filter out background noise and enhance the relevant audio signals. With regard to objective olfactory analysis, there have been several recent studies from a TCM perspective. A recent study introduced a new odor map with the ability to characterize odor quality that was comparable to that of highly skilled human "sniffers" [13]. The algorithms of these odor detection tools have significant potential for quantifying olfactory diagnosis in TCM. AI algorithms are then applied to data generated by these devices to identify specific volatile organic compound profiles that are associated with different health conditions. This is a complex task given the vast number of potential volatile organic compounds and their concentrations; however, machine learning models have shown the ability to handle this complexity and provide objective data for diagnosis. Therefore, the primary focus should be on building an AI odor monitoring system. Such a detection system can be developed by selecting specific biomarker reagents [14]. The development of diagnostic molecular biomarkers for olfaction diagnosis in TCM is also a substantial task. These biomarkers should be capable of quantifying the olfactory diagnostic process in TCM more accurately.

However, there are still some challenges in the application of AI to auscultation and olfaction, including challenges regarding data quality and standardization. The collection of auditory and olfactory data requires highly accurate sensors and devices. Data quality and standardization are critical for training accurate AI models. Inaccurate or inconsistent data can lead to misjudgments by the AI system. The second challenge involves the recognition of extremely complex sound and smell patterns that can be perceived differently among individuals. AI needs to be able to recognize and understand these complexities, which places high demands on the design and training of algorithms. We need to explore and develop more advanced sensor technologies to improve the accuracy and consistency of data collection and use deep learning techniques to improve the ability of AI models to recognize complex sound and odor patterns.

Inquiry

Interrogation diagnosis (or inquiry diagnosis) directly asks patients questions about various physiological and psychological feelings. This methodology includes gathering information about the patient's family history, primary complaints, living conditions, dietary habits, sleep patterns, and other physical condition characteristics. This process allows the practitioner to gain a comprehensive understanding of the patient's overall health, including factors that may contribute to their current condition. A thorough understanding of a patient can also avoid the influence of a previous medical history on treatment. The inquiry aims to provide a holistic view of the patient in

consideration of not only physical symptoms but also lifestyle and environmental factors that could impact their health. The content of TCM inquiries is mainly based on the "Ten Brief Inquiries"; however, at present, TCM inquiries also incorporate past history, allergy history, and family history from modern medical records [15]. The GatorTron system, which was developed by Yang et al [16], enhances the use of clinical narratives in the creation of various medical AI systems, thus ultimately leading to better health care delivery and health outcomes. However, electronic health record (EHR) analysis for TCM inquiry is not yet well developed and primarily relies on natural language extraction techniques to extract electronic medical record data, which are then used to establish a knowledge repository for traditional Chinese clinical cases. For example, AI systems are capable of identifying TCM-specific symptoms such as "fatigue" and "dry mouth" from patient narratives, thus correlating these symptoms with associated internal organ imbalances. This sophisticated recognition aids physicians in assessing patients' constitutions and developing personalized treatment plans. For individuals with chronic conditions, AI facilitates a more in-depth analysis by sifting through extensive health records to forecast disease progression, thereby providing physicians with a solid foundation for accurate diagnoses. Moreover, AI extends its support to patients who require ongoing care by offering tailored advice on diet and exercise, thus significantly contributing to the enhancement of their quality of life and the mitigation of relapse risks. The advent of smart wearables has further empowered AI by enabling real-time health data collection, which is swiftly relayed to AI for analysis. This system proactively notifies health care providers and patients about emerging health concerns, thus exemplifying the potential of AI in diagnostics and proactive patient care within the framework of TCM.

In the future, it will be essential to confirm the accuracy of large language models (LLMs) such as GPT-3.5 and GPT-4 in TCM diagnosis [17]. This process requires a nuanced approach that acknowledges the complexity and richness of TCM terminology. The first step is to collect comprehensive patient data, including symptoms, medical history, lifestyle factors, and any other relevant information. These data must be preprocessed to ensure that they are suitable for AI analysis. AI models, especially LLMs, are trained in neurolinguistic programming to understand and interpret human language. In the context of TCM, this involves training models to recognize and analyze specific terminology that is used in patient inquiries. Afterward, AI models must be trained to understand the context in which TCM terms are used. This includes recognizing relationships between different symptoms and their implications with regard to overall health according to TCM principles. However, TCM is practiced worldwide, and patient inquiries may also be in various languages or dialects. AI models need to be trained on diverse data sets to ensure that they can handle different languages and cultural interpretations of TCM terms. A significant amount of labeled data and expert input are subsequently required for validation. Collaborations with TCM practitioners to annotate and validate data can improve the model's accuracy. In summary, although AI with LLMs shows significant promise for managing EHRs, TCM inquiry demonstrates a unique knowledge system. The fine-tuning of LLMs is essential for transforming these general-purpose models into specialized models that are adept at handling TCM EHRs [18]. Future efforts should entail constructing a knowledge system for TCM diagnosis. It will then be necessary to fine-tune LLMs for TCM diagnosis based on existing LLM data models, thus providing AI tools for case analysis in TCM diagnosis. The integration of AI into the TCM inquiry process is a complex task that requires the careful consideration of unique aspects of TCM terminology and practice. With the right approach, including ongoing research and collaboration with TCM experts, AI can be effectively used to analyze patient data and enhance the diagnostic process in TCM.

Palpation

Pulse diagnosis is one of the 4 main pillars of TCM assessment. By palpating the pulse at 3 specific positions on the wrists ("cun," "guan," and "chi"), practitioners can gain a comprehensive understanding of a person's overall health and the state of specific organs. TCM pulse diagnosis consists of approximately 29 different pulse types that encompass a range of descriptors, including floating pulses and scattered pulses [19]. The intersection of pulse diagnosis and AI presents 2 main challenges. TCM pulse detection has historically relied on manually palpating the arteries beneath the skin to detect the pulse, thus lacking objective standards. In the process of AI-driven traditional Chinese pulse diagnosis, 2 critical issues need to be addressed: the development of pulse measurement devices and the standardization of pulse detection data. Lan et al [20] created a sensing device that features a multipoint sensor to measure pulse. Due to the complexity of pulse detection, previous methods that have primarily relied on multipoint sensors have only offered a limited scope of information. The development of pulse measurement devices has led to significant technological advances in recent years, and these advances are mainly reflected in innovations in sensor technology and the application of AI algorithms. Photovoltaic volumetric pulse wave sensors, which are based on photoplethysmography, are among the most common types of sensors used in pulse measurement devices. Pulse waves are measured by detecting the flow of blood in the microvasculature to obtain physiological parameters such as heart rate [21]. Photoplethysmographic sensors, such as smartwatches and fitness trackers, are widely used in consumer electronics. Some devices use pressure sensors to measure pulse waves, especially in continuous blood pressure monitoring. These sensors are often embedded in wearable devices that can monitor changes in blood pressure over time. To address the challenge of normalizing pulse data, AI algorithms preprocess the data before analysis, including filtering, denoising, and normalization, to ensure data quality. AI technology that is currently under development is working to improve the cross-device compatibility of algorithms so that data from devices from different manufacturers and models can be consistently analyzed, thus promoting data standardization and interoperability. The standardization of TCM pulse diagnosis is key to promoting the use of AI technology in TCM pulse measurements. It is necessary to establish unified pulse data and diagnostic standards along with integrating more diagnostic methods such as tongue diagnosis and diagnostic observation, from which we can develop an integrated TCM

diagnostic platform and improve the comprehensiveness and accuracy of diagnosis.

In the future, more powerful multipoint sensing devices and multimodal detection devices will be needed to comprehensively examine pulse data and achieve better quantification. A challenge still remains in determining whether pulse data from these detectors can adequately reflect the characteristics of pulse diagnosis and in improving the classification of pulse patterns. To address the challenge of enhancing the precision of AI in interpreting pulse data for future research and development, noncontact pulse measurement techniques have demonstrated significant advancements. These methods eliminate the need for physical contact with the patient, which is particularly crucial for monitoring in unique or sensitive situations. For instance, the polarized multispectral imaging technique for noncontact heart rate measurement has refined the accuracy of data acquisition by pioneering new methods for extracting pulse waves from the palm [22]. This innovation contributes to the establishment of a standardized framework for pulse data, thus facilitating seamless data sharing and comparisons across various devices and systems. However, the attainment of high-quality data hinges on precise labeling, which is a process that can be both labor intensive and costly. In the context of electrocardiogram data annotation, the requirement for specialized physicians introduces variability, in which different medical professionals may offer conflicting assessments. This reality compounds the complexity and challenges associated with data preprocessing. To overcome these obstacles, it is imperative to refine data annotation protocols and invest in the development of more efficient and accurate labeling tools. By doing so, we can ensure that AI systems are trained on the most reliable data, thereby improving their diagnostic capabilities and contributing to the advancement of AI in health care. In summary, the development of detection methods and quantification of detection-based data are bottlenecks in the process of AI-driven pulse diagnosis.

AI-Powered Tuina Massage Robot

Tuina massage (also known as Chinese medical massage) is a traditional hands-on manipulation treatment that is guided by the principles of TCM. It is widely used to treat various ailments, such as knee osteoarthritis, chronic neck pain, and insomnia [23-25]. The tuina massage serves 3 primary functions: facilitating the circulation of meridians, harmonizing qi and blood circulation, and augmenting the immune system [26-28]. These functions are essential for disease prevention and treatment and overall well-being. The integration of AI into tuina massage therapy is in its early stages. Efforts are underway to develop highly intelligent massage equipment and robotics based on TCM tuina to enhance its effectiveness, with a focus on improving the comfort, intelligence, and safety of massage robots [29]. Vibration and percussion are the 2 main types of tuina massage robotics. These devices offer acupressure techniques; however, manual massage from experienced physiotherapists provides additional popular movements, such as light stroking, stretching, and advanced kneading techniques, that machines cannot replicate. Therefore, the development of massage robots is a significant research focus for greater health care demands. Challenges mainly exist in their control, structure, and path planning; however, ongoing efforts aim to optimize their design and functionality. For example, the incorporation of a series-parallel hybrid structure may enhance flexibility while maintaining stiffness and precision [30]. Future research should focus on ergonomics to design high-performance massage robots that integrate advanced AI technologies for better control, sensing, and essential functions.

In current TCM tuina practice, the Expert Manipulative Massage Automation (EMMA) electronic massager, which was developed by AiTreat Pte Ltd in Singapore, is widely used. To deliver precise and effective massage based on muscle feedback, EMMA uses advanced sensor-based technology to identify focus points and adjust pressure levels. By detecting stiffness and resistance, EMMA can pinpoint muscle knots and tension points, thus applying varying pressure levels based on feedback and user preferences. In addition, EMMA incorporates Internet of Things technology for remote control, programming, and updates, thus enhancing its functionality in “green” Internet of Things applications. In EMMA technology, machine learning algorithms (especially convolutional neural networks in deep learning) are used to identify and analyze muscle tension patterns, acupuncture point locations, and physiological responses of patients. Through training, these algorithms are able to identify specific treatment points from sensor data to provide a customized massage solution for the patient. This pattern recognition capability allows the robotic massage therapist to pinpoint the area to be treated, thus mimicking the diagnostic process of an experienced massage therapist. The robotic masseur is able to adjust the intensity and speed of the massage based on real-time feedback from the patient. For example, if the sensors detect that a patient is experiencing discomfort at a certain pressure level, then the AI system can immediately adjust the intensity to ensure the comfort and effectiveness of the treatment. Through its advanced data analytics and learning capabilities, the AI application in EMMA technology is able to accurately identify treatment points and adjust massage intensity based on the patient’s real-time feedback. The EMMA massager has garnered high levels of acceptability and satisfaction among healthy volunteers, thus demonstrating its feasibility [31]. Nonetheless, research on massage robots still faces challenges, particularly regarding their clinical effectiveness. In addition, massage robots are categorized as class-I medical devices that do not require Food and Drug Administration approval for marketing in the United States [32]. Traditional medical device classification focuses on physical and biological characteristics, whereas the functionality of AI devices relies more on software and algorithms. Therefore, new classification criteria need to be developed that consider the specificities and potential risks of AI technologies. In the future, there will be a need for more standardized regulations to oversee research on massage robots [33]. Medical devices process and analyze large amounts of patient data, which requires regulations to include stringent requirements for data security and privacy protection. Medical device regulations need to incorporate specifications for data collection, storage, processing, and transmission to ensure the security and confidentiality of patients’ information, including the validation and clinical testing of AI algorithms. The use of AI medical devices involves the collection and analysis of large

amounts of personal health information, which can threaten patients' privacy. Regulations must ensure that the collection and use of patient information comply with privacy protection standards to prevent unauthorized access and data breaches. This will entail validating the functionality and therapeutic efficacy of medical devices to guarantee their safety and efficacy for users.

We propose the following perspective on the development of tuina robots. First, the overall stability of the tuina technique involves the stability of variable mechanical parameters and resulting morphological changes in mechanical effects during technique operation. These factors include mechanical characteristics such as force; speed; frequency; displacement; and kinematic features such as limb range of motion, joint angles, and overall movement amplitude. For example, the dexterity of the robotic arm is key to achieving an accurate simulation of a human masseur's maneuvers; however, it requires sophisticated mechanical design, including joint flexibility, end-effector versatility, and overall structural stability. The robot arm's control system also needs to process large amounts of data and make decisions in real time. This includes trajectory planning, motion control, and complex algorithms for force and position control. To ensure safety, collision detection and response mechanisms also need to be implemented. Consequently, tuina has significant limitations and subjectivity, thus making it difficult to objectively quantify and accurately assess efficacy. AI offers unique advantages in addressing this issue, which is primarily manifested in the digitization of tuina techniques (ie, the development of precision and flexibility in massage robots). Addressing the accuracy of the tuina technique is a prominent issue that may require more diverse AI algorithms to digitize massage techniques and analyze the clinical effects of different massage methods. The accuracy of Chinese massage largely depends on the precise positioning of acupoints. Researchers are developing a human body model based on the mechanism of "bone degree and minutes" in Chinese medicine, which realizes the calculation of 3D coordinate values of acupoints through robotics as well as identifying and tracking human body features by using such sensor technologies as depth cameras, thus realizing the precise positioning of acupoints. Second, another advantage of AI includes personalized health care services. The personalized parameter settings of massage robots are core parameters for future tuina robots. There are significant differences in individuals' sensitivity and tolerance to pressure. The comfort and pain thresholds of people can vary, thus significantly affecting their experience with massage robots. AI can analyze the user's physical condition, health data, and personal preferences to design a personalized massage program. Through the integration of advanced intelligent sensors, massage robots can monitor users' physiological responses, such as muscle tension and body temperature, in real time. According to these data, massage robots can adjust their massage strength, speed, and focus area to overcome the "subhealth pain problem" of accurate positioning and efficient massage. At present, there are few studies on the clinical effectiveness of AI nudging robots, and patient self-reported changes in pain level, duration of pain relief, reduction in drug dependence, and objective measures of mobility (such as gait analysis) will be important indicators for evaluating their effectiveness in the future. When considering factors such as safety and comfort, AI data recording and analysis can also be used to measure the clinical efficacy of tuina robots.

AI-Directed Acupuncture Manipulation

Acupuncture, which is a therapeutic technique in TCM that has been practiced for thousands of years, has gained widespread global acceptance and demonstrated significant efficacy for various chronic diseases, particularly pain-related conditions. This therapeutic approach involves stimulating specific areas, known as acupoints, on the patient's body, thus eliciting sensations such as soreness, numbness, fullness, or heaviness, which is commonly referred to as "De Qi" or achieving qi [34]. Due to the inherent subjectivity and reliance on experience in traditional acupuncture practices, there is growing interest in parameter-based electroacupuncture to address these limitations [35]. By setting different parameters using an electroacupuncture device, clinical efficacy can be enhanced, thus facilitating further research. However, the efficacy of acupuncture is still not universally recognized [36], possibly for 2 main reasons. First, the inadequate design and implementation of past clinical research methods have led to a lack of clinical evidence. Second, the mechanism of acupuncture remains unclear, thus necessitating more high-quality evidence to elucidate its biological mechanisms for informed clinical decision-making [37]. The integration of AI and acupuncture shows great potential for substantially improving the precision of acupuncture prescriptions and treatment techniques. A bibliometric study by Zhou et al [38] demonstrated substantial progress in AI research within the acupuncture field over the past 2 decades, with significant contributions from the United States and China. However, the application of AI in acupuncture lacks a clear framework, with a scarcity of systematic research and a lack of organization of relevant technologies and application approaches.

Given the unique characteristics of AI and the importance of data mining in clinical acupuncture practice and manipulation, further research is needed [39]. Clinical trials are costly and limited, and most articles that analyze the safety and efficacy of acupuncture are of low quality and lack comprehensive analyses. There is still a lack of standardized acupuncture point selection protocols for many diseases [40]. Therefore, future efforts should focus on standardizing TCM while improving the quality of randomized controlled trials on acupuncture to obtain more and higher-quality clinical data, thus providing a foundation for AI-based clinical data mining. AI can analyze a patient's symptoms, signs, and physiological data and compare them to a large body of medical knowledge. Through machine learning and pattern recognition algorithms, AI can help clinicians interpret diagnostic data and provide potential pathological patterns or disease classifications that can help acupuncturists in developing treatment strategies. AI can then be used to analyze large amounts of clinical data and research the literature to determine the most effective point selection for a particular condition or disease situation. A recent study "linked" original studies and 332 systematic evaluations of evidence in 20 disease areas by using AI analysis techniques to

comprehensively improve clinical evidence for acupuncture therapy in the Epistemonikos database, which constructed a total of 77 evidence matrices [41]. This will facilitate the development of a machine learning framework to predict the efficacy of acupuncture and patient prognosis. Acupuncture manipulation techniques are crucial components of acupuncture therapy, and their efficacy is paramount [42]. However, the determination of the optimal stimulation intensity in clinical research is often challenging because of technique selection, treatment duration, needling speed, and force [43-45]. Therefore, the quantification and standardization of acupuncture manipulation, such as needle insertion force, duration, and direction, are essential for achieving clinical efficacy and AI-guided acupuncture manipulation. In response to the problems in standardizing operation techniques, AI technologies, especially machine learning models and sensor technologies, are being used to capture and analyze the nuances of manual needling operations. Acupuncture robots that are currently under development can accurately gauge the location of acupuncture points by measuring a person's height and sebum thickness and use ultrasound sensors to control the depth and speed of needling. These sensors and machine learning models are able to identify key parameters such as the needling force, speed, and angle to ensure the standardization and consistency of treatment. The application of sensor technology in acupuncture focuses on the precise control and measurement of the depth, force, and speed of needling. This robot uses an ultrasound sensor to control the depth and speed of needling. By emitting ultrasonic waves and receiving their echoes, the ultrasonic sensor can accurately measure the distance between the tip of the needle and the surface of the tissue to ensure that the depth of the needles is appropriate and avoid unnecessary injury to the patient. Through built-in mechanical sensors, the robot can also automatically adjust the needle insertion process according to changes in needle insertion resistance to ensure safe needle insertion.

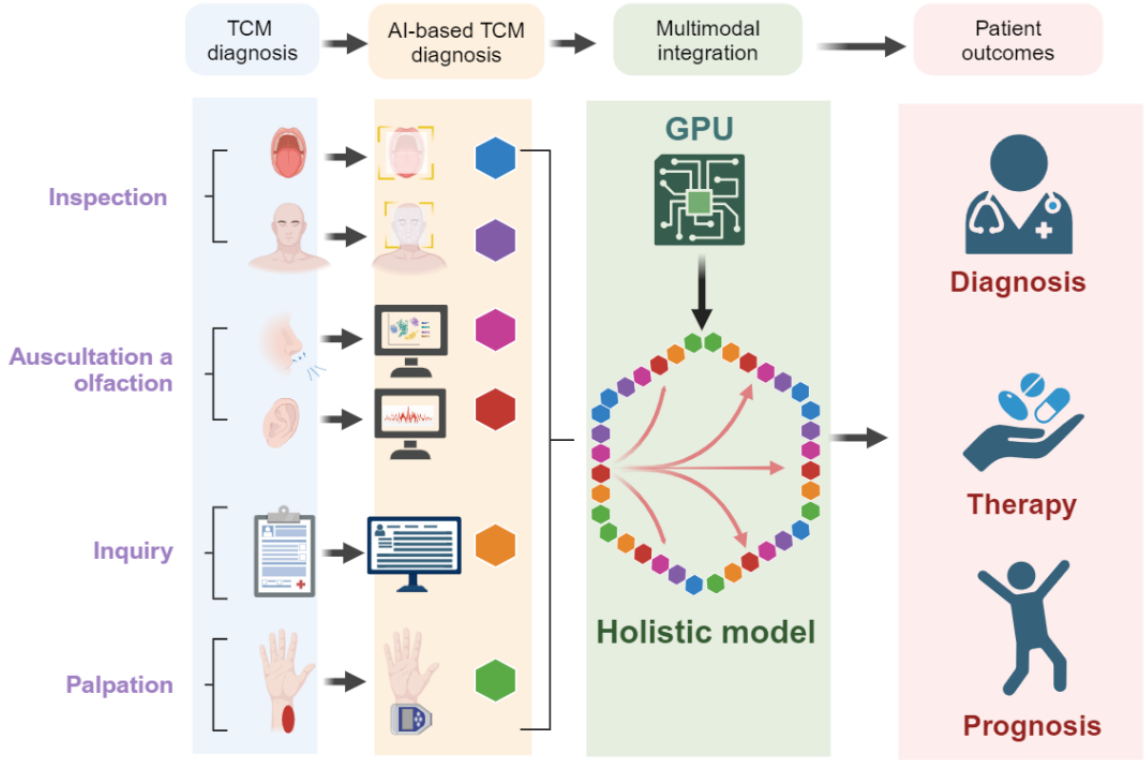
The standardization of acupuncture manipulation forms the basis of the use of AI in acupuncture. With AI technology, we propose three different ways to help standardize acupuncture: (1) imaging recognition–based standardization of acupuncture practitioners' techniques, (2) analysis of parameters derived from acupuncture practitioners' lifting and thrusting techniques using neural network image analysis systems, and (3) extraction of spatiotemporal features from video images of acupuncture operations by using computer vision technology [46]. In addition, a hybrid model that combines 3D convolutional neural networks and neural networks is used to recognize and classify dynamic hand gestures in acupuncture operation videos, thus enabling quantitative analyses and technical inheritance research for various techniques. Another approach involves recording acupuncture practitioners' movements and mechanical parameters during acupuncture procedures by using 3-axis posture sensors. Davis et al [47] developed force and motion sensor technology (acusensors) to quantify the linear and rotational movements of acupuncture needles and the force and torque that are generated during manual needle manipulation. A standardized TCM acupuncture manipulation database was established for the quantification of motion and force patterns. These data serve as a crucial tool for future AI applications in acupuncture. Finally, acupuncture parameters based on other electrophysiological signals have been recorded, thus showing significant differences in electrophysiological signals between acupuncture points and nearby nonacupuncture points and highlighting the electrical specificity of acupoints [48,49]. This finding serves as compelling evidence for TCM theory and provides parameters for the standardization of acupuncture stimulation. In addition, collaboration between AI experts, acupuncturists, and biomedical engineers is essential for developing and improving acupuncture-related technologies. The data analysis and intelligent algorithms that are provided by AI experts can help acupuncturists better understand treatment effects and optimize treatment plans. The clinical experience and theoretical knowledge of acupuncturists can guide AI experts in developing intelligent systems that better meet clinical needs. Moreover, there are some prominent conditions or events existing outside of normal circumstances that exist beyond the abilities of AI. In such cases, acupuncturists can make judgments based on their own experience and knowledge. Technical support from biomedical engineers subsequently ensures that these intelligent systems can be effectively applied in practice. Close collaboration among the 3 factors is the key to promoting technological innovation in acupuncture, improving treatment outcomes, and standardizing and popularizing acupuncture techniques. Through this interdisciplinary collaboration, modern technology can be better used to enhance the value and impact of traditional acupuncture medicine. AI technology can be used to simulate acupuncture operations and provide support for learning and training. For example, through virtual reality and augmented reality technologies, AI can create simulated acupuncture treatment scenarios that allow learners to practice acupuncture techniques and decision-making processes in a virtual environment. This approach improves learning efficiency and reduces risks in actual practice.

Outlook

The future is promising for the integration of AI into TCM diagnosis (Figure 1). The core of TCM diagnosis involves 4 fundamental methods: inspection, auscultation and olfaction, inquiry, and palpation. First, through inspection, we can explore the application of AI in facial and tongue diagnosis. By leveraging image processing techniques, we anticipate groundbreaking innovations in these areas. Due to the low contrast and noisy nature of ultrasound images, which require extensive knowledge of tongue structure and ultrasound data interpretation, it can be challenging for medical professionals to accurately diagnose various conditions. To overcome these challenges, researchers have proposed new deep neural networks, referred to as BowNet models [50]. These models leverage the global prediction capabilities of encoder-decoder fully convolutional neural networks and the high-resolution extraction features of dilated convolutions. The models are designed to automatically extract and track the tongue contour in real-time ultrasound data. By providing more accurate and objective tongue contour tracking, AI can assist TCM practitioners in making more informed diagnoses and treatment decisions. However, the digitization of valuable empirical data

from TCM facial and tongue diagnosis remains a challenge, although unsupervised machine learning may provide a solution. First, addressing the challenge of digitizing TCM experience data often requires seamless collaboration among diverse teams, which encompasses TCM practitioners, AI experts, and health care institutions. This initiative involves partnering with medical facilities to meticulously gather and curate a rich array of TCM clinical case data, thus encompassing detailed patient complaints, symptomatic expressions, tongue diagnoses, and pulse assessments. Subsequently, synergy between TCM professionals and AI researchers has extended to the intricate task of knowledge structuring. The translation of the profound insights of TCM theories, herbal formulas, and medicinal properties into a web-based knowledge graph can greatly enhance the accessibility and utility of this ancient medical wisdom. This graph facilitates a nuanced understanding of intricate relationships within TCM, underpins the development of intelligent recommendation systems, and assists in informed decision-making within the realm of TCM. Through these collaborative efforts, the rich tapestry of TCM experience data is meticulously incorporated into accessible, digital formats, thereby bridging the gap between ancient wisdom and modern technological advancements. These initiatives have propelled the modernization and globalization of TCM and significantly contributed to the broader landscape of health care innovation. Second, TCM diagnosis through auscultation and olfaction provides new avenues for development. We can develop specialized AI tools for auditory and olfactory systems that can complement TCM diagnosis. Third, the analysis of inquiry in TCM diagnosis presents an exciting frontier. There has been significant progress in Western medicine in this area with regard to LLMs, and TCM EHRs contain unique knowledge graphs that are specific to TCM diagnosis. Therefore, fine-tuning is necessary for the analysis of TCM EHRs. Fourth, pulse diagnosis is a cornerstone of TCM diagnosis and requires the development of more advanced pulse detection tools. By obtaining substantial pulse data, we can standardize objective pulse analysis. By leveraging machine learning and classification techniques, we can align these data with TCM pulse patterns, thus ultimately achieving AI-driven pulse diagnosis. Quantifiable metrics or benchmarks are instrumental in ensuring the high performance of AI systems, establishing unambiguous objectives, and measuring advancements in the integration of AI technologies. For example, metrics such as diagnostic accuracy and generalization capability can assess AI systems' proficiency in managing patient data across varying regions, age groups, and sexes, thereby ensuring their efficacy across a wide array of populations. By juxtaposing outcomes that are generated by AI systems with diagnoses that are rendered by seasoned TCM experts, the precision of AI systems in pinpointing specific conditions can be quantified. This quantification is facilitated through the computation of statistical indicators such as sensitivity (true positive rate), specificity (true negative rate), precision, and the F_1-score. These benchmarks serve as a foundation for the continuous refinement of AI systems, thus ensuring their optimal integration and application within the sphere of TCM. These directions of development have the potential to revolutionize TCM diagnosis, thus enhancing its accuracy and efficiency.

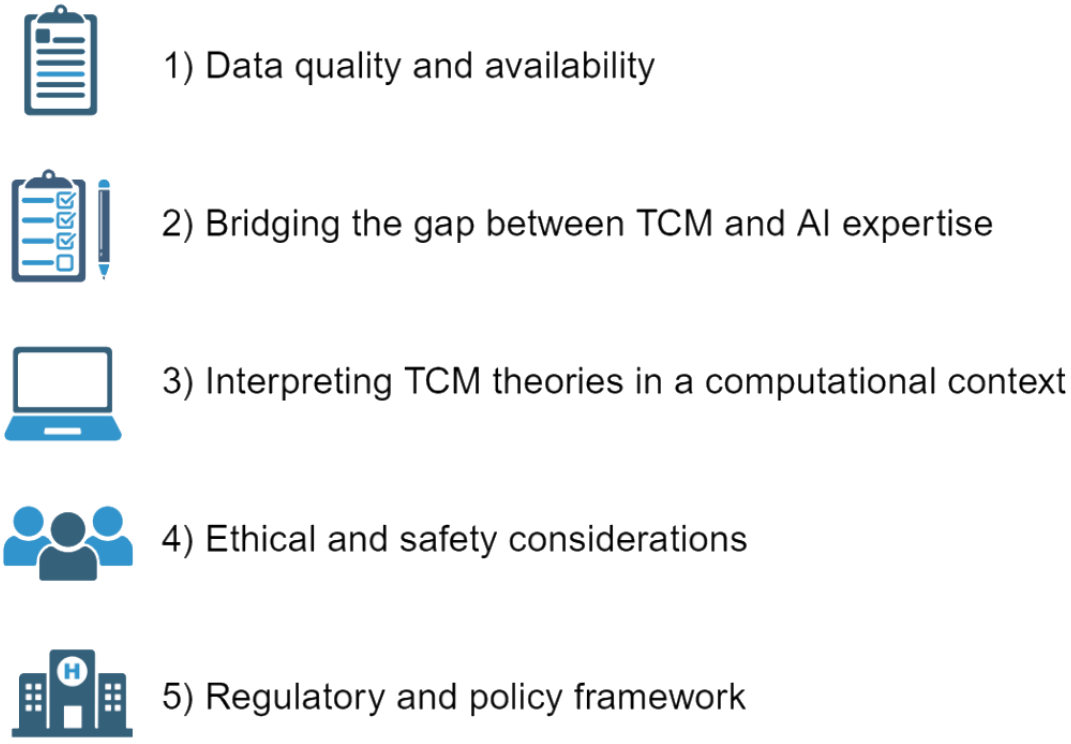
Figure 1. Overview of artificial intelligence (AI) development strategies based on traditional Chinese medicine (TCM) diagnosis. The acquisition and standardization of unimodal data through TCM diagnostic techniques is followed by the integration of multimodal data using a comprehensive model. This approach aids in enhancing predictions and supports TCM diagnoses for treatment and prognosis. GPU: graphics processing unit.



Challenges

There are unique challenges to the use of AI in TCM (Figure 2). First, regarding data quality and availability, the successful implementation of AI in TCM relies on access to reliable and standardized data sets. High-quality data can potentially promote clinical diagnosis and treatment in precision TCM. However, data collection and digitization efforts in TCM can be challenging, and the quality and interoperability of existing data sets may vary. Varied interpretations of identical conditions among TCM practitioners can result in divergent diagnostic criteria and terminological applications. Such disparities, coupled with the potential for errors and biases in manually entered data, can significantly impact the learning efficacy of AI systems. Consequently, the establishment of standardized TCM diagnostic criteria and terminology glossaries, in addition to the implementation of uniform data entry protocols for TCM practitioners, is essential for mitigating interpractitioner discrepancies. Furthermore, the use of advanced automated data collection techniques, including image recognition and natural language processing, is instrumental in enhancing the quality and precision of the collected data. These measures collectively contribute to the refinement of the analytical capabilities of AI within the realm of TCM, thus ensuring a more accurate and reliable diagnostic process. Second, to bridge the gap between TCM and AI expertise, the integration of AI technologies into TCM requires collaboration and communication among TCM practitioners and AI experts. Bridging the gap between these domains is crucial for developing AI algorithms that align with TCM principles and meet specific clinical needs. Measures could be taken to promote collaboration between TCM organizations and technology companies, as well as higher-education institutions, to facilitate the development of AI-driven TCM diagnostic and therapeutic tools. These collaborations will foster innovation and create platforms for TCM students and practitioners to perform internships in the technology industry. In addition, the provision of scholarships and research grants is critical for incentivizing and sustaining interdisciplinary scholarship. By allowing students and researchers to delve into the convergence of TCM and AI, we can accelerate the digitization of TCM knowledge. Third, TCM theories must be interpreted in a computational context. TCM theories are often complex and based on holistic and individualized perspectives. The translation of this knowledge into AI algorithms and computational models is a significant challenge that requires careful consideration of cultural, philosophical, and theoretical aspects. Many concepts in Chinese medicine, such as qi, yin and yang, and the 5 elements, are abstract and ambiguous, and it is difficult to describe these concepts in precise mathematical language; therefore, ambiguous logic can be used to address ambiguous concepts in TCM, and Bayesian networks can be used to simulate causality and uncertainty in the theory of Chinese medicine. Fourth, there are notable ethical and safety considerations regarding this scenario, as with any implementation of AI in health care [51]. Ensuring patient privacy, data security, and transparency in algorithm decision-making is essential for building trust and ethical practices in AI-supported TCM. Informed patient consent must be obtained before collecting and using patient data. This includes a full explanation of the purpose of data collection, how the data will be used, how long they will be stored, and potential risks. To protect patients' privacy, all the data sets that are used for machine learning should be anonymized by removing or encrypting any personally identifiable information. During the development and deployment of AI systems, ethical review committees need to be established to ethically review the design, implementation, and evaluation of AI systems to ensure that all the activities meet ethical standards. Fifth, the integration of AI into TCM necessitates clear regulatory frameworks and policies that govern its implementation, including issues related to data protection, algorithm validation, and clinical decision-making. Sixth, the function and efficacy of TCM are broadly accepted worldwide; however, the underlying mechanism has remained enigmatic, thus limiting people's confidence in TCM and precision therapy that is learned by AI. In summary, the process of implementing and validating AI tools in a clinical setting requires careful planning and rigorous execution. Representative and actionable clinical environments are selected to develop pilot projects. These projects should focus on specific TCM diagnostic tasks, such as tongue analysis, pulse recognition, and symptom assessment. Clinical trials need to be designed and executed to evaluate the performance of AI tools in real clinical settings. This includes randomized controlled trials and prospective cohort studies to assess the impact of AI tools on patient outcomes. Finally, the results and experiences will be published and shared through academic journals and conferences to promote communication and learning within the industry.

Figure 2. Summary of the challenges of integrating artificial intelligence (AI) into traditional Chinese medicine (TCM) diagnosis.



Conclusions

In conclusion, the integration of AI into TCM exhibits immense promise for improving diagnosis, including inspection, auscultation and olfaction, inquiry, and palpation. The successful integration of AI into TCM is evident through advancements in areas such as image analysis for tongue diagnosis, the development of intelligent tuina massage systems, and the application of machine learning to refine treatment protocols based on individual patient data. Addressing the challenges of data quality, the standardization of data sets, interdisciplinary collaboration, the interpretation of TCM theories, ethical considerations, and regulatory frameworks is crucial for the successful and responsible implementation of AI in TCM. By overcoming these challenges, we can leverage the power of AI to enhance patient care, personalize treatments, and advance our understanding of TCM. Moreover, we can develop more precise AI models that are tailored to TCM, thus creating a positive cycle of problem-solving and progress that ultimately leads to better patient care. By combining the wisdom of TCM with the power of AI technology, we can improve patient outcomes and promote the integration of TCM into modern health care systems. It is imperative to conduct more research into AI's ability to decode complex diagnostic patterns that are inherent to TCM. The validation of AI-enhanced TCM treatment methods through clinical trials is essential to ensure their safety and efficacy, thus providing empirical support for their widespread adoption. As we advance the integration of AI into TCM, it is vital to uphold ethical standards that prioritize patient rights, cultural integrity, and data privacy. The responsible use of AI will ensure that technological advancements align with the principles and practices of TCM, thus safeguarding the well-being of patients and respecting the cultural significance of this ancient medical system. The fusion of AI with TCM has the potential to bridge traditional and modern medical practices, enrich global health, and foster cultural exchange. By integrating these 2 domains, we can create a more comprehensive health care system that is both innovative and respectful of historical practices.

Finally, a call to action is made to all stakeholders (practitioners, researchers, policy makers, and investors) to collaborate and support the integration of AI into TCM. Through collective efforts, we can harness AI to transform patient care, broaden our understanding of TCM on a global scale, and identify new horizons in health care that are both deeply rooted in tradition and boldly futuristic.

Acknowledgments

This research was funded by grants from the Ministry of Science and Technology of China and the National Key R&D Program of China (2023YFC2506802). The images in Figures 1 and 2 were created using BioRender [52].

Authors' Contributions

All the authors participated in the conceptualization, methodology, validation, and writing of the manuscript.

Conflicts of Interest

None declared.

References

1. Wang Y, Shi X, Li L, Efferth T, Shang D. The impact of artificial intelligence on traditional Chinese medicine. Am J Chin Med 2021;49(6):1297-1314. [doi: 10.1142/S0192415X21500622] [Medline: 34247564]
2. Zhang S, Wang W, Pi X, He Z, Liu H. Advances in the application of traditional Chinese medicine using artificial intelligence: a review. Am J Chin Med 2023;51(5):1067-1083. [doi: 10.1142/S0192415X23500490] [Medline: 37417927]
3. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 2019 Mar;69(2):127-157 [FREE Full text] [doi: 10.3322/caac.21552] [Medline: 30720861]
4. Bondi E, Maggioni E, Brambilla P, Delvecchio G. A systematic review on the potential use of machine learning to classify major depressive disorder from healthy controls using resting state fMRI measures. Neurosci Biobehav Rev 2023 Jan;144:104972. [doi: 10.1016/j.neubiorev.2022.104972] [Medline: 36436736]
5. Sun J, Dong QX, Wang SW, Zheng YB, Liu XX, Lu TS, et al. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian J Psychiatr 2023 Sep;87:103705. [doi: 10.1016/j.ajp.2023.103705] [Medline: 37506575]
6. Wang Z, Wang D, Liu W, Wang Z. Traditional Chinese medicine diagnosis and treatment based on systematics. iLIVER 2023 Dec;2(4):181-187. [doi: 10.1016/j.iliver.2023.08.004]
7. Maciocia G. Diagnosis in Chinese Medicine: A Comprehensive Guide. Amsterdam, The Netherlands: Churchill Livingstone; 2004.
8. Zhang YH, Lv J, Gao W, Li J, Fang JQ, He LY, et al. Practitioners' perspectives on evaluating treatment outcomes in traditional Chinese medicine. BMC Complement Altern Med 2017 May 18;17(1):269 [FREE Full text] [doi: 10.1186/s12906-017-1746-8] [Medline: 28521826]
9. Lu LM, Chen X, Xu JT. Determination methods for inspection of the complexion in traditional Chinese medicine: a review [Article in Chinese]. Zhong Xi Yi Jie He Xue Bao 2009 Aug;7(8):701-705. [doi: 10.3736/jcim20090801] [Medline: 19671406]
10. Huang Z, Miao J, Chen J, Zhong Y, Yang S, Ma Y, et al. A traditional Chinese medicine syndrome classification model based on cross-feature generation by convolution neural network: model development and validation. JMIR Med Inform 2022 Apr 06;10(4):e29290 [FREE Full text] [doi: 10.2196/29290] [Medline: 35384854]
11. Liu Q, Li Y, Yang P, Liu Q, Wang C, Chen K, et al. A survey of artificial intelligence in tongue image for disease diagnosis and syndrome differentiation. Digit Health 2023 Aug 06;9:20552076231191044 [FREE Full text] [doi: 10.1177/20552076231191044] [Medline: 37559828]
12. Chiu CC, Chang HH, Yang CH. Objective auscultation for traditional Chinese medical diagnosis using novel acoustic parameters. Comput Methods Programs Biomed 2000 Jun;62(2):99-107. [doi: 10.1016/s0169-2607(00)00055-9] [Medline: 10764936]
13. Lee BK, Mayhew EJ, Sanchez-Lengeling B, Wei JN, Qian WW, Little KA, et al. A principal odor map unifies diverse tasks in olfactory perception. Science 2023 Sep;381(6661):999-1006. [doi: 10.1126/science.ade4401] [Medline: 37651511]
14. Fu W, Xu L, Yu Q, Fang J, Zhao G, Li Y, et al. Artificial intelligent olfactory system for the diagnosis of Parkinson's disease. ACS Omega 2022 Jan 26;7(5):4001-4010 [FREE Full text] [doi: 10.1021/acsomega.1c05060] [Medline: 35155895]
15. Li M, Wen G, Zhong J, Yang P. Personalized intelligent syndrome differentiation guided by TCM consultation philosophy. J Healthc Eng 2022 Nov 07;2022:6553017 [FREE Full text] [doi: 10.1155/2022/6553017] [Medline: 36389107]
16. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
17. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023 Nov 22;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
18. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
19. Wang YY, Wang SH, Jan MY, Wang WK. Past, present, and future of the pulse examination (mài zhěn). J Tradit Complement Med 2012 Jul;2(3):164-185 [FREE Full text] [doi: 10.1016/s2225-4110(16)30096-7] [Medline: 24716130]
20. Lan KC, Litscher G, Hung TH. Traditional Chinese medicine pulse diagnosis on a smartphone using skin impedance at acupoints: a feasibility study. Sensors (Basel) 2020 Aug 17;20(16):4618 [FREE Full text] [doi: 10.3390/s20164618] [Medline: 32824477]
21. Liu Z, Zhang L, Wu J, Zheng Z, Gao J, Lin Y, et al. Machine learning-based classification of circadian rhythm characteristics for mild cognitive impairment in the elderly. Front Public Health 2022 Oct 28;10:1036886 [FREE Full text] [doi: 10.3389/fpubh.2022.1036886] [Medline: 36388285]

22. Yao Y, Zhou S, Alastruey J, Hao L, Greenwald SE, Zhang Y, et al. Estimation of central pulse wave velocity from radial pulse wave analysis. Comput Methods Programs Biomed 2022 Jun;219:106781. [doi: 10.1016/j.cmpb.2022.106781] [Medline: 35378395]
23. Liu Y, Bai X, Zhang H, Zhi X, Jiao J, Wang Q, et al. Efficacy and safety of tuina for senile insomnia: a protocol for systematic review and meta-analysis. Medicine (Baltimore) 2022 Feb 25;101(8):e28900 [FREE Full text] [doi: 10.1097/MD.0000000000028900] [Medline: 35212294]
24. Zhu Q, Li J, Fang M, Gong L, Sun W, Zhou N. [Effect of Chinese massage (Tui Na) on isokinetic muscle strength in patients with knee osteoarthritis]. J Tradit Chin Med 2016 Jun;36(3):314-320 [FREE Full text] [doi: 10.1016/s0254-6272(16)30043-7] [Medline: 27468545]
25. Cheng ZJ, Zhang SP, Gu YJ, Chen ZY, Xie FF, Guan C, et al. Effectiveness of Tuina therapy combined with Yijinjing exercise in the treatment of nonspecific chronic neck pain: a randomized clinical trial. JAMA Netw Open 2022 Dec 01;5(12):e2246538 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.46538] [Medline: 36512354]
26. Liu D, Zhang YQ, Yu TY, Han SL, Xu YJ, Guan Q, et al. Regulatory mechanism of the six-method massage antipyretic process on lipopolysaccharide-induced fever in juvenile rabbits: a targeted metabolomics approach. Heliyon 2023 Dec 08;10(1):e23313 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e23313] [Medline: 38148795]
27. Lei LM, Wu GF, Qiu SY. Literature analysis of tuina in regulating subhealth status in recent 10 years. J Acupunct Tuina Sci 2014 Feb 1;12:60-66. [doi: 10.1007/s11726-014-0749-y]
28. Liu Y, Cao L, Liu J, Zhang Z, Fan P, Zhu Y, et al. Increased hippocampal glucocorticoid receptor expression and reduced anxiety-like behavior following Tuina in a rat model with allergic airway inflammation. J Manipulative Physiol Ther 2022 Oct;45(8):586-594. [doi: 10.1016/j.jmpt.2023.04.008] [Medline: 37294215]
29. Zhang Y, Zhang W. Recent patents on Chinese massage robot. Recent Pat Eng 2017 Dec 01;11(3):156-161. [doi: 10.2174/2212797610666170127171713]
30. Pang Z, Zhang B, Yu J, Sun Z, Gong L. Design and analysis of a Chinese medicine based humanoid robotic arm massage system. Appl Sci 2019 Oct 12;9(20):4294. [doi: 10.3390/app9204294]
31. Yang J, Croghan IT, Fokken SC, Johnson DE, Calva JJ, Do A, et al. Satisfaction and feasibility evaluation of an electronic massager-expert manipulative massage automation (EMMA): a pilot study. J Prim Care Community Health 2023;14:21501319231199010 [FREE Full text] [doi: 10.1177/21501319231199010] [Medline: 37698255]
32. Johnson JA. FDA regulation of medical devices. Congressional Research Service. 2016 Sep 14. URL: https://crsreports.congress.gov/product/pdf/R/R42130 [accessed 2024-06-19]
33. Sijia L. New law sparks the expectation over the future of traditional Chinese medicine: can TCM law effectively promote the development of TCM industry in China? Med Law 2018 Mar;37(1):193-228 [FREE Full text]
34. Chen T, Zhang WW, Chu YX, Wang YQ. Acupuncture for pain management: molecular mechanisms of action. Am J Chin Med 2020;48(4):793-811. [doi: 10.1142/S0192415X20500408] [Medline: 32420752]
35. Zhang YY, Chen QL, Wang Q, Ding SS, Li SN, Chen SJ, et al. Role of parameter setting in electroacupuncture: current scenario and future prospects. Chin J Integr Med 2022 Oct;28(10):953-960. [doi: 10.1007/s11655-020-3269-2] [Medline: 32691284]
36. Yang C, Hao Z, Zhang LL, Guo Q. Efficacy and safety of acupuncture in children: an overview of systematic reviews. Pediatr Res 2015 Aug;78(2):112-119. [doi: 10.1038/pr.2015.91] [Medline: 25950453]
37. Ee C, Xue C, Chondros P, Myers SP, French SD, Teede H, et al. Acupuncture for menopausal hot flashes: a randomized trial. Ann Intern Med 2016 Feb 02;164(3):146-154. [doi: 10.7326/M15-1380] [Medline: 26784863]
38. Zhou Q, Zhao T, Feng K, Gong R, Wang Y, Yang H. Artificial intelligence in acupuncture: a bibliometric study. Math Biosci Eng 2023 Apr 27;20(6):11367-11378 [FREE Full text] [doi: 10.3934/mbe.2023504] [Medline: 37322986]
39. Zhao S, Huang T. Application of artificial intelligence in acupuncture and moxibustion. Int J Clin Acupunct 2022;31(3):224. [doi: 10.3103/S1047197922030061]
40. Peixun Y, Bing G, Yujun X. Meta-analysis of the effect of distal or local point selection on acupuncture efficacy. World J Acupunct Moxibustion 2018 Jun;28(2):114-120. [doi: 10.1016/j.wjam.2018.05.005]
41. Lu L, Zhang Y, Tang X, Ge S, Wen H, Zeng J, et al. Evidence on acupuncture therapies is underused in clinical practice and health policy. BMJ 2022 Feb 25;376:e067475 [FREE Full text] [doi: 10.1136/bmj-2021-067475] [Medline: 35217525]
42. Stux G, Berman B, Pomeranz B. Basics of Acupuncture, Fifth Edition. Berlin, Heidelberg: Springer; 2003.
43. Xing W, Li Q. Effects of different manipulations of acupuncture on electrical activity of stomach in humans. J Tradit Chin Med 1998 Mar;18(1):39-42. [Medline: 10437261]
44. Huang T, Huang X, Zhang W, Jia S, Cheng X, Litscher G. The influence of different acupuncture manipulations on the skin temperature of an acupoint. Evid Based Complement Alternat Med 2013;2013:905852 [FREE Full text] [doi: 10.1155/2013/905852] [Medline: 23476709]
45. Tang W, Yang H, Liu T, Gao M, Xu G. Study on quantification and classification of acupuncture lifting-thrusting manipulations on the basis of motion video and self-organizing feature map neural network. Shanghai J Acupunct Moxibustion 2017(12):1012-1020.

46. Zhu M, Liu DM, Pei J, Zhan YJ, Shen HY. An acupuncture manipulation classification system based on three-axis attitude sensor and computer vision. Zhen Ci Yan Jiu 2023 Dec 25;48(12):1274-1281. [doi: 10.13702/j.1000-0607.20221145] [Medline: 38146251]
47. Davis RT, Churchill DL, Badger GJ, Dunn J, Langevin HM. A new method for quantifying the needling component of acupuncture treatments. Acupunct Med 2012 Jun;30(2):113-119 [FREE Full text] [doi: 10.1136/acupmed-2011-010111] [Medline: 22427464]
48. Hamed Mozaffari M, Lee WS. Encoder-decoder CNN models for automatic tracking of tongue contours in real-time ultrasound data. Methods 2020 Jul 01;179:26-36. [doi: 10.1016/j.ymeth.2020.05.011] [Medline: 32450205]
49. Zhou Q, Gai S, Lin N, Zhang J, Zhang L, Yu R, et al. Power spectral differences of electrophysiological signals detected at acupuncture points and non-acupuncture points. Acupunct Electrother Res 2014;39(2):169-181. [doi: 10.3727/036012914x14054537750508] [Medline: 25219030]
50. Kim M, Seo HD, Sawada K, Ishida M. Study of biosignal response during acupuncture points stimulations. In: Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2008 Presented at: IEMBS 2008; August 20-25, 2008; Vancouver, BC. [doi: 10.1109/iembs.2008.4649246]
51. Feng C, Zhou S, Qu Y, Wang Q, Bao S, Li Y, et al. Overview of artificial intelligence applications in Chinese medicine therapy. Evid Based Complement Alternat Med 2021 Mar 17;2021:6678958 [FREE Full text] [doi: 10.1155/2021/6678958] [Medline: 33815559]
52. BioRender. URL: https://www.biorender.com/ [accessed 2024-06-24]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
EMMA: Expert Manipulative Massage Automation
LLM: large language model
TCM: traditional Chinese medicine

Edited by G Eysenbach, A Castonguay; submitted 18.03.24; peer-reviewed by J Sun, X Zhang; comments to author 05.04.24; revised version received 10.05.24; accepted 31.05.24; published 28.06.24.

Please cite as:
Lu L, Lu T, Tian C, Zhang X
AI: Bridging Ancient Wisdom and Modern Innovation in Traditional Chinese Medicine
JMIR Med Inform 2024;12:e58491
URL: https://medinform.jmir.org/2024/1/e58491
doi:10.2196/58491
PMID:38941141

Viewpoint

Considerations for Quality Control Monitoring of Machine Learning Models in Clinical Practice

Louis Faust[1], PhD; Patrick Wilson[1], MPH; Shusaku Asai[1], MS; Sunyang Fu[2], PhD; Hongfang Liu[2], PhD; Xiaoyang Ruan[2], PhD; Curt Storlie[1], PhD

[1]Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, United States

[2]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Louis Faust, PhD
Robert D and Patricia E Kern Center for the Science of Health Care Delivery
Mayo Clinic
Mayo Clinic, 200 First St. SW
Rochester, MN, 55905
United States
Phone: 1 (507) 284 2511
Email: Faust.Louis@mayo.edu

Abstract

Integrating machine learning (ML) models into clinical practice presents a challenge of maintaining their efficacy over time. While existing literature offers valuable strategies for detecting declining model performance, there is a need to document the broader challenges and solutions associated with the real-world development and integration of model monitoring solutions. This work details the development and use of a platform for monitoring the performance of a production-level ML model operating in Mayo Clinic. In this paper, we aimed to provide a series of considerations and guidelines necessary for integrating such a platform into a team's technical infrastructure and workflow. We have documented our experiences with this integration process, discussed the broader challenges encountered with real-world implementation and maintenance, and included the source code for the platform. Our monitoring platform was built as an R shiny application, developed and implemented over the course of 6 months. The platform has been used and maintained for 2 years and is still in use as of July 2023. The considerations necessary for the implementation of the monitoring platform center around 4 pillars: feasibility (what resources can be used for platform development?); design (through what statistics or models will the model be monitored, and how will these results be efficiently displayed to the end user?); implementation (how will this platform be built, and where will it exist within the IT ecosystem?); and policy (based on monitoring feedback, when and what actions will be taken to fix problems, and how will these problems be translated to clinical staff?). While much of the literature surrounding ML performance monitoring emphasizes methodological approaches for capturing changes in performance, there remains a battery of other challenges and considerations that must be addressed for successful real-world implementation.

(JMIR Med Inform 2024;12:e50437) doi:10.2196/50437

KEYWORDS

artificial intelligence; machine learning; implementation science; quality control; monitoring; patient safety

Introduction

As machine learning (ML) models integrate into clinical practice, ensuring their continued efficacy becomes a critical task. A pervasive limitation in ML is the inability of most models to adapt to changes in their environment over time. As a result, a model that may have performed exceptionally in its development environment can become gradually or immediately less accurate while in production [1,2]. This problem has been well studied by the ML community, with current literature offering invaluable methodological strategies for the detection of declining model performance and the ethical implications of such declines [3-7]. However, the proper choice of monitoring algorithm is only one step in the larger series of problems and considerations surrounding the sustained maintenance of these models in a real-world scenario. While some authors address the wider set of problems encountered in the long-term maintenance strategy of a deployed model, it is typically only an acknowledgment of these problems, rather than the personal experiences and solutions developed to solve them [8,9]. As

such, we aimed to supplement current literature with an alternative approach in which we provided an in-depth review of the experiences and challenges encountered when integrating our ML monitoring solution into clinical practice.

This paper focuses on an ML model implemented into Mayo Clinic's practice in 2018. The model, known as "Control Tower," is a fully integrated health care delivery model that predicts the need for inpatient palliative care through modeling palliative care consultation. The model runs automatically on all inpatients at Mayo Clinic's St Marys and Methodist Hospitals in Rochester, Minnesota, with patient scores monitored by the palliative care practice [10]. The approach was to treat the palliative care consult as a time-to-event outcome. Some of the features used are static (patient demographics and prior history), while others are time varying or dynamic (such as laboratory values, vitals, and in-hospital events). To capture the time-varying nature of these covariates, we used a heterogeneous Poisson process. Furthermore, it was crucial to account for nonlinearity and interactions; as a result, we used a gradient boosting machine. The model was validated through a clinical trial conducted from 2019 to 2022 to assess real-world effectiveness and is still in use by the palliative care practice as of July 2023 [11,12]. The study by Murphree et al [10] provides a complete methodological overview of the ML model and validation procedure. The Control Tower monitoring platform was developed and implemented over the course of 6 months. The platform has been used and maintained for 2 years and is still in use as of July 2023.

This paper provides a series of guidelines for developing and integrating ML performance monitoring into a team's workflow. Guidelines were developed from real-world experiences and challenges encountered throughout this process by a data science team at Mayo Clinic. In addition, a comprehensive overview of the developed monitoring platform is provided, as well as the accompanying source code for demonstration purposes (Multimedia Appendix 1). Overall, this paper serves as a primer for considerations that must be made when implementing and maintaining a model-monitoring system in a clinical setting, coupled with the corresponding solutions that our team had used.

Development of the Model Monitoring Platform

Overview

Traditionally, guidelines are developed through expert-driven processes, such as the Delphi method that seeks to provide standards through initial conceptions followed by several rounds of revisions until ultimately converging to an agreed-upon set [13]. However, in emerging areas where expertise is sparse, expert-driven approaches are often costly when seeking consensus of multiple experts through multiple rounds of responses [14]. An alternative to the expert-driven approach is experience-driven methodologies, which emphasize the personal experiences and observations of individuals who have directly encountered the phenomena. Normally these methodologies focus on practical knowledge through the explication of the "real world." Our team opted to derive a set of guidelines based on our specific real-world experiences and the challenges faced when designing, implementing, and integrating the Control Tower monitoring platform. Our specific methodologies used throughout this process are documented here and later generalized into a series of guidelines in the *Design Considerations* section.

Establishing the Team and Responsibilities

When planning the phases of Control Tower, it was decided that the role of monitoring the model would remain with the model development team. The task of monitoring was divided among 4 team members, rotating the responsibility of monitoring, monthly. This approach ensured monitoring would not significantly inhibit the bandwidth of any 1 team member. Monitoring responsibilities did not fall to the team member who developed the model, as their primary task in monitoring would be to retrain the model when necessary. The monitoring platform was checked biweekly, Mondays and Thursdays, to balance coverage and analyst time. The Monday check ensured immediate response to any issues that may have occurred during the previous weekend, and the Thursday check provided enough time before the upcoming weekend to identify and resolve any errors that may have occurred during the week. Typically, a single-model monitoring session would take approximately 5 to 10 minutes, assuming no problems were encountered.

Platform Development

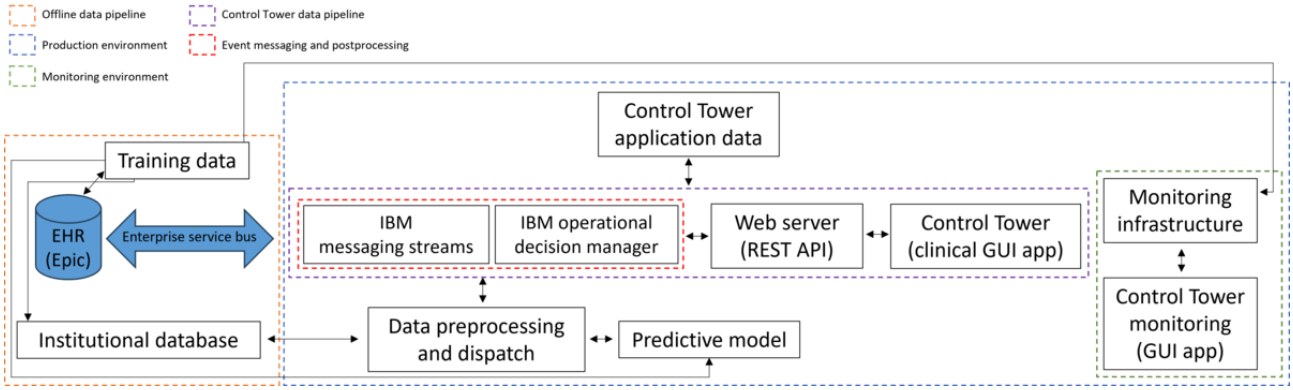
Overview

Performance monitoring of Control Tower was accomplished through the development of an R Shiny web application that comprised data visualizations and interactive tables. The goal was to create a centralized, user-friendly platform for all team members to check model performance. The platform consisted of 5 different tabs addressing different types of data shift, providing multiple degrees of granularity depending on the depth of investigation required. The model used a set of 126 features, measured daily, and was called an average of 80,000 times per day. Daily metrics collected for performance monitoring included mean and scale covariate shifts per feature, predicted probabilities, and the number of daily predictions made by the model. The resulting data size of these collected performance monitoring metrics was trivial; however, capturing patient-generated data resulted in data creation on the order of GB per day, requiring a dedicated storage space.

Figure 1 provides an overview of the system architecture for the Control Tower model and monitoring platform. The figure details the offline data pipeline used for the initial training of the Control Tower model; the components of the broader production environment and pipelines necessary for the predictive model and clinical graphical user interface (GUI) app; and finally, the components necessary for monitoring the performance of the Control Tower model. A more detailed visualization and comprehensive description of the system architecture is provided by Murphree et al [10]. Briefly, they outlined our deployment strategy which integrates a Representational State Transfer application programming interface within a Docker container, enabling the integration of

predictive models into the Control Tower GUI. The data ingestion and preprocessing pipeline, integrated with IBM Streams and Operational Decision Manager, facilitates real-time prediction processing triggered by updates to institutional health records (Health Level Seven messages by our electronic health record). The Control Tower GUI application is built with Angular (Google LLC).

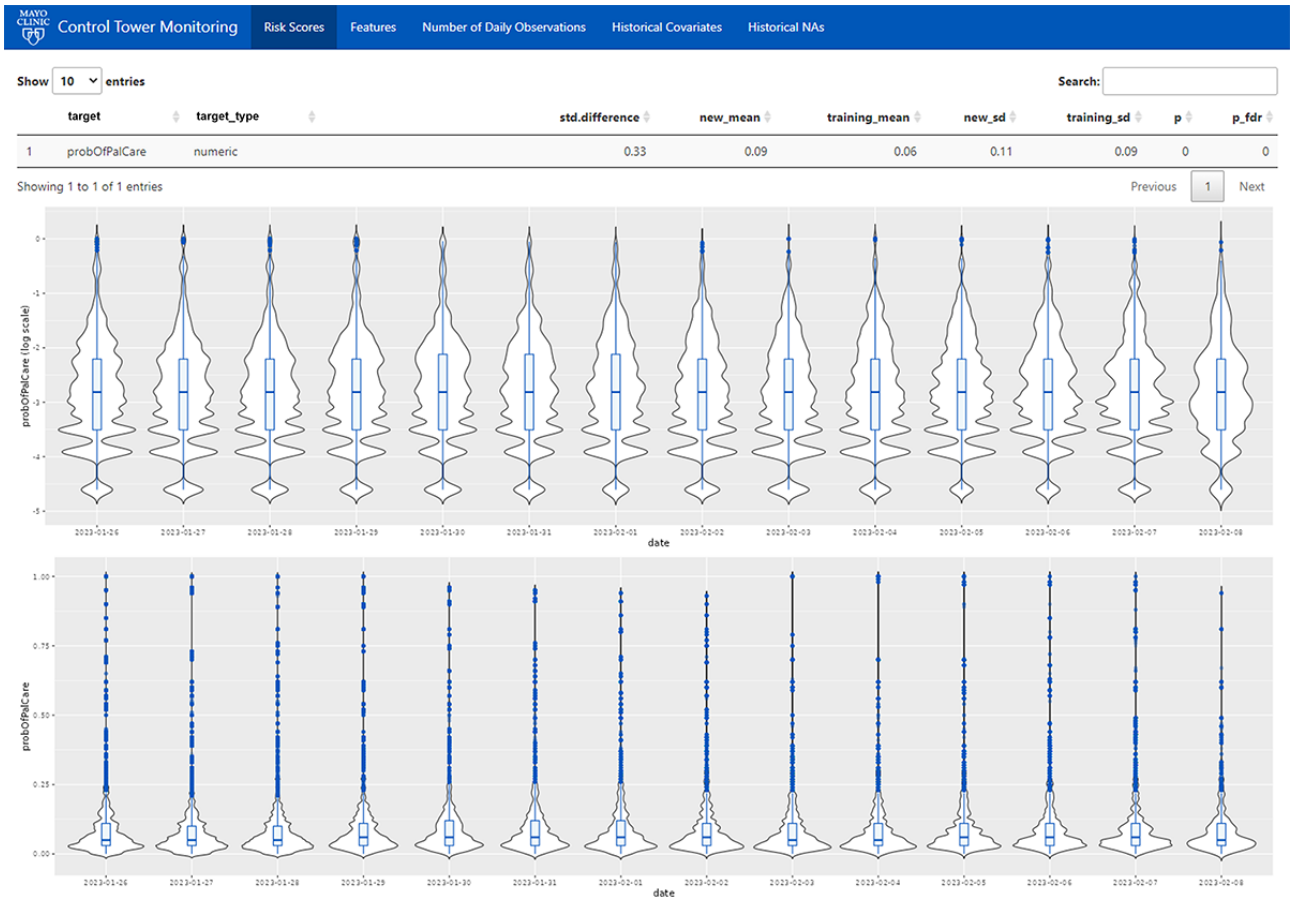
Figure 1. System architecture for Control Tower. For the Control Tower monitoring platform, we have 3 parent processes (training, production, and monitoring) that constitute our deployment. Child processes include the orchestration of the streams, events, and the prediction pipeline, which sends scores to the graphical user interface (GUI). EHR: electronic health record; REST API: Representational State Transfer application programming interface.



Monitoring Model Probabilities

In the absence of ground truth labels, predicted probabilities from the model were monitored as an alternative to evaluating model performance. The platform visualizes these predicted probabilities as distributions of daily risk scores (Figure 2). Distributions are plotted on probability and logarithmic scales, allowing for easier detection of shifts when most predicted probabilities are low, considering that most patients will not be "high risk." Historical daily distributions extend back 2 weeks, which is considered an optimal amount of time to notice shifts without overwhelming the user with data. Alongside these visualizations, several statistics are presented for comparing the prior 2 weeks data against the original training data. Means and SDs for the incoming and training distributions, the standard difference, and a *P* value for the Kolmogorov-Smirnov test of differences between the 2 distributions are provided. These statistics allow the user to detect gradual, more long-term changes that may go unnoticed when surveying 2 weeks of historic data. Overall, the tab shown in Figure 2 provides an overview of model predictions, allowing the user to quickly gauge whether a sudden or gradual probability shift has occurred.

Figure 2. The startup screen of the Control Tower performance monitoring platform. This screen provides the user a quick overview of the model's predicted probabilities over the past 2 weeks. The table near the top provides several statistics comparing the distribution of predicted probabilities over the last 2 weeks with the predicted probabilities on the training data. The 2 graphs contain a series of violin plots featuring the daily distribution of predicted probabilities. Given that the predicted probabilities cluster near 0, the distributions are also displayed on the log scale for easier visual inspection.



Monitoring Covariate Shift

Covariate shift was addressed in Control Tower by creating an interactive table containing all features included in the model (Figure 3) [4]. The table lists feature names and type, that is, continuous or discrete, and displays different statistical tests and plots, dependent on the feature type. To assess the impact of a feature with drift, the team included global feature importance scores from the originally trained model, in this case, the gradient boosting machine's relative influence rank statistic. Providing a ranking of features based on the extent of error reduction in the model enables the user to triage different drifts. All other things being equal, a drifting feature with higher importance to the model than another feature would indicate a higher priority need of a fix. Similar to the predicted probability tab, the previous 2 weeks of incoming data are compared with the training data, with standard differences, means, and SDs provided. To accommodate for the discrete variables present, the distributional Kolmogorov-Smirnov test is changed to the chi-square test. The user can sort the table by column, allowing them to quickly pinpoint features, for example, with high standard difference. Clicking on a feature's row in the table generates 2 plots underneath the table: the first is a line graph visualizing the daily standard differences, spanning back 2 weeks, and the second plot is dependent on the feature type. For continuous variables, the plot compares the feature's daily distributions over the past 2 weeks with the distribution of the training data, using box plots. For discrete variables, bar plots are displayed in a similar fashion indicating the percentage of patients where the discrete feature was present or "True." In tandem with the interactive table, these plots provide an efficient means of investigating a feature's historic values at a glance.

When a deeper investigation into a feature is necessary, the 2-week "look-back" may be insufficient. Therefore, the platform also keeps a log of the full historic feature trends, spanning back to when the model was deployed (Figure 4). Feature plots are sorted by the model's global feature importance and color-coded "green" or "red" to indicate whether the feature significantly drifted from the initial training distribution. Significance was determined via a nonparametric test developed by Capizzi and Masarotto [15], using a *P* value of .05. A nonparametric model was used because a moderate number of features were highly skewed, making traditional methods that assume normal distributions unworkable. Each feature contains plots for the location (level) and dispersion (scale) of the distribution. Overall, this tab, in addition to serving as a historical reference, provides a simple way to spot check for gradual shifts. Finally, an additional tab (Figure 5) is provided to assess the proportion of missing values over time, using the same visualizations and tests.

The final tab of the platform provides a simple line graph displaying the number of daily calls made to the model within the previous 2 weeks (Figure 6). Monitoring the number of daily

calls can provide quick insight into whether the model is performing appropriately. For example, an abnormal number of model calls in a day, such as 0, may indicate an error in the data pipeline or model environment.

Figure 3. The "Features" screen of the platform details the distributions of all features used by the model. The distribution of each feature based on the last 2 weeks of data is compared with the feature's distribution from the training data. These comparisons are provided via statistics in the table near the top, which can be sorted by each statistic to quickly find features with potential drift. Clicking on a feature populates 2 graphs, which are displayed below the table. The first graph displays the standardized difference between the feature's distribution for that day against the distribution from the training data. Below this graph, one of the 2 graphs will be displayed depending on whether the selected feature was binary or continuous. These graphs display the daily distributions of the feature, using bar graphs for binary features (red outline) or box plots for continuous features (yellow outline).

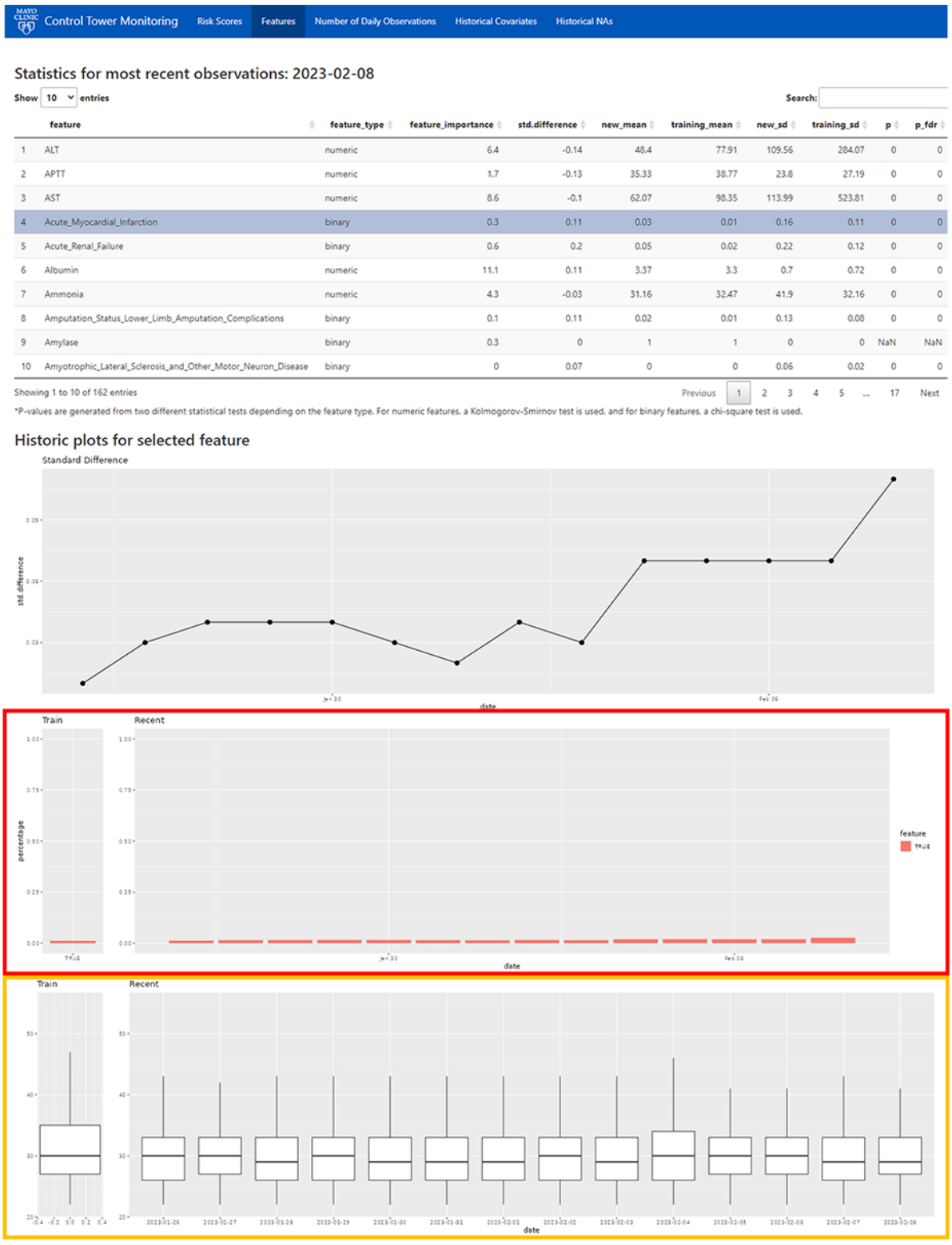


Figure 4. The "Historical Covariates" screen of the platform visualizes each feature's daily distribution, beginning with the training data and then spanning from the day the model was deployed and onward. Each feature contains plots for the location (level) and dispersion (scale) of the nonparametric distribution. Each feature's graph is color-coded "green" or "red" to indicate whether the feature's distribution has significantly drifted from the initial training distribution, with red indicating significant drift.

Figure 5. The "Historical NA's" screen of the platform visualizes each feature's historical missingness, beginning with the training data and then spanning from the day the model was deployed and onward. Each feature contains plots for the location (level) and dispersion (scale) of the nonparametric distribution. Each feature's graph is color-coded "green" or "red" to indicate whether the feature's missingness has significantly drifted from the initial training distribution, with red indicating significant drift. NA: not available.

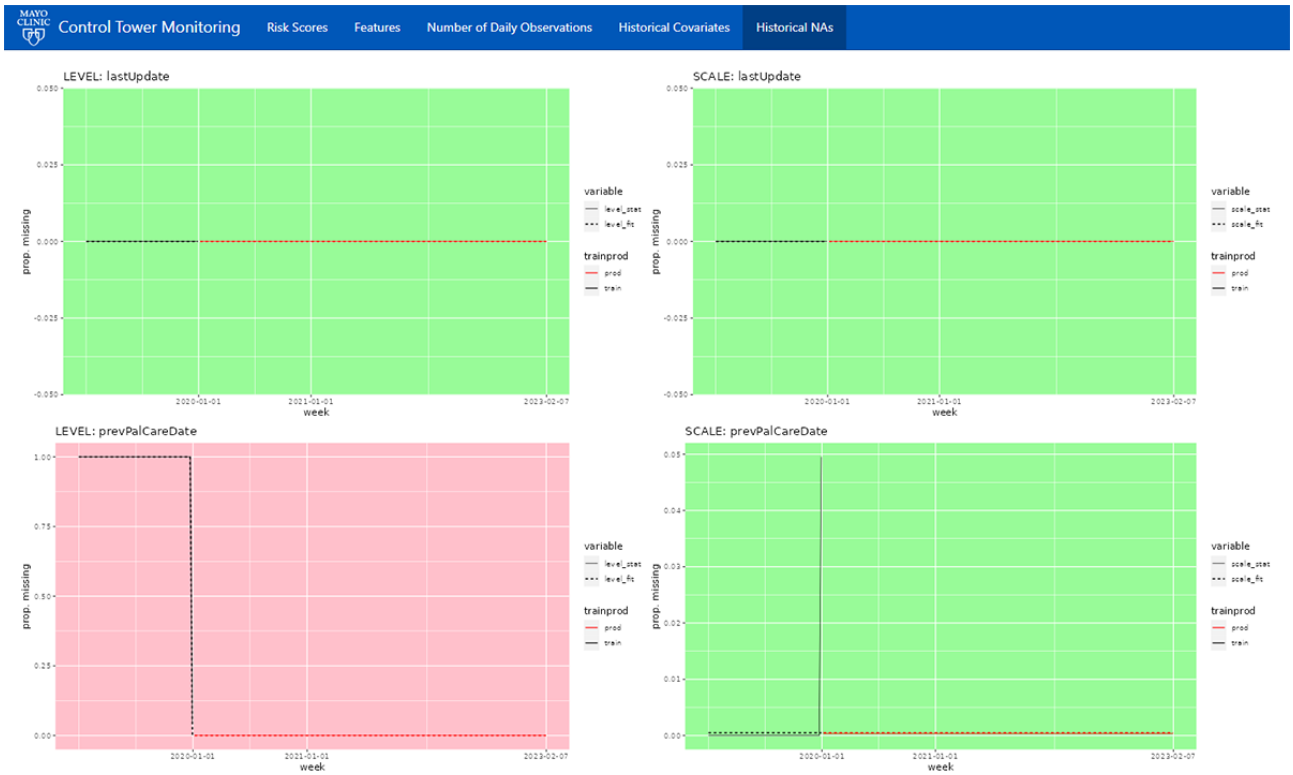
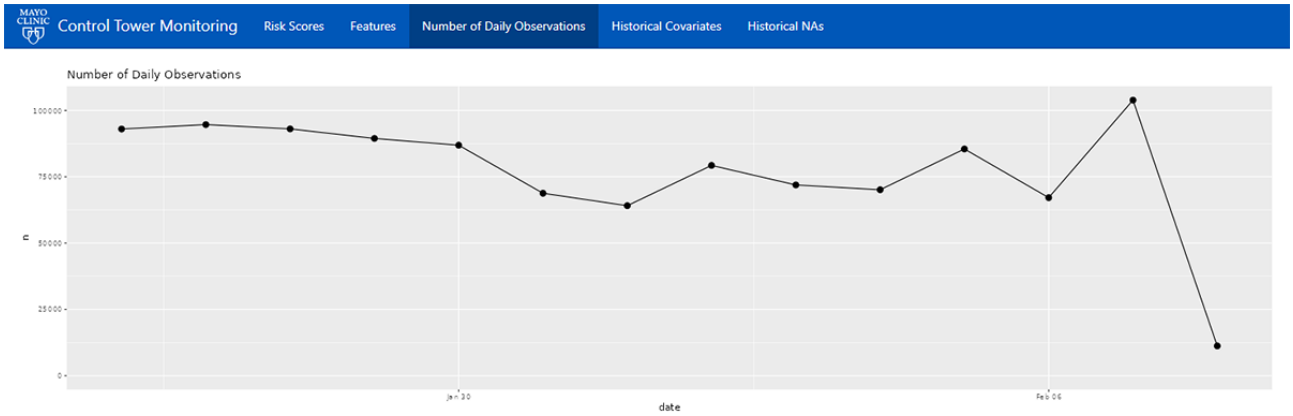


Figure 6. The “Number of Daily Observations” screen visualizes the number of model calls or predictions made each day over the past 2 weeks.



Error Classification

Any production-level model is susceptible to various errors, and Control Tower was no exception. Most errors primarily revolved around technical infrastructure, particularly issues with databases being inaccessible due to nightly processing or a high influx of requests. In Textbox 1, a sample of encountered errors while monitoring Control Tower is presented. Although some errors were seemingly random occurrences, such as server reboots or expired certificates, others were more frequent and persistent. For instance, every night at specific hours, the database that supplies data to Control Tower, called Clarity, became unresponsive due to data updates. On January 5, 2022, this process was delayed and caused errors in the morning scores. In addition, updates to our electronic health record (Epic, Epic Systems Corporation), often resulted in Clarity being temporarily unavailable. In such cases, most issues were resolved on the same day, requiring no further action besides acknowledging the possibility of outdated or missing scores. However, a few errors necessitated intervention. On November 7, 2022, a data mart containing diagnosis codes underwent structural changes, breaking a Control Tower query. Furthermore, the team identified a covariate shift where they observed a gradual decrease in troponin blood tests. This error was traced back to a change in laboratory codes used for troponin; the clinical practice had adopted a new laboratory code that was not present in the training data. To address this, the error was rectified by associating the new codes with the “Troponin” feature on the platform’s back end.

Textbox 1. Error logging: A convenience sample of encountered errors while monitoring Control Tower is presented. This was constructed through email chains of discussions between IT personnel who oversaw the Control Tower system and the data scientists who oversaw model delivery.

Date and error

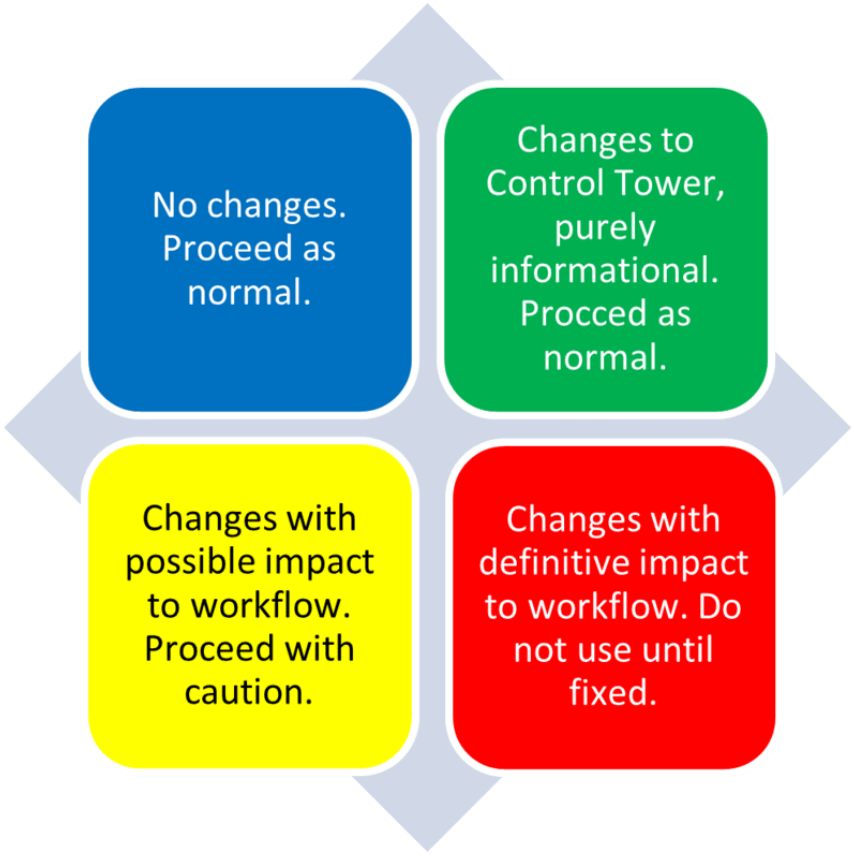
- August 26, 2019
  - Multiple errors in logs. It looks like calls were during 1:45 AM to 6:00 AM. During the period 1:45 AM to 6:00 AM, all messages are failing due to Clarity Refresh.
- November 27, 2019
  - Server reboot schedule, Control Tower team was not notified of schedule leading to unexpected downtime.
- July 24, 2020
  - Increase in FHIR (Fast Healthcare Interoperability Resources) API (application programming interface) for real-time observation calls leading to timeouts of model predictions.
- February 15, 2021
  - Generic FHIR API error call: “HTTP error code: 500.”
- April 8, 2021
  - Model errors after Epic upgrade.
- July 30, 2021
  - Troponin issues fixed causing covariate drift in model scores.
- September 15, 2021
  - Production system competed for resources requiring scale back of Control Tower scores updates. Errors created and schedule has now been updated for processing.
- January 5, 2022
  - Nightly Clarity Database delay causing morning score errors.
- May 16, 2022
  - IBM queue server certificate update causing server errors.
- November 7, 2022
  - Data mart for diagnoses codes update causing pipeline to break down.
- November 8, 2022
  - Issues with Clarity Database slowing down Control Tower queues.
- March 21, 2023
  - Control Tower FHIR API for real-time unit changes failing for a single request, causing payload slowdown.
- April 5, 2023
  - JSON structure changing causing model error (unintended repo change).
- May 1, 2023
  - An unplanned issue impacting Enterprise API Services, who manages real-time data feeds, resulting in internal server error.

Monitoring Protocol

Actions prompted by model monitoring feedback were synthesized into a protocol to communicate model failures and downtimes with clinical staff. The Control Tower protocol used a triage system, consisting of 4 stages in which each stage prompts a message to the clinical team, as outlined in Figure 7. The first stage (blue) was reserved for when everything was operating as expected. Stage 2 (green) indicated a minor change in the layout of the tool, such as a user interface change. These first 2 stages delivered informational prompts to the user, notifying them of the tool’s status and requiring no action from the user. The third stage (yellow) indicated possible performance degradation, such as when patient scores from the model were

not up-to-date, that is, a day old. The day-old scores can still provide evidence for action, but the clinical team may need to be cautious, as updated scores can change the patient's risk. As such, users were notified of these issues and asked to use the tool at their discretion. The last stage (red) indicated a significant error within the tool, such as a covariate completely missing from the model's input. This stage would notify staff not to use the tool until a fix was implemented.

Figure 7. Communication triage protocol for Control Tower. The protocol's stages are color-coded to signify different statuses and recommendations: blue for normal operation, green for minor layout changes, yellow for potential performance issues, and red for significant errors.



Design Considerations

Overview

From our experience of developing and implementing the Control Tower monitoring platform, we have derived a series of broader considerations necessary for model monitoring to serve as generalized guidelines for future implementations. Central to these guidelines arose themes of feasibility, design, implementation, and policy. While existing frameworks have proven successful in managing long-term IT infrastructure projects in health care, ML models are experimental and inherently open systems, entailing costly development and maintenance. As such, they demand additional considerations due to their reliance not solely on technical data but also on statistical and clinical assessments to identify errors. Consequently, there is no unequivocal, predetermined signal that can be provided to an IT group lacking clinical or data science expertise to detect these errors. Identifying them often necessitates the accumulation of statistical data over time. While readily available and accessible data may be used to identify some errors, others may require weeks or months of new data collection to draw inferences. Furthermore, in traditional software operations, refinements can be implemented more quickly. Responding to user feedback can often be met with bug fixes or minor feature requests. However, implementing refinements to an ML model often requires a longer development cycle, as many changes will require a complete retraining of the model or the acquisition of novel data. Such complications in model maintenance underscore the need for input from multiple teams, alongside established structures and policies, to ensure effective orchestration of model maintenance.

Feasibility: Are the Resources to Facilitate Productive Monitoring Available?

Successful real-world implementation of models must consider the present and *future* bandwidth limitations of a team. In an ideal scenario, a team would be provided ongoing dedicated time or personnel to ensure the upkeep of deployed models. However, this is not always feasible, and the responsibility of maintenance competes with a team's constant stream of new projects and tasks. As such, it is important to first determine how long-term monitoring of a model (or eventually, *models*) will be integrated into the team's workflow. For example, who will check in on the model and with what frequency? If and when the model requires retraining, who will perform the retraining, and how will they be guaranteed the flexibility to shift from their current projects to accomplish this?

In addition to personnel, computational resources must also be considered for long-term monitoring. Regarding storage requirements, the amount of data produced from monitoring

depends on a variety of factors, including the model's feature set, the frequency of calls made to the model, and the number of feature and performance metrics that will be tracked. For example, a model that delivers constant, real-time feedback in a critical care setting may, in turn, require constant monitoring to ensure performance does not suddenly degrade, resulting in performance metrics and feature distribution logs needing to be generated continuously.

When assessing the feasibility of long-term monitoring, an attractive option to consider is automated monitoring: developing models that detect whether a significant change has occurred in a deployed model's predictions or its incoming data. While our team chose a "hands-on" approach, others have found success in implementing an additional model for performance monitoring to notify the team of data shifts and, in some cases, automatically retrain the original model [16]. Ultimately, the decision to use an automated monitoring model comes down to the type of model deployed, the bandwidth of the team, and the reliability of the data to support automatically retraining the deployed model. In any case, even an automated monitoring model will still require some human monitoring as well. Our team is currently working on an automated monitoring system for this purpose.

Design: Deciding What and How to Monitor?

Overview

The design of an investigative platform to facilitate model monitoring may range from dynamic and interactive interfaces to static reports, the choice of which is dependent on feasibility factors and nuances of the particular scenario, but design should ultimately enable rapid and comprehensive assessment. Furthermore, there are standard functionalities each platform should feature to appropriately assess long-term model performance.

Deployed models encounter performance declines through distributional and relational shifts in the underlying data [17]. These shifts are the crux of why postdeployment monitoring is necessary, and no model is immune to them, regardless of how well it performed in its testing environment [18]. This impediment has received a wealth of attention from the ML community and has been synthesized into 3 types of distinct shifts.

Prior Probability Shift

The distribution of the target variable Y changes between the training data and incoming data, but not P (X|Y) [19]. This can occur when the prevalence of a disease changes over time in the target population; however, the underlying factors that cause the disease remain constant, for example, a spike in influenza rates during the influenza season. Prior probability shift is assessed by monitoring the distribution of the target variable over time, measuring for any sudden or gradual changes.

Covariate Shift

Distributions of input data diverge between training data and new incoming data [4]. Such shifts may occur in the clinical setting, for example, when diagnostic screenings are updated. This procedural change may decrease the specific laboratory values, which are heavily relied upon by the model. Conversely, diagnostic variables that were initially infrequent may become more prevalent over time. This can result in situations where the model, which had limited instances of these variables during training, struggles to fully capture, and therefore use, their predictive signals.

Concept Drift

The relationship between the incoming input data and target variable changes over time, drifting from the original relationship captured in the training data [20]. The COVID-19 pandemic provided a real-world example of concept drift, as hospital census models were affected by admissions that drastically moved toward higher-risk patients due to increases in complications from the COVID-19 disease and decreases in hospital use among people with a milder spectrum of illness [21].

Usability

A successful UI will take into consideration the professional backgrounds of those using the platform. However, when the responsibilities of monitoring are handed to a different group, the new group's level of familiarity with the model should guide the design. For example, guidelines for what is acceptable variance should be established and implemented. One method for accomplishing this may be through using control charts, allowing the modeling team to prespecify a simple and visual approximation of how much drift is tolerable before action must be taken [22].

Implementation: How Will the Platform Be Built and Sustained?

When implementing a monitoring platform, it is necessary to consider how the back end of the platform will process and store the necessary data elements. The efficiency of this task is critical and must accommodate the model's scale and responsiveness. Data can amass quickly as large feature sets are monitored, and the model may be called frequently to predict on many patients throughout the day. Furthermore, the back end must be capable of efficiently parsing, formatting, and, if necessary, compressing the data into clean data sets for the platform to analyze and visualize. For Control Tower, many of these data storage requirements were already in place for capturing and storing the necessary patient elements. This will likely be the case for many clinical scenarios, as patient data must be securely and efficiently housed. Instead, implementation efforts are more apt to center around ensuring these data elements are efficiently piped to the monitoring platform.

Using a web application for model monitoring provides a dynamic interface, allowing any user with log-in permissions to view the real-time status of the model and the surrounding data. This investigation mechanism can eliminate potential confusion, which may arise from a routine generation and sharing of static technical reports, such as accidentally referencing outdated documents. When selecting a programming language to build the app, preference should be given to those languages that facilitate efficient app development. For Control Tower, R Shiny was used given the team's previous experience with the package and strong background in R. The R package

provides a user-friendly environment for quickly creating, testing, and publishing web applications. Similar web application tools exist across multiple programming languages including Python and Java, and as such, teams are likely to find a web application package in a language they are familiar with.

When coding the app, modular coding practices should be adopted to ensure flexibility and scalability. Such adoption promotes versatility of the app to incorporate additional statistical measures or visualizations and allows the app to be easily translated for other monitoring use cases. Leveraging modular coding practices at the onset of app development allows for future additions, revisions, and ports to be made with minimal effort. For Control Tower, modular coding practices were primarily used to better facilitate development across multiple team members. This practice allowed for functions to be easily repurposed by other team members to avoid duplication of work and to allow the app to be easily extended to other ML models within Mayo Clinic.

The number of programming languages used in the data pipeline plays a significant role in shaping the development process and the overall efficiency of the monitoring. To facilitate this, minimizing the number of programming languages used across the various tasks can streamline development and maintenance through ease of interpretability and integration. This can reduce maintenance costs and overhead by reducing interoperability concerns and decreasing the learning curve for new team members. Minimizing these ongoing costs is a necessity when considering the model will ideally be in production long term. However, if the development team is proficient in multiple languages, leveraging the strengths of each may have its advantages, such as reducing bottlenecks in development or data transfer, while increasing the flexibility of a system. In the case of Control Tower, R (R Foundation for Statistical Computing), Python (Python Software Foundation), and shell scripts were used, favoring R for app development, Python for data processing, and shell scripts for scheduling various model and platform tasks.

In addition, upstream problems will inevitably manifest; therefore, implementing a notification system for these errors can proactively address disruptions, minimizing the downtime of the pipeline. One method for accomplishing this is to incorporate error logging and alerts into cron jobs, which can immediately notify the team of any failures. Such notifications are critical for model monitoring, as some errors may be undetectable to the end user, resulting in the continued use of inaccurate information. As such, it is vital for monitoring teams to identify, communicate, and resolve errors as soon as possible.

Finally, integrating regular checking of the platform into the team workflow allowed the team to not only stay abreast of model performance but also maintain an intuitive sense of potential broader complications surrounding the model. For example, monitoring the probability distributions of the model ultimately provided the team with a sense of whether further investigations into the model would be necessary. However, investigations into the distributions of the individual features allowed for potential diagnoses as to why the model may begin to degrade in performance, as well as alluded to data pipeline errors that may be present. By maintaining a sense of these wider issues, shifts in the outcome could be more easily prevented and diagnosed

Policy: What Is the Response to Platform Feedback?

Overview

Once the monitoring platform is deployed and available, the next stage of considerations surrounds how knowledge provided by the platform will be used. A set of policies must be developed to determine which actions will be taken based on monitoring feedback, addressing such questions as “At what point is a data shift significant enough to prompt retraining?” and “How will errors be communicated with technical and clinical staff?” Generally, such a policy should cover error designation and response, when to retrain, and how to communicate failures. In addition, a well-defined policy allows for the task of monitoring to more easily be extended across various teams and roles.

Error Designation and Response

It is essential to establish and define a process that determines when a specific degree of shift or drift in the model qualifies as an error warranting a response. The question *“How much drift is necessary to take action?”* represents one of the more subjective aspects of model monitoring. In scenarios where multiple team members are tasked with overseeing model performance or possess limited familiarity with the model, substantial interrater variability becomes a concern. For example, one team member might observe a 5% shift in the distribution of a feature and consider it inconsequential, while another member might view it as a reason for immediate action. To address this variability, the Control Tower team would send email updates to other team members detailing any shifts that were noticed; this would allow for a collective discussion on whether to take action as well as allow for a convenient forum to keep all team members updated on the model’s status. Regardless of the criteria used to identify shifted covariates or outcomes, team members must communicate and establish agreement on the minimal drift threshold requiring action, while ensuring that utmost priority is placed on maintaining optimal model accuracy.

Even with consensus on the magnitude of a shift, several contextual factors can influence the team’s risk tolerance toward these shifts. Significant changes may occur without sustained trends, indicating a regression to the mean. Alternatively, a dramatic shift might happen for a variable with minimal contribution to the risk score. While predefined cutoff points could be considered to standardize investigations, these benchmarks may still necessitate ongoing human review and could vary for each feature, making it impractical to define for every feature in large feature sets.

Even if an error is defined with a certain level of risk in mind, there are considerations in the response to the error and the amount of time one needs to allocate for remediation. A deployed model is prone to errors from a variety of sources, ranging from data shifts to IT scheme modifications. Given the diversity of potential errors, an effective policy will include guidelines for the categorization of errors along with the

appropriate responses to each. The errors encountered with Control Tower fell broadly into 4 categories.

1. Technical infrastructure: database issues, expiration of certificates, and password updates often causing the pipeline to fail
2. Explained shift: a significant data shift with an identified root cause
3. Unexplained shift: a significant data shift with an unidentified root cause
4. Performance loss: a decrease in the model's performance metrics, which may manifest with or without data shift

Categorizing errors for appropriate response is crucial, as it establishes a standardized knowledge base for reporting, ultimately enhancing the efficiency of troubleshooting. Categorization often leads to the discovery of similar strategies for mitigating similar error types. For instance, errors related to database refreshing or password expiration typically do not require immediate intervention, while performance losses in accuracy or calibration often necessitate retraining of the model. Appropriate categorization also offers the advantage of reducing risk tolerance while enhancing response efficiency. Having encountered an error previously increases the likelihood of streamlining investigations, enabling the examination of lower-risk shifts or drifts.

When ongoing outcomes data are available, performance loss can be detected by looking for significant shifts across a variety of classification performance metrics including area under the receiver operating characteristic, area under the precision-recall curve, calibration, subgroup differences, and so on. When such data are absent, as in the case of Control Tower, performance loss can only be inferred by looking for significant shifts in the distribution of predicted probabilities of the model. To supplement assessing predicted probabilities, potential performance loss may also be identified by looking for significant shifts in the features of the model. While significant shifts may occur in these features without significant shifts in the model's output, drifts in feature distributions can signal other potential problems necessary to address. While performance loss may be resolved or mitigated through upstream pipeline errors, some instances may require the model to be retrained.

Model Retraining

The circumstances for when to retrain a model will vary across teams and platforms, often dependent on the cost of retraining. As such, it is necessary for a platform policy to clearly state when, and when not, to retrain. For example, many errors will not require model retraining such as simple pipeline errors or data shifts due to changes in medical coding, requiring only a small update to the pipeline. Therefore, it is important to first identify and fix any upstream errors before considering retraining. There are even instances of significant shifts that do not warrant retraining. For example, one could have several shifted covariates in the model with trivial importance scores, effectively having no impact on predictive performance. From the perspective of model importance, one may bin covariates that have little impact and essentially treat them as nuisance variables.

Assuming no upstream errors are present, a model should always be retrained when significant and sustained performance loss is encountered. Defining *significant* and *sustained* will be specific to each scenario, depending on the algorithm and health care delivery model. However, it is incumbent upon the team to define an appropriate window for performance to vary, with a lower limit triggering retraining.

It is important to note that retraining does not have to be used sparingly, assuming the bandwidth is available. When feasible, it may be good practice to routinely retrain the model with the expectation that updated data are more current with clinical practice. Such versioning of the model would allow for new features, incremental improvements, and technical debt management. For Control Tower, versioning allowed us to spot potential bugs or fixes and investigate new features.

Communicating With Clinical Staff

The clinical team using the model's outputs must be consistently informed about the model's status due to its significance to their workflow and overall trust in the model. The model's standard operating procedure outlines how the clinical team should use the model and details communication protocols between IT, data science, and the clinical users. Protocols should consist of dedicated contacts for various issues and plans for how to operate during model performance shifts and downtime.

Discussion

Principal Findings

As ML models require consistent monitoring to ensure sustained accuracy, a series of decisions must be made for how best to integrate model monitoring into a team's workflow. Problems, considerations, and solutions that arise from this process can vary greatly depending on the setting, nature of the model, and available bandwidth, both from the technical team and their computational resources. While prior work has established the importance of monitoring and corresponding statistical solutions, this paper provides specific considerations and solutions derived from the real-world implementation and day-to-day use [23,24]. Throughout the integration of Control Tower, our team found that these considerations centered around 4 phases that serve as a road map when planning a long-term modeling strategy: feasibility, design, implementation, and policy.

Experiences

Development and implementation of the platform faced several obstacles, which we attribute to the inherent realities of integrating real-world applications. First, the team was unable to complete the platform by the time the associated clinical trial for the Control Tower model began recruitment. This required the team to omit crucial features from the platform, such as monitoring for concept drift. Monitoring concept drift required collecting ground truth outcomes, that is, whether patients actually received palliative care. Collecting these patient outcomes required building a separate data pipeline, which was the team's original intent, but as the team took on additional tasks, the pipeline was passed over in favor of monitoring predicted probabilities. While omitted from Control Tower in scenarios where outcomes data are tracked, we direct the readers

to literature providing a more comprehensive understanding of concept drift [25-27].

The original intention for Control Tower was to have the model run any time a patient's laboratory values were updated, ensuring that the patient's risk scores were always reflective of current data. While the model was originally deployed using this dynamic system, it, unfortunately, proved too taxing for the IT infrastructure in which it was hosted. To alleviate this problem, the model and platform were switched to running on a batch schedule, updating patient risk scores and the monitoring platform every 4 hours. While this delivery schedule proved more manageable, model calls made between these 4-hour updates ran the risk of using outdated patient data, potentially impacting performance. Given that the workload imposed by the original schedule was infeasible, this was considered a fair compromise. Finally, the implementation of the platform occurred during the COVID-19 pandemic, which affected staffing and resulted in IT furloughs. Unfortunately, this meant that technical infrastructure problems, which could typically be fixed by IT on the same day, instead, took up to 1 week to fix, resulting in prolonged downtime for the model.

Despite these challenges, there were several positive experiences to highlight. First, a significant amount of collaboration occurred within the data science team in order to have the monitoring platform in a usable state by the time of the model's clinical trial. This required analysts to tend to a variety of tasks, often on a moment's notice. Following deployment, there was also sufficient bandwidth from the team members to continue monitoring the platform as they took on additional projects. Second, IT furloughs as a result of the pandemic were resolved within 6 months, allowing routine technical infrastructure issues to once again be resolved on the day of occurrence, resulting in less model downtime. Finally, the model's predicted probabilities remained, for the most part, consistent, making for a stable tool throughout the documented 2 years of use. Using a simple linear regression model, we examined the relationship between daily predicted probabilities (dependent variable) and time since deployment (independent variable), observing a slope of 0.005 at a P value of <.001, suggesting a statistically significant, but functionally small trend, with the mean probability increasing .005% each day.

Limitations

Despite a thorough detailing of our experiences, it is important to note that this paper covers only a single implementation. While we have recounted the challenges and considerations necessary for Control Tower, this is not an exhaustive list, and other teams and platforms may encounter challenges foreign to ours.

Future Considerations

The Control Tower platform allowed our team to successfully monitor performance and maintain our deployed model for 2 years. Moving forward, our team is planning to automate parts of this task, for example, by implementing an automated email notification system, notifying the team when the number of model calls, predicted probabilities, and incoming data streams shift beyond their respective significance thresholds. While this modification is not intended to outright replace the manual checks of the platform, it will allow the team to check the platform at a lesser frequency. This system will serve as a placeholder while the team develops a new model to monitoring the performance of Control Tower, leveraging supervised learning to detect shifts in the probability and multivariate covariate distributions [28,29].

The team also considered an online or continuous learning model to automatically address data drift. In continuous learning, the algorithm would update its predictions as new data come in and alleviates the need to manually retrain the data [30]. Although appealing, an automated system, in this sense, would require more policy changes and would bring with it a number of issues. First, there are several cost and computing issues that could make an implementation difficult, as entire systems would need to know when to train and to do it without interrupting the current pipeline, as well as a validation step to ensure sustained, if not improved, accuracy. Second, the algorithm must remain trustworthy for clinicians. Did the algorithm *unlearn* anything important? Did it learn anything irrelevant or incorrect? As an example, if a covariate shift occurred due to a missing laboratory code, resulting in increasingly missing values of that laboratory, we would not want the model to learn a new relationship with the missingness; instead, we would make an update to the data pipeline to resolve the missingness. Finally, all continuous learning models require ready access to the gold-standard outcome, which might not be feasible in all cases.

Conclusions

Once an ML model has been successfully developed and deployed, it must be continuously monitored to ensure its efficacy amidst an ever-evolving practice and stream of patients. While a variety of methods have been proposed to statistically monitor the performance of models, this is only one factor to consider when implementing a long-term modeling strategy. By disseminating the broader experiences of integrating ML monitoring platforms into clinical practice, readers will be better equipped for the considerations and challenges encountered during their own implementations.

Acknowledgments

The research reported in this paper was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (award number R01EB019403). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to concerns of patient identification but are available from the corresponding author on reasonable request. Such requests may require separate approval by the Mayo Clinic Institutional Review Board.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Source code for a demo of the Control Tower Model Monitoring R Shiny application.
[ZIP File (Zip Archive), 26937 KB - medinform_v12i1e50437_app1.zip]

References

1. Allen B, Dreyer K, Stibolt RJ, Agarwal S, Coombs L, Treml C, et al. Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: try it, buy it, check it. J Am Coll Radiol 2021 Dec;18(11):1489-1496. [doi: 10.1016/j.jacr.2021.08.022] [Medline: 34599876]
2. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021 Aug 01;181(8):1065-1070 [FREE Full text] [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]
3. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
4. Schwaighofer A, Quinonero-Candela J, Sugiyama M, Lawrence ND. Dataset Shift in Machine Learning. New York, NY: Penguin Random House LLC; 2008.
5. Klinkenberg R, Joachims T. Detecting concept drift with support vector machines. In: Proceedings of the Seventeenth International Conference on Machine Learning. 2000 Presented at: ICML 2000; June 29-July 02, 2000; Stanford, CA. [doi: 10.1007/978-1-4615-0907-3_3]
6. Huang R, Geng A, Li Y. On the importance of gradients for detecting distributional shifts in the wild. arXiv Preprint posted online October 1, 2021.
7. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021 Jun 28. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2023-07-05]
8. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. In: Proceedings of the Software Engineering for Machine Learning. 2014 Presented at: SE4ML 2014; December 13, 2014; Montreal, QC. [doi: 10.7551/mitpress/12440.003.0011]
9. Pruneski JA, Williams RJ3, Nwachukwu BU, Ramkumar PN, Kiapour AM, Martin RK, et al. The development and deployment of machine learning models. Knee Surg Sports Traumatol Arthrosc 2022 Dec;30(12):3917-3923. [doi: 10.1007/s00167-022-07155-4] [Medline: 36083354]
10. Murphree DH, Wilson PM, Asai SW, Quest DJ, Lin Y, Mukherjee P, et al. Improving the delivery of palliative care through predictive modeling and healthcare informatics. J Am Med Inform Assoc 2021 Jul 12;28(6):1065-1073 [FREE Full text] [doi: 10.1093/jamia/ocaa211] [Medline: 33611523]
11. Wilson PM, Philpot LM, Ramar P, Storlie CB, Strand J, Morgan AA, et al. Improving time to palliative care review with predictive modeling in an inpatient adult population: study protocol for a stepped-wedge, pragmatic randomized controlled trial. Trials 2021 Oct 16;22(1):635 [FREE Full text] [doi: 10.1186/s13063-021-05546-5] [Medline: 34530871]
12. Wilson PM, Ramar P, Philpot LM, Soleimani J, Ebbert JO, Storlie CB, et al. Effect of an artificial intelligence decision support tool on palliative care referral in hospitalized patients: a randomized clinical trial. J Pain Symptom Manage 2023 Jul;66(1):24-32. [doi: 10.1016/j.jpainsymman.2023.02.317] [Medline: 36842541]
13. Dalkey NC, Brown BB, Cochran S. The Delphi Method: An Experimental Study of Group Opinion. Santa Monica, CA: RAND Corporation; 1969.
14. Barrett D, Heale R. What are Delphi studies? Evid Based Nurs 2020 Jul 19;23(3):68-69. [doi: 10.1136/ebnurs-2020-103303] [Medline: 32430290]
15. Capizzi G, Masarotto G. Phase I distribution-free analysis of univariate data. J Qual Technol 2017 Nov 21;45(3):273-284. [doi: 10.1080/00224065.2013.11917938]
16. Shayesteh B, Fu C, Ebrahimzadeh A, Glitho RH. Automated concept drift handling for fault prediction in edge clouds using reinforcement learning. IEEE Trans Netw Serv Manage 2022 Jun;19(2):1321-1335. [doi: 10.1109/tnsm.2022.3153279]
17. Nestor B, McDermott MB, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. arXiv Preprint posted online August 02, 2019 [FREE Full text]

18. Fu S, Wen A, Schaeferle GM, Wilson PM, Demuth G, Ruan X, et al. Assessment of data quality variability across two EHR systems through a case study of post-surgical complications. AMIA Jt Summits Transl Sci Proc 2022;2022:196-205 [FREE Full text] [Medline: 35854735]
19. Dockès J, Varoquaux G, Poline JB. Preventing dataset shift from breaking machine-learning biomarkers. Gigascience 2021 Oct 28;10(9):giab055 [FREE Full text] [doi: 10.1093/gigascience/giab055] [Medline: 34585237]
20. Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TW, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Sci Rep 2021 Dec 26;11(1):23017 [FREE Full text] [doi: 10.1038/s41598-021-02481-y] [Medline: 34837021]
21. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. BMJ Open 2021 Mar 16;11(3):e045343 [FREE Full text] [doi: 10.1136/bmjopen-2020-045343] [Medline: 33727273]
22. Woodall WH, Spitzner DJ, Montgomery DC, Gupta S. Using control charts to monitor process and product quality profiles. J Qual Technol 2018 Feb 16;36(3):309-320. [doi: 10.1080/00224065.2004.11980276]
23. Petersen C, Smith J, Freimuth RR, Goodman KW, Jackson GP, Kannry J, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. J Am Med Inform Assoc 2021 Mar 18;28(4):677-684 [FREE Full text] [doi: 10.1093/jamia/ocaa319] [Medline: 33447854]
24. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based Software as a Medical Device (SaMD) - discussion paper and request for feedback. U.S. Food & Drug Administration. URL: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf [accessed 2023-07-05]
25. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. IEEE Trans Knowl Data Eng 2018 Oct 18;31(12):2346-2363. [doi: 10.1109/tkde.2018.2876857]
26. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv 2014 Mar;46(4):1-37. [doi: 10.1145/2523813]
27. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Japkowicz N, Stefanowski J, editors. Big Data Analysis: New Algorithms for a New Society. Cham, Switzerland: Springer; 2016.
28. Hwang W, Runger G, Tuv E. Multivariate statistical process control with artificial contrasts. IIE Transact 2007 Mar 22;39(6):659-669. [doi: 10.1080/07408170600899615]
29. Deng H, Runger G, Tuv E. System monitoring with real-time contrasts. J Qual Technol 2017 Nov 21;44(1):9-27. [doi: 10.1080/00224065.2012.11917878]
30. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health 2020 Jul;2(6):e279-e281 [FREE Full text] [doi: 10.1016/S2589-7500(20)30102-3] [Medline: 33328120]

Abbreviations

GUI: graphical user interface
ML: machine learning

Edited by C Lovis; submitted 07.07.23; peer-reviewed by H Joo, C Yu, M Bjelogrlic, H Mueller; comments to author 03.08.23; revised version received 22.08.23; accepted 04.05.24; published 28.06.24.

Please cite as:
Faust L, Wilson P, Asai S, Fu S, Liu H, Ruan X, Storlie C
Considerations for Quality Control Monitoring of Machine Learning Models in Clinical Practice
JMIR Med Inform 2024;12:e50437
URL: https://medinform.jmir.org/2024/1/e50437
doi:10.2196/50437
PMID:38941140

Viewpoint

Bridging Real-World Data Gaps: Connecting Dots Across 10 Asian Countries

Guilherme Silva Julian[1], MSc; Wen-Yi Shau[2], MD, PhD; Hsu-Wen Chou[3], PhD; Sajita Setia[4], MD

[1]Pfizer Brazil, São Paulo, Brazil

[2]Pfizer Corporation Hong Kong Limited, Hong Kong, China (Hong Kong)

[3]Pfizer Ltd, Taipei, Taiwan

[4]Executive Office, Transform Medical Communications Limited, Wanganui, New Zealand

Corresponding Author:
Sajita Setia, MD
Executive Office, Transform Medical Communications Limited
184 Glasgow Street
Wanganui, 4500
New Zealand
Phone: 64 276175433
Email: sajita.setia@transform-medcomms.com

Abstract

The economic trend and the health care landscape are rapidly evolving across Asia. Effective real-world data (RWD) for regulatory and clinical decision-making is a crucial milestone associated with this evolution. This necessitates a critical evaluation of RWD generation within distinct nations for the use of various RWD warehouses in the generation of real-world evidence (RWE). In this article, we outline the RWD generation trends for 2 contrasting nation archetypes: “Solo Scholars”—nations with relatively self-sufficient RWD research systems—and “Global Collaborators”—countries largely reliant on international infrastructures for RWD generation. The key trends and patterns in RWD generation, country-specific insights into the predominant databases used in each country to produce RWE, and insights into the broader landscape of RWD database use across these countries are discussed. Conclusively, the data point out the heterogeneous nature of RWD generation practices across 10 different Asian nations and advocate for strategic enhancements in data harmonization. The evidence highlights the imperative for improved database integration and the establishment of standardized protocols and infrastructure for leveraging electronic medical records (EMR) in streamlining RWD acquisition. The clinical data analysis and reporting system of Hong Kong is an excellent example of a successful EMR system that showcases the capacity of integrated robust EMR platforms to consolidate and produce diverse RWE. This, in turn, can potentially reduce the necessity for reliance on numerous condition-specific local and global registries or limited and largely unavailable medical insurance or claims databases in most Asian nations. Linking health technology assessment processes with open data initiatives such as the Observational Medical Outcomes Partnership Common Data Model and the Observational Health Data Sciences and Informatics could enable the leveraging of global data resources to inform local decision-making. Advancing such initiatives is crucial for reinforcing health care frameworks in resource-limited settings and advancing toward cohesive, evidence-driven health care policy and improved patient outcomes in the region.

(JMIR Med Inform 2024;12:e58548) doi:10.2196/58548

KEYWORDS

Asia; electronic medical records; EMR; health care databases; health technology assessment; HTA; real-world data; real-world evidence

Introduction

Real-world data (RWD) in medical and health research describes data relating to patients’ health status or the delivery of health care in an environment outside of conventional clinical trials. This includes data routinely collected for treatment and disease registries, electronic medical records (EMRs), insurance claims, and other health databases that collect information reported by patients or health care professionals [1,2]. Extending from this concept of RWD is real-world evidence (RWE), which is the analyses produced from appropriate, well-designed studies using RWD [3].

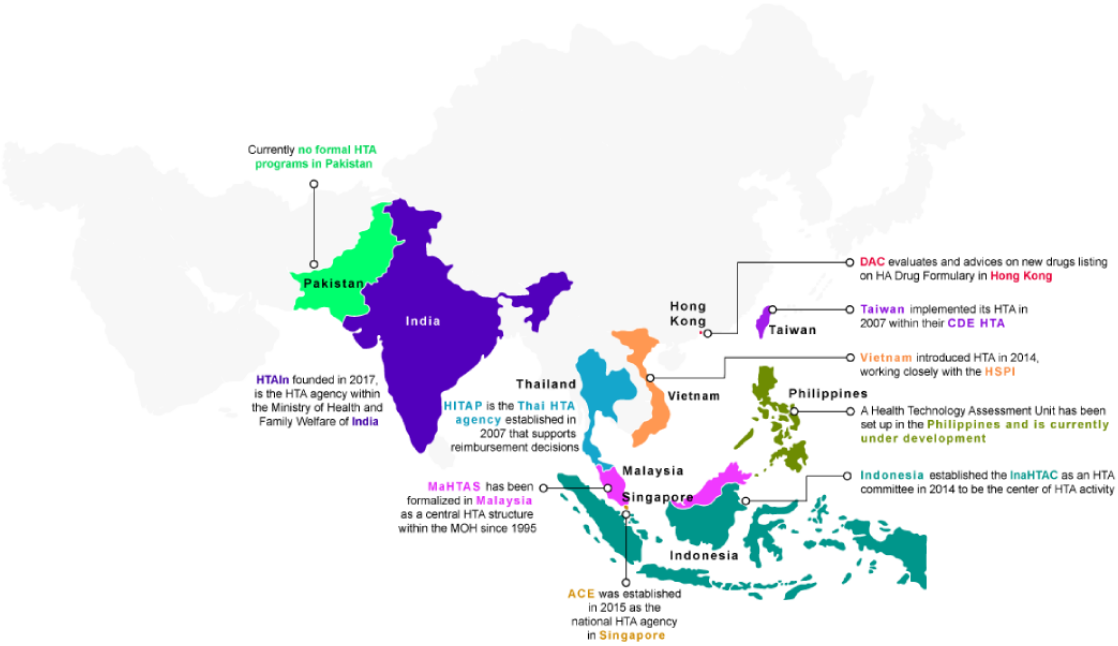
In randomized controlled trials, patients with severe forms of disease or multiple comorbidities are typically excluded from

participating due to potential risks or adverse events [2]. This leads to a large number of underrepresented patients, reducing the generalizability of interventions in the population, particularly for those who may need it most. RWD can provide the external validity needed when exploring different interventions and treatment strategies across the health care system [2]. To provide external validity for different interventions and disease management strategies, different study styles are used for RWD generation. These can be broadly categorized as comparative effectiveness research (CER) and descriptive studies (non-CER) [4]. CER can be defined as studies that primarily compare interventions and strategies, while non-CER are observational studies that aim to provide a descriptive overview of factors such as disease prevalence or treatment patterns [4,5].

Two scoping reviews have been conducted to explore the spectrum of RWE from linked databases in Asia and their possible impact on health care evolution, strategy, and policy [6,7]. The literature search for these reviews was conducted on PubMed in September 2022 and May 2023, and analysis of RWD publications across these countries assumed a linear distribution of studies [6,7]. In the initial scoping review, the trends and research warehouses for RWD-published studies from 3 countries in Asia with varying health care systems, Taiwan, India, and Thailand, were evaluated. The second scoping review continued this explorative research in 7 other diverse Asian health care systems using the established protocol from the prior review [4,7]. This study aimed to understand the evolving landscape of RWD use and its implications across Hong Kong, Indonesia, Malaysia, Pakistan, the Philippines, Singapore, and Vietnam. The number of total publications and single-country studies (SCS) and cross-country collaboration studies (CCCS) from all the countries in these scoping reviews was used to archetype them as either “Solo Scholars” if they published predominantly SCS with relatively higher number of studies published in last 5 years or “Global Collaborators” if they published less numbers and predominantly CCCS [7]. Hong Kong, India, Malaysia, Singapore, Taiwan, and Thailand were categorized as Solo Scholars, and Indonesia, Pakistan, the Philippines, and Vietnam were categorized as Global Collaborators (Multimedia Appendix 1).

Using these archetypes generated from the 2 scoping reviews, this viewpoint review intends to evaluate evolving trends and patterns in the use of various RWD warehouses and contextualize these findings within the broader health care and economic trends. Studying these evolving trends is crucial as it helps understand how RWD warehouses are generating RWE, which in turn informs health care policy and economic decisions [8]. Health technology assessment (HTA) organizations evaluate the clinical and economic aspects of medicines and health care technologies and recommend reimbursement and other policy criteria. Country HTAs are increasingly incorporating RWE for crucial complementary evidence, particularly in cost-effectiveness analyses [8,9]. However, data accessibility issues, lack of knowledge on robust methodologies, and a shortage of qualified researchers limit the generation of RWE for regulatory and reimbursement decision-making [10]. Our selection of countries for this scoping review was strategically based on the contrasting spectrum of HTA expertise at different timelines of development, from relatively mature systems in Taiwan, Singapore, Hong Kong, Malaysia, and Thailand to emerging frameworks in India, Indonesia, the Philippines, and Vietnam, and nascent systems in Pakistan (Figure 1).

Figure 1. Timeline of HTA development across selected 10 countries in Asia. ACE: Agency for Care Effectiveness; CDE: Centre for Drug Evaluation; DAC: Drug Advisory Committee; HA: Hospital Authority; HITAP: Health Intervention and Technology Assessment Program; HSPI: Health Strategy and Planning Institute; HTA: health technology assessment; HTAIn: Health Technology India; InaHTAC: Indonesian Health Technology and Assessment Committee; MaHTAS: Malaysian Health Technology Assessment; MOH: Ministry of Health; UHC: universal health coverage.



Country Trends and Key Identified Databases for 10 Asian Countries

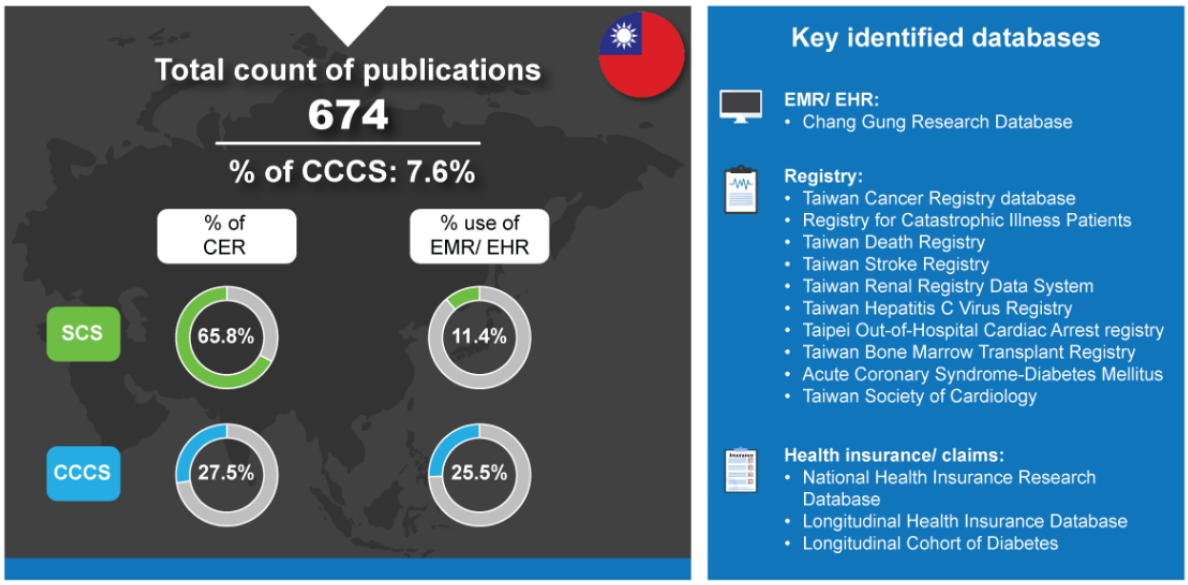
Harnessing RWD is pivotal for improving health care outcomes, and understanding country-specific trends and the databases that underpin this research is crucial [11,12]. The following section provides a granular view of the health care and RWD landscape across 10 Asian countries, dissecting the trends in research publications and the key linked databases that have been pivotal in generating RWE (Multimedia Appendix 2 [13-20]). Countries are listed in order of highest to lowest counts of generated RWD publications in the last few years [6,7].

Solo Scholars

Taiwan

Taiwan is a thriving democracy with a population of 23.3 million and a prosperous market economy. Taiwan's gross domestic product (GDP) grew by 2.45% in 2022 [21]. Development of the digital infrastructure in health care is one of the Taiwan government's major goals in the next few years [22]. Although there seems a lower overall usage of EMRs or electronic health records (EHRs) in SCS for Taiwan (71/623, 11.4%) (Figure 2), there was a noticeable increase in their usage from 2.8% (1/36) to 19.4% (20/103) from 2017 to 2022 [6]. As shown in Figure 2, Taiwan has a high percentage of CER in their SCS (410/623, 65.8%). Inversely, the trend for higher CER is not seen in the CCCS (14/51, 27.5%).

Figure 2. Real-world data landscape for Taiwan (2017-2022). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



Taiwan has a robust health care system, which is reflected in the percentages describing their RWD publications. Following the creation of the National Health Insurance (NHI) in 1995, Taiwan has set a benchmark in the region with a successful universal health coverage (UHC) scheme in their health care system [23]. Enrollment in the NHI provides coverage for essential medical care and pharmaceuticals for all residents [24]. The success of this system is exemplified in Taiwan's favorable result on the world's health and health systems ranking in 2023 [25], on which they were placed fourth (based on the Legatum Prosperity Index that also accounts for economic and social well-being). Furthermore, most health care facilities in Taiwan upload the claims data of each visit to NHI, including patient visits, drug prescriptions, surgeries, and examinations [23]. However, Taiwan's NHI has changed since its creation in 1995 and continues to evolve [26]. In 2007, Taiwan implemented its HTA within their Centre for Drug Evaluation to evaluate the financial suitability and clinical effectiveness of new drugs for reimbursement decisions [27]. This evolved into the formation of the Division of HTA in 2008 [28]. The high proportion of NHI drug expenses was a huge burden on Taiwan's health care system, and there were general recommendations to have a reduction in drug expenditure [29]. As such, with this goal of reducing drug expenditures, some behind-the-counter products were delisted from Taiwan's national Drug Reimbursement Scheme [23]. Along with this, coverage change was aimed at reshaping patient expectations and attitudes toward health. Taiwan's vision for 2030 includes "precision health," which places an important focus on health promotion. This concept of precision health involves using data to gather sufficient information that can be used to predict health risks and prevent diseases at the population level [30]. This vision relies on adopting descriptive RWD to produce RWE that can aid in the refinement of specific health policies that contribute to precision health and less comparative research. This also partially explains why there was a decrease in the publication of CER studies most notably from 75% in 2018 to 60% in 2022 [6]. It is worth noting that the literature search on the publication of RWE from linked databases in Taiwan was conducted in September 2022 on PubMed, and the analysis assumed a linear distribution for studies in 2017 and 2022.

Taiwan stands out as a significant contributor to RWD studies in Asia, predominantly through SCS, because it employs a diverse array of databases and holds a wealth of insurance claims data.

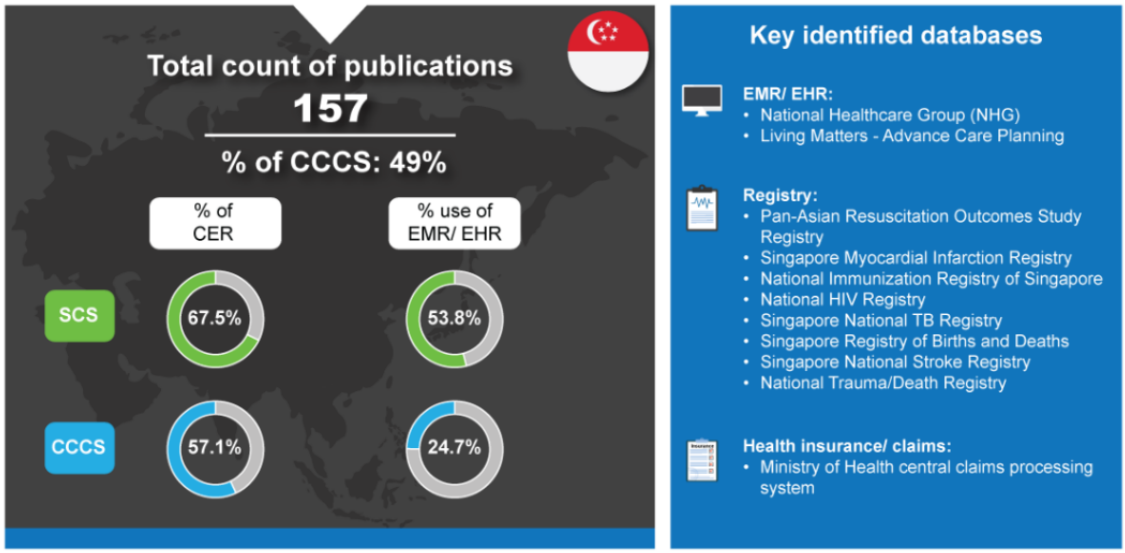
The key identified databases contributing to RWD publications included the following [6]:

- *Chang Gung Research Database (CGRD)*: CGRD is a multi-institutional EMR database collected from the Chang Gung Memorial Hospital system, which is the largest medical system in Taiwan [31]. Except for 2 municipal hospitals, all EMR data from Chang Gung Memorial Hospital are included in the CGRD. The database includes approximately 6%-20% outpatient and 10%-12% inpatient claims records [32,33].
- *Taiwan Cancer Registry (TCR) Database*: TCR was developed to monitor cancer incidence at the national level. The TCR holds basic information on patients with newly diagnosed cancer from hospitals with >50 beds throughout the country [34].
- *Registry for Catastrophic Illness Patients*: It is associated with the National Health Insurance Research Database (NHIRD) and is used to identify patients with catastrophic illness in the Taiwan health insurance system [35].
- *Taiwan Death Registry*: It contains information on accurate causes of death and dates for all residents of Taiwan [36].
- *Taiwan Stroke Registry*: This is a nationwide hospital-based registry that enrolls patients with stroke from 19 academic medical centers, 37 regional hospitals, and 8 district hospitals. The data are collected prospectively by trained neurologists and study nurses [37].
- *Taiwan Renal Registry Data System*: It collects patients' clinical and laboratory information from all dialysis units in Taiwan [38].
- *Taiwan Hepatitis C Virus Registry*: This registry program is a nationwide hepatitis C virus platform implemented by the Taiwan Association for the Study of the Liver [39].
- *Taipei Out-of-Hospital Cardiac Arrest Registry*: It includes prehospital and hospital information on patients with out-of-hospital cardiac arrest in Taipei [40].
- *Taiwan Bone Marrow Transplant Registry*: This registry holds clinical data from consecutive allogeneic hematopoietic cell transplant recipients from 15 transplant centers in Taiwan [41].
- *Acute Coronary Syndrome-Diabetes Mellitus Registry*: It is a nationwide registry of patients with acute coronary syndrome in Taiwan that collects real-world clinical practices and outcomes data [42].
- *Taiwan Society of Cardiology Registry*: This registry collects data from patients at 21 medical centers or teaching hospitals in Taiwan from patients who are hospitalized with acute new-onset heart failure or acute decompensated chronic heart failure with a reduced ejection fraction [43].
- *NHIRD*: It covers >99.6% of the Taiwanese population and collects claims data from outpatient and inpatient hospital care settings [26].
- *Longitudinal Health Insurance Database*: This database is derived from the NHI system and includes the registration files and original reimbursement claims of a million randomly selected beneficiaries, under the NHI program [44].
- *Longitudinal Cohort of Diabetes*: This is a de-identified subset of data from the NHIRD [45].

Singapore

Singapore is a prosperous nation with one of the highest GDP per capita in Asia. In 2022, Singapore's GDP grew by 3.6% [46]. In 2023, the Singapore government announced the plan to spend US $2.5 billion on info-communications technology with the redevelopment of major hospitals [47]. Compared with Taiwan, Singapore relies on higher use of EMRs or EHRs for real-world SCS (43/80, 53.8%), which is balanced with their use of clinical registries (46/80, 57.5%) [7]. Although Singapore SCS was predominantly in CER style (54/80, 67.5%), unlike Taiwan, the percentage of CCCS publications remained relatively high (57/77, 57.1%; Figure 3).

Figure 3. Real-world data landscape for Singapore (2018-2023). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies; TB: tuberculosis.



In 2023, Singapore was regarded as top-ranking in the world's health and health systems, 3 places ahead of Taiwan [25]. Despite emerging issues surrounding health care services and the affordability of health care, Singapore is able to achieve good health outcomes with a health index score of 86.9 through a hybrid public and private health care model that is dominated by public hospitals in the hospital sector and private clinics in the outpatient sector [25,48]. Trusted Research and Real World Data Utilization or TRUST platform stands as the cornerstone for national data exchange in Singapore by facilitating the secure and anonymized sharing of health-related RWD between public and private health care sectors [49]. The Ministry of Health (MOH) aimed to increase health care capacity to enhance health care affordability in the Healthcare 2020 Masterplan [50]. Within the MOH, the Agency for Care Effectiveness was established in 2015 as the national HTA agency with the aim of evaluating drugs for subsidization and medical technologies [51,52]. This may partially explain Singapore's relatively high percentage of CER publications in both SCS and CCCS, particularly as these types of studies enable a comparison of different medications and health management strategies. This is further magnified by the observation that the percentage of SCS CER publications from Singapore increased from 2018 to 2023 (analysis of these publications assumed a linear distribution for studies in 2018 and 2023) [7].

The key identified databases contributing to RWD publications from Singapore included the following [7]:

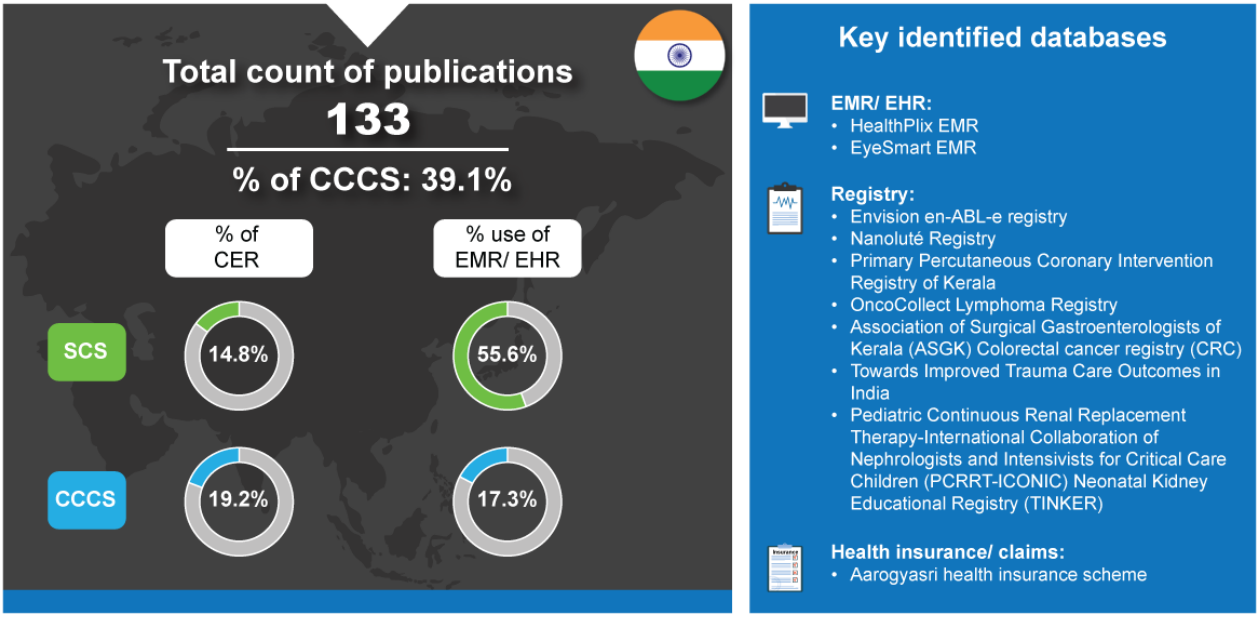
- *National Healthcare Group*: This record links key administrative and clinical information from a group of National Healthcare Group health care institutions for patients with chronic diseases such as diabetes mellitus, hypertension, dyslipidemia, stroke, cardiovascular diseases, and chronic renal disease [53].
- *Living Matters—Advance Care Planning (ACP)*: This framework provides a comprehensive web-based resource about ACP, including options for documenting patients' care preferences, particularly valuable in emergencies or when making critical care decisions. ACP details are also integrated into EMR or EHR, and all health care providers involved in patient care can easily access and understand a patient's end-of-life care preferences [54].
- *Pan-Asian Resuscitation Outcomes Study Registry*: The Pan-Asian Resuscitation Outcomes Study registry contains information from dispatch centers, ambulances, and hospitals from 7 countries in the Asia Pacific region (Japan, South Korea, Taiwan, Thailand, United Arab Emirates-Dubai, Singapore, and Malaysia) [55].
- *Singapore Myocardial Infarction Registry*: It is an island-wide registry that is being managed by the National Registry of Diseases Office that contains epidemiological data on acute myocardial infarction cases diagnosed in public and private sector hospitals and some out-of-hospital deaths certified by medical practitioners in Singapore [56].
- *National Immunization Registry of Singapore*: This registry collects and maintains accurate, complete, and current vaccination records of children and adults living in Singapore [57].
- *National Human Immunodeficiency Virus Registry*: This registry is a name - based system that holds identifiable data for known human immunodeficiency virus cases in Singapore, as well as contacts of cases that were reported [58].
- *Singapore National Tuberculosis Registry*: This registry is electronically linked to the 2 mycobacterial culture laboratories in Singapore and captures all positive Mycobacterium tuberculosis complex culture results in the country [59].
- *Singapore Registry of Births and Deaths*: This death registry is maintained by the Ministry of Home Affairs Immigration and Checkpoints Authority. It collects data on the cause and date of death of all Singaporeans and permanent residents in Singapore [60].
- *Singapore National Stroke Registry*: This is a countrywide registry of risk factors, stroke subtypes, management, and outcomes of incident and recurrent stroke in Singapore [61].
- *National Trauma/Death Registry*: This is a registry that contains anatomical injury codes, indicators of physiological response to injury, and patient demographics [62].
- *MOH's Central Claims Processing System*: It is used in Singapore to process the patient's MediSave and MediShield claims [63].

India

India is the largest lower-middle-income country in the world and accounts for around 18% of the total population [64]. According to Nexdigm, the Indian health care industry is expected to reach more than US $610 billion by 2026, as there is a growing demand for specialized and higher-quality health care facilities [65]. Hospitals, clinical trials, telemedicine, medical tourism, medical devices, medical and diagnostic equipment, and health insurance are among the key products and services that would drive this growth [66].

Compared with Taiwan and Singapore, India has a less extensive health care system [25], which is also reflected in its comparatively lower percentage of SCS (12/81, 14.8%) and CCCS (10/52, 19.2%; Figure 4). India attracts a growing medical tourism market with a growth of 22%-25% from 2014 and advocacy for adoption of EMR across the country [67]. Among RWD databases, the studies primarily use EHR or EMR (45/81, 55.6%), with an increasing trend from 20% in 2017 to 48% in 2022 [6]. Its usage is expected to continue to rise with the National Digital Health Mission, since 2020 [68]. India has made commitments toward achieving UHC and has been producing policies and institutional changes that are directed toward increasing health coverage and access to health services [69]. Government-funded health care sector is the provider of health care to lower-income populations; however, the private health care sector is the dominant health care provider [70]. In 2020, 70% of hospital market share was controlled by private sector providers, and 63% of hospital beds were provided by the hospital sector [71,72].

Figure 4. Real-world data landscape for India (2017-2022). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



India launched Ayushman Bharat—one of the most ambitious health missions ever to achieve UHC in 2018 [73]. Ayushman Bharat encompasses 2 complementary schemes, Health and Wellness Centres and the National Health Protection Scheme [69]. Along with this, there have been other initiatives to achieve UHC [69]. The Pradhan Mantri Jan Arogya Yojana scheme provides secondary and tertiary hospital care insurance to the bottom 40% of the population [68]. Previously called the "Medical Technology Assessment Board," HTAIn, founded in 2017, is the HTA agency within the Ministry of Health and Family Welfare tasked with developing an HTA system to aid in decision-making for resource allocation at the national and state levels [74,75]. Reflecting these policy aims, the proportion of CER for SCS and CCCS was low in India but increased from 2017 to 2022 (12/81, 14.8% and 10/52, 19.2%, respectively, Figure 4), assuming a linear distribution for studies in 2017 and 2022 [6].

The key identified databases contributing to RWD publications from India included the following [6]:

- *HealthPlix EMR*: This is an artificial intelligence–powered electronic medical software system in India [76].
- *EyeSmart EMR*: This is an EMR and Hospital Management System in India that integrates the clinical, surgical, financial, and operational functions of the LV Prasad Eye Institute on a single platform [77].
- *Envision en-ABL-e Registry*: This was a multicenter registry that enrolled 2500 patients treated with 3286 Abluminus DES (Envision Scientific, Surat, India) across 31 centers across the country from June 2012 to December 2018 [78].
- *Nanoluté Registry*: Nanoluté registry was used to observe the clinical performance of a novel sirolimus-coated balloon (Concept Medical Research Private Limited, India) for treating coronary de novo and restenotic lesions [79].
- *Primary Percutaneous Coronary Intervention Registry of Kerala*: This registry is a large multicenter primary percutaneous coronary intervention registry from Kerala, India. It reports long‐term outcomes of patients presenting with ST‐segment–elevation myocardial infarction to percutaneous coronary intervention–capable hospitals or facilities [80].
- *OncoCollect Lymphoma Registry*: This registry was set up in 2017 as a collaborative group effort to evaluate current practices for managing diffuse large B-cell lymphoma in middle-income countries [81].
- *Association of Surgical Gastroenterologists of Kerala Colorectal Cancer Registry*: This registry collects demographics and perioperative outcomes of colorectal cancer in Kerala from volunteer members of the Association of Surgical Gastroenterologists of Kerala [82].
- *Towards Improved Trauma Care Outcomes in India*: It is a multicenter trauma registry that contains data on trauma patients admitted to 4 public university hospitals in Mumbai, Delhi, and Kolkata from October 1, 2013, to September 30, 2015 [83].
- *The Indian Pediatric Continuous Renal Replacement Therapy-International Collaboration of Nephrologists and Intensivists for Critical Care Children Neonatal Kidney Educational Registry*: This registry is a database of all admitted neonates ≤28 days who received intravenous fluids for at least 48 hours [84].
- *Aarogyasri Health Insurance Scheme*: This is a social insurance scheme with a private-public partnership model to deal with the problems of medical expenditures for poor households [85].

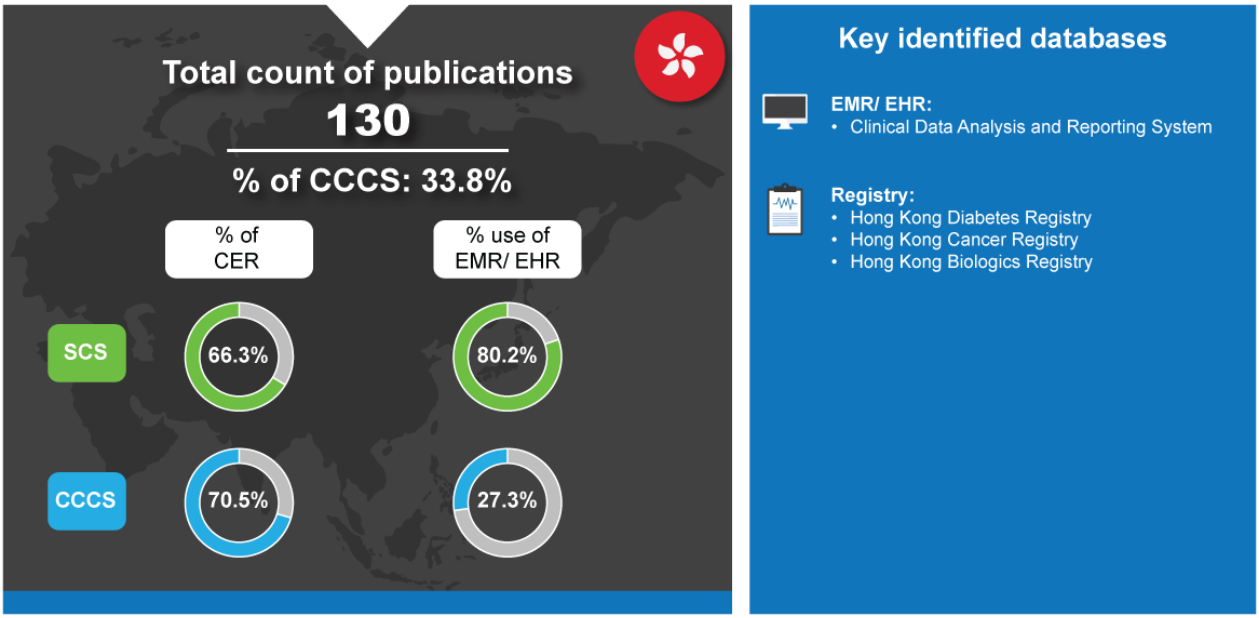
Hong Kong

Health care services in Hong Kong are provided by both private and government-funded public sectors, with public medical services provided by the Department of Health and the Hong Kong Hospital Authority (HA) [86]. Hong Kong has a well-developed health care system that ranked 14th among the health and health systems ranking of countries worldwide in 2023 [25]. Hong Kong's system is a parallel, segmented public

and private financed and provided health care, where the public sector accounted for 51% of total health expenditure and the private sector accounted for 49% in 2017/18 [87]. With the government's commitment to provide UHC, the Drug Advisory Committee is used to evaluate and advise on new drugs listed on the Hospital Authority Drug Formulary, which is the largest public health care service provider in Hong Kong. One of the missions of the Drug Advisory Committee is to ensure equal access of patients to cost-effective drugs with proven safety and efficacy [88].

As shown in Figure 5, CER publications made up 66.3% (57/86) of SCS publications and 70.5% (31/44) of CCCS. To put these high percentages into context, the drug appraisal and review process involves an assessment of clinical evidence and health economic evidence that is then used to make recommendations on reimbursement of new drugs [89]. In Hong Kong, there has been an increase in the requirement for a systematic HTA, placing emphasis on the value and comparison of emerging and conventional drugs [88]. In line with this, there is an upward trend in the percentage of SCS CER publications for Hong Kong from 2018 to 2023, assuming a linear distribution for studies in 2018 and 2023 [7].

Figure 5. Real-world data landscape for Hong Kong (2018-2023). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



The key identified databases contributing to RWD publications from Hong Kong included the following [7]:

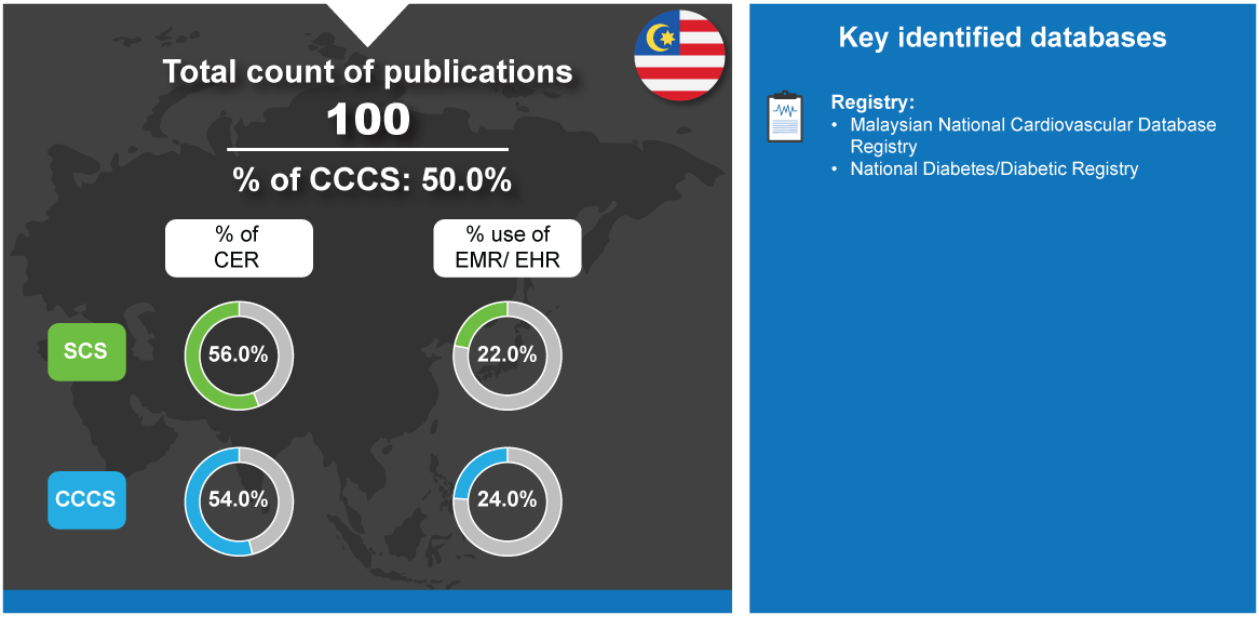
- *Clinical Data Analysis and Reporting System (CDARS)*: CDARS is a centralized database developed for research and audit purposes. It is managed by the HA. It collects anonymized records of demographics, admission, prescription, diagnosis, procedure, laboratory test, and death information [90].
- *Hong Kong Diabetes Registry*: This is a diabetes registry used for quality assurance and risk stratification to facilitate subsequent triage of patients to different clinic settings [91].
- *Hong Kong Cancer Registry*: This is a population-based registry committed to collecting and conducting analyses on data from all cancer cases in Hong Kong [92].
- *Hong Kong Biologics Registry*: This is a registry that was established in December 2005 by The Hong Kong Society of Rheumatology to capture efficacy and safety data regarding the use of biological agents for the treatment of rheumatic disease [93].

Malaysia

Malaysia is an upper-middle-income country with a population of about 34 million [94]. It is located at the heart of Southeast Asia and represents one of the regional hubs for information and communication technology and medical travel [95]. Digitalization of information forms across major industrial sectors will help secure Malaysia's role in the future global economy. As of 2023, Malaysia was ranked 42nd on the health and health systems of countries worldwide [25]. Malaysia's health care system consists of tax-funded and highly subsidized government-led services, with a fast-growing private sector [96]. Their dual-tier system consists of a government-led public sector with an existing private sector, and in 2019, their public health expenditure amounted to 52% of their total health expenditure.

Malaysia's 2023 Health White Paper outlines a transformative masterplan for health management information and data system [97]. They aim to improve the health outcomes and well-being of Malaysians and the use of the Malaysian Health Technology Assessment, also termed as MaHTAS. HTA has been formalized in Malaysia as a central structure within the MOH since 1995 and is a trusted medical evidence source [97]. Their generation of RWD publication used registries, with no, or low contributions from other database types [7] (Figure 6). In line with other countries that use HTA, 56% (28/50) of SCS were CER, as were 54% (27/50) of the CCCS [7].

Figure 6. Real-world data landscape for Malaysia (2018-2023). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



The key identified databases contributing to RWD publications from Malaysia included the following [7]:

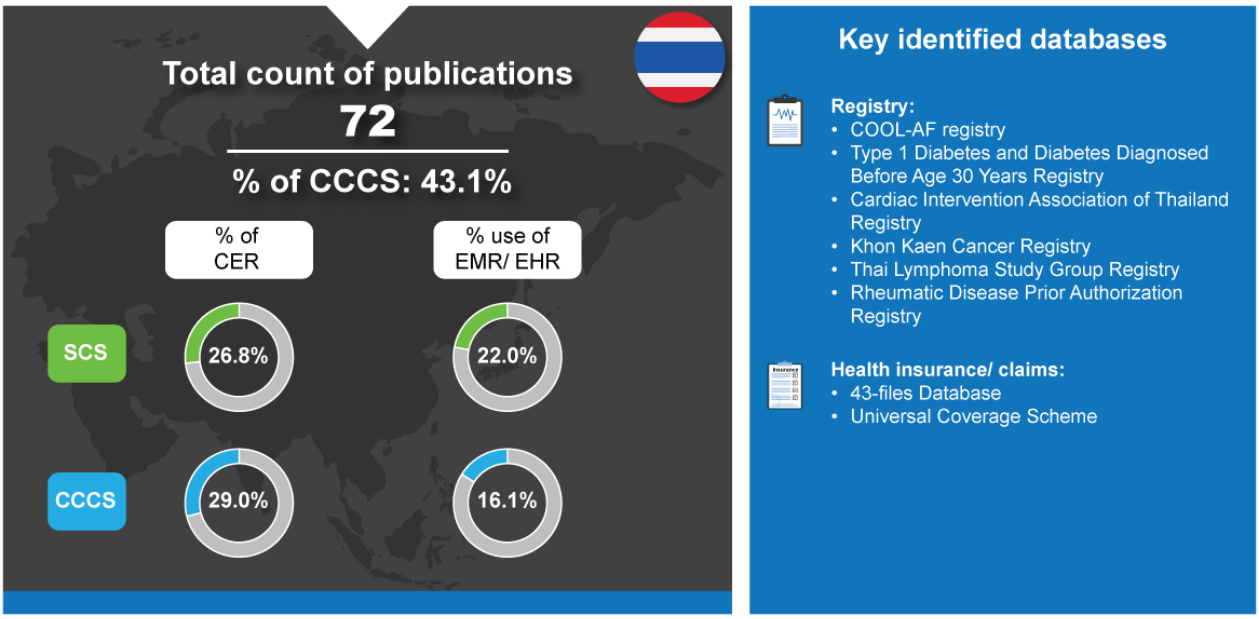
- *Malaysian National Cardiovascular Database Registry*: This is a service supported by the Malaysian MOH to collect information about cardiovascular diseases [98].
- *National Diabetes/Diabetic Registry*: This is a registry used to enable tracking of glycemic control and clinical outcomes of patients with diabetes managed at MOH health clinics [99].

Thailand

Thailand is also an upper-middle-income country but with the second largest economy (after Indonesia) in the Association of Southeast Asian Nations (ASEAN). Its GDP in 2022 was US $526 billion [100]. Thailand has the fourth greatest number of Joint Commission International accreditation–certified hospitals after Saudi Arabia, the United Arab Emirates, and Brazil. Yet the medical expenses in Thailand are 50%-80% lower than those in Europe, the United States, and Canada [101].

On the health and health systems ranking of countries worldwide in 2023, Thailand was ranked 13th [25]. In 2020, the private health expenditure in Thailand accounted for about 71.2% of the total health expenditure [102]. Furthermore, Thailand has become internationally known for its success with UHC policy since its development in 2002 [103]. The National List of Essential Medicines is a drug reimbursement list for the public health insurance schemes in Thailand, and the Health Intervention and Technology Assessment Program is a Thai HTA agency established in 2007 that supports reimbursement decisions [104]. Thailand used a relatively low percentage of EHR or EMR warehouses in their RWD publications, with only 22% contribution in SCS publications and 16.1% in CCCS publications (Figure 7). Surprisingly, they also had a relatively low percentage of CER publications with 26.5% of their SCS being CER publications and 29% of the CCCS being CER publications [6].

Figure 7. Real-world data landscape for Thailand (2017-2022). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; COOL-AF: Cohort of Antithrombotic Use and Optimal International Normalized Ratio Levels in Patients with Atrial Fibrillation; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



The key identified databases contributing to RWD publications from Thailand included the following [6]:

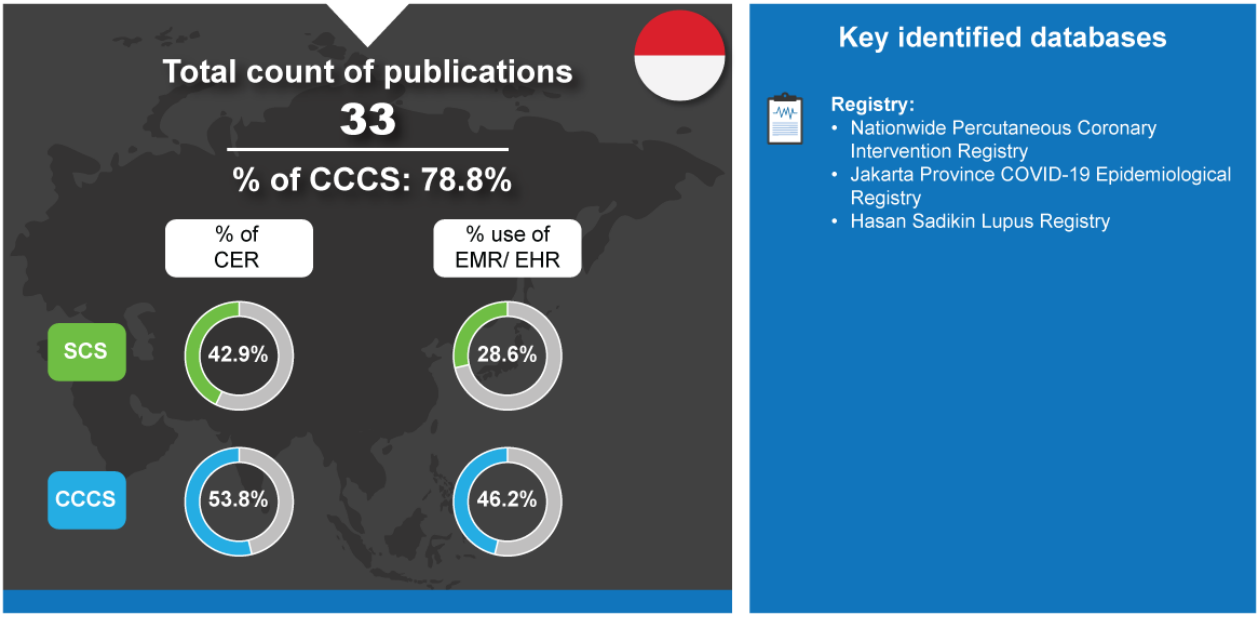
- *COOL-AF Registry*: The "Cohort of Antithrombotic Use and Optimal International Normalized Ratio Levels in Patients with Atrial Fibrillation (COOL-AF)" registry is a database of patients with atrial fibrillation in Thailand [105].
- *Type 1 Diabetes and Diabetes Diagnosed Before Age 30 Years Registry*: This registry was established in 2014 and involves 31 hospitals to evaluate glycemic control and complications in patients with type 1 diabetes [106].
- *Cardiac Intervention Association of Thailand Registry*: This nationwide registry was an initiative of the Cardiac Intervention Association of Thailand. All cardiac catheterization laboratories in Thailand were invited to participate [107].
- *Khon Kaen Cancer Registry*: This is a population-based cancer registry of Khon Kaen that covers 26 districts in Northeastern Thailand [108].
- *Thai Lymphoma Study Group Registry*: This is a web-based nationwide lymphoma registry from the Thai Lymphoma Study Group [109].
- *Rheumatic Disease Prior Authorization Registry*: This is a national registry used for government reimbursement in the Rheumatic Disease Prior Authorization system. This registry contains data on patients' demographic and clinical characteristics at baseline, data-related disease activity, and type of biologic medication prescribed [110].
- *43-Files Database*: These are administrative data collected by the Thai Ministry of Public Health for the purpose of reimbursement and health service use [111].
- *Universal Coverage Scheme*: This health coverage scheme consists of 3 main public insurance schemes that offer full-service coverage. The 3 schemes are the Civil Servants Medical Benefits Scheme for civil servants and their dependents, Social Health Insurance for private sector employees, and the Universal Coverage Scheme that covers 70% of the population in Thailand [112,113].

Global Collaborators

Indonesia

Indonesia is a country of 279.5 million people and is South East Asia's largest economy with a GDP of ~US $1.32 trillion in 2022. Ranked 97th in the 2023 health and systems ranking, Indonesia is progressing their UHC through the expansion of NHI (Jaminan Kesehatan Nasional scheme) [25,114,115]. The Indonesian government established the Indonesian Health Technology and Assessment Committee as an HTA committee in 2014 to be the center of HTA activity, and as a starting point for the development of HTA [116]. Similar to the other Solo Scholars in this study, Indonesia was responsible for a relatively low number of RWD publications [7]. Of their SCS, 28.6% (2/7) used EMR and 42.9% (3/7) were CER (Figure 8).

Figure 8. Real-world data landscape for Indonesia (2018-2023). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



The key identified databases contributing to RWD publications from Indonesia included the following [7]:

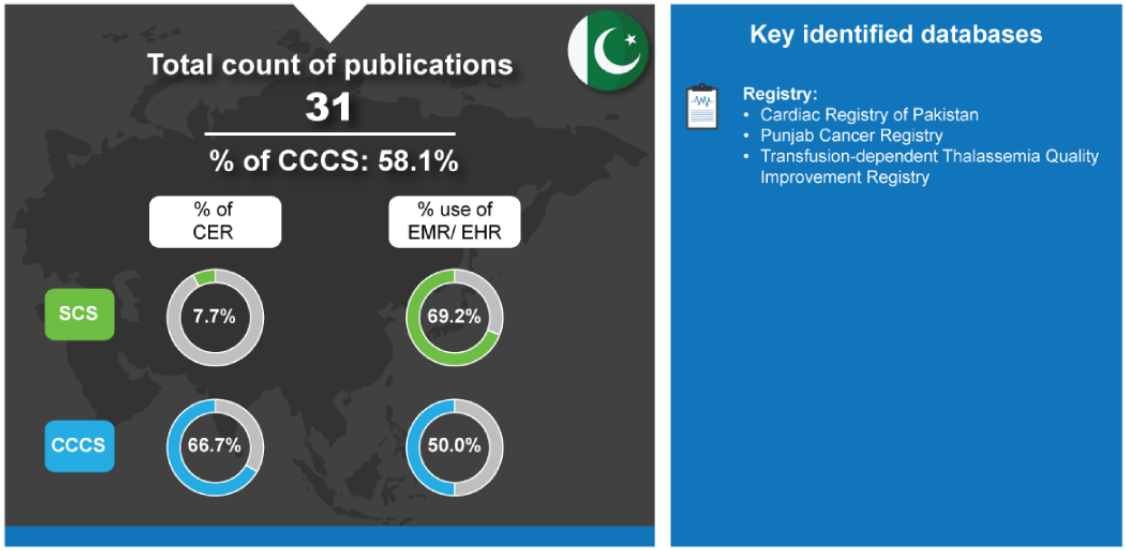
- *Nationwide Percutaneous Coronary Intervention Registry*: This is a multicenter registry of interventional cardiology projects involving 9 centers across Indonesia [117].
- *Jakarta Province COVID-19 Epidemiological Registry*: This is a database of patients with confirmed COVID-19 cases from Jakarta province [118].
- *Hasan Sadikin Lupus Registry*: This is a registry created in January 2016 that reports the medical records of patients with systemic lupus erythematosus from the Dr Hasan Sadikin General Hospital [119].

Pakistan

Health care has been identified as one of the best industry prospects for Pakistan, a country with a population of more than 240 million and a GDP growth rate of 0.29% in 2023 [120]. Ranked 124th on the health and health systems ranking of countries in 2023, Pakistan's health care comprises a 3-tier system of primary, secondary, and tertiary levels. The public and private sectors work together to provide the best possible care, but there have been tremendous concerns about the failure of the delivery of quality care due to various factors, ranging from inadequate infrastructure to inequitable distribution of health care facilities [121]. Of note, to our knowledge, there are currently no formal HTA programs in Pakistan [122].

These challenges are reflected in their low number of publications and, more specifically, their low number of SCS. Among the SCS, only 7.7% (1/13) were CER studies. Alternatively, due to their inadequate infrastructure, more CER was conducted in CCCS (12/18, 66.7%) (Figure 9).

Figure 9. Real-world data landscape for Pakistan (2018-2023). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



The key identified databases contributing to RWD publications from Pakistan included the following [7]:

- *Cardiac Registry of Pakistan*: This is a cardiac registry that includes 25 facilities from 3 provinces of Pakistan. Standardized data are collected every 3 weeks for this database [123].
- *Punjab Cancer Registry*: This registry collects population-level cancer statistics in Pakistan [124].
- *Transfusion-Dependent Thalassemia Quality Improvement Registry*: It is a database comprising patients with transfusion-dependent thalassemia from 4 centers in Karachi, Pakistan [125].

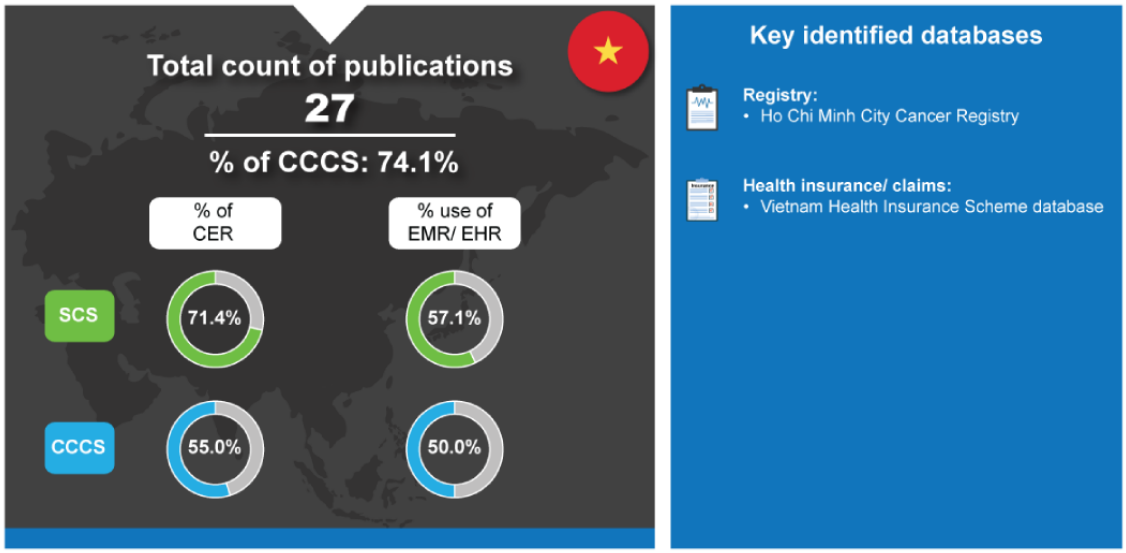
Vietnam

In the 2022 ASEAN Business Outlook Survey, AmCham Singapore members indicated Vietnam as the top Asia Pacific country (30%), where companies are considering expansion, followed by Malaysia (25%), Thailand (24%), and Indonesia (23%) [126]. The per capita GDP for Vietnam was US $4086 in 2022 and is meant to increase to at least US $18,000 by 2035. A large population of almost 100 million (half of which are younger than 30 years), consistent strong economic growth, and ongoing reform have created a dynamic and rapidly evolving commercial environment in Vietnam [126].

Vietnam was ranked 44th on the health and health systems ranking of countries in 2023 [25]. The Vietnamese MOH manages 3 levels of health service delivery: primary level in districts and communes, secondary level in provinces, and tertiary level in national institutions under central government control. Working within this hierarchical system, Vietnam aims to build a solid health care infrastructure from the grassroots [127]. Despite the significant progress, the health care system still faces challenges to keeping ahead of the country's escalating population [127]. Vietnam is making significant progress toward achieving UHC and is committed to delivering UHC [127]. In 2014, Vietnam introduced the HTA, working closely with the Health Strategy and Planning Institute to facilitate its institutionalization [128].

Similar to the other Global Collaborators, Vietnam generated a low percentage of publications. However, 71.4% (5/7) of SCS and 55% (11/20) of CCCS were CER (Figure 10). Based on the literature search of RWE publications from linked databases in Vietnam conducted in May 2023 on PubMed where the analysis assumed a linear distribution for studies in 2019 and 2023, Vietnam also exhibited the most significant increase with a growth rate of 24.5% in RWD publications in the last 5 years [7].

Figure 10. Real-world data landscape for Vietnam (2018-2023). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



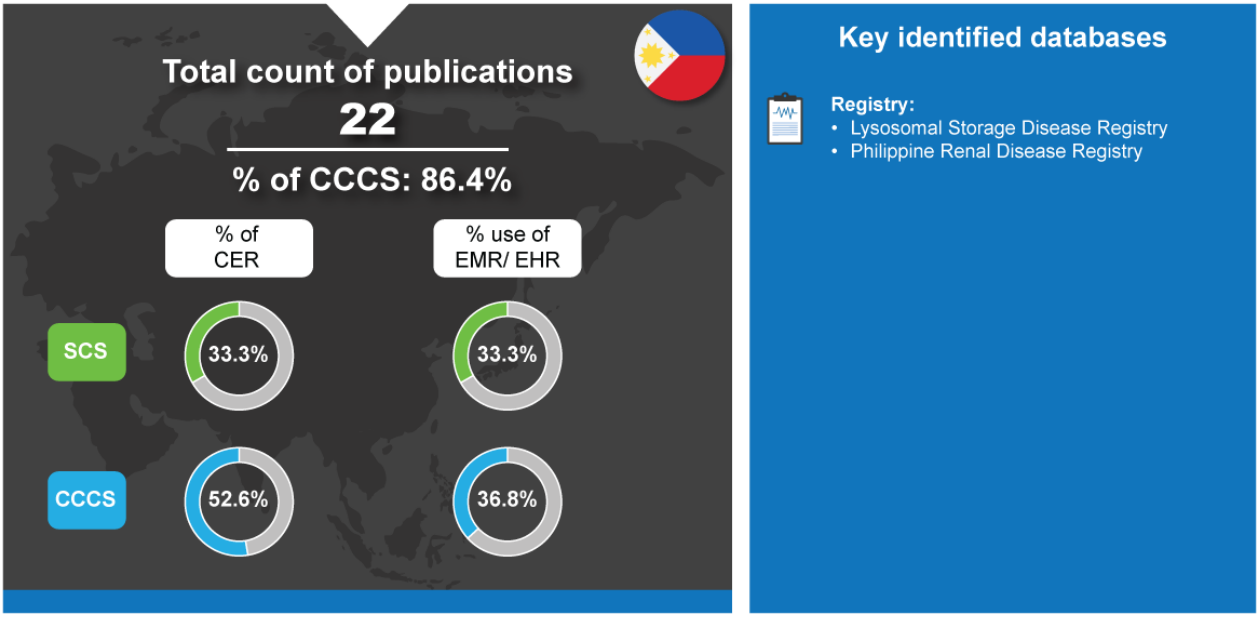
The key identified databases contributing to RWD publications from Vietnam included the following [7]:

- *Ho Chi Minh City Cancer Registry*: This registry documents all diagnosed cancer cases in Ho Chi Minh City [129].
- *Vietnam Health Insurance Scheme Database*: This claims database is managed by the Vietnam Social Security Service and it contains medical examinations and care, preventive care, rehabilitation, maternity care, and prescribed medications [130].

Philippines

The health care system is shared between the public and private sectors in Philippines and was ranked 96th on the health and health systems ranking of countries worldwide in 2023 [25]. To respond to health financing gaps, the Philippines made major reforms including the passage of the Universal Health Care Law in 2019; however, there are ongoing issues with expenditure [131]. Following this passage, there were increased efforts to institutionalize HTA in the Philippines and as a result, a health technology assessment unit was set up and is under development [52]. During the time period of 2018-2023, the majority of the Philippines' RWD publications were CCCS (19/22, 86.4%) and only 33.3% (1/3) each of SCS were CER and used EMR (Figure 11). Although the adoption of EMR for research is low and the publication count has been scanty in the last 5 years [7], the Philippines is estimated to spend approximately US $4.4 billion on digital infrastructure over the next 6 years [132].

Figure 11. Real-world data landscape for Philippines (2018-2023). CCCS: cross-country collaboration studies; CER: comparative effectiveness research; EHR: electronic health record; EMR: electronic medical record; SCS: single-country studies.



The key identified databases contributing to RWD publications from the Philippines included the following [7]:

- *Lysosomal Storage Disease Registry*: It is a registry of Filipinos diagnosed with any lysosomal storage disease [133].
- *Philippine Renal Disease Registry*: This is a registry consisting of 2 major components: The Chronic Renal Disease Registry and the End Stage Renal Disease Registry. The Chronic Renal Disease Registry is composed of the Renal Biopsy Registry, and the End Stage Renal Disease Registry is composed of Haemodialysis, Peritoneal Dialysis, and Transplant Registries [134].

Discussion

RWD is used to inform policies in different capacities across the countries. Focusing on the types and sources of RWD, we noted key differences across the countries. For example, Taiwan has a strong market economy [21], with a robust and successful health care system [25]. It was noted that they had fewer publications using EHR or EMR and more so used health insurance or administrative claims for their numerous publications. They also had many identified databases and, as such, were able to access several data sources to generate RWE (Figure 1). Taiwan exemplifies a country with the infrastructure and resources to comfortably produce RWE. A trend in Solo Scholar countries such as Taiwan was that they are prolific in the publication of SCS, with a general decrease in CER studies over time [6,7]. Typically, Global Collaborators have less robust economies and health care systems. These were countries that had less variety in data sources within their health system and a lack of standardized or integrated databases. These countries published more CCCS with a general increasing trend in the percentage of CER studies [7].

For Global Collaborators, CCCS are important for the generation of RWE. This highlights the necessity for data-sharing strategies and a need for the development of data-sharing frameworks, particularly to aid countries that have weaker health care systems and less established infrastructure for RWD [7]. It also underscores the potential benefits of opening data initiatives, where governments play a pivotal role [135]. Opening these data resources to a broader spectrum of stakeholders, including academic institutions, can spur innovation, improve public health outcomes, and foster a more collaborative ecosystem [136]. Such initiatives not only democratize access to valuable health data but also pave the way for a more inclusive approach to tackling global health challenges. This reflects a growing recognition of the importance of transparency and collaboration between the public sector and external entities in enhancing health care delivery and research [137]. Indeed, many nations such as Vietnam leverage regional and global studies conducted in a CER style to match their requirements for HTA. HTA agencies across Asia could contribute to these types of developments as this may be beneficial to them. In 2011, the HTAsiaLink Network was set up as a collaborative network for HTA agencies in Asia [138]. There is also the International Network of Agencies for Health Technology Assessment, which is an international HTA community, though to our knowledge, Taiwan, Singapore, and Malaysia are the only countries discussed that are part of this network [139]. Cross-sectoral partnerships through HTAsiaLink Network and International Network of Agencies for Health Technology Assessment could help address challenges related to RWE generation, which are much more common in lesser developed countries, and support country HTA and regulatory decision makers in decision-making [140,141]. However, there is also a need to improve database identification and integration across countries as there is no consistent naming of the databases, resulting in difficulty in identifying and analyzing RWD publications across Asia. Outside of reimbursement decisions, these improvements would be helpful for a range of groups including health professionals and researchers, health system managers, researchers, policy makers, and investors. Along with these improvements, there

is a need for focused use of EMR to streamline the process of generating RWD for disease surveillance and management [142].

In recent years, an increasing number of Asian countries have begun to recognize the importance of HTA in making informed pricing and reimbursement decisions for health care technologies [143]. This shift toward evidence-based health care policy is critical to ensure that decisions around health care technologies are transparent and equitable [144]. To further enhance the transparency and effectiveness of the HTA process, it is crucial to involve all stakeholders at every step of the implementation. In this context, global observational networks such as the Observational Health Data Sciences and Informatics (OHDSI) program offer a promising avenue for enriching HTA processes with robust RWE. Countries such as Taiwan, Singapore, and India already participate in OHDSI, although the global presence remains limited [145]. By aligning governmental initiatives with the Observational Medical Outcomes Partnership Common Data Model used by OHDSI, there is a unique opportunity to rapidly scale up the generation and use of RWE [146]. Such collaboration broadens the evidence base for HTA and facilitates international cooperation and benchmarking in health care technology assessment. Moreover, linking HTA processes with Observational and Medical Outcomes Partnerships or OHDSI could serve as a catalyst for developing sophisticated and evidence-driven health care policies.

In the Hong Kong government's 2021-2022 budget, an increase of US $480 million was allotted to the HA to meet increasing demand for health care services [86]. One such service is the CDARS, which emerged as the key EMR database for Hong Kong [7]. It is a database developed by HA, which is readily used in Hong Kong for RWD [90]. Among all 10 countries, Hong Kong leveraged EMR for most of its RWD SCS (80.2%) (Figure 4). Streamlined use of EMRs has the potential to collect a variety of RWD, which reduces the need to set up multiple disease-specific clinical registries for generating non–cost-related RWE. The CDARS database is a good example of this as it has been established for research purposes and contains necessary medical information, including patient demographics, details on admission, diagnosis, and prescription and laboratory tests [90].

Conclusions

RWD plays a significant role in informing policy decisions across Asia. There are differing trends and patterns in the use of databases for published RWD across Asia and clear gaps in usage of warehouses across countries. However, based on the economic and health care development trends, there seems a great potential in generating fit-for-purpose RWE across distinct health care systems. Global Collaborators demonstrate a reliance on international partnerships for CCCS, due to a strategic drive to overcome infrastructural limitations. The review stresses the necessity for enhanced data-sharing strategies and more robust database integration, which are critical for countries with limited health care system resources. Furthermore, the consistent naming and use of databases, especially EMRs, are pivotal for advancing disease surveillance and RWD generation across Asia. The findings call for joint efforts by HTA agencies and stakeholders to fortify RWD frameworks, which would not only aid reimbursement decisions but also support the broader spectrum of health care stakeholders.

Acknowledgments

Desktop research and medical writing for this article were funded by Pfizer and were conducted by Transform Medical Communications Limited (Transform Medcomms), New Zealand. The authors thank Ms Dansoa Tabi-Amponsah from Transform Medcomms for editorial assistance.

Data Availability

Data generated or analyzed during this study are included in Multimedia Appendices 1 and 2.

Authors' Contributions

All the authors were involved in the idea's conception, design, and interpretation of the facts and data. SS led the manuscript writing, and all authors were engaged in revising it for scientific content and final approval before submission for publication.

Conflicts of Interest

GSJ, H-WC, and W-YS declare that while being employees of Pfizer, there is no conflict of interest in relation to the work presented in this article. The views and opinions expressed herein are solely those of the author(s) and do not reflect the views or positions of their employers.

Multimedia Appendix 1
Methodology for plotting Solo Scholars and Global Collaborators archetypes.
[PDF File (Adobe PDF File), 316 KB - medinform_v12i1e58548_app1.pdf]

Multimedia Appendix 2
Database types used in real-world studies across 10 target Asian countries.

[PDF File (Adobe PDF File), 138 KB - medinform_v12i1e58548_app2.pdf]

References

1. Naidoo P, Bouharati C, Rambiritch V, Jose N, Karamchand S, Chilton R, et al. Real-world evidence and product development: opportunities, challenges and risk mitigation. Wien Klin Wochenschr 2021;133(15-16):840-846 [FREE Full text] [doi: 10.1007/s00508-021-01851-w] [Medline: 33837463]
2. Shau W, Setia S, Shinde S, Santoso H, Furtner D. Generating fit-for-purpose real-world evidence in Asia: how far are we from closing the gaps? Perspect Clin Res 2023;14(3):108-113 [FREE Full text] [doi: 10.4103/picr.picr_193_22] [Medline: 37554247]
3. Bhatt A. Conducting real-world evidence studies in India. Perspect Clin Res 2019;10(2):51-56 [FREE Full text] [doi: 10.4103/picr.PICR_8_19] [Medline: 31008069]
4. Shau WY, Setia S, Shinde SP, Santoso H, Furtner D. Contemporary databases in real-world studies regarding the diverse health care systems of India, Thailand, and Taiwan: protocol for a scoping review. JMIR Res Protoc 2022;11(12):e43741 [FREE Full text] [doi: 10.2196/43741] [Medline: 36512386]
5. Comparative effectiveness research. National Library of Medicine, National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/mesh/?term=comparative+effectiveness+research/ [accessed 2024-02-12]
6. Shau WY, Setia S, Chen YJ, Ho TY, Prakash Shinde S, Santoso H, et al. Integrated real-world study databases in 3 diverse Asian health care systems in Taiwan, India, and Thailand: scoping review. J Med Internet Res 2023;25:e49593 [FREE Full text] [doi: 10.2196/49593] [Medline: 37615085]
7. Shau WY, Santoso H, Jip V, Setia S. Integrated real-world data warehouses across 7 evolving Asian health care systems: scoping review. J Med Internet Res 2024;26:e56686 [FREE Full text] [doi: 10.2196/56686] [Medline: 38749399]
8. Verkerk K, Voest EE. Generating and using real-world data: a worthwhile uphill battle. Cell 2024;187(7):1636-1650. [doi: 10.1016/j.cell.2024.02.012] [Medline: 38552611]
9. Kc S, Lin LW, Bayani DBS, Zemlyanska Y, Adler A, Ahn J, et al. What, where, and how to collect real-world data and generate real-world evidence to support drug reimbursement decision-making in Asia: a reflection into the past and a way forward. Int J Health Policy Manag 2023;12:6858 [FREE Full text] [doi: 10.34172/ijhpm.2023.6858] [Medline: 37579427]
10. Zisis K, Pavi E, Geitona M, Athanasakis K. Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. J Pharm Pharm Sci 2024;27:12302 [FREE Full text] [doi: 10.3389/jpps.2024.12302] [Medline: 38481726]
11. Lee Y, Lee YJ, Ha IH. Real-world data analysis on effectiveness of integrative therapies: a practical guide to study design and data analysis using healthcare databases. Integr Med Res 2023;12(4):101000 [FREE Full text] [doi: 10.1016/j.imr.2023.101000] [Medline: 37953753]
12. Solà-Morales O, Sigurðardóttir K, Akehurst R, Murphy LA, Mestre-Ferrandiz J, Cunningham D, et al. Data governance for real-world data management: a proposal for a checklist to support decision making. Value Health 2023;26(4S):32-42 [FREE Full text] [doi: 10.1016/j.jval.2023.02.012] [Medline: 36870678]
13. Agency for Healthcare Research and Quality. AHRQ methods for effective health care. In: Gliklich RE, Dreyer NA, Leavy MB, editors. Registries for Evaluating Patient Outcomes: A User's Guide. Rockville, MD: Agency for Healthcare Research and Quality; 2014.
14. Rumbold JM, Pierscionek B. The effect of the general data protection regulation on medical research. J Med Internet Res 2017 Feb 24;19(2):e47 [FREE Full text] [doi: 10.2196/jmir.7108] [Medline: 28235748]
15. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013 Jan 01;20(1):117-121 [FREE Full text] [doi: 10.1136/amiajnl-2012-001145] [Medline: 22955496]
16. Lee PC, Kao FY, Liang FW, Lee YC, Li ST, Lu TH. Existing data sources in clinical epidemiology: the Taiwan National Health Insurance laboratory databases. Clin Epidemiol 2021;13:175-181 [FREE Full text] [doi: 10.2147/CLEP.S286572] [Medline: 33688263]
17. Sung SF, Hsieh CY, Hu YH. Two decades of research using Taiwan's National Health Insurance Claims Data: bibliometric and text mining analysis on PubMed. J Med Internet Res 2020 Jun 16;22(6):e18457 [FREE Full text] [doi: 10.2196/18457] [Medline: 32543443]
18. Tham TY, Tran TL, Prueksaritanond S, Isidro JS, Setia S, Welluppillai V. Integrated health care systems in Asia: an urgent necessity. Clin Interv Aging 2018;13:2527-2538 [FREE Full text] [doi: 10.2147/CIA.S185048] [Medline: 30587945]
19. Lin LW, Ahn J, Bayani DBS, Chan K, Choipel D, Isaranuwatchai W. Use of real-world data and real-world evidence to support drug reimbursement decision-making in Asia: a non-binding guidance document prepared by the REAL World Data In ASia for HEalth Technology Assessment in Reimbursement (REALISE) working group. Health Intervention and Policy Evaluation Research (HIPER), National University of Singapore. 2020. URL: https://hiper.nus.edu.sg/realise-guidance/ [accessed 2024-02-10]
20. Raman SR, O'Brien EC, Hammill BG, Nelson AJ, Fish LJ, Curtis LH, et al. Evaluating fitness-for-use of electronic health records in pragmatic clinical trials: reported practices and recommendations. J Am Med Inform Assoc 2022 Apr 13;29(5):798-804 [FREE Full text] [doi: 10.1093/jamia/ocac004] [Medline: 35171985]

21. Market overview. Taiwan—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/taiwan-market-overview [accessed 2024-02-10]
22. Medical devices. Taiwan—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/taiwan-medical-devices [accessed 2024-07-20]
23. Hsu JC, Lu CY. The evolution of Taiwan's national health insurance drug reimbursement scheme. Daru 2015;23(1):15 [FREE Full text] [doi: 10.1186/s40199-014-0080-7] [Medline: 25889754]
24. Cheng TM. Reflections on the 20th anniversary of Taiwan's single-payer National Health Insurance System. Health Aff (Millwood) 2015;34(3):502-510. [doi: 10.1377/hlthaff.2014.1332] [Medline: 25732502]
25. Health and health systems ranking of countries worldwide in 2023. Health, Pharma & Medtech. 2023. URL: https://www.statista.com/statistics/1376359/health-and-health-system-ranking-of-countries-worldwide/ [accessed 2024-02-10]
26. Lin LY, Warren-Gash C, Smeeth L, Chen PC. Data resource profile: the national health insurance research database (NHIRD). Epidemiol Health 2018;40:e2018062 [FREE Full text] [doi: 10.4178/epih.e2018062] [Medline: 30727703]
27. Chiu WT, Pwu RF, Gau CS. Affordable health technology assessment in Taiwan: a model for middle-income countries. J Formos Med Assoc 2015;114(6):481-483 [FREE Full text] [doi: 10.1016/j.jfma.2015.01.016] [Medline: 26009485]
28. Center for Drug Evaluation, Taiwan. Health technology assessment center for drug evaluation, Taiwan. URL: https://www.cde.org.tw/eng/HTA/ [accessed 2024-02-26]
29. Chung CH. Trends in pharmaceutical expenditure in the Taiwan national health insurance database at different hospital levels. J Comp Eff Res 2023;12(2):e220162 [FREE Full text] [doi: 10.2217/cer-2022-0162] [Medline: 36511826]
30. Hsiao WW, Lin JC, Fan CT, Chen SS. Precision health in Taiwan: a data-driven diagnostic platform for the future of disease prevention. Comput Struct Biotechnol J 2022;20:1593-1602 [FREE Full text] [doi: 10.1016/j.csbj.2022.03.026] [Medline: 35495110]
31. Huang YT, Chen YJ, Chang SH, Kuo CF, Chen MH. Discharge status validation of the Chang Gung Research database in Taiwan. Biomed J 2022;45(6):907-913 [FREE Full text] [doi: 10.1016/j.bj.2021.12.006] [Medline: 34971827]
32. Shao SC, Chan YY, Kao Yang YH, Lin SJ, Hung MJ, Chien RN, et al. The Chang Gung Research Database—a multi-institutional electronic medical records database for real-world epidemiological studies in Taiwan. Pharmacoepidemiol Drug Saf 2019;28(5):593-600. [doi: 10.1002/pds.4713] [Medline: 30648314]
33. Tsai MS, Lin MH, Lee CP, Yang YH, Chen WC, Chang GH, et al. Chang Gung Research Database: a multi-institutional database consisting of original medical records. Biomed J 2017;40(5):263-269 [FREE Full text] [doi: 10.1016/j.bj.2017.08.002] [Medline: 29179881]
34. Chiang CJ, You SL, Chen CJ, Yang YW, Lo WC, Lai MS. Quality assessment and improvement of nationwide cancer registration system in Taiwan: a review. Jpn J Clin Oncol 2015;45(3):291-296. [doi: 10.1093/jjco/hyu211] [Medline: 25601947]
35. Chang GH, Tsai MS, Liu CY, Lin MH, Tsai YT, Hsu CM, et al. End-stage renal disease: a risk factor of deep neck infection—a nationwide follow-up study in Taiwan. BMC Infect Dis 2017;17(1):424 [FREE Full text] [doi: 10.1186/s12879-017-2531-5] [Medline: 28610562]
36. Hu SH, Huang MY, Chen CY, Hsieh HM. Treatment patterns of targeted and nontargeted therapies and survival effects in patients with locally advanced head and neck cancer in Taiwan. BMC Cancer 2023;23(1):567 [FREE Full text] [doi: 10.1186/s12885-023-11061-4] [Medline: 37340424]
37. Yeh HL, Hsieh FI, Lien LM, Kuo WH, Jeng JS, Sun Y, Taiwan Stroke Registry Investigators, et al. Patient and hospital characteristics associated with do-not-resuscitate/do-not-intubate orders: a cross-sectional study based on the Taiwan stroke registry. BMC Palliat Care 2023;22(1):138 [FREE Full text] [doi: 10.1186/s12904-023-01257-7] [Medline: 37715158]
38. Su PC, Zheng CM, Chen CC, Chiu LY, Chang HY, Tsai MH, et al. Effect of dialysis modalities on all-cause mortality and cardiovascular mortality in end-stage kidney disease: a Taiwan renal registry data system (TWRDS) 2005-2012 study. J Pers Med 2022;12(10):1715 [FREE Full text] [doi: 10.3390/jpm12101715] [Medline: 36294854]
39. Lu MY, Huang CF, Hung CH, Tai CM, Mo LR, Kuo HT, TACR Study Group. Artificial intelligence predicts direct-acting antivirals failure among hepatitis C virus patients: a nationwide hepatitis C virus registry program. Clin Mol Hepatol 2024 Jan;30(1):64-79 [FREE Full text] [doi: 10.3350/cmh.2023.0287] [Medline: 38195113]
40. Chi CY, Chen YP, Yang CW, Huang CH, Wang YC, Chong KM, et al. Characteristics, prognostic factors, and chronological trends of out-of-hospital cardiac arrests with shockable rhythms in Taiwan—a 7-year observational study. J Formos Med Assoc 2022;121(10):1972-1980 [FREE Full text] [doi: 10.1016/j.jfma.2022.01.024] [Medline: 35216883]
41. Lee CC, Hsu TC, Kuo CC, Liu MA, Abdelfattah AM, Chang CN, et al. Validation of a post-transplant lymphoproliferative disorder risk prediction score and derivation of a new prediction score using a national bone marrow transplant registry database. Oncologist 2021;26(11):e2034-e2041 [FREE Full text] [doi: 10.1002/onco.13969] [Medline: 34506688]
42. Wei CC, Shyu KG, Cheng JJ, Lo HM, Chiu CZ. Diabetes and adverse cardiovascular outcomes in patients with acute coronary syndrome—data from Taiwan's acute coronary syndrome full spectrum data registry. Acta Cardiol Sin 2016;32(1):31-38 [FREE Full text] [doi: 10.6515/acs20150322a] [Medline: 27171367]
43. Wang CC, Chang HY, Yin WH, Wu YW, Chu PH, Wu CC, et al. TSOC-HFrEF registry: a registry of hospitalized patients with decompensated systolic heart failure: description of population and management. Acta Cardiol Sin 2016;32(4):400-411 [FREE Full text] [doi: 10.6515/acs20160704a] [Medline: 27471353]

44. Keller JJ, Kang JH, Lin HC. Association between osteoporosis and psoriasis: results from the longitudinal health insurance database in Taiwan. Osteoporos Int 2013;24(6):1835-1841. [doi: 10.1007/s00198-012-2185-5] [Medline: 23052942]
45. Cheng SW, Wang CY, Chen JH, Ko Y. Healthcare costs and utilization of diabetes-related complications in Taiwan: a claims database analysis. Medicine (Baltimore) 2018;97(31):e11602 [FREE Full text] [doi: 10.1097/MD.0000000000011602] [Medline: 30075532]
46. Market overview. Singapore—Country Commercial Guide. International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/singapore-market-overview [accessed 2024-02-10]
47. Market opportunities. Singapore—Country Commercial Guide. International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/knowledge-product/singapore-market-opportunities [accessed 2024-02-10]
48. Tan CC, Lam CSP, Matchar DB, Zee YK, Wong JEL. Singapore's health-care system: key features, challenges, and shifts. Lancet 2021;398(10305):1091-1104. [doi: 10.1016/S0140-6736(21)00252-X] [Medline: 34481560]
49. Improving Health Outcomes Through Trusted Data Exchange. Precision Health Research Singapore. URL: https://trustplatform.sg/ [accessed 2024-02-26]
50. Better health, better future for all. Ministry of Health Singapore. URL: https://extranet.who.int/countryplanningcycles/sites/default/files/planning_cycle_repository/singapore/singapore_healthcare_masterplan_2020.pdf [accessed 2024-02-10]
51. Segar V, Ang PK, Foteff C, Ng K. A review of implementation frameworks to operationalize health technology assessment recommendations for medical technologies in the Singapore setting. Int J Technol Assess Health Care 2021;37(1):e56. [doi: 10.1017/S0266462321000222] [Medline: 33843519]
52. Sharma M, Teerawattananon Y, Dabak SV, Isaranuwatchai W, Pearce F, Pilasant S, et al. A landscape analysis of health technology assessment capacity in the association of South-East Asian Nations region. Health Res Policy Syst 2021;19(1):19 [FREE Full text] [doi: 10.1186/s12961-020-00647-0] [Medline: 33573676]
53. Heng BH, Sun Y, Cheah JT, Jong M. The Singapore national healthcare group diabetes registry—descriptive epidemiology of type 2 diabetes mellitus. Ann Acad Med Singap 2010;39(5):348-352 [FREE Full text] [Medline: 20535422]
54. How CH, Koh LH. PILL series. Not that way: advance care planning. Singapore Med J 2015;56(1):19-21; quiz 22 [FREE Full text] [doi: 10.11622/smedj.2015005] [Medline: 25640095]
55. Doctor NE, Ahmad NS, Pek PP, Yap S, Ong ME. The Pan-Asian Resuscitation Outcomes Study (PAROS) clinical research network: what, where, why and how. Singapore Med J 2017;58(7):456-458 [FREE Full text] [doi: 10.11622/smedj.2017057] [Medline: 28741005]
56. Ho AFW, Loy EY, Pek PP, Wah W, Tan TXZ, Liu N, et al. Emergency medical services utilization among patients with ST-segment elevation myocardial infarction: observations from the Singapore myocardial infarction registry. Prehosp Emerg Care 2016;20(4):454-461. [doi: 10.3109/10903127.2015.1128032] [Medline: 26986553]
57. National Immunisation Registry. Programme NI. 2016. URL: https://www.nir.hpb.gov.sg/nirp/eservices/aboutUs [accessed 2024-02-06]
58. Ho ZJM, Huang F, Wong CS, Chua L, Ma S, Chen MI, et al. Using a HIV registry to develop accurate estimates for the HIV care cascade—the Singapore experience. J Int AIDS Soc 2019;22(7):e25356 [FREE Full text] [doi: 10.1002/jia2.25356] [Medline: 31347260]
59. Gan SH, KhinMar KW, Ang LW, Lim LKY, Sng LH, Wang YT, et al. Recurrent tuberculosis disease in Singapore. Open Forum Infect Dis 2021;8(7):ofab340 [FREE Full text] [doi: 10.1093/ofid/ofab340] [Medline: 34307732]
60. Ho AFW, Lim MJR, Earnest A, Blewer A, Graves N, Yeo JW, Singapore PAROS Investigators. Long term survival and disease burden from out-of-hospital cardiac arrest in Singapore: a population-based cohort study. Lancet Reg Health West Pac 2023;32:100672 [FREE Full text] [doi: 10.1016/j.lanwpc.2022.100672] [Medline: 36785853]
61. Venketasubramanian N, Chang HM, Chan BPL, Young SH, Kong KH, Tang KF, Singapore Stroke Registry. Countrywide stroke incidence, subtypes, management and outcome in a multiethnic Asian population: the Singapore Stroke Registry—methodology. Int J Stroke 2015;10(5):767-769. [doi: 10.1111/ijs.12472] [Medline: 25753306]
62. Wui LW, Shaun GE, Ramalingam G, Wai KMS. Epidemiology of trauma in an acute care hospital in Singapore. J Emerg Trauma Shock 2014;7(3):174-179 [FREE Full text] [doi: 10.4103/0974-2700.136860] [Medline: 25114427]
63. Exchange of data pursuant to CCPS (Singapore). Documentation S. URL: https://tinyurl.com/3uncjww8 [accessed 2024-07-19]
64. Sharma MG, Popli H. Challenges for lower-middle-income countries in achieving universal healthcare: an Indian perspective. Cureus 2023;15(1):e33751 [FREE Full text] [doi: 10.7759/cureus.33751] [Medline: 36655151]
65. Market overview. India—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/knowledge-product/exporting-india-market-overview [accessed 2024-02-10]
66. Market opportunities. India—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/india-healthcare-and-life-science [accessed 2024-02-10]
67. Ghia C, Rambhad G. Implementation of equity and access in Indian healthcare: current scenario and way forward. J Mark Access Health Policy 2023;11(1):2194507 [FREE Full text] [doi: 10.1080/20016689.2023.2194507] [Medline: 36998432]
68. Lahariya C, Sahoo KC, Sundararaman T, Prinja S, Rajsekhar K, Pati S. Universal health coverage in India and health technology assessment: current status and the way forward. Front Public Health 2023;11:1187567 [FREE Full text] [doi: 10.3389/fpubh.2023.1187567] [Medline: 37333525]

69. Zodpey S, Farooqui HH. Universal health coverage in India: Progress achieved and the way forward. Indian J Med Res 2018;147(4):327-329 [FREE Full text] [doi: 10.4103/ijmr.IJMR_616_18] [Medline: 29998865]
70. Saxena SG, Godfrey T. India's opportunity to address human resource challenges in healthcare. Cureus 2023;15(6):e40274 [FREE Full text] [doi: 10.7759/cureus.40274] [Medline: 37448434]
71. Breakdown of hospital services in India in financial year 2020, by public and private sector. Statista. 2023. URL: https://www.statista.com/statistics/1252917/india-breakdown-of-hospital-services-by-public-and-private/ [accessed 2024-02-27]
72. Kumar A. The transformation of the Indian healthcare system. Cureus 2023;15(5):e39079 [FREE Full text] [doi: 10.7759/cureus.39079] [Medline: 37378105]
73. Cabinet approves ayushman bharat-national health protection mission. Press Information Bureau, Government of India. URL: https://pib.gov.in/newsite/PrintRelease.aspx?relid=176049#:~:text=These%20centres%20will%20provide%20comprehensive,crore%20for%20this%20flagship%20programme [accessed 2024-02-10]
74. MacQuilkan K, Baker P, Downey L, Ruiz F, Chalkidou K, Prinja S, et al. Strengthening health technology assessment systems in the global south: a comparative analysis of the HTA journeys of China, India and South Africa. Glob Health Action 2018;11(1):1527556 [FREE Full text] [doi: 10.1080/16549716.2018.1527556] [Medline: 30326795]
75. Gheorghe A, Mehndiratta A, Baker P, Culyer A, Prinja S, Kar SS, et al. Health technology assessment in India in the next decade: reflections on a vision for its path to maturity and impact. BMJ Evid Based Med 2024:112491 [FREE Full text] [doi: 10.1136/bmjebm-2023-112491] [Medline: 38290800]
76. The HealthPlix story. Technologies H. 2023. URL: https://healthplix.com/about/ [accessed 2024-02-06]
77. eyeSmart. LV Prasad Eye Institute. URL: https://eyesmartemr.com/about/ [accessed 2024-02-10]
78. Sharma K, Dani S, Desai D, Kumar P, Bhalani N, Vasavada A, et al. Two-year safety and efficacy of Indigenous Abluminus Sirolimus Eluting Stent. Does it differ amongst diabetics?—Data from en-ABLe—REGISTRY. J Cardiovasc Thorac Res 2021;13(2):162-168 [FREE Full text] [doi: 10.34172/jcvtr.2021.31] [Medline: 34326971]
79. Dani S, Shah D, Sojitra P, Parikh K, Shetty R, di Palma G, et al. A novel nanocarrier sirolimus-coated balloon for coronary interventions: 12-month data from the Nanoluté Registry. Cardiovasc Revasc Med 2019;20(3):235-240. [doi: 10.1016/j.carrev.2018.06.003] [Medline: 30196029]
80. Jabir A, Mathew A, Zheng Y, Westerhout C, Viswanathan S, Sebastian P, et al. Procedural volume and outcomes after primary percutaneous coronary intervention for ST-segment-elevation myocardial infarction in Kerala, India: report of the cardiological society of India-Kerala primary percutaneous coronary intervention registry. J Am Heart Assoc 2020;9(12):e014968 [FREE Full text] [doi: 10.1161/JAHA.119.014968] [Medline: 32476563]
81. Nair R, Bhurani D, Rajappa S, Kapadia A, Reddy Boya R, Sundaram S, et al. Diffuse large b-cell lymphoma: clinical presentation and treatment outcomes from the lymphoma registry. Front Oncol 2021;11:796962 [FREE Full text] [doi: 10.3389/fonc.2021.796962] [Medline: 35186714]
82. Krishnan P, Kurumboor P, Varma D, Mallick S, Vayoth SO, Gopalakrishnan U, et al. Voluntary perioperative colorectal cancer registry from Kerala-An initial overview. Indian J Gastroenterol 2020;39(3):243-252. [doi: 10.1007/s12664-019-00998-9] [Medline: 32936377]
83. Roy N, Gerdin M, Ghosh S, Gupta A, Kumar V, Khajanchi M, et al. 30-Day in-hospital trauma mortality in four urban university hospitals using an Indian trauma registry. World J Surg 2016;40(6):1299-1307. [doi: 10.1007/s00268-016-3452-y] [Medline: 26911610]
84. Sethi SK, Wazir S, Sahoo J, Agrawal G, Bajaj N, Gupta NP, et al. Risk factors and outcomes of neonates with acute kidney injury needing peritoneal dialysis: results from the prospective TINKER (The Indian PCRRT-ICONIC Neonatal Kidney Educational Registry) study. Perit Dial Int 2022;42(5):460-469. [doi: 10.1177/08968608221091023] [Medline: 35574693]
85. Reddy S, Mary I. Rajiv aarogyasri community health insurance scheme in Andhra Pradesh, India: a comprehensive analytic view of private public partnership model. Indian J Public Health 2013;57(4):254-259 [FREE Full text] [doi: 10.4103/0019-557X.123264] [Medline: 24351388]
86. Market overview. Hong Kong—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/hong-kong-healthcare [accessed 2024-02-10]
87. Fung VL, Lai AH, Yam CH, Wong EL, Griffiths SM, Yeoh E. Healthcare vouchers for better elderly services? Input from private healthcare service providers in Hong Kong. Health Soc Care Community 2022;30(2):e357-e369. [doi: 10.1111/hsc.13203] [Medline: 33128419]
88. Wong CKH, Wu O, Cheung BMY. Towards a transparent, credible, evidence-based decision-making process of new drug listing on the Hong Kong hospital authority drug formulary: challenges and suggestions. Appl Health Econ Health Policy 2018;16(1):5-14 [FREE Full text] [doi: 10.1007/s40258-017-0339-5] [Medline: 28702874]
89. O'Rourke B, Oortwijn W, Schuller T, International Joint Task Group. The new definition of health technology assessment: a milestone in international collaboration. Int J Technol Assess Health Care 2020;36(3):187-190. [doi: 10.1017/S0266462320000215] [Medline: 32398176]
90. Sing CW, Woo YC, Lee AC, Lam JK, Chu JK, Wong IC, et al. Validity of major osteoporotic fracture diagnosis codes in the clinical data analysis and reporting system in Hong Kong. Pharmacoepidemiol Drug Saf 2017;26(8):973-976. [doi: 10.1002/pds.4208] [Medline: 28371079]

91. Chan JCN, Lim LL, Luk AOY, Ozaki R, Kong APS, Ma RCW, et al. From Hong Kong diabetes register to JADE program to RAMP-DM for data-driven actions. Diabetes Care 2019;42(11):2022-2031. [doi: 10.2337/dci19-0003] [Medline: 31530658]
92. Hospital Authority. Hong Kong Cancer Registry. 2023. URL: https://www3.ha.org.hk/cancereg/ [accessed 2024-02-06]
93. Mok CC, Chan KY, Lee KL, Tam LS, Lee KW, Hong Kong Society of Rheumatology. Factors associated with withdrawal of the anti-TNFα biologics in the treatment of rheumatic diseases: data from the Hong Kong Biologics Registry. Int J Rheum Dis 2014;17 Suppl 3:1-8. [doi: 10.1111/1756-185X.12264] [Medline: 24382315]
94. Market overview. Malaysia—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/malaysia-market-overview [accessed 2024-02-10]
95. Market opportunities. Malaysia—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/malaysia-market-opportunities [accessed 2024-07-20]
96. Roza S, Junainah S, Izzuna MMG, Nurhasni KARK, Yusof MAM, Noormah MD, et al. Health technology assessment in Malaysia: past, present, and future. Int J Technol Assess Health Care 2019;35(6):446-451. [doi: 10.1017/S0266462319000023] [Medline: 30864531]
97. Awang S, Agins B, Ujang IRM, Narayanan DN, Zulkifli NW, Hamidi N. Development of the national policy for quality in healthcare for Malaysia. Health Res Policy Syst 2023;21(1):119 [FREE Full text] [doi: 10.1186/s12961-023-01063-w] [Medline: 37964336]
98. National Cardiovascular Disease Database. What is the national cardiovascular disease database?. URL: https://www.acrm.org.my/ncvd/faq.htm#:~:text=What%20is%20NCVD%3F,and%20treatment%20in%20the%20country [accessed 2024-02-06]
99. National diabetes registry report 2020. Ministry of Health Malaysia. URL: https://www.researchgate.net/publication/354238285_National_Diabetes_Registry_Report_2020 [accessed 2024-07-20]
100. Medical devices. Thailand—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/thailand-medical-devices-and-technology [accessed 2024-02-10]
101. Market opportunities. Thailand—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/knowledge-product/thailand-market-opportunities [accessed 2024-02-10]
102. Proportion of public health expenditure as a share of total health expenditure in Thailand from 2011 to 2020. Statista. 2023. URL: https://www.statista.com/statistics/1424663/thailand-share-of-public-health-expenditure/ [accessed 2024-02-27]
103. Tangcharoensathien V, Witthayapipopsakul W, Panichkriangkrai W, Patcharanarumol W, Mills A. Health systems development in Thailand: a solid platform for successful implementation of universal health coverage. Lancet 2018;391(10126):1205-1223 [FREE Full text] [doi: 10.1016/S0140-6736(18)30198-3] [Medline: 29397200]
104. Butani D, Faradiba D, Dabak SV, Isaranuwatchai W, Huang-Ku E, Pachanee K, et al. Expanding access to high-cost medicines under the universal health coverage scheme in Thailand: review of current practices and recommendations. J Pharm Policy Pract 2023;16(1):138 [FREE Full text] [doi: 10.1186/s40545-023-00643-z] [Medline: 37936171]
105. Bucci T, Shantsila A, Romiti GF, Teo WS, Chao TF, Shimizu W, et al. External validation of COOL-AF scores in the Asian Pacific Heart Rhythm Society Atrial Fibrillation Registry. JACC Asia 2024;4(1):59-69 [FREE Full text] [doi: 10.1016/j.jacasi.2023.09.011] [Medline: 38222252]
106. Dejkhamron P, Santiprabhob J, Likitmaskul S, Deerochanawong C, Rawdaree P, Tharavanij T, Thai Type 1 Diabetes Diabetes Diagnosed Before Age 30 Years Registry, Care, Network (T1DDAR CN). Type 1 diabetes management and outcomes: a multicenter study in Thailand. J Diabetes Investig 2021;12(4):516-526 [FREE Full text] [doi: 10.1111/jdi.13390] [Medline: 32815278]
107. Sansanayudh N, Chandavimol M, Srimahachota S, Limpijankit T, Hutayanon P, Kiatchoosakun S, et al. Patient characteristics, procedural details, and outcomes of contemporary percutaneous coronary intervention in real-world practice: insights from Nationwide Thai PCI Registry. J Interv Cardiol 2022;2022:5839834 [FREE Full text] [doi: 10.1155/2022/5839834] [Medline: 35935123]
108. Saenrueang T, Promthet S, Kamsa-Ard S, Pengsaa P. Cervical cancer in khon kaen, Thailand: analysis of 1990-2014 incidence data and prediction of future trends. Asian Pac J Cancer Prev 2019;20(2):369-375 [FREE Full text] [doi: 10.31557/APJCP.2019.20.2.369] [Medline: 30803194]
109. Rattanathammethee T, Norasetthada L, Bunworasate U, Wudhikarn K, Julamanee J, Noiperm P, et al. Outcomes of polatuzumab vedotin-containing regimens in real-world setting of relapsed and or refractory diffuse large B-cell lymphoma patients: a matched-control analysis from the thai lymphoma study group (TLSG). Ann Hematol 2023;102(7):1887-1895. [doi: 10.1007/s00277-023-05273-8] [Medline: 37202499]
110. Narongroeknawin P, Chevaisrakul P, Kasitanon N, Kitumnuaypong T, Mahakkanukrauh A, Siripaitoon B, Thai Rheumatism Association. Drug survival and reasons for discontinuation of the first biological disease modifying antirheumatic drugs in Thai patients with rheumatoid arthritis: analysis from the Thai Rheumatic Disease Prior Authorization registry. Int J Rheum Dis 2018;21(1):170-178. [doi: 10.1111/1756-185X.12937] [Medline: 28737837]
111. Barua P, Narattharaksa K. Association between health insurance and incidence of death in stateless children in Tak province, Thailand. J Health Manage 2020;22(3):348-362. [doi: 10.1177/0972063420937930]

112. Marshall AI, Witthayapipopsakul W, Chotchoungchatchai S, Wangbanjongkun W, Tangcharoensathien V. Contracting the private health sector in Thailand's universal health coverage. PLOS Glob Public Health 2023;3(4):e0000799 [FREE Full text] [doi: 10.1371/journal.pgph.0000799] [Medline: 37115744]
113. Tangcharoensathien V, Patcharanarumol W, Greetong T, Suwanwela W, Kesthom N, Viriyathorn S, et al. Defining the benefit package of Thailand universal coverage scheme: from pragmatism to sophistication. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7182149/ [accessed 2024-07-20]
114. Asante A, Cheng Q, Susilo D, Satrya A, Haemmerli M, Fattah RA, et al. The benefits and burden of health financing in Indonesia: analyses of nationally representative cross-sectional data. Lancet Glob Health 2023;11(5):e770-e780 [FREE Full text] [doi: 10.1016/S2214-109X(23)00064-5] [Medline: 37061314]
115. Wasir R, Irawati S, Makady A, Postma M, Goettsch W, Buskens E, et al. Use of medicine pricing and reimbursement policies for universal health coverage in Indonesia. PLoS One 2019;14(2):e0212328 [FREE Full text] [doi: 10.1371/journal.pone.0212328] [Medline: 30779809]
116. Chavarina KK, Faradiba D, Sari EN, Wang Y, Teerawattananon Y. Health economic evaluations for Indonesia: a systematic review assessing evidence quality and adherence to the Indonesian health technology assessment (HTA) guideline. Lancet Reg Health Southeast Asia 2023;13:100184 [FREE Full text] [doi: 10.1016/j.lansea.2023.100184] [Medline: 37383554]
117. Alkatiri AA, Firman D, Haryono N, Yonas E, Pranata R, Fahri I, et al. Comparison between radial versus femoral percutaneous coronary intervention access in Indonesian hospitals, 2017-2018: a prospective observational study of a national registry. Int J Cardiol Heart Vasc 2020;27:100488. [doi: 10.1016/j.ijcha.2020.100488] [Medline: 32154360]
118. Harbuwono DS, Handayani DOTL, Wahyuningsih ES, Supraptowati N, Ananda, Kurniawan F, Rebekka, et al. Impact of diabetes mellitus on COVID-19 clinical symptoms and mortality: Jakarta's COVID-19 epidemiological registry. Prim Care Diabetes 2022;16(1):65-68 [FREE Full text] [doi: 10.1016/j.pcd.2021.11.002] [Medline: 34857490]
119. Hamijoyo L, Candrianita S, Rahmadi AR, Dewi S, Darmawan G, Suryajaya BS, et al. The clinical characteristics of systemic lupus erythematosus patients in Indonesia: a cohort registry from an Indonesia-based tertiary referral hospital. Lupus 2019;28(13):1604-1609. [doi: 10.1177/0961203319878499] [Medline: 31566078]
120. Market opportunities. Pakistan—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/pakistan-market-opportunities [accessed 2024-02-10]
121. Muhammad Q, Eiman H, Fazal F, Ibrahim M, Gondal MF. Healthcare in Pakistan: navigating challenges and building a brighter future. Cureus 2023;15(6):e40218 [FREE Full text] [doi: 10.7759/cureus.40218] [Medline: 37435262]
122. Sivalal S. Health technology assessment in the Asia pacific region. Int J Technol Assess Health Care 2009;25 Suppl 1:196-201. [doi: 10.1017/S0266462309090631] [Medline: 19534841]
123. Peerwani G, Khan SM, Khan MD, Bashir F, Sheikh S, Ramsey DJ, et al. Gender differences in clinical outcomes after percutaneous coronary intervention-analysis of 15,106 patients from the cardiac registry of Pakistan database. Am J Cardiol 2023;188:61-67. [doi: 10.1016/j.amjcard.2022.11.020] [Medline: 36473306]
124. Punjab Cancer Registry. 2022. URL: http://punjabcancerregistry.org.pk/ [accessed 2024-02-06]
125. Hoodbhoy Z, Ehsan L, Alvi N, Sajjad F, Asghar A, Nadeem O, et al. Establishment of a thalassaemia major quality improvement collaborative in Pakistan. Arch Dis Child 2020;105(5):487-493. [doi: 10.1136/archdischild-2018-315743] [Medline: 30737261]
126. Vietnam—Country Commercial Guide. The International Trade Administration, US Department of Commerce. Market opportunities. URL: https://www.trade.gov/country-commercial-guides/vietnam-market-opportunities [accessed 2024-02-10]
127. Quan NK, Taylor-Robinson AW. Vietnam's evolving healthcare system: notable successes and significant challenges. Cureus 2023;15(6):e40414 [FREE Full text] [doi: 10.7759/cureus.40414] [Medline: 37456482]
128. Lee H, Nguyen TT, Park S, Hoang VM, Kim W. Health technology assessment development in Vietnam: a qualitative study of current progress, barriers, facilitators, and future strategies. Int J Environ Res Public Health 2021;18(16):8846 [FREE Full text] [doi: 10.3390/ijerph18168846] [Medline: 34444597]
129. Pham DX, Ho TH, Bui TD, Ho-Pham LT, Nguyen TV. Trends in breast cancer incidence in Ho Chi Minh City 1996-2015: a registry-based study. PLoS One 2021;16(2):e0246800 [FREE Full text] [doi: 10.1371/journal.pone.0246800] [Medline: 33566857]
130. Ng JYS, Ramadani RV, Hendrawan D, Duc DT, Kiet PHT. National health insurance databases in Indonesia, Vietnam and the Philippines. Pharmacoecon Open 2019;3(4):517-526 [FREE Full text] [doi: 10.1007/s41669-019-0127-2] [Medline: 30859490]
131. Amit AML, Pepito VCF, Dayrit MM. Advancing universal health coverage in the Philippines through self-care interventions. Lancet Reg Health West Pac 2022;26:100579 [FREE Full text] [doi: 10.1016/j.lanwpc.2022.100579] [Medline: 36105555]
132. Market opportunities. Philipines—Country Commercial Guide. The International Trade Administration, US Department of Commerce. URL: https://www.trade.gov/country-commercial-guides/philippines-market-opportunities [accessed 2024-02-10]
133. Racoma MJC, Calibag MKKB, Cordero CP, Abacan MAR, Chiong MAD. A review of the clinical outcomes in idursulfase-treated and untreated Filipino patients with mucopolysaccharidosis type II: data from the local lysosomal storage disease registry. Orphanet J Rare Dis 2021;16(1):323 [FREE Full text] [doi: 10.1186/s13023-021-01875-5] [Medline: 34289859]

134. Renal disease control program (ReDCoP). Institute NKaT. URL: https://elibrary.judiciary.gov.ph/thebookshelf/showdocs/10/48965 [accessed 2024-02-06]
135. Varhol RJ, Norman R, Randall S, Lee CMY, Trevenen L, Boyd JH, et al. Public preference on sharing health data to inform research, health policy and clinical practice in Australia: a stated preference experiment. PLoS One 2023;18(11):e0290528 [FREE Full text] [doi: 10.1371/journal.pone.0290528] [Medline: 37972118]
136. Holmgren AJ, Esdar M, Hüsers J, Coutinho-Almeida J. Health information exchange: understanding the policy landscape and future of data interoperability. Yearb Med Inform 2023;32(1):184-194 [FREE Full text] [doi: 10.1055/s-0043-1768719] [Medline: 37414031]
137. Chao K, Sarker MNI, Ali I, Firdaus RBR, Azman A, Shaed MM. Big data-driven public health policy making: potential for the healthcare industry. Heliyon 2023;9(9):e19681 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e19681] [Medline: 37809720]
138. Teerawattananon Y, Luz K, Yothasmutra C, Pwu R, Ahn J, Shafie AA, et al. Historical development of the HTAsiaLINK network and its key determinants of sucess. Int J Technol Assess Health Care 2018;34(3):260-266 [FREE Full text] [doi: 10.1017/S0266462318000223] [Medline: 29911515]
139. INAHTA members. INAHTA. 2020. URL: https://www.inahta.org/members/ [accessed 2024-02-27]
140. Falkowski A, Ciminata G, Manca F, Bouttell J, Jaiswal N, Farhana Binti Kamaruzaman H, et al. How least developed to lower-middle income countries use health technology assessment: a scoping review. Pathog Glob Health 2023;117(2):104-119 [FREE Full text] [doi: 10.1080/20477724.2022.2106108] [Medline: 35950264]
141. Teerawattananon Y, Rattanavipapong W, Lin LW, Dabak SV, Gibbons B, Isaranuwatchai W, et al. Landscape analysis of health technology assessment (HTA): systems and practices in Asia. Int J Technol Assess Health Care 2019;35(6):416-421 [FREE Full text] [doi: 10.1017/S0266462319000667] [Medline: 31594553]
142. Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. J Med Syst 2018;42(11):214 [FREE Full text] [doi: 10.1007/s10916-018-1075-6] [Medline: 30269237]
143. Teerawattananon Y, Teo YY, Dabak S, Rattanavipapong W, Isaranuwatchai W, Wee H, et al. Tackling the 3 big challenges confronting health technology assessment development in Asia: a commentary. Value Health Reg Issues 2020;21:66-68 [FREE Full text] [doi: 10.1016/j.vhri.2019.07.001] [Medline: 31655465]
144. Liu G, Wu EQ, Ahn J, Kamae I, Xie J, Yang H. The development of health technology assessment in Asia: current status and future trends. Value Health Reg Issues 2020;21:39-44 [FREE Full text] [doi: 10.1016/j.vhri.2019.08.472] [Medline: 31634795]
145. Databases which have been converted to the OMOP CDM. Observational Health Data Sciences and Informatics. 2021. URL: https://www.ohdsi.org/web/wiki/doku.php?id=resources:2020_data_network [accessed 2024-02-26]
146. The OHDSI research network. Observational Health Data Sciences and Informatics. 2017. URL: https://www.ohdsi.org/web/wiki/doku.php?id=welcome [accessed 2024-02-26]

Abbreviations

ACP: advance care planning
ASEAN: Association of Southeast Asian Nations
CCCS: cross-country collaboration studies
CDARS: clinical data analysis and reporting system
CER: comparative effectiveness research
CGRD: Chang Gung Research Database
COOL-AF: Cohort of Antithrombotic Use and Optimal International Normalized Ratio Levels in Patients with Atrial Fibrillation
EHR: electronic health record
EMR: electronic medical record
GDP: gross domestic product
HA: Hospital Authority
HTA: health technology assessment
MOH: Ministry of Health
NHI: National Health Insurance
NHIRD: National Health Insurance Research Database
OHDSI: Observational Health Data Sciences and Informatics
RWD: real-world data
RWE: real-world evidence
SCS: single-country studies
TCR: Taiwan Cancer Registry
UHC: universal health coverage

Edited by G Eysenbach, A Benis; submitted 18.03.24; peer-reviewed by CY Chen; comments to author 30.05.24; revised version received 17.06.24; accepted 19.07.24; published 15.08.24.

Please cite as:
Julian GS, Shau WY, Chou HW, Setia S
Bridging Real-World Data Gaps: Connecting Dots Across 10 Asian Countries
JMIR Med Inform 2024;12:e58548
URL: https://medinform.jmir.org/2024/1/e58548
doi:10.2196/58548
PMID:39026427

Original Paper

Applying the Non-Adoption, Abandonment, Scale-up, Spread, and Sustainability Framework Across Implementation Stages to Identify Key Strategies to Facilitate Clinical Decision Support System Integration Within a Large Metropolitan Health Service: Interview and Focus Group Study

Manasha Fernando[1], MPH; Bridget Abell[1], PhD; Steven M McPhail[1,2], PhD; Zephanie Tyack[1], PhD; Amina Tariq[1], PhD; Sundresan Naicker[1], PhD

[1]Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health and Social Work, Faculty of Health, Queensland University of Technology, Brisbane, Australia

[2]Digital Health and Informatics Directorate, Metro South Health, Brisbane, Australia

Corresponding Author:
Sundresan Naicker, PhD
Australian Centre for Health Services Innovation and Centre for Healthcare Transformation
School of Public Health and Social Work, Faculty of Health
Queensland University of Technology
Q Block, 60 Musk Avenue
Brisbane, 4059
Australia
Phone: 61 3138 6454
Fax: 61 0449 876 034
Email: sundresan.naicker@qut.edu.au

Abstract

Background: Computerized clinical decision support systems (CDSSs) enhance patient care through real-time, evidence-based guidance for health care professionals. Despite this, the effective implementation of these systems for health services presents multifaceted challenges, leading to inappropriate use and abandonment over the course of time. Using the Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework, this qualitative study examined CDSS adoption in a metropolitan health service, identifying determinants across implementation stages to optimize CDSS integration into health care practice.

Objective: This study aims to identify the theory-informed (NASSS) determinants, which included multiple CDSS interventions across a 2-year period, both at the health-service level and at the individual hospital setting, that either facilitate or hinder the application of CDSSs within a metropolitan health service. In addition, this study aimed to map these determinants onto specific stages of the implementation process, thereby developing a system-level understanding of CDSS application across implementation stages.

Methods: Participants involved in various stages of the implementation process were recruited (N=30). Participants took part in interviews and focus groups. We used a hybrid inductive-deductive qualitative content analysis and a framework mapping approach to categorize findings into barriers, enablers, or neutral determinants aligned to NASSS framework domains. These determinants were also mapped to implementation stages using the *Active Implementation Framework stages* approach.

Results: Participants comprised clinical adopters (14/30, 47%), organizational champions (5/30, 16%), and those with roles in organizational clinical informatics (5/30, 16%). Most determinants were mapped to the organization level, technology, and adopter subdomains. However, the study findings also demonstrated a relative lack of long-term implementation planning. Consequently, determinants were not uniformly distributed across the stages of implementation, with 61.1% (77/126) identified in the exploration stage, 30.9% (39/126) in the full implementation stage, and 4.7% (6/126) in the installation stages. Stakeholders engaged in more preimplementation and full-scale implementation activities, with fewer cycles of monitoring and iteration activities identified.

Conclusions: These findings addressed a substantial knowledge gap in the literature using systems thinking principles to identify the interdependent dynamics of CDSS implementation. A lack of sustained implementation strategies (ie, training and longer-term, adopter-level championing) weakened the sociotechnical network between developers and adopters, leading to communication barriers. More rigorous implementation planning, encompassing all 4 implementation stages, may, in a way, help in addressing the barriers identified and enhancing enablers.

(JMIR Med Inform 2024;12:e60402) doi:10.2196/60402

KEYWORDS

medical informatics; adoption and implementation; behavior; health systems

Introduction

Background

The integration of digital technologies in health care services, especially the implementation of computerized clinical decision support systems (CDSSs), promises to enhance the quality, safety, and efficiency of patient care [1]. Evidence continues to build in favor of implementing CDSS for the optimization of clinical management decisions [2], thereby enabling more effective risk-based decision-making and the delivery of personalized care within the realm of acute health care [3]. Noteworthy applications include the adoption of computerized provider order entry systems [4], the deployment of point-of-care alerts to enhance patient safety [5], and the integration of electronic health record data and artificial intelligence (AI) for decision support [6,7].

Indeed, CDSSs are increasingly incorporating AI and machine learning to realize several critical benefits to clinical decision-making [8]. AI and machine learning techniques facilitate rapid analysis of extensive clinical data, including patient records and medical literature, surpassing traditional rule-based systems in swiftly generating insights and recommendations [1,8]. These technologies excel in identifying intricate patterns and relationships within data, enhancing diagnostic accuracy and treatment recommendations [1]. The relevance of AI and machine learning to CDSS adopters, such as health care providers, lies in the ability to tailor recommendations and predictions to specific scenarios [9]. For instance, machine learning can enhance diagnostic accuracy for common medical conditions by learning from large datasets, which directly benefits the adopters by providing more precise and actionable insights [10]. Although AI-enhanced CDSSs promise more precise, timely, and personalized clinical support, they are designed to complement rather than replace human judgment, necessitating careful consideration of implementation risks and limitations to optimize patient care and outcomes [11].

Despite these considerable potential benefits, the effective implementation of these systems within health services presents multifaceted challenges [1,12]. These include managing diverse stakeholder expectations, emergent clinical and sociopolitical contexts, and changing strategic priorities [13,14]. Failing to address these challenges may give rise to unanticipated outcomes related to low adoption rates [15,16], inappropriate use [17,18], unforeseen consequences [19,20], and long-term technology abandonment [21].

CDSSs can be valuable tools in health care, offering guidance to professionals; however, these systems can face limitations in practice. For example, CDSSs may struggle to accommodate patients with complex comorbidities, potentially leading to treatment recommendations that inadvertently worsen certain conditions such as prescribing heart disease medication that could harm kidneys [1]. Furthermore, CDSSs often do not incorporate patient preferences, cultural beliefs, or financial constraints as these data sources are not considered, necessitating personalized adjustments by clinicians to ensure treatment adherence and efficacy [1]. In acute care settings, CDSS may not keep pace with rapidly evolving clinical conditions, requiring clinicians to rely on their real-time assessments rather than potentially outdated CDSS guidance [1,9]. Moreover, there are high costs associated with the implementation, adoption, and maintenance of CDSSs [22,23]. These limitations of CDSSs underscore the critical role of effective implementation so that the right CDSS can provide the right information at the right time to the right patient [1].

Examining the factors that drive the successful adoption of CDSSs and those that impede its progress contributes to a deeper understanding of the dynamic interplay between technology and health care delivery [24-26]. This can be approached systematically [1,12-21] with guidance from well-established theories within the discipline of implementation science, which also accounts for organizational complexity [27,28]. A recent scoping review found that models, rather than theories or frameworks (18/42, 43% of the included studies), were most frequently used to guide CDSS adoption and evaluation strategies [29]. Unlike frameworks, models can be limited in examining the complexity of sustained implementation, acceptability, and adoption of technology across organizational and system levels [29,30]. The Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework uses a complex systems approach, which encapsulates the determinants NASSS of technological adoption overtime in health care settings [24]. This framework serves as a conceptual lens through which technological interventions are viewed as part of a complex system consisting of many processes and components. The utility of the NASSS framework lies in its ability to identify contextually appropriate determinants and inform implementation strategies, thereby shedding light on the factors that impact the success of digital health implementations [31-33]. Contextually informed implementation strategies tailored to influence clinician behavior could have a greater influence on CDSS adoption than technological design and content features [15,24-26,34].

Objectives

As such, this research was characterized by 2 principal aims. First, we aimed to identify the NASSS theory-informed determinants that either facilitate or hinder the application of CDSSs within a metropolitan health service. These include multiple CDSS interventions at the health-service level and at a single hospital setting spanning for 2 years. Second, this study aimed to map these determinants into specific stages of the implementation process, thereby developing a systems-level understanding of CDSS application across implementation stages. A stage-specific mapping approach allows for more nuanced and tailored strategies for CDSS integration, ensuring that the unique challenges and opportunities associated with each implementation stage are addressed [35-37].

Methods

Ethical Considerations

The Metro South Health Human Research Ethics Committee granted ethical clearance for this research (HREC/2020/QMS/64807). All participants provided written and verbal informed consent before participating in the study. Participation was voluntary and participants could withdraw at any time. All data were deidentified and handled in accordance with the Metro South Human Research Ethics Committee guidelines.

Study Design and Theoretical Framework

This qualitative study used a hybrid inductive-deductive approach [38]. The deductive approach was informed by the NASSS framework to identify contextually specific determinants associated with the use of CDSS technology in a large metropolitan health service across discrete implementation cycles during a 2-year period. The NASSS framework positions technological interventions as part of a complex system and has been used to guide implementation efforts and identify factors that influence technology implementation success in health services [24,31-33].

Study Setting

This study was conducted in a metropolitan health service comprising 5 hospitals, which serve a large catchment area (3856 km^2) in Australia. Consequently, those employed by the health service within the digital health and informatics portfolio were recruited to take part in the study. It must be noted that this department operates at both a health-service level and a facility level. In addition, clinicians who worked at either of the 2 largest hospitals within the health service were also recruited. With a facility of 1033 beds, hospital 1 is the largest teaching and training university hospital and is equipped with all major medical specialties except maternity services and pediatrics. Hospital 2 comprises most medical specialties, with 459 beds, including maternity services and pediatrics. Both hospitals (and the health service at large) used the same integrated electronic medical record system that had been implemented before the commencement of the study.

Participant Recruitment and Sample

The participants sampled in this study included staff from the clinical informatics unit, which operated at the health-service level. In addition, clinical staff who worked within the 2 largest (defined according to the number of available beds) hospitals within this health service were also recruited to participate in the study. In this study, CDSSs were defined as any electronic system or interface designed to provide health system users with tailored information to inform decision-making within a particular context or situation [1,31]. The participants who met the following selection criteria were recruited for this study: experience with decision-making, governance, purchasing, design, and implementation of CDSS initiatives within the health service. This could include those with roles in informatics, governance, and management, as well as frontline clinical staff. The researchers used purposive sampling throughout the study, seeking representativeness of participants who were involved in using or implementing CDSS. The researchers also sought representativeness across a range of implementer roles, that is involved in making decisions or engaging in CDSS procurement, rollout, and upgrades, and adopter roles, that is clinician users of the CDSS [39,40]. To obtain this representativeness, the researchers estimated a sample size of 30 to 40 participants.

Acting as a knowledge broker, SM, an academician who is also embedded in the health service, used a knowledge brokering process [41] to identify potential participants during informal discussions with health-service staff. Those identified were then formally invited to the study using internal memos and emails sent from SM, with participants given a week to respond. Nonresponders were followed up one more time within a fortnight of the initial email. Saturation was deemed to be achieved when no new concepts or understanding were identified after 3 consecutive interviews following purposive sampling [39].

Study Materials and Data Collection

The reporting of findings was guided by the Standards for Reporting Qualitative Research checklist (Multimedia Appendix 1). A semistructured, NASSS framework–informed interview guide, which focused on the availability, development, and perceptions of CDSS within the participants' health system (Multimedia Appendix 2), was developed. Questions also explored decisions around the implementation of CDSS as well as their adoption, use, and sustainment in the course of time.

Data collection was conducted between March 2021 and March 2023. On the basis of the participants' preferences and availability, data were collected through one-on-one interviews or through focus group discussions among groups consisting of 6 participants. This was done through in-person and web-based (Teams; Microsoft Corporation) techniques. Interviews and focus group discussions were led by SM, an experienced digital health and health services researcher (male, PhD qualified, embedded in health service, and familiar to a few participants), and SN, an experienced mixed methods health services researcher (male and PhD qualified). The interview duration was approximately 1 hour, while focus group discussions were conducted for up to 2 hours. The interviews and focus group discussions were audio recorded and transcribed

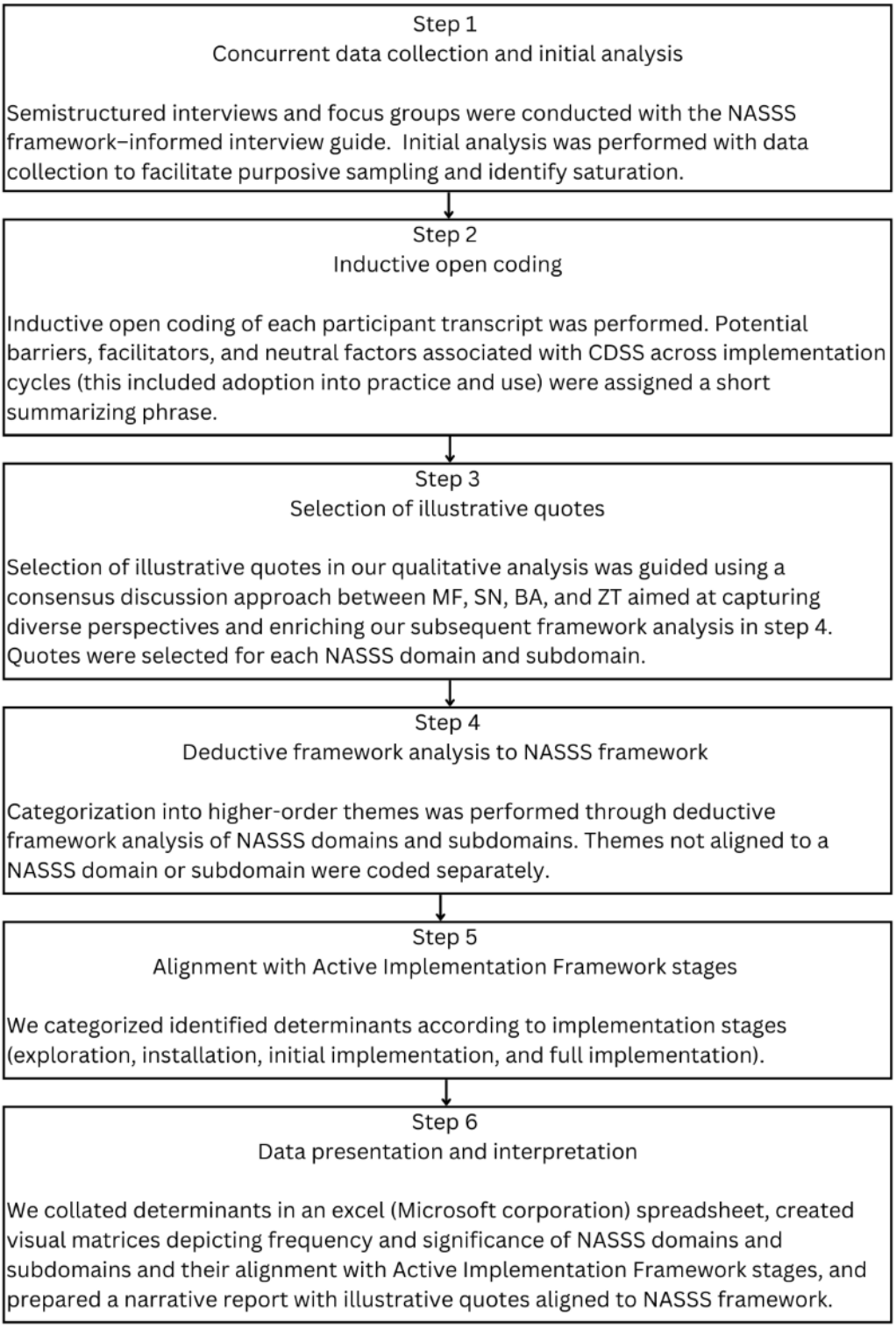
verbatim. Preliminary findings were monitored and discussed with the research team. This information was used to guide recruitment until saturation was reached [33,42].

Obtaining Saturation

Initially, SN analyzed the transcripts thematically and concurrently with data collection to facilitate purposive sampling and identify saturation. Early analysis of transcripts allowed us to identify preliminary themes and gaps, informing subsequent interviews and ensuring that a diverse range of perspectives were explored [33,42,43]. This iterative process of concurrent analysis and interviewing was integral to refining our understanding and ensuring the trustworthiness of the findings [42,44]. At the completion of the interviews and focus group discussions, MF and SN analyzed data using qualitative content analysis [31] and framework analysis [31,45-47]. Framework analysis is a systematic qualitative analysis widely used in health and social care research to organize and analyze large volumes of textual data, enabling comparison by case and theme [48].

The data analysis process is outlined in Figure 1. First, barriers, facilitators, and neutral factors associated with CDSS across implementation cycles were assigned a short summarizing phrase [31,42]. This was done through inductive open coding of each transcript. These initial phrases were discussed and revised, with a few being discarded, amended, or subsumed to create higher-order codes [31]. Through an initial inductive analysis of codes, the researchers aimed to capture the context-specific elements of the data without prematurely applying a predefined framework [38,46]. This approach enabled the discovery of preliminary themes and patterns, which were subsequently aligned with the NASSS domains and subdomains during the deductive phase of the analysis [33,38,46].

Figure 1. Analytic process used in data analysis. CDSS: clinical decision support system; NASSS: Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability.



Alignment With the NASSS and Active Implementation Stages Frameworks

As a part of this process, where appropriate, illustrative quotes were extracted for each of these codes [42]. The selection of illustrative quotes during qualitative analysis was guided using a consensus approach after discussion with MF, SN, BA, and ZT. The selection of quotes aimed to capture diverse perspectives and enrich subsequent mapping to the NASSS framework to showcase the variations within each theme and subtheme, ensuring that findings were both grounded in the data collected and demonstrative of the range of perspectives expressed by the participants [33,45,46,48-50]. The outcome of this process was to reflexively map each of the higher-order themes onto ≥1 of the NASSS domains and subdomains with which they were aligned [31]. This analysis was primarily

deductive, with the intent to align barriers, facilitators, and neutral factors identified in our inductive analysis with preexisting domains within the NASSS framework [31]. It is important to note that, in the analysis, no restriction was placed on the number of domains or subdomains with which an individual-coded barrier, facilitator, or neutral factor could align [31]. This was not only to categorize using the NASSS framework but to reflect on the role of the theoretical framework in capturing the complexity of CDSS implementation within a real-world health service [31,51]. Themes not aligned to an NASSS domain or subdomain were coded separately [46]. Finally, the Active Implementation Framework [35,36] was used to categorize and display identified determinants according to the stages for digital health implementation. These stages include exploration, which is focused on the feasibility and organizational readiness, installation, which is centered on the organizational preparation, initial implementation, which covers implementation-initiation techniques such as training, and full implementation, which emphasizes sustainment [35-37].

Data Presentation and Interpretation

A descriptive numerical summary of the identified determinants was collated in an Excel (Microsoft Corporation) spreadsheet, mapping determinants to the NASSS framework domains and the Active Implementation Framework stages [31,42]. To illustrate the alignment of barriers, facilitators, and neutral factors with the NASSS framework and the Active Implementation Framework, the researchers developed visual matrices to depict the frequency and salience of the identified themes and their association with various implementation phases [31]. This visual and numerical representation aimed to demonstrate common findings across the NASSS domains relating to CDSS implementation and to identify gaps in the implementation of CDSS across different phases.

Trustworthiness

Throughout the analysis, the researchers maintained trustworthiness and reflexivity by reflecting on the research process through discussion as a research team and considering how their perspectives and the chosen frameworks may have influenced data interpretation [45-47,49]. This study provided a narrative report of the framework mapping to each NASSS domain and subdomain supported by direct quotes from the interviews and focus groups [33,45,46] and discussed the implications for practice.

Results

Participant Characteristics and Implementation Roles

A total of 30 participants, including implementers, decision makers, and CDSS end users, across several departments took part in this study. Table 1 provides comprehensive demographic information about the participants.

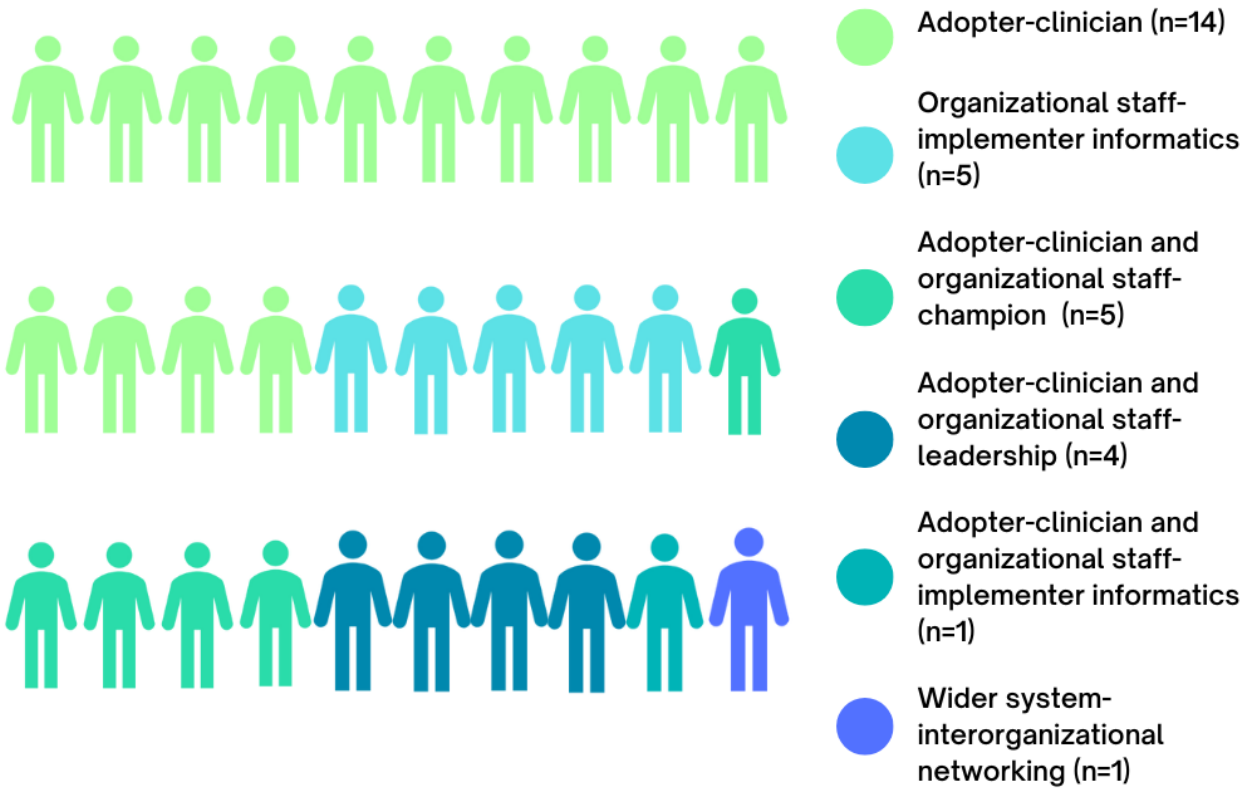
Table 1. Participant demographics (N=30).

Position and categories	Participants, n (%)
Sex	
Female	13 (43)
Male	17 (57)
Portfolio[a]	
Medical specialty (ie, radiology, pharmacy, and cardiology)	16 (53)
Digital health and informatics directorate	8 (27)
Nursing and allied health	6 (20)
Job role	
Physician (ie, residents, junior physicians, consultants, specialists, and directors)	18 (60)
Clinical informatics (ie, project manager, senior management, and data analyst)	6 (20)
Allied health and nursing (ie, physiotherapist and triage nurse)	6 (20)
Implementation role	
Adopter-clinician (ie, physicians, allied health staff, and nursing staff)	14 (47)
Organizational staff–implementer informatics	5 (17)
Adopter-clinician and organizational staff champion	5 (17)
Adopter-clinician and organizational staff leadership	4 (13)
Adopter-clinician and organizational staff-implementer informatics	1 (3)
Wider system–interorganizational networker	1 (3)

[a]Portfolio refers to the health service department associated with the participant, recognizing that some participants had a clinical background, that is, physician, but were experienced at a department level with decision-making, governance, purchasing, design, and/or implementation of clinical decision support system initiatives within this health service.

Participants were mapped to NASSS-informed implementation roles, highlighting their level of decision-making within the hospital systems of interest, as shown in Figure 2. Of the 30 participants, 14 (47%) clinical staff identified solely as adopters of CDSS tools. In addition, 17% (5/30) of clinical staff took on the role of organizational staff champions, that is, clinicians who advocate for the technology and its use [52,53]. Overall, 17% (5/30) of participants were informatics professionals, 13% (4/30) of participants acted in dual roles of organizational staff leadership and clinical adopters, that is, as clinical directors. Only 1 (3%) participant encompassed adopter-clinician and organizational staff–implementer informatics roles, that is, as a clinician who worked within the digital health and informatics directorate portfolio. Only 1 (3%) participant was considered a wider system–interorganizational networker.

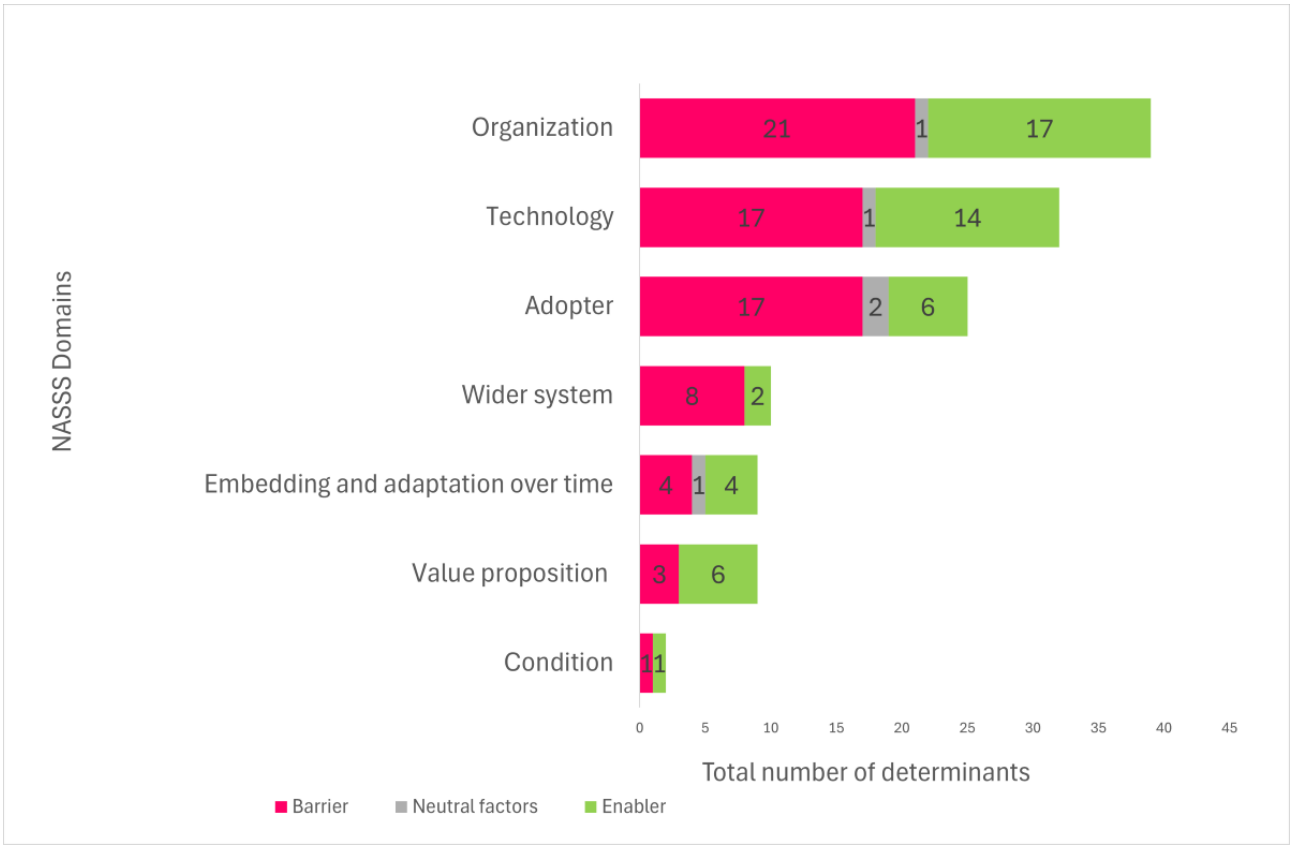
Figure 2. Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework–informed implementation roles of participants (N=30).



Key Determinants Associated with CDSS Implementation

The identified determinants encompass specific elements pertinent to CDSS implementation. These elements were classified into barriers, enablers, and neutral factors. For a comprehensive breakdown of these determinants, refer to Multimedia Appendix 3. The findings were mapped to the NASSS domains and subdomains, as shown in Figure 3. Barriers (71/126, 56.3%), enablers (50/126, 39.6%), and neutral factors (5/126, 3.9%) associated with CDSS implementation were identified in this study. Most determinants were mapped to the organization (39/126, 30.9%), technology (32/126, 25.3%), and adopter (25/126, 19.8%) domains.

Figure 3. Key determinants associated with clinical decision support system (CDSS) implementation mapped to the Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) domains.



Framework Analysis of NASSS Domains and Subdomains

Organizational Domain

Key organizational level determinants were mapped to the following NASSS subdomains. The barriers included limited implementation capacity stemming from a lack of longer-term implementation planning know-how:

> *That's something that's a little bit vague for me, do we plough on? With a state that shows it works. Then you bring in your usability. Or do you bring in your app and then do your study?...And so that, that design of the next phase sort of up here, but I don't know the best way to do it would be...how I'd explain it.* [Clinical adopter, senior management]

The participant highlights a concern on whether to continue with existing approaches that show some amount of success (eg, *state that shows it works*) or to introduce new technologies (eg, apps) and evaluate their impact. This decision-making process ties into limited implementation capacity, underscoring the importance of strategic planning and expertise in guiding successful implementation efforts. This may have flow-on effects, resulting in inappropriate resource prioritization, which is identified as a barrier that needs to be worked on:

> *Generally, they will provide support for when they're rolling out something new. But often you don't actually discover issues or problems with that new process or software until you're using it.* [Clinical adopter]

Furthermore, this impacted *technological readiness* as a barrier, as stated by a participant:

> *I think [Health Service] really hasn't accommodated us because...if you're on a VPN connection, you're throttled down to three or four megabits per second.* [Clinical adopter]

While not every clinician in the hospital may always be on a virtual private network (VPN), VPN throttling affects system performance in hospitals by limiting bandwidth, thereby hindering the speed and efficiency of accessing critical systems or data. This can negatively impact workflow efficiency and user satisfaction for those required to use a VPN. This quote also reflects participant's concerns about organizational readiness to adopt new technologies effectively in comparison to existing organization infrastructure and practices. Conversely, the organization appeared well- equipped to support initial or early-stage CDSS implementation activities, including training:

> *Champions are trained, they've done validation, then it's a case of, OK, now we go live with that product.* [Informatics staff, senior management]

The organization has also established a highly skilled clinical informatics workforce, as stated by a participant:

> *They're the only group that has a clinical informatics solution. So, you can do a training...in clinical informatics. So, it's nice because they have some*

> *champions now integrated.* [Clinical adopter, digital consultant]

This workforce ensures ongoing maintenance and support in navigating the digital ecosystem:

> *There is a department in place to help you navigate all the things you need to do. Whether that's from an IT procurement perspective, whether or not that's from a cybersecurity perspective...* [Informatics staff]

These enablers were mapped to the subdomain of work needed to plan, implement and monitor change. It must also be noted that the inclusion of organizational champions across clinical contexts was a significant enabler in the adoption of several CDSSs across this hospital system. This was a key factor in enabling the initial work needed to plan, implement, and monitor change.

The informatics team has a close familial culture due to the team members working together for a long time:

> *All of us have been working together for 10 plus years, some of them 15. Very close-knit team. So that is what holds us together.* [Informatics staff]

The tight-knit nature of the informatics team can be seen as part of the organizational culture, impacting how the team works together, their ability to innovate, and their readiness to implement changes in the organization. Although mapped to the organizational domain of the NASSS framework, given that the primary focus of this culture was not strictly covered by standard organizational structures or leadership theories associated with the NASSS framework's *general capacity to innovate* [54] but instead highlighted the unique interpersonal dynamics within the team, we coded this separately from the NASSS framework.

Technology Domain

In the hospital system, the most common technological barrier was associated with the interoperability of CDSSs, which was mapped to the material properties subdomain:

> *Sometimes, the lack of interface between the systems...they run a parallel system.* [Clinical adopter]

Participants also shared that they were not always fully aware of CDSS modifications:

> *There used to be a feature in the IEMR where you could taper medication doses...I don't know what ever happened to it and it just wasn't there anymore.* [Clinical adopter]

In addition, users were not always aware of the best ways to optimize using CDSS technology in this hospital system:

> *I know as someone who works in digital health and informatics, there is...a little button to the side of that that says, Click to see the 'information. But I don't think a lot of people realize that...* [Clinical adopter, digital consultant]

However, when participants were able to experience how CDSS supported effective practice, they were more likely to engage and use the technology:

> *If you're not too familiar with this medication or you've forgot what the maximum dose should be or you were distracted by something else, the power plan [CDSS] has it there and says you shouldn't be going above this...so I think that's good from a patient safety perspective.* [Clinical adopter]

> *I would imagine ease of use for a start...if it's too clunky, they won't go anywhere, you know, even introducing a new app into the system if it...takes 10 minutes to set it up.* [Clinical adopter, senior management]

> *Ease of use drives clinicians' confidence.* [Clinical adopter, senior management]

These quotes are illustrative of how *ease of use* can enhance clinicians' overall confidence and proficiency in using the technology effectively. This aspect minimizes the learning curve and allows users to operate the CDSS with minimal training or interruption. This can lead to increased confidence among clinicians because they are able to interact with the system more efficiently and focus more on patient care rather than struggling with the technology. Therefore, this enabler mapped to the NASSS subdomains of knowledge generated by the technology and knowledge to use it.

Adopter and Condition

Despite some participants viewing technology as supportive of effective practice, others expressed concerns with machine learning found in modern CDSSs. These concerns were about limiting the capacity for critical thinking, professional autonomy, and personal legal ramifications:

> *The threat of machine learning would be that you could become reliant on, you know, what an algorithm is telling you and directing you to do, and you might lose that art of being able to, like, go well...I recognise this issue because I've seen it before.* [Clinical adopter]

> *This is where I think it's something we're going to have to have to work out and develop some protections around it...I want to make sure this thing is safe.* [Clinical adopter]

Conversely, adopters expressed positive perceptions of CDSS technology when there was a clear lived experience of improved workflow:

> *From a workflow perspective, it reduces the number of clicks and the number of individual actions you need to do to...So I think it makes things a lot easier in that respect.* [Clinical adopter]

This was also the case for patient-centered care observed through improved outcomes:

> *I think that's...the excitement of machine learning is learning more...individualising the dose for a patient.* [Clinical adopter]

Effective adaptation to the local clinical context was also an enabler, as stated by a participant:

> *One sort of springs to mind that we do use as a screening tool...It's helpful to do it, sort of pops up for every patient during your initial clinical assessment...it's just like three little questions that you ask...it is an incentive for me to do a complete set of vitals each time...I feel like I do it more often because I've got that incentive to get that score.* [Clinical adopter]

Participants emphasized the importance of adapting CDSSs to local clinical environments. CDSS was viewed positively when demonstratively relevant adaptations were made in contrast to a clear lack of clinically contextual adaptation:

> *Sometimes it can be a bit too rigid, and if you're wanting to do something that falls out of the power plan, they sometimes can make the job harder...* [Clinical adopter]

> *The warfarin one works well and I think it's a better example of how our plans would be used.* [Clinical adopter]

Consequently, this determinant was also mapped to the condition domain.

Value Proposition

CDSS using AI was seen as valuable to adopters by enabling efficiency in protocolized tasks with predefined workflows:

> *So, it's really saving time on things that can very well be done by sort of an artificial intelligence...it won't replace the reporting.* [Clinical adopter, senior management]

In contrast, when considering the relationship between demand-side value (to adopters) and supply-side value (to developers), different priorities can create misalignment regarding the value proposition of a CDSS initiative. This misalignment may lead to communication barriers:

> *I think our clients and customers, they think they're things really important. Because they have no visibility of the strokes that we're managing at the moment, yeah so. For them, it's hard to say what do you mean. My initiative in one tiny little ward, which means so much to me...you can't help me. And I think, because truthfully, they don't have visibility of all the other things that we're doing.* [Informatics staff]

Wider System

The lack of clear and consistent governance surrounding the application of CDSSs within the hospital context was perceived as a major wider system barrier by most participants:

> *But it's also really raised a lot or highlighted a lot of the areas where there could be more maturity...I guess, operating in terms of making sure that whatever processes need to be in place for governance and decision-making, ethics, liability.* [Informatics staff, senior management]

However, participants also noted that the experience of the COVID-19 pandemic showed that wider system decision-making to facilitate the appropriate use of digital technology could be streamlined in an effective and efficient manner:

> *I think COVID certainly showed us that we can streamline a lot of our decision making...* [Clinical adopter, digital consultant]

Embedding and Adaption With Time

Participants noted the resilience of this hospital system in adapting to and accommodating the application of technology as a key strength. This stemmed from an innovation culture and recognition of the importance of evaluating technology across implementation cycles, even if they did not currently have the resources or exact know-how to do so:

> *I'm a massive champion of that whole point of we do go through a cycle, right? Where you plan your budget, you deliver, you maintain most systems...go to that next step, which is to evaluate and more importantly, evaluate does this thing still deliver the same value statement that we thought it would at the start.* [Informatics staff]

A noted barrier was the acknowledgment that adapting complex CDSSs, particularly AI, to contextual changes in the system in the course of time was specialized and labor intensive:

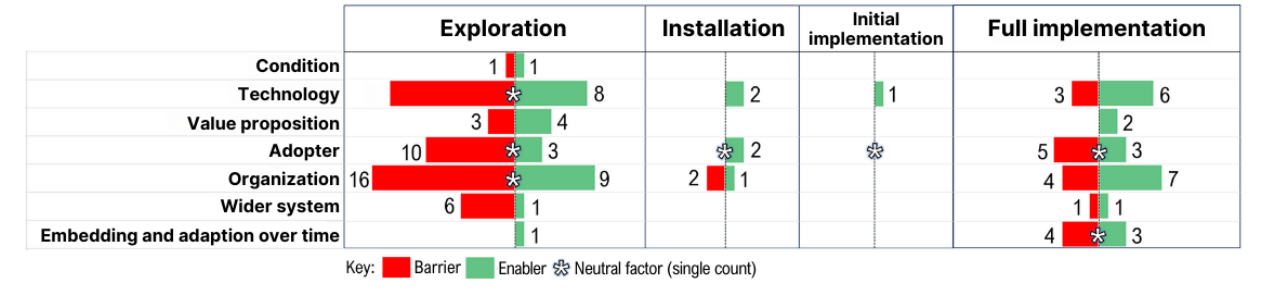
> *So, we need to optimise it...how do we then make sure that the tool stays accurate over time? No one really has, I think, quite worked out, I think "that's going to be the challenge."* [Clinical adopter, senior management]

Mapping NASSS-Informed Determinants to Implementation Stages Across the Health Service

Overview

It must be noted that when mapped to Active Implementation Framework stages [35-37], there was an uneven distribution of determinants across implementation stages, with most determinants falling in the scope of the exploration and full implementation stages. This highlighted a tendency of stakeholders to engage in the frequent preimplementation or exploration activity and full-scale adoption activity. In contrast, fewer determinants were associated with identifying contextual drivers, developing adopter readiness, and facilitating contextual capacity building to sustain ongoing adaption within an established feedback and monitoring strategy. Figure 4 illustrates the uneven distribution of the 126 identified determinants; Out of the identified determinants, 77 (61%) were solely associated with the *exploration* stage, 39 (31%) with the *full implementation* stage, and 6 (5%) with the *middle installation* stage.

Figure 4. A tabular diagram illustrating the Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) determinants mapped to implementation stages.



Exploration

Most determinants described in this study were identified during the preimplementation stage, which is associated with assessing organizational readiness and technological procurement (exploration stage). Enablers in this stage included mapping CDSS procurement to the clinical context and recognizing the importance of involving diverse stakeholders. Conversely, there were several barriers relating to the organizational, technological, and wider system within the exploration stage. This included a lack of streamlined executive decision-making for CDSS procurement and adaption and navigating complex governance processes.

Installation and Initial Implementation

These stages describe actions associated with identifying contextual factors, which may facilitate the integration of a CDSS into a system to inform an initial smaller-scale rollout or pilot. The organization appeared to have engaged in fewer installation-stage piloting activities; however, the key enabling determinants were identified. Among these adopter-level factors associated with enhancing clinician trust in CDSSs through education, training, and leveraging peer champions were noted. Conversely, barriers in this stage included a lack of targeted resourcing to facilitate full-cycle CDSS implementation planning.

Full Implementation

This stage represents a state of full technological integration across an entire system or organization. This involves activities associated with ongoing monitoring and iteration, scalability, and replicability. Several enablers were identified within this phase, particularly at an organizational and technological level, when considered in the context of adaption and embedment in the course of time. This included decision makers recognizing the need for ongoing investment in capacity building to sustain a digitally informed workforce and the need for ongoing iteration of CDSSs to support effective practice in the course of time. The barriers identified included the lack of sustained organizational champions postadoption and insufficient training and information on technology upgrades which occurred over time.

Multistage Implementation Activities

Four determinants spanned multiple stages. Organizational-level training of champions supported CDSS adoption in both exploration and installation stages. A notable barrier was the lack of knowledge and guidance for implementation planning, affecting organizational readiness. Trust in CDSS was recognized as varying among clinicians and was not systemically evaluated across all implementation stages. Inbuilt tracking mechanisms to measure uptake and patterns of CDSS were seen as beneficial for transparency and user fidelity throughout implementation cycles but were not uniformly explored or applied.

Discussion

Principal Findings

By mapping NASSS-informed determinants influencing CDSS implementation cycles using the Active Implementation Framework stages [35-37], our research provides pragmatic insights to inform tailored integration strategies of these technologies into large hospital systems. Despite encountering limitations in implementation capacity and planning know-how, the institution demonstrated strength through a well-equipped clinical informatics workforce and early-stage training of its adopter workforce. This finding aligns with existing literature, emphasizing the critical role of organizational culture and support structures in successful technology adoption [13,31]. Second, CDSS implementation faced technological hurdles, including interoperability and interface issues, a frequently reported CDSS engagement barrier [15,22,55]. However, positive user experiences emerged as a significant factor influencing CDSS use, particularly when adopters could directly experience improved efficiency and better patient outcomes for themselves. This underscores the importance of user-centric design and showcases the practical impact of CDSS on health care workflows, aligning with the broader pattern of emphasizing end user perspectives in technology adoption [12,15,17]. The presence of peer champions among clinicians emerged as a significant enabler. Peer-to-peer support and advocacy played a crucial role in enhancing clinician confidence and acceptance of CDSSs, contributing to a smoother initial implementation process. Moreover, the study reveals wider systemic barriers, such as a lack of clear governance for CDSS applications, aligning with the ongoing discourse on the necessity for robust regulatory frameworks in the broader implementation of digital health care technologies [25,56]. This finding emphasizes the need for a systemic approach to address governance gaps and ensure the effective integration of CDSSs into health care systems. These findings strongly align with our recent scoping review, which mapped reported CDSS

implementation barriers and enablers to NASSS domains [31]. Although further empirical evidence from successful implementation is required, this further highlights the reliability of the NASSS framework as a pragmatic tool to identify meaningful domain-level determinants associated with the implementation of technology within health systems [24,54].

When determinants were mapped to the *Active Implementation Framework stages* [35-37], a clear pattern emerged highlighting implementation activities most notable in the early stage of *exploration* and the later stage of *full implementation*. The absence of planned piloting and process evaluation was not unexpected, given a lack of evaluation and implementation planning *know-how* was identified as a key organizational readiness barrier.

This may be addressed through the application of theory-informed approaches to implementation planning [27,30]. This finding aligns with the gaps identified in our recent scoping review examining the use of theories, models, and frameworks to support CDSS implementation within hospital systems [29]. This review [29] reported an inconsistent application of systematic approaches to implementation planning within hospital systems in several countries.

Furthermore, while organizational-level training of champions was crucial and supported CDSS adoption in both the exploration and installation stages, this was not sustained throughout the implementation life cycle. This weakening of the informatics-adopter sociotechnical network during the course of time may have contributed to the communication issues adopters frequently highlighted. This study identified 3 key communication issues in the implementation of CDSSs. First is the challenges in effectively communicating and coordinating within the complicated web of governance structures, potentially impeding the progression of CDSS implementation. Second, the limitations in ongoing communication about system updates and advancements may lead to adopters being unaware of improvements, impacting their ability to fully use the CDSS. Third, trust in CDSSs was recognized as varying among clinicians, indicating a potential communication challenge in conveying the benefits and reliability of the system uniformly across all CDSS adopters. More rigorous implementation planning, encompassing all 4 stages, may go some way in addressing the identified communication gaps [27,29,30,35,36].

Furthermore, this study identified the need for ongoing adopter support and planned evaluation strategies encompassing the full implementation cycle [36,56]. Evidence indicates that sustained cycles of monitoring and iteration may lead to sustained integration of CDSS in health care systems [21,34,57]. For example, planned mechanisms for user tracking allow for real-time assessment of uptake, which may impact the choice and delivery of implementation strategies to facilitate the continuous adoption of the innovation in practice [10,58].

Strengths and Limitations

The study's strengths lie in its holistic systems thinking approach, which provides a comprehensive understanding of the challenges and enablers within each implementation stage and across implementation cycles for complex systems such as CDSS. The researchers acknowledge that the final phase, defined as *full implementation* within the *Active Implementation Framework*, does not fully address all components of sustainability, which can include factors such as changing legislation, government (federal and state) budgets, and workforce capacity [35,36,57,59]. However, this framework does acknowledge factors that bridge the policy gap within an organization, including clear governance pathways and longer-term resourcing, in addition to recognizing embedded behavior change and ongoing monitoring and iteration as parts of full implementation cycles [35].

It must be noted that while this study was intentionally designed for a 2-year period, there may have been specific events (including the COVID-19 pandemic) that may have influenced participant perceptions. However, it was not within the study's scope to contextualize a comprehensive list of events that could have impacted perceptions. Nonetheless, our questions were designed to probe reasons behind participant perceptions when interviewed; early analysis of transcripts allowed us to identify preliminary themes and gaps, ensuring that a diverse range of perspectives were explored and stopped when no new insights were gained [33,42,44]. The participants' roles can influence their perceptions and experiences with CDSS implementation. Clinical adopters might focus more on usability and patient impact, whereas senior management or informatics professionals may prioritize technical challenges and integration issues [60]. Findings may not fully generalize to other health care settings with different organizational structures, levels of technological maturity, or cultural contexts [10]. It must be noted that patient perspectives were outside the study scope, which focused on clinicians as end users within an acute-care (hospital) setting. Recommendations and insights derived from this study may need to be interpreted cautiously in broader contexts and may apply more directly to the study settings where similar participant distributions and roles are prevalent [10,61]. Future work should also consider examining patient perspectives to enhance insights.

The study was conducted within a specific health care system, potentially limiting the generalizability of some findings. However, the use of the NASSS framework has identified multilevel determinants that align with trends in contemporary research findings across the health care system [31,62,63]. Moreover, the study's participant population represents a broad spectrum of stakeholders involved in CDSS implementation, including clinical staff adopters, organizational champions, and informatics professionals. This diversity ensured the representativeness of perspectives needed to understand the broader contextual factors applicable beyond this study's settings [63,64]. By including individuals from various departments and decision-making levels, the study captures a holistic view of the challenges and facilitators influencing CDSS adoption, contributing to a nuanced understanding of implementation dynamics within the hospital system. Furthermore, this qualitative study was exploratory in nature and enabled us to unpack the intersectionality of multiple determinants in influencing a range of implementation outcomes.

Future research could extend the use of the NASSS framework to the application of CDSS in different health care environments

or medical conditions such as in rural hospitals or mental health care [31,34]. Furthermore, investigations might integrate a systems-level framework, such as the NASSS framework, using behavioral assessments, standardized psychological tests, or multimodal methodologies to explore individual emotional reactivity for CDSS adoption [13,65]. Such endeavors may deepen insights into health care professionals' emotional responses, such as stress levels, satisfaction, and confidence [13], to CDSS use. The insights gained could contribute to broader knowledge about technology implementation across diverse health care contexts, informing tailored strategies for improved adoption, long-term viability, and best practices to enhance CDSS adoption in various health care settings [31,66]. Additional quantitative data and further empirical application of the NASSS framework may be beneficial in further exploring specific qualitative findings identified in this study. Future research could explore the longitudinal aspects of CDSS adoption to capture changes and adaptations over extended periods by conducting trials with integral process evaluations to test the identified implementation factors across diverse populations and settings [53,60,63].

Conclusions

These findings address a significant knowledge gap in the literature using system thinking principles to identify the interdependent dynamics of CDSS implementation. In moving forward, this study serves as a catalyst for informed decision-making in CDSS implementation, offering actionable insights for practitioners and researchers alike. As technology continues to evolve and health care landscapes transform, the lessons gleaned from this study provide a foundation for refining CDSS implementation strategies and advancing patient-centered, efficient, and ethically sound health care practices.

Acknowledgments

The authors would like to gratefully acknowledge the study participants who gave their valuable time to participate in this study.

Data Availability

The dataset (which includes individual transcripts) is not publicly available due to confidentiality policies.

Authors' Contributions

MF, SM, and SN were involved in the conceptualization of the study. MF and SN were involved in the methodology, and MF, BA, and SN were involved in the validation. Formal analysis was performed by MF and SN, visualization was performed by MF and BA, and investigation was performed by SM and SN. The study was supervised by BA, SM, ZT, and SN. Writing the original draft involved MF and SN, and the writing was reviewed and edited by MF, BA, SM, ZT, AT, and SN.

Conflicts of Interest

MF reports that financial support and paper publishing charges were provided by the Digital Health Cooperative Research Centre (DHCRC-0058). SM reports that financial support was provided by fellowships administered by the National Health and Medical Research Council administered (#1181138). Other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Multimedia Appendix 1
Standards for Reporting Qualitative Research checklist.
[DOCX File , 20 KB - medinform_v12i1e60402_app1.docx]

Multimedia Appendix 2
Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability (NASSS) framework–informed interview guide.
[DOCX File , 436 KB - medinform_v12i1e60402_app2.docx]

Multimedia Appendix 3
Supplementary breakdown of mapping.
[XLSX File (Microsoft Excel File), 99 KB - medinform_v12i1e60402_app3.xlsx]

References

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
2. Kwok R, Dinh M, Dinh D, Chu M. Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: implementation of a dynamic and integrated electronic decision support system. Emerg Med Australas 2009 Feb;21(1):31-37 [FREE Full text] [doi: 10.1111/j.1742-6723.2008.01149.x] [Medline: 19254310]

3. Barrett M, Boyne J, Brandts J, Brunner-La Rocca HP, de Maesschalck L, de Wit K, et al. Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care. EPMA J 2019 Dec 22;10(4):445-464 [FREE Full text] [doi: 10.1007/s13167-019-00188-9] [Medline: 31832118]
4. Moja L, Polo Friz H, Capobussi M, Kwag K, Banzi R, Ruggiero F, et al. Effectiveness of a hospital-based computerized decision support system on clinician recommendations and patient outcomes: a randomized clinical trial. JAMA Netw Open 2019 Dec 02;2(12):e1917094 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.17094] [Medline: 31825499]
5. Bertsche T, Pfaff J, Schiller P, Kaltschmidt J, Pruszydlo MG, Stremmel W, et al. Prevention of adverse drug reactions in intensive care patients by personal intervention based on an electronic clinical decision support system. Intensive Care Med 2010 Apr 9;36(4):665-672. [doi: 10.1007/s00134-010-1778-8] [Medline: 20143221]
6. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. Implementation of artificial intelligence-based clinical decision support to reduce hospital readmissions at a regional hospital. Appl Clin Inform 2020 Aug 02;11(4):570-577 [FREE Full text] [doi: 10.1055/s-0040-1715827] [Medline: 32877943]
7. Uslu A, Stausberg J. Value of the electronic medical record for hospital care: update from the literature. J Med Internet Res 2021 Dec 23;23(12):e26323 [FREE Full text] [doi: 10.2196/26323] [Medline: 34941544]
8. Elhaddad M, Hamam S. AI-driven clinical decision support systems: an ongoing pursuit of potential. Cureus 2024 Apr;16(4):e57728 [FREE Full text] [doi: 10.7759/cureus.57728] [Medline: 38711724]
9. Moazemi S, Vahdati S, Li J, Kalkhoff S, Castano LJ, Dewitz B, et al. Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: a systematic review. Front Med (Lausanne) 2023 Mar 31;10:1109411 [FREE Full text] [doi: 10.3389/fmed.2023.1109411] [Medline: 37064042]
10. Chen Z, Liang N, Zhang H, Li H, Yang Y, Zong X, et al. Harnessing the power of clinical decision support systems: challenges and opportunities. Open Heart 2023 Nov 28;10(2):e002432 [FREE Full text] [doi: 10.1136/openhrt-2023-002432] [Medline: 38016787]
11. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. Med Law Rev 2023 Nov 27;31(4):501-520 [FREE Full text] [doi: 10.1093/medlaw/fwad013] [Medline: 37218368]
12. Laka MA, Milazzo A, Merlin T. Factors that impact the adoption of clinical decision support systems (CDSS) for antibiotic management. Int J Environ Res Public Health 2021 Feb 16;18(4):1901 [FREE Full text] [doi: 10.3390/ijerph18041901] [Medline: 33669353]
13. Ackerhans S, Huynh T, Kaiser C, Schultz C. Exploring the role of professional identity in the implementation of clinical decision support systems-a narrative review. Implement Sci 2024 Feb 12;19(1):11 [FREE Full text] [doi: 10.1186/s13012-024-01339-x] [Medline: 38347525]
14. Schwartz JM, Moy AJ, Rossetti SC, Elhadad N, Cato KD. Clinician involvement in research on machine learning-based predictive clinical decision support for the hospital setting: a scoping review. J Am Med Inform Assoc 2021 Mar 01;28(3):653-663 [FREE Full text] [doi: 10.1093/jamia/ocaa296] [Medline: 33325504]
15. Kouri A, Yamada J, Lam Shin Cheung J, van de Velde S, Gupta S. Do providers use computerized clinical decision support systems? A systematic review and meta-regression of clinical decision support uptake. Implement Sci 2022 Mar 10;17(1):21 [FREE Full text] [doi: 10.1186/s13012-022-01199-3] [Medline: 35272667]
16. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
17. Lugtenberg M, Pasveer D, van der Weijden T, Westert GP, Kool RB. Exposure to and experiences with a computerized decision support intervention in primary care: results from a process evaluation. BMC Fam Pract 2015 Oct 16;16(1):141 [FREE Full text] [doi: 10.1186/s12875-015-0364-0] [Medline: 26474603]
18. Klarenbeek SE, Schuurbiers-Siebers OC, van den Heuvel MM, Prokop M, Tummers M. Barriers and facilitators for implementation of a computerized clinical decision support system in lung cancer multidisciplinary team meetings-a qualitative assessment. Biology (Basel) 2020 Dec 25;10(1):9 [FREE Full text] [doi: 10.3390/biology10010009] [Medline: 33375573]
19. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc 2007 Oct 11;2007:26-30 [FREE Full text] [Medline: 18693791]
20. Stone EG. Unintended adverse consequences of a clinical decision support system: two cases. J Am Med Inform Assoc 2018 May 01;25(5):564-567 [FREE Full text] [doi: 10.1093/jamia/ocx096] [Medline: 29036296]
21. Cresswell KM, Bates DW, Williams R, Morrison Z, Slee A, Coleman J, et al. Evaluation of medium-term consequences of implementing commercial computerized physician order entry and clinical decision support prescribing systems in two 'early adopter' hospitals. J Am Med Inform Assoc 2014 Oct;21(e2):e194-e202 [FREE Full text] [doi: 10.1136/amiajnl-2013-002252] [Medline: 24431334]
22. Liberati E, Ruggiero F, Galuppo L, Gorli M, González-Lorenzo M, Maraldi M, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017 Sep 15;12(1):113 [FREE Full text] [doi: 10.1186/s13012-017-0644-2] [Medline: 28915822]
23. Devaraj S, Sharma SK, Fausto DJ, Viernes S, Kharrazi H. Barriers and facilitators to clinical decision support systems adoption: a systematic review. J Bus Adm Res 2014 Jul 24;3(2):36. [doi: 10.5430/jbar.v3n2p36]

24. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
25. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
26. Greenhalgh T, Russell J. Why do evaluations of eHealth programs fail? An alternative set of guiding principles. PLoS Med 2010 Nov 02;7(11):e1000360 [FREE Full text] [doi: 10.1371/journal.pmed.1000360] [Medline: 21072245]
27. Nilsen P, Reed J, Nair M, Savage C, Macrae C, Barlow J, et al. Realizing the potential of artificial intelligence in healthcare: learning from intervention, innovation, implementation and improvement sciences. Front Health Serv 2022 Sep 15;2:961475 [FREE Full text] [doi: 10.3389/frhs.2022.961475] [Medline: 36925879]
28. Bauer MS, Kirchner J. Implementation science: what is it and why should I care? Psychiatry Res 2020 Jan;283:112376 [FREE Full text] [doi: 10.1016/j.psychres.2019.04.025] [Medline: 31036287]
29. Fernando M, Abell B, Tyack Z, Donovan T, McPhail SM, Naicker S. Using theories, models, and frameworks to inform implementation cycles of computerized clinical decision support systems in tertiary health care settings: scoping review. J Med Internet Res 2023 Oct 18;25:e45163 [FREE Full text] [doi: 10.2196/45163] [Medline: 37851492]
30. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215 [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]
31. Abell B, Naicker S, Rodwell D, Donovan T, Tariq A, Baysari M, et al. Identifying barriers and facilitators to successful implementation of computerized clinical decision support systems in hospitals: a NASSS framework-informed scoping review. Implement Sci 2023 Jul 26;18(1):32 [FREE Full text] [doi: 10.1186/s13012-023-01287-y] [Medline: 37495997]
32. Thomas EE, Chambers R, Phillips S, Rawstorn JC, Cartledge S. Sustaining telehealth among cardiac and pulmonary rehabilitation services: a qualitative framework study. Eur J Cardiovasc Nurs 2023 Dec 14;22(8):795-803. [doi: 10.1093/eurjcn/zvac111] [Medline: 36468293]
33. Nordmann K, Sauter S, Redlich MC, Möbius-Lerch P, Schaller M, Fischer F. Challenges and conditions for successfully implementing and adopting the telematics infrastructure in German outpatient healthcare: a qualitative study applying the NASSS framework. Digit Health 2024;10:20552076241259855 [FREE Full text] [doi: 10.1177/20552076241259855] [Medline: 39070890]
34. Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. Int J Med Inform 2020 May;137:104072. [doi: 10.1016/j.ijmedinf.2019.104072] [Medline: 32200295]
35. Blanchard C, Livet M, Ward C, Sorge L, Sorensen TD, McClurg MR. The active implementation frameworks: a roadmap for advancing implementation of comprehensive medication management in primary care. Res Social Adm Pharm 2017 Sep;13(5):922-929. [doi: 10.1016/j.sapharm.2017.05.006] [Medline: 28549800]
36. Metz A, Bartley L, Ball H, Wilson D, Naoom S, Redmond P. Active implementation frameworks for successful service delivery: Catawba county child wellbeing project. Res Soc Work Pract 2014 Jul 28;25(4):415-422. [doi: 10.1177/1049731514543667]
37. Fixsen DL, Blase KA, Naoom SF, Wallace F. Core implementation components. Res Soc Work Pract 2009 May 27;19(5):531-540. [doi: 10.1177/1049731509335549]
38. Proudfoot K. Inductive/deductive hybrid thematic analysis in mixed methods research. J Mix Methods Res 2022 Sep 20;17(3):308-326. [doi: 10.1177/15586898221126816]
39. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
40. Campbell S, Greenwood M, Prior S, Shearer T, Walkem K, Young S, et al. Purposive sampling: complex or simple? Research case examples. J Res Nurs 2020 Dec 18;25(8):652-661 [FREE Full text] [doi: 10.1177/1744987120927206] [Medline: 34394687]
41. Churruca K, Ludlow K, Taylor N, Long JC, Best S, Braithwaite J. The time has come: embedded implementation research for health care improvement. J Eval Clin Pract 2019 Jun 10;25(3):373-380. [doi: 10.1111/jep.13100] [Medline: 30632246]
42. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep 11;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
43. Saunders B, Sim J, Kingstone T, Baker S, Waterfield J, Bartlam B, et al. Saturation in qualitative research: exploring its conceptualization and operationalization. Qual Quant 2018;52(4):1893-1907 [FREE Full text] [doi: 10.1007/s11135-017-0574-8] [Medline: 29937585]
44. Mazaheri M, Eriksson LE, Heikkilä K, Nasrabadi AN, Ekman S, Sunvisson H. Experiences of living with dementia: qualitative content analysis of semi-structured interviews. J Clin Nurs 2013 Nov 02;22(21-22):3032-3041. [doi: 10.1111/jocn.12275] [Medline: 23815315]

45. Musson D, Buchanan H, Nolan M, Asimakopoulou K. Barriers and facilitators to using an objective risk communication tool during primary care dental consultations: a theoretical domains framework (TDF) informed qualitative study. J Dent 2024 Mar;142:104853. [doi: 10.1016/j.jdent.2024.104853] [Medline: 38244908]
46. Moffat A, Cook EJ, Chater AM. Examining the influences on the use of behavioural science within UK local authority public health: qualitative thematic analysis and deductive mapping to the COM-B model and theoretical domains framework. Front Public Health 2022 Oct 20;10:1016076 [FREE Full text] [doi: 10.3389/fpubh.2022.1016076] [Medline: 36339139]
47. Skillman M, Cross-Barnet C, Friedman Singer R, Rotondo C, Ruiz S, Moiduddin A. A framework for rigorous qualitative research as a component of mixed method rapid-cycle evaluation. Qual Health Res 2019 Jan;29(2):279-289. [doi: 10.1177/1049732318795675] [Medline: 30175660]
48. Collaço N, Wagland R, Alexis O, Gavin A, Glaser A, Watson EK. Using the framework method for the analysis of qualitative dyadic data in health research. Qual Health Res 2021 Jul 13;31(8):1555-1564 [FREE Full text] [doi: 10.1177/10497323211011599] [Medline: 33980102]
49. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. Qual Psychol 2022 Feb 13;9(1):3-26. [doi: 10.1037/qup0000196]
50. Lingard L. Beyond the default colon: effective use of quotes in qualitative research. Perspect Med Educ 2019 Dec 22;8(6):360-364 [FREE Full text] [doi: 10.1007/s40037-019-00550-7] [Medline: 31758490]
51. Lynch EA, Mudge A, Knowles S, Kitson AL, Hunter SC, Harvey G. "There is nothing so practical as a good theory": a pragmatic guide for selecting theoretical approaches for implementation projects. BMC Health Serv Res 2018 Nov 14;18(1):857 [FREE Full text] [doi: 10.1186/s12913-018-3671-z] [Medline: 30428882]
52. Gui X, Chen Y, Zhou X, Reynolds TL, Zheng K, Hanauer DA. Physician champions' perspectives and practices on electronic health records implementation: challenges and strategies. JAMIA Open 2020 Apr;3(1):53-61 [FREE Full text] [doi: 10.1093/jamiaopen/ooz051] [Medline: 32607488]
53. Sequeira L, Almilaji K, Strudwick G, Jankowicz D, Tajirian T. EHR "SWAT" teams: a physician engagement initiative to improve Electronic Health Record (EHR) experiences and mitigate possible causes of EHR-related burnout. JAMIA Open 2021 Apr;4(2):ooab018 [FREE Full text] [doi: 10.1093/jamiaopen/ooab018] [Medline: 33898934]
54. Greenhalgh T, Abimbola S. The NASSS framework - a synthesis of multiple theories of technology implementation. Stud Health Technol Inform 2019 Jul 30;263:193-204. [doi: 10.3233/SHTI190123] [Medline: 31411163]
55. Westerbeek L, Ploegmakers KJ, de Bruijn GJ, Linn A, van Weert JC, Daams J, et al. Barriers and facilitators influencing medication-related CDSS acceptance according to clinicians: a systematic review. Int J Med Inform 2021 Aug;152:104506 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104506] [Medline: 34091146]
56. Lehoux P, Rivard L, de Oliveira RR, Mörch CM, Alami H. Tools to foster responsibility in digital solutions that operate with or without artificial intelligence: a scoping review for health and innovation policymakers. Int J Med Inform 2023 Feb;170:104933 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104933] [Medline: 36521423]
57. Cresswell KM, Lee L, Mozaffar H, Williams R, Sheikh A, NIHR ePrescribing Programme Team. Sustained user engagement in health information technology: the long road from implementation to system optimization of computerized physician order entry and clinical decision support systems for prescribing in hospitals in England. Health Serv Res 2017 Oct 07;52(5):1928-1957 [FREE Full text] [doi: 10.1111/1475-6773.12581] [Medline: 27714800]
58. Bunger AC, Powell BJ, Robertson HA, MacDowell H, Birken SA, Shea C. Tracking implementation strategies: a description of a practical approach and early findings. Health Res Policy Syst 2017 Feb 23;15(1):15 [FREE Full text] [doi: 10.1186/s12961-017-0175-y] [Medline: 28231801]
59. Trinkley KE, Kahn MG, Bennett TD, Glasgow RE, Haugen H, Kao DP, et al. Integrating the practical robust implementation and sustainability model with best practices in clinical decision support design: implementation science approach. J Med Internet Res 2020 Oct 29;22(10):e19676 [FREE Full text] [doi: 10.2196/19676] [Medline: 33118943]
60. Laka M, Carter D, Milazzo A, Merlin T. Challenges and opportunities in implementing clinical decision support systems (CDSS) at scale: interviews with Australian policymakers. Health Policy Technol 2022 Sep;11(3):100652. [doi: 10.1016/j.hlpt.2022.100652]
61. Polit DF, Beck CT. Generalization in quantitative and qualitative research: myths and strategies. Int J Nurs Stud 2010 Nov;47(11):1451-1458. [doi: 10.1016/j.ijnurstu.2010.06.004] [Medline: 20598692]
62. Abimbola S, Patel B, Peiris D, Patel A, Harris M, Usherwood T, et al. The NASSS framework for ex post theorisation of technology-supported change in healthcare: worked example of the TORPEDO programme. BMC Med 2019 Dec 30;17(1):233 [FREE Full text] [doi: 10.1186/s12916-019-1463-x] [Medline: 31888718]
63. Burchett HE, Kneale D, Blanchard L, Thomas J. When assessing generalisability, focusing on differences in population or setting alone is insufficient. Trials 2020 Mar 20;21(1):286 [FREE Full text] [doi: 10.1186/s13063-020-4178-6] [Medline: 32197623]
64. Hogg HD, Al-Zubaidy M, Technology Enhanced Macular Services Study Reference Group, Talks J, Denniston AK, Kelly CJ, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res 2023 Jan 10;25:e39742 [FREE Full text] [doi: 10.2196/39742] [Medline: 36626192]
65. Panicker R, George A. Adoption of automated clinical decision support system: a recent literature review and a case study. Arch Med Health Sci 2023;11(1):86. [doi: 10.4103/amhs.amhs_257_22]

66. Shulha M, Hovdebo J, D'Souza V, Thibault F, Harmouche R. Integrating explainable machine learning in clinical decision support systems: study involving a modified design thinking approach. JMIR Form Res 2024 Apr 16;8:e50475 [FREE Full text] [doi: 10.2196/50475] [Medline: 38625728]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
NASSS: Non-Adoption, Abandonment, Scale-Up, Spread, and Sustainability
VPN: virtual private network

Edited by C Lovis; submitted 09.05.24; peer-reviewed by A Delaforce, NJ Haque; comments to author 16.06.24; revised version received 09.08.24; accepted 17.08.24; published 17.10.24.

Please cite as:
Fernando M, Abell B, McPhail SM, Tyack Z, Tariq A, Naicker S
Applying the Non-Adoption, Abandonment, Scale-up, Spread, and Sustainability Framework Across Implementation Stages to Identify Key Strategies to Facilitate Clinical Decision Support System Integration Within a Large Metropolitan Health Service: Interview and Focus Group Study
JMIR Med Inform 2024;12:e60402
URL: https://medinform.jmir.org/2024/1/e60402
doi:10.2196/60402
PMID:

Viewpoint

Data Ownership in the AI-Powered Integrative Health Care Landscape

Shuimei Liu[1], JSD; L Raymond Guo[2], MA, MS

[1]School of Juris Master, China University of Political Science and Law, Beijing, China

[2]College of Health and Human Sciences, Northern Illinois University, Dekalb, IL, United States

Corresponding Author:
Shuimei Liu, JSD
School of Juris Master
China University of Political Science and Law
25 Xitucheng Rd
Hai Dian Qu
Beijing, 100088
China
Phone: 1 (734) 358 3970
Email: shuiliu0802@alumni.iu.edu

Abstract

In the rapidly advancing landscape of artificial intelligence (AI) within integrative health care (IHC), the issue of data ownership has become pivotal. This study explores the intricate dynamics of data ownership in the context of IHC and the AI era, presenting the novel Collaborative Healthcare Data Ownership (CHDO) framework. The analysis delves into the multifaceted nature of data ownership, involving patients, providers, researchers, and AI developers, and addresses challenges such as ambiguous consent, attribution of insights, and international inconsistencies. Examining various ownership models, including privatization and communization postulates, as well as distributed access control, data trusts, and blockchain technology, the study assesses their potential and limitations. The proposed CHDO framework emphasizes shared ownership, defined access and control, and transparent governance, providing a promising avenue for responsible and collaborative AI integration in IHC. This comprehensive analysis offers valuable insights into the complex landscape of data ownership in IHC and the AI era, potentially paving the way for ethical and sustainable advancements in data-driven health care.

(JMIR Med Inform 2024;12:e57754) doi:10.2196/57754

KEYWORDS

data ownership; integrative healthcare; artificial intelligence; AI; ownership; data science; governance; consent; privacy; security; access; model; framework; transparency

Introduction

Integrative health care (IHC), which emphasizes a holistic approach to patient well-being [1], increasingly incorporates artificial intelligence (AI) to enhance health care delivery. Within this intersection, questions regarding data ownership become pivotal [2]. The wealth of patient data created during the practice of IHC and processed by AI includes medical history, social determinants of health (SDOHs), lifestyle factors, and treatment responses. These underscore the importance of clarifying ownership and control over this sensitive information. Varied legal and ethical perspectives on health data ownership have emerged. These include individuals retaining certain rights and interests in the data, leading to the dilemma of data ownership in health care [3], especially in the age of AI. For instance, in common law countries like the United States, health care providers typically own physical patient records, not patients. While patients have access rights under privacy or freedom of information laws (and health information laws in some regions), these do not equate to ownership. Similarly, government agencies (eg, disease registry administrators) own patient data stored in their databases. In this context, we focus on a detailed exploration of data ownership in the IHC and AI era, discussing its implications and challenges in the United States.

Many countries grapple with data ownership frameworks, particularly in health care. Some, like those with national electronic health records, prioritize centralized accessibility [4,5]. Others, with federalized health care systems, navigate the division of responsibility between national and regional

governments [6]. This complexity underscores the need to examine the US context. Within the United States itself, data ownership can also vary significantly. Just like the health care example, some regions might have more centralized data access, while others operate with a more fragmented system. This internal diversity emphasizes the importance of exploring such frameworks across the US landscape.

This study aims to evaluate common data ownership models that apply to the IHC setting, scrutinizing their appropriateness based on the purpose of law and ethics. Through this exploration, we intend to contribute to the ongoing dialogue surrounding the responsible integration of AI in IHC. The study sheds light on the significance of patient-centric data governance, providing insights into the legal and ethical implications and considerations for ensuring responsible and transparent AI implementation within integrative health (IH).

IH Model

IHC embraces a holistic approach to wellness and centers on the interconnectedness of the mind, body, and spirit, advocating for comprehensive healing that addresses all facets of an individual's health [1]. IHC often adopts an interdisciplinary modality, fostering collaboration among practitioners from diverse fields to deliver optimal care. This integrative team may include medical doctors, nurses, acupuncturists, chiropractors, and other health care professionals [7]. The IHC model has gained significant momentum in recent years. The United States National Institutes of Health established a dedicated center, the National Center for Complementary and Integrative Health, for IHC research. Additionally, the Veteran's Health Administration's Whole Health System [8] exemplifies its implementation within large health care systems. This growing recognition positions IHC as a key component of the learning health system, aiming to continuously improve patient-centered care through data-driven insights [9].

At the core of IHC is integrating conventional and complementary approaches, forming a coordinated health care ecosystem. This approach emphasizes multimodal interventions, combining conventional health care practices (medication, physical rehabilitation, and psychotherapy) with complementary health approaches (such as acupuncture, yoga, massage, lifestyle coaching, and so on) [1]. These tailored interventions address the entire individual rather than focusing solely on a specific organ system. IHC seeks to deliver comprehensive health care that addresses an individual's well-being by promoting well-coordinated care among diverse providers and institutions.

Incorporating integrative approaches to health and wellness is gaining momentum within health care settings in the United States [10]. Researchers are actively investigating the potential benefits of IHC in diverse contexts, such as pain management for military personnel and veterans, symptom alleviation for patients with cancer and survivors, and programs promoting healthy behaviors [11]. These ongoing investigations seek to shed light on the transformative potential of IHC in enhancing patients' overall well-being, and patient and provider-generated data often appear through the processes.

An IH practice process typically involves the following steps:

1. The patient consults with an IH practitioner to discuss their health concerns and goals.
2. The IH practitioner assesses the patient's health and devises a treatment plan that may incorporate a blend of conventional and IH practices.
3. The IH practitioner may recommend the patient to other providers, such as chiropractors, acupuncturists, or nutritionists.
4. The patient collaborates with these additional providers to develop a more comprehensive treatment plan.
5. The health care providers work together to offer the patient the best possible care, focusing on the specific condition and overall well-being and long-term health outcomes improvement.

For instance, a patient experiencing chronic pain might consult an IH practitioner who crafts a treatment plan encompassing physical therapy, acupuncture, and herbal supplements. The practitioner may also recommend a nutritionist to guide the patient in making dietary changes, supporting their healing, and using food as medicine [12] to ease the mental response to the body changes. The IH model is increasingly embraced in health care, reflecting a growing interest in holistic approaches to address health concerns. The interdisciplinary collaboration among IHC practitioners enables patients to receive comprehensive and effective care. However, it is worth noting that the data generated by patients, providers, and other stakeholders throughout this process can be substantial on a large scale, covering a wide range of data fields and categories (Table 1).

In addition to the 3 main categories mentioned earlier, IH clinical practice may also generate data on the following:

1. Patient-reported outcomes (PROs): Include assessments such as pain scales, mood evaluations, and sleep diaries
2. Clinical laboratory tests: Involve procedures like blood work and imaging studies
3. Patient engagement and adherence to treatment plans: Monitoring how actively and consistently patients participate in and follow their treatment plans
4. Cost-effectiveness of IH interventions: Evaluating IH approaches' economic efficiency and value
5. Adverse events or side effects: Tracking any negative reactions or undesirable effects resulting from IH therapies

IHC represents a paradigm shift in the health care model, moving from a disease-centered approach to a holistic one that prioritizes patient's well-being by incorporating a wider range of data sources, fostering interdisciplinary collaboration, and potentially leveraging AI for personalized insights. IHC's strength lies in its interdisciplinary nature, where various providers (doctors, therapists, and nutritionists) collaborate and continuously learn from patient data. This collaborative approach, while fostering innovation and new interventions, creates a complex data landscape. Since IHC collects a wide range of data—medical history, behaviors, SDOHs, and even patient-developed wellness plans (and often uses AI to analyze it), the ownership of these combined datasets and the potential new knowledge derived from them becomes unclear.

Furthermore, the very process of AI analysis creates additional complexities. As AI identifies patterns and trends within this rich patient data, new knowledge may be generated. Who owns these "derived data"? Does the ownership lie with the patient who provided the original information or with the platform that developed the AI creating the insights?

These uncertainties around data ownership can discourage patients from fully engaging with IH programs, fearing a loss of control over their personal health information. Clear data ownership policies and legal frameworks are essential to navigate these complexities. The health care stakeholders, including patients, caregivers, providers, and health care systems, need to understand how their data are being used, who has access to it, and for what purposes. Only then can IHC unlock its full potential for holistic and personalized medicine and improved health outcomes while ensuring patient trust and privacy. Next, we will dissect the data ownership issues from the practice perspective.

Table 1. Data types generated in the clinical practice.

Category	Examples
Patient information	This type of data can include demographic details, medical history, family history, lifestyle factors, and SDOHs[a].
Treatment details	This encompasses integrative health therapies provided, including acupuncture, massage therapy, herbal medicine, nutritional counseling, and mindfulness practices. It also includes information on these treatments' frequency, duration, and intensity.
Patient outcomes	This involves assessing the impact of integrative health interventions on patient health and well-being. Relevant outcome measures include symptom management, quality of life, functional status, and overall satisfaction with care.

[a]SDOH: social determinant of health.

Data Ownership Issues Rooted in the IHC Practice

Data ownership issues can arise at all stages of health care, including the IHC practice process, from the initial assessment to developing a treatment plan to monitoring the patient's progress.

Assessment Phase

During the assessment phase, the IHC practitioner may collect data from the patient's electronic medical records (EMRs), diagnostic tests, and questionnaires. Like other medical fields, data collected from EMRs and tests are often used to plan treatment options. In the IHC settings, PRO data, or real-world data, have more essential roles than traditional models, as PRO data may assist shared decision-making [13] and are associated with the enrollment of IHC approaches [14]. However, from the data ownership perspective, these data may contain sensitive information about the patient's health condition and lifestyle.

For example, an IHC practitioner may ask patients about their diet, exercise habits, sleep patterns, stress levels, and use of herbs and other supplements that the local policy or law may not regulate. The sensitivity of this information may burden some providers when certain patients have situations that they may not want to share with other stakeholders to avoid potential troubles. For instance, a patient might be hesitant to disclose a history of substance abuse or mental health concerns, fearing discrimination from employers or insurers. This is especially concerning in the context of IHC, where expanded access to data raises the stakes for patients who could face negative consequences for past medical decisions, such as declining to seek treatment or noncompliance. These data are essential for the practitioner to develop an accurate and effective treatment plan for the patient Determining how these data can be used for medical diagnosis, how patients can authorize providers to access them, and how patients can control or participate in data transfer and usage raises significant challenges. However, it is also important to note that these data are sensitive and confidential. The patient has a right to know how their data are being used and to control who has access to it.

Treatment Plan Development

Once the practitioner has assessed the patient's health, they will develop a treatment plan. This plan may include a combination of conventional and IH practices, as discussed above. For example, an IHC practitioner may recommend that a patient with chronic pain take a combination of over-the-counter pain relievers, acupuncture, and yoga. The practitioner may also suggest dietary changes and stress management techniques.

The treatment plan may require the patient to share additional data with the practitioner, such as their treatment response or progress. Many of these data may not be categorized or can be innovative exploratory work that no other medical field has touched, which reflects the whole person-based care model. These data can be generated through new interventions, while the treatment plan outcomes formulated between patients and interdisciplinary providers may not have standard end point criteria. Another challenging part here is understanding the different data levels related to a person's life. Beyond that, much data can be associated with SDOH [15] and other measurements that have yet to be invented or discovered, which could pose future privacy challenges. For instance, the integration of genetic data or continuous environmental monitoring could create more detailed profiles, raising new questions about who has access, how it can be used, and the potential for discrimination [16].

Monitoring Phase

During the monitoring phase, the practitioner will track the patient's progress and adjust the treatment plan as needed. This may involve collecting additional data from the patient, such as their symptoms, quality of life, and satisfaction with the treatment.

These data are essential for the practitioner to ensure that the patient receives the best possible care. However, it is also

important to note that these data are personal, sensitive, and confidential; like the data generated during the treatment phase, they can be collected by multiple providers. It may also contain health data collected from the patient's family member and caregiver, including qualitative data that may have the patient's family member's personal information. Thus, the patient has a right to know how their data are being used and to control who has access to it.

AI in the IH Setting

AI technologies have proven to be powerful tools in many health care fields, leading a shift in health care delivery focusing on the patient and their overall well-being [17]. AI holds immense potential to revolutionize the IHC [2] by enhancing patient outcomes, boosting efficiency, and transforming health care delivery [18]. One of AI's key contributions lies in enabling personalized medicine. By analyzing patient data encompassing medical history, genetics, lifestyle, and environmental factors, AI can assist IHC providers in tailoring treatment plans to individual needs, ensuring optimal care [19-21]. In addition, AI can analyze patient data to identify patterns or markers indicative of underlying health conditions, facilitating earlier and more accurate diagnoses [22]. AI's capabilities extend to improving treatment outcomes by providing real-time analysis of patient data and presenting relevant treatment options, empowering IHC practitioners to make informed decisions [23]. Furthermore, AI can automate routine tasks, such as scheduling appointments and managing patient records, alleviating the administrative burden on practitioners and allowing them to devote more time to patient care [24,25].

Significant breakthroughs in AI and IHC primarily focus on optimizing therapeutic models, including AI-assisted acupuncture [26], traditional Chinese medicine diagnoses through tongue and lip analysis [19], and traditional Chinese medicine syndrome identifications [27]. In addition, the research explores the use of AI in mindfulness practices [28] and medication adherence [29], leveraging EMR and natural language processing to improve the syndrome pattern diagnosis of lung diseases in integrative medicine [19]. While these clinical trials and applications demonstrate promising progress, other endeavors strive to leverage AI's advantages in IHC beyond improving existing models, such as patient education and AI-powered symptom analysis. To summarize, the potential applications of AI in IHC can be followed by the 3 phases of IHC—assessment, treatment, and monitoring.

Assessment Phase

Personalized Risk Assessment

AI can analyze vast patient data, including genetic, lifestyle, and environmental factors, to identify individuals at higher risk of developing chronic diseases or adverse health outcomes [30]. This personalized risk assessment can guide preventive health care measures and early interventions.

Symptom Analysis and Pattern Recognition

AI-powered tools can analyze patient-reported symptoms, medical history, and clinical data to identify patterns and potential underlying conditions [31]. This can help clinicians make more accurate diagnoses and tailor treatment plans accordingly.

Mental Health Assessment and Screening

AI-based chatbots and virtual assistants can engage in conversations with patients to assess their mental health status and identify potential signs of depression, anxiety, or other mental health concerns. This can facilitate early intervention and support [32].

Treatment Phase

Personalized Treatment Planning

AI can analyze patient data and clinical guidelines to generate personalized treatment plans considering individual factors, including genetic predispositions, past treatments, and coexisting conditions [30,33]. This can optimize treatment efficacy and minimize side effects.

Drug Dosage Optimization

AI can analyze patient data and medication profiles to determine the optimal dosage for prescribed medications, reducing the risk of adverse drug reactions and improving treatment outcomes [34,35].

Nutritional Guidance and Meal Planning

AI-powered tools can analyze individual dietary needs, preferences, and health goals to provide personalized nutritional guidance and meal planning recommendations, supporting a healthy lifestyle and disease management [36].

Monitoring Phase

Real-Time Remote Monitoring

AI-enabled wearable devices and sensors can continuously collect patient data, such as vital signs, activity levels, and sleep patterns, and transmit it to health care providers for real-time monitoring [37]. This allows for early detection of potential health concerns and timely interventions.

Predictive Analytics for Disease Exacerbation

AI can analyze patient data and identify patterns that predict potential disease exacerbations or adverse health events, enabling proactive interventions and preventing complications [24,30].

Patient Engagement and Adherence support

AI-powered chatbots and virtual assistants can engage with patients, provide reminders, and offer personalized support to improve medication adherence and lifestyle modifications, enhancing treatment outcomes [38,39].

Data Generated in AI-Incorporated IHC

With the potential applications of AI in IH clinical practice, the types of data generated can expand beyond traditional patient information and treatment details. Here are some examples of data that can be created with AI integration:

1. Patient-generated health data: AI can analyze data from wearable devices, fitness trackers, and patient-reported

symptom trackers to provide insights into patient lifestyle, sleep patterns, and overall health status. These data can be used to personalize treatment plans and monitor patient progress [40,41].
2. Real-time biofeedback data: AI can analyze biofeedback data from devices that measure heart rate variability, skin conductance, and other physiological signals [42]. These data can be used to assess patient stress levels, anxiety, and pain, allowing for real-time adjustments to IH interventions.
3. Genomic and proteomic data: AI can analyze genetic and protein expression data to identify individual variations in drug metabolism, disease susceptibility, and response to IH therapies [43]. This information can tailor treatment plans and predict potential adverse reactions.
4. Predictive analytics: AI can analyze historical data and patient characteristics to predict the likelihood of future health events or treatment outcomes [44]. This information can be used to proactively identify patients at risk and tailor preventive care or treatment plans.
5. AI-driven treatment recommendations: AI can analyze patient data and clinical guidelines to provide personalized treatment recommendations, including the type, dosage, and frequency of IH therapies [30,45]. This can streamline treatment planning and improve patient adherence.
6. AI-powered clinical decision support: AI can provide real-time clinical decision support to health care providers, suggesting appropriate IH therapies based on patient data and evidence-based guidelines. This can enhance clinical decision-making and improve patient care [30,46,47].
7. AI-powered research and clinical trials: AI can facilitate the design, analysis, and interpretation of clinical trials in IH, leading to faster advancements in evidence-based practice [48,49].

Data Ownership Issues in the IHC or AI Setting

IHC and AI data collection raise complex ownership concerns due to several factors. First, individual contributions to these systems are often intertwined, making it unclear who truly "owns" the resulting data. Second, machine-generated data and AI-derived insights introduce new questions about who holds rights to these intellectual creations. Finally, traditional legal frameworks like copyright and privacy struggle to adapt to the unique dynamics of IHC and AI, leaving ownership ambiguous and potentially sparking disputes. Furthermore, there are several other challenges:

1. Ambiguity around consent and control: Individuals interacting with IHC and AI systems may struggle to understand how their data are collected, used, and shared. Consent mechanisms might be opaque, leaving users unsure if they retain any control over their information [50].
2. Difficulty attributing authorship and creativity: As AI systems increasingly contribute to data generation and analysis, it becomes challenging to determine who deserves credit for the resulting insights [51-53]. Is it the human who provided the initial data, the developer who created the AI, or the AI itself?
3. Balancing individual rights with collective benefits: While data collected through IHC and AI can offer societal benefits like improved health care or personalized services, these advantages can come at the cost of individual privacy and autonomy. Striking a balance between these competing interests remains a significant challenge [50,54].
4. Exploitation and bias risks: Unethical actors might exploit data ownership ambiguities to manipulate or discriminate against individuals [55]. Biased algorithms trained on skewed datasets can further perpetuate such injustices.
5. International complexities: Data ownership laws and regulations vary significantly across jurisdictions, creating challenges for global IHC and AI projects [56]. This can lead to confusion and hinder responsible data governance.

Addressing these complex issues requires ongoing collaboration between technology developers, policy makers, legal experts, and the public. We can ensure equitable data ownership and responsible AI development that benefits all through open dialogue and innovative solutions.

Who Has the Right to Own the Data?

Data are more critical than ever in the AI and machine learning era. This is especially true in IHC, where AI or machine learning can be used to develop new treatments, improve patient care, and conduct research. However, the ownership of IHC data is a complex issue. There are several stakeholders who may claim ownership of IHC data.

First, patients may argue that they own their data, including data derived from their medical records and diagnostic tests. Traditionally speaking, patients have limited control over the data once the deidentified data are shared with a broader audience. Due to legislation mandating patient privacy, health care providers, institutions, and governing bodies establish policies and practices that determine patients' ability to access and control their personal health information [57].

Second, IHC providers may argue that they own data derived from their patient interactions, such as data from clinical notes and patient portals. Furthermore, a provider can claim the data ownership if collected from a new intervention or clinical trial. The interaction between patients and providers is also meaningful, as patients can refuse to share the data in any research capacity. While data privacy laws allow patients to control who has their data and how they are used, limited knowledge about existing data holdings creates an informational asymmetry, hindering their ability to fully exercise these rights [58]. Thus, health care providers may proactively opt patients out of broad research programs at the outset to address limited patient control over data reuse [59]. While opting out can prevent future data collection, it does not necessarily erase existing data held by researchers, government agencies, or private entities. This creates a situation where patients may struggle to exercise their data privacy rights due to limited knowledge of which entities hold their data.

Third, researchers may argue that they own data from their research, including data derived from IHC patient data and interventions. Many clinical providers are involved in linear

research activities, which often follow the health care stream from identifying diseases to developing interventions. When considering the application of AI in clinical practice, the researchers may claim ownership of the developed algorithm (through a patent); however, in some circumstances, they may claim ownership of the data being used in the research trials and projects.

Fourth, AI and machine learning developers may argue that they own the data to train AI or machine learning algorithms, including IHC patient data, as some deidentified personal data can be purchased or licensed for research depending on its availability and approval processes. However, access to other sensitive health data, particularly from IHC settings, is often more restricted. These datasets may require specific approval from research ethics committees before access is granted [60,61].

Finally, technology companies may argue that they own data collected through their wearable devices and other health-tracking apps, including IHC patient data. Multiple companies can claim over the same data collected at the same time.

Data Ownership Models Related to IHC

Data ownership models in health care are a complex and evolving topic. There are a variety of different models, each with its advantages and disadvantages. Some of the most common data ownership models in health care are described in this section.

Privatization Postulate

Originating from John Locke's natural rights theory, the privatization postulate in health care data ownership asserts that data are a valuable private asset owned and controlled by individuals or organizations [62,63]. Within the context of IHC and AI collaboration, this model raises concerns about private entities' potential monetization of IHC data. This practice could hinder interdisciplinary collaboration, as apprehensions regarding data protection might limit information sharing among health care providers. Furthermore, developing AI under the privatization postulate may lead to proprietary algorithms, restricting their accessibility and hindering the collective advancement of IHC treatments. The focus on individual or organizational ownership may create barriers to the seamless sharing of insights and innovations, impeding the collaborative potential of IHC in the era of AI.

Despite these challenges, the privatization postulate does offer advantages. It recognizes the economic value of health care data, potentially incentivizing individuals, and organizations to invest in data collection and analysis. This could lead to advancements in personalized health care solutions and tailored treatment plans. However, the drawbacks lie in the potential negative impact on collaboration, data accessibility, and collective progress in the IHC landscape. Striking a balance between recognizing the value of data as an asset and fostering collaborative efforts is crucial for successfully integrating AI in health care under this ownership model.

Communization Postulate

Unlike the privatization postulate, the communization postulate views data as a public good to be shared openly, and data can be used simultaneously and legally [63-65]. In the context of IHC and AI, this model emphasizes collaboration and coordination among interdisciplinary health care providers, researchers, and patients. The concept of shared data platforms and open-source AI aligns to improve resource use and, consequently, patient outcomes. Challenges may arise while this model envisions a more inclusive and collective approach to health care data. Concerns about responsible and ethical AI use and the equitable sharing of benefits necessitate careful consideration. Achieving a balance between open collaboration and addressing ethical concerns becomes imperative to realize the positive outcomes envisioned fully under the communization postulate.

The advantages of the communization postulate lie in its potential to break down data silos, promoting seamless data sharing and accessibility among health care providers. This collaborative environment can foster innovation, leading to more effective IHC treatments. However, the model also raises ethical considerations, such as ensuring data are used responsibly and equitably [63]. Striking this balance is crucial for successfully implementing the communization postulate in IHC, ensuring that the benefits of shared data extend to all stakeholders while upholding ethical standards.

Intellectual Property

Ownership of health data can be both tangible and intangible property [66]. Regarding tangible property rights, the answer is sometimes for sure. For example, it is likely to be said that medical providers, rather than patients, typically own physical medical records in the United States [3]. Meanwhile, health data are intangible information. Relevant stakeholders can own health information based on different types of laws in the field of intellectual property, including patent law, copyright law and copyright in databases, trademark law, and trade secrets. However, such ownership protection must meet various criteria, leading to clarity and incomplete or partial ownership protection of health data [66]. For example, health data should be patent eligible in order to enjoy patent protection [67]. Also, trade secrets or relevant confidential information laws apply to limited types of health data and several questions are still open concerning ownership of health data [68]. Furthermore, conferring ownership rights through intellectual property law is even more complicated in the AI background, such as AI's capacity to claim intellectual property rights, determining contributions between humans and AI, and so forth. Therefore, answering ownership questions concerning AI-generated health data in the context of intellectual property law is highly complex and uncertain.

Next, we will discuss the current data ownership models in health care data that are related to IHC. Understanding the advantages and disadvantages of these models can help to address the rising issues and conflicts in the data ownership of IHC in the AI era.

Distributed Access Control Model

The distributed access control (DAC) [69] model presents a decentralized approach to data ownership, providing individual health care providers or organizations more control over their data, especially in the context of IHC and AI. This model addresses critical concerns related to privacy and security in health care data. By allowing entities to control access to their data through mechanisms such as role-based access control or attribute-based access control, the DAC model aims to safeguard sensitive patient information. However, the emphasis on individual control may lead to challenges in data sharing between providers, resulting in fragmented care and potentially hindering medical research progress within the interdisciplinary landscape of IHC.

While the DAC model helps mitigate privacy and security concerns, it introduces complexities related to data silos and barriers to efficient information exchange. The fragmented nature of data ownership under DAC can hinder collaborative efforts in IHC, limiting the comprehensive understanding of patient health and potentially compromising the effectiveness of treatments. In addition, difficulties in research access may arise, as researchers need permission from each provider or organization that owns the data. Balancing individual control with the need for seamless collaboration and research access becomes essential in implementing the DAC model effectively within IHC, particularly in the era of AI.

Data Trusts

Data trusts, as legal entities holding data for multiple stakeholders, offer an alternative to the DAC and communization models in the landscape of IHC and AI [26]. In this context, data trusts provide increased control over data for stakeholders, promoting responsible and ethical use. Establishing a neutral and trusted third party to manage data helps address data ownership, privacy, and security concerns. However, challenges persist, particularly regarding the complexity and cost of setting up and maintaining data trusts in IHC. The intricacies involved in creating and maintaining these legal entities may not be feasible for all IHC providers, limiting the universal adoption of this model.

Despite the potential advantages, such as improved data sharing and collaboration, data trusts may face difficulties aligning the interests of the trust and stakeholders. Conflicts over data ownership and use could arise, highlighting the importance of establishing clear guidelines and frameworks for the functioning of data trusts in the realm of IHC and AI. In addition, holding data trusts accountable for their actions may prove challenging due to their complex and opaque nature. Striking a balance between the benefits and challenges of data trusts becomes crucial for their effective integration into the IHC landscape, ensuring that they contribute positively to data management and use in the era of AI.

Blockchain Technology

Blockchain technology emerges as a promising solution for secure and transparent data ownership records, particularly in the context of IHC and AI. In health care, blockchain could enhance transparency, accountability, and data sharing, reducing the risk of breaches and other security incidents [70]. However, concerns persist about the scalability and reliability of blockchain technology, mainly when applied to manage large amounts of health care data within the interdisciplinary collaboration inherent in IHC. The relatively new nature of blockchain introduces uncertainties about its widespread implementation and integration into existing health care systems.

The advantages of blockchain in IHC include its potential to create an immutable and tamper-proof ledger, ensuring the integrity of health care data. This can be particularly beneficial in maintaining accurate patient records and supporting collaborative efforts among health care providers. However, the complexity and cost of implementing blockchain technology may pose challenges, especially for smaller IHC providers with limited resources. The lack of a clear regulatory framework adds to the complexity, introducing uncertainties about data ownership and usage within the IHC landscape. Furthermore, data privacy laws, both common law and civil, are often incompatible with public blockchains because anyone can see the information stored on them. This transparency can be a major issue for sensitive data. To help developers navigate this challenge, the National Institute of Standards and Technology has created a flowchart to identify suitable blockchain use cases. Striking a balance between leveraging the benefits of blockchain and addressing the challenges is essential for its successful integration into the evolving landscape of IHC and AI.

Rethinking the Data Ownership Framework of IHC Practice in the AI Era

While IH offers a promising shift toward holistic patient well-being through collaboration and AI-powered insights, the complex data landscape it creates necessitates a robust data-sharing model. The current lack of clarity around ownership of combined datasets and AI-derived knowledge discourages patient participation and hinders progress. Addressing these concerns involves more than just technical solutions; it requires a data ownership model to safeguard data ownership rights.

Further clarification and specification of data ownership from a legal perspective is essential to consider the recommendations and legal action. Property is not the object itself but rather the ability to assert control through aggregated legal interests as McGuire et al [3] pointed out [71]. Ownership encompasses legal rights, including but not limited to possession, access, and control. Despite opposing views on establishing property rights in data for reasons such as public good, lack of market failure, fundamental rights, and transaction costs [64], it is crucial to promptly assign health data ownership. This step is necessary to actively incentivize the high-quality, efficient generation, dissemination, and use of medical data, thus energizing AI development in IH settings.

Simultaneously, the bestowed ownership of health data should be limited to strike a balance among various stakeholders' interests and appropriately reduce transaction costs. Limiting ownership for different entities aligns with the law's purpose of promoting societal progress and maintaining balance. As discussed earlier, patients, IHC providers, researchers, and AI

or machine learning developers all contribute to IHC practice data in the AI era. Co-ownership among these interested parties is essential, but granting full ownership rights to each stakeholder could significantly increase transactional costs, potentially hindering the application of health data in clinical settings, research, and AI fields. Due to the complexity involving numerous rights and interest holders, providing recommendations is challenging.

One suggested framework is to grant patients ownership only over their personal health data. Since health data without personal information have less connection to patients and dealing with numerous patients would dramatically increase transactional costs, restricting ownership to personal health data is a prudent choice. Furthermore, specific legal rules for ownership rights concerning patients' personal data can be explored with reference to Articles 5-22 of the General Data Protection Regulation (GDPR) [72]. Under GDPR, patients can request access to their own data and have some level of control (eg, delete the data or ask not to be shared with third parties). However, they do not have data ownership, which is similar to other privacy frameworks. Another crucial aspect is to establish limitations or exceptions regarding health data ownership for stakeholders like patients, IHC providers, researchers, and AI or machine learning developers. These limitations and exceptions help balance the interests among stakeholders, drawing inspiration from various fair use models in intellectual property law.

Defining ownership rights and limitations for stakeholders in the context of IHC in the AI age is challenging, especially with the emergence of AI. Identifying and allocating data ownership rights, such as determining ownership based on proportion or primary contributors, present ongoing challenges. Consideration must be given to patients' rights and privacy, physicians' efforts, AI practitioners' involvement and time commitment, and public interests. Only through this comprehensive approach can medical development and the balance of various interests be promoted, aligning with the purpose and spirit of legal regulations. While this is just 1 aspect of many possible recommendations, it is crucial as legal clarity on these issues will directly impact the establishment and implementation of other suggestions to address data ownership in IHC practice in the AI era.

In summary, we propose the Collaborative Healthcare Data Ownership (CHDO) framework. The CHDO emphasizes the collective power of data when stakeholders work together. It acknowledges that various parties contribute valuable insights to health care data, from patients to providers, researchers, and AI developers. The CHDO framework addresses this by proposing three key features:

1. Shared ownership: The CHDO framework goes beyond traditional ownership models, where one entity holds exclusive rights. Instead, it advocates for co-ownership, granting stakeholders specific rights and responsibilities over the data based on their contributions. This fosters trust and incentivizes collaboration, unlocking the full potential of data for research, development, and personalized care.
2. Defined access and control: The CHDO framework advocates the establishment of clear guidelines for accessing and using data. Patients retain control over their personal health information, while other stakeholders can access anonymized or aggregated data for approved purposes. This balance ensures individual privacy while enabling collective advancements in health care.
3. Fair and transparent governance: The CHDO framework recognizes the need for robust governance structures. Transparent policies and procedures ensure equitable access, prevent misuse, and address potential conflicts. This fosters trust and accountability among all stakeholders, creating a sustainable environment for data-driven health care progress.

Based on the analysis of various data ownership models in the context of IHC and AI presented in the previous sections, the CDHO co-ownership model offers several advantages over other frameworks. It addresses the concerns raised by the privatization postulate regarding the impact on public interest by granting patients ownership over their personal health data. In addition, it mitigates the ethical issues that may arise from the communization postulate and the stand-alone DAC model, such as privacy and security concerns. Furthermore, the CDHO model avoids limited ownership protections based on patent law, trade secrets or relevant confidential information laws, copyright law, and trademark law in the context of IHC and AI. It is important to note that the other 2 models, data trusts and blockchain technology, are primarily concerned with the management and storage of health care data and can incur significant costs. In these models, conflicts between stakeholders can persist in the absence of clear ownership rights, and there is no clear guidance for resolving them. Whether trusts or blockchains are used, a prerequisite for their establishment is the clear identification of the party with ownership rights to establish the trust or blockchain. The CHDO model effectively addresses this issue by clearly defining and balancing the interests of all parties, ensuring individual privacy and security, and promoting the realization of public interest.

These advantages are particularly significant in the context of IHC and AI. IHC's collaborative nature and focus on patient-centered care necessitate a data ownership model that fosters trust and incentivizes collaboration (shared ownership). Furthermore, the need to balance individual privacy with the potential of data for research and development aligns well with the CHDO framework's defined access and control mechanisms. Finally, the CHDO framework's emphasis on fair and transparent governance is crucial for navigating the complex ethical considerations surrounding AI use in health care. By implementing these principles, the CHDO framework unlocks a new era of collaboration, empowering stakeholders and fostering a healthier future for all, ultimately propelling the health care industry toward a data-driven and AI-integrated future that prioritizes both individual rights and collective progress.

Conclusion

While a universal solution may be elusive, navigating data ownership challenges in AI-powered IHC requires tailoring approaches to specific stakeholder needs and regulations. By ensuring that data use benefits individual patients, adheres to legal frameworks, and contributes to societal well-being, this collaboration can unlock the full potential of AI in IHC while mitigating legal risks. In essence, addressing data ownership in IHC can pave the way for a more streamlined, effective, and ethical integration of AI in health care, ultimately amplifying its benefits for the whole society.

Conflicts of Interest

None declared.

References

1. Gannotta R, Malik S, Chan AY, Urgun K, Hsu F, Vadera S. Integrative medicine as a vital component of patient care. Cureus 2018;10(8):e3098. [doi: 10.7759/cureus.3098] [Medline: 30338174]
2. Cramer H. Artificial intelligence, complementary and integrative medicine: a paradigm shift in health care delivery and research? J Integr Complement Med 2023;29(3):131-133. [doi: 10.1089/jicm.2023.0040] [Medline: 36920088]
3. McGuire AL, Roberts J, Aas S, Evans BJ. Who owns the data in a medical information commons? J Law Med Ethics 2019;47(1):62-69. [doi: 10.1177/1073110519840485] [Medline: 30994077]
4. Khairat S, Coleman GC, Russomagno S, Gotz D. Assessing the status Quo of EHR accessibility, usability, and knowledge dissemination. EGEMS (Wash DC) 2018;6(1):9. [doi: 10.5334/egems.228] [Medline: 30094281]
5. Zharima C, Griffiths F, Goudge J. Exploring the barriers and facilitators to implementing electronic health records in a middle-income country: a qualitative study from South Africa. Front Digit Health 2023;5:1207602 [FREE Full text] [doi: 10.3389/fdgth.2023.1207602] [Medline: 37600481]
6. Alderwick H, Hutchings A, Briggs A, Mays N. The impacts of collaboration between local health care and non-health care organizations and factors shaping how they work: a systematic review of reviews. BMC Public Health 2021;21(1):753 [FREE Full text] [doi: 10.1186/s12889-021-10630-1] [Medline: 33874927]
7. Boon H, Verhoef M, O'Hara D, Findlay B. From parallel practice to integrative health care: a conceptual framework. BMC Health Serv Res 2004;4(1):15 [FREE Full text] [doi: 10.1186/1472-6963-4-15] [Medline: 15230977]
8. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, Committee on Transforming Health Care to Create Whole Health: Strategies to Assess, Scale, and Spread the Whole Person Approach to Health. In: Meisnere M, South-Paul J, Krist AH, editors. Achieving Whole Health: A New Approach for Veterans and the Nation. Washington DC: National Academies Press; 2013:26854.
9. Guo L, Reddy KP, van Iseghem T, Pierce WN. Enhancing data practices for whole health: strategies for a transformative future. Learn Health Syst 2024;8(Suppl 1):e10426 [FREE Full text] [doi: 10.1002/lrh2.10426] [Medline: 38883871]
10. Complementary, alternative, or integrative health: what's in a name? NCCIH. URL: https://www.nccih.nih.gov/health/complementary-alternative-or-integrative-health-whats-in-a-name [accessed 2023-10-08]
11. Reed DE, Bokhour BG, Gaj L, Barker AM, Douglas JH, DeFaccio R, et al. Whole health use and interest across veterans with co-occurring chronic pain and PTSD: an examination of the 18 VA medical center flagship sites. Glob Adv Health Med 2022;11:21649561211065374 [FREE Full text] [doi: 10.1177/21649561211065374] [Medline: 35174004]
12. Downer S, Berkowitz SA, Harlan TS, Olstad DL, Mozaffarian D. Food is medicine: actions to integrate food and nutrition into healthcare. BMJ 2020;369:m2482 [FREE Full text] [doi: 10.1136/bmj.m2482] [Medline: 32601089]
13. Dusek JA, JaKa M, Wallerius S, Fairchild S, Victorson D, Rivard RL, et al. Rationale for routine collection of patient reported outcomes during integrative medicine consultation visits. Complement Ther Med 2018;37:43-49. [doi: 10.1016/j.ctim.2018.01.012] [Medline: 29609936]
14. Elwy AR, Taylor SL, Zhao S, McGowan M, Plumb DN, Westleigh W, et al. Participating in complementary and integrative health approaches is associated with veterans' patient-reported outcomes over time. Med Care 2020;58(Suppl 2 9S):S125-S132. [doi: 10.1097/MLR.0000000000001357] [Medline: 32826782]
15. Misawa J, Ichikawa R, Shibuya A, Maeda Y, Hishiki T, Kondo Y. Social determinants affecting the use of complementary and alternative medicine in Japan: an analysis using the conceptual framework of social determinants of health. PLoS One 2018;13(7):e0200578 [FREE Full text] [doi: 10.1371/journal.pone.0200578] [Medline: 30011303]
16. Clayton EW, Evans BJ, Hazel JW, Rothstein MA. The law of genetic privacy: applications, implications, and limitations. J Law Biosci 2019;6(1):1-36 [FREE Full text] [doi: 10.1093/jlb/lsz007] [Medline: 31666963]
17. Seaver LH, Khushf G, King NM, Matalon DR, Sanghavi K, Vatta M, et al. Points to consider to avoid unfair discrimination and the misuse of genetic information: a statement of the American college of medical genetics and genomics (ACMG). Genet Med 2022;24(3):512-520 [FREE Full text] [doi: 10.1016/j.gim.2021.11.002] [Medline: 35253645]
18. Geng W, Qin X, Yang T, Cong Z, Wang Z, Kong Q, et al. Model-based reasoning of clinical diagnosis in integrative medicine: real-world methodological study of electronic medical records and natural language processing methods. JMIR Med Inform 2020;8(12):e23082 [FREE Full text] [doi: 10.2196/23082] [Medline: 33346740]

19. Wang ZY, Guo ZH. Intelligent Chinese medicine: a new direction approach for integrative medicine in diagnosis and treatment of cardiovascular diseases. Chin J Integr Med 2023;29(7):634-643. [doi: 10.1007/s11655-023-3639-7] [Medline: 37222830]
20. Feng C, Zhou S, Qu Y, Wang Q, Bao S, Li Y, et al. Overview of artificial intelligence applications in Chinese medicine therapy. Evid Based Complement Alternat Med 2021;2021:6678958 [FREE Full text] [doi: 10.1155/2021/6678958] [Medline: 33815559]
21. Ahuja AS, Polascik BW, Doddapaneni D, Byrnes ES, Sridhar J. The digital metaverse: applications in artificial intelligence, medical education, and integrative health. Integr Med Res 2023;12(1):100917 [FREE Full text] [doi: 10.1016/j.imr.2022.100917] [Medline: 36691642]
22. Ng JY, Cramer H, Lee MS, Moher D. Traditional, complementary, and integrative medicine and artificial intelligence: novel opportunities in healthcare. Integr Med Res 2024;13(1):101024 [FREE Full text] [doi: 10.1016/j.imr.2024.101024] [Medline: 38384497]
23. Yelne S, Chaudhary M, Dod K, Sayyad A, Sharma R. Harnessing the power of AI: a comprehensive review of its impact and challenges in nursing science and healthcare. Cureus 2023;15(11):e49252. [doi: 10.7759/cureus.49252] [Medline: 38143615]
24. Lu SC, Delaney CW, Tracy MF, Chi C, Monsen KA. Informatics and artificial intelligence approaches that promote use of integrative health therapies in nursing practice: a scoping review. OBM ICM 2020;5(1):1-22 [FREE Full text]
25. Wang Y, Shi X, Efferth T, Shang D. Artificial intelligence-directed acupuncture: a review. Chin Med 2022;17(1):80 [FREE Full text] [doi: 10.1186/s13020-022-00636-1] [Medline: 35765020]
26. Chu H, Moon S, Park J, Bak S, Ko Y, Youn BY. The use of artificial intelligence in complementary and alternative medicine: a systematic scoping review. Front Pharmacol 2022 Mar 31;13:826044 [FREE Full text] [doi: 10.3389/fphar.2022.826044] [Medline: 35431917]
27. Sturgill R, Martinasek M, Schmidt T, Goyal R. A novel artificial intelligence-powered emotional intelligence and mindfulness app (Ajivar) for the college student population during the COVID-19 pandemic: quantitative questionnaire study. JMIR Form Res 2021;5(1):e25372 [FREE Full text] [doi: 10.2196/25372] [Medline: 33320822]
28. Babel A, Taneja R, Mondello Malvestiti F, Monaco A, Donde S. Artificial intelligence solutions to increase medication adherence in patients with non-communicable diseases. Front Digit Health 2021;3:669869 [FREE Full text] [doi: 10.3389/fdgth.2021.669869] [Medline: 34713142]
29. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
30. Khalifa M, Albadawy M. Artificial intelligence for clinical prediction: exploring key domains and essential functions. Comput Methods Programs Biomed Update 2024;5:100148. [doi: 10.1016/j.cmpbup.2024.100148]
31. Thakkar A, Gupta A, de Sousa A. Artificial intelligence in positive mental health: a narrative review. Front Digit Health 2024;6:1280235 [FREE Full text] [doi: 10.3389/fdgth.2024.1280235] [Medline: 38562663]
32. Krishnan G, Singh S, Pathania M, Gosavi S, Abhishek S, Parchani A, et al. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm. Front Artif Intell 2023;6:1227091 [FREE Full text] [doi: 10.3389/frai.2023.1227091] [Medline: 37705603]
33. Vora LK, Gholap AD, Jetha K, Thakur RRS, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics 2023;15(7):1916 [FREE Full text] [doi: 10.3390/pharmaceutics15071916] [Medline: 37514102]
34. Khan O, Parvez M, Kumari P, Parvez S, Ahmad S. The future of pharmacy: how AI is revolutionizing the industry. Intelligent Pharmacy 2023;1(1):32-40. [doi: 10.1016/j.ipha.2023.04.008]
35. Theodore Armand TP, Nfor KA, Kim JI, Kim HC. Applications of artificial intelligence, machine learning, and deep learning in nutrition: a systematic review. Nutrients 2024;16(7):1073 [FREE Full text] [doi: 10.3390/nu16071073] [Medline: 38613106]
36. Shajari S, Kuruvinashetti K, Komeili A, Sundararaj U. The emergence of AI-based wearable sensors for digital health technology: a review. Sensors (Basel) 2023;23(23):9498 [FREE Full text] [doi: 10.3390/s23239498] [Medline: 38067871]
37. Jadczyk T, Wojakowski W, Tendera M, Henry TD, Egnaczyk G, Shreenivas S. Artificial intelligence can improve patient management at the time of a pandemic: the role of voice technology. J Med Internet Res 2021;23(5):e22959 [FREE Full text] [doi: 10.2196/22959] [Medline: 33999834]
38. Clark M, Bailey S. Chatbots in Health Care: Connecting Patients to Information: Emerging Health Technologies. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health; 2024 Jan. URL: https://www.ncbi.nlm.nih.gov/books/NBK602381/ [accessed 2024-11-14]
39. Khatiwada P, Yang B, Lin JC, Blobel B. Patient-generated health data (PGHD): understanding, requirements, challenges, and existing techniques for data security and privacy. J Pers Med 2024 Mar 03;14(3):282 [FREE Full text] [doi: 10.3390/jpm14030282] [Medline: 38541024]
40. Patient-generated health data. HealthIT.gov. URL: https://www.healthit.gov/topic/scientific-initiatives/pcor/patient-generated-health-data-pghd [accessed 2024-06-21]

41. Alneyadi M, Drissi N, Almeqbaali M, Ouhbi S. Biofeedback-based connected mental health interventions for anxiety: systematic literature review. JMIR Mhealth Uhealth 2021;9(4):e26038 [FREE Full text] [doi: 10.2196/26038] [Medline: 33792548]
42. Vilhekar RS, Rawekar A. Artificial intelligence in genetics. Cureus 2024;16(1):e52035. [doi: 10.7759/cureus.52035] [Medline: 38344556]
43. Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. J Intern Med 2020;288(1):62-81 [FREE Full text] [doi: 10.1111/joim.13030] [Medline: 32128929]
44. Karalis VD. The integration of artificial intelligence into clinical practice. Applied Biosciences 2024 Jan 01;3(1):14-44 [FREE Full text] [doi: 10.3390/applbiosci3010002]
45. Elhaddad M, Hamam S. AI-driven clinical decision support systems: an ongoing pursuit of potential. Cureus 2024;16(4):e57728. [doi: 10.7759/cureus.57728] [Medline: 38711724]
46. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. Health Serv Res Manag Epidemiol 2024;11:23333928241234863 [FREE Full text] [doi: 10.1177/23333928241234863] [Medline: 38449840]
47. Chopra H, Annu, Shin DK, Munjal K, Priyanka, Dhama K, et al. Revolutionizing clinical trials: the role of AI in accelerating medical breakthroughs. Int J Surg 2023;109(12):4211-4220. [doi: 10.1097/JS9.0000000000000705] [Medline: 38259001]
48. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci 2019;40(8):577-591 [FREE Full text] [doi: 10.1016/j.tips.2019.05.005] [Medline: 31326235]
49. Jackson BR, Ye Y, Crawford JM, Becich MJ, Roy S, Botkin JR, et al. The ethics of artificial intelligence in pathology and laboratory medicine: principles and practice. Acad Pathol 2021;8:2374289521990784 [FREE Full text] [doi: 10.1177/2374289521990784] [Medline: 33644301]
50. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 2019;7:e7702 [FREE Full text] [doi: 10.7717/peerj.7702] [Medline: 31592346]
51. Bellaiche L, Shahi R, Turpin MH, Ragnhildstveit A, Sprockett S, Barr N, et al. Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. Cogn Res Princ Implic 2023;8(1):42. [doi: 10.1186/s41235-023-00499-6] [Medline: 37401999]
52. Sadek M, Kallina E, Bohné T, Mougenot C, Calvo RA, Cave S. Challenges of responsible AI in practice: scoping review and recommended actions. AI & Soc 2024:1-18. [doi: 10.1007/s00146-024-01880-9]
53. Lewis B. Navigating health data privacy in AI—balancing ethics and innovation. Loeb & Loeb LLP. 2023 Oct. URL: https://www.loeb.com/en/insights/publications/2023/10/navigating-health-data-privacy-in-ai-balancing-ethics-and-innovation [accessed 2024-06-22]
54. Belenguer L. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI Ethics 2022;2(4):771-787 [FREE Full text] [doi: 10.1007/s43681-022-00138-8] [Medline: 35194591]
55. Morley J, Murphy L, Mishra A, Joshi I, Karpathakis K. Governing data and artificial intelligence for health care: developing an international understanding. JMIR Form Res 2022;6(1):e31623 [FREE Full text] [doi: 10.2196/31623] [Medline: 35099403]
56. Pormeister K. Genetic data and the research exemption: is the GDPR going too far? Int Data Priv Law 2017;7(2):137-146.
57. Edemekong PF, Annamaraju P, Haydel MJ. Health Insurance Portability and Accountability Act. StatPearls. Treasure Island, FL: StatPearls Publishing URL: https://www.ncbi.nlm.nih.gov/books/NBK500019/ [accessed 2024-11-14]
58. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. J Med Ethics 2015 May;41(5):404-409 [FREE Full text] [doi: 10.1136/medethics-2014-102374] [Medline: 25617016]
59. Hodgkins AJ, Mullan J, Mayne DJ, Boyages CS, Bonney A. Australian general practitioners' attitudes to the extraction of research data from electronic health records. Aust J Gen Pract 2020;49(3):145-150 [FREE Full text] [doi: 10.31128/AJGP-07-19-5024] [Medline: 32113209]
60. Allen J, Adams C, Flack F. The role of data custodians in establishing and maintaining social licence for health research. Bioethics 2019;33(4):502-510. [doi: 10.1111/bioe.12549] [Medline: 30657596]
61. Mirchev M, Mircheva I, Kerekovska A. The academic viewpoint on patient data ownership in the context of big data: scoping review. J Med Internet Res 2020;22(8):e22214 [FREE Full text] [doi: 10.2196/22214] [Medline: 32808934]
62. Piasecki J, Cheah PY. Ownership of individual-level health data, data sharing, and data governance. BMC Med Ethics 2022;23(1):104 [FREE Full text] [doi: 10.1186/s12910-022-00848-y] [Medline: 36309719]
63. Hall MA, Schulman KA. Ownership of medical information. JAMA 2009;301(12):1282-1284. [doi: 10.1001/jama.2009.389] [Medline: 19318657]
64. Thouvenin F, Tamò-Larrieux A. Data Ownership and Data Access Rights: Meaningful Tools for Promoting the European Digital Single Market?. Cambridge, England: Cambridge University Press; 2021:316-339.
65. Liddell K, Simon D, Lucassen A. Patient data ownership: who owns your health? J Law Biosci 2021;8(2):lsab023 [FREE Full text] [doi: 10.1093/jlb/lsab023] [Medline: 34611493]
66. Association for Molecular Pathology v. Myriad Genetics, Inc. 569 U.S 576. Justia U.S. Supreme Court. 2013. URL: https://supreme.justia.com/cases/federal/us/569/576/ [accessed 2024-11-11]

67. Bovenberg JA, Almeida M. Patients v. Myriad or the GDPR access right v. the EU database right. Eur J Hum Genet 2019;27(2):211-215. [doi: 10.1038/s41431-018-0258-4] [Medline: 30262921]
68. Hu VC, Richard Kuhn D, Ferraiolo DF. Access control for emerging distributed systems. Computer (Long Beach Calif) 2018;51:10.1109/MC.2018.3971347. [doi: 10.1109/MC.2018.3971347] [Medline: 31092952]
69. Chen HS, Jarrell JT, Carpenter KA, Cohen DS, Huang X. Blockchain in healthcare: a patient-centered model. Biomed J Sci Tech Res 2019;20(3):15017-15022. [Medline: 31565696]
70. Yaga D, Mell P, Roby N, Scarfone K. Blockchain technology overview. National institute of Standards and Technology. URL: https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8202.pdf [accessed 2024-11-05]
71. Art. 94 GDPR: repeal of directive 95/46/EC. Intersoft Consulting. URL: https://gdpr-info.eu/art-94-gdpr/ [accessed 2024-02-26]
72. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artificial Intelligence in Healthcare 2020:25-60.

Abbreviations

AI: artificial intelligence
CHDO: Collaborative Healthcare Data Ownership
DAC: distributed access control
EMR: electronic medical record
GDPR: General Data Protection Regulation
IH: integrative health
IHC: integrative health care
PRO: patient-reported outcomes
SDOH: social determinant of health

Edited by A Benis; submitted 27.02.24; peer-reviewed by S McLennan, J Scheibner; comments to author 26.03.24; revised version received 22.06.24; accepted 24.10.24; published 19.11.24.

Please cite as:
Liu S, Guo LR
Data Ownership in the AI-Powered Integrative Health Care Landscape
JMIR Med Inform 2024;12:e57754
URL: https://medinform.jmir.org/2024/1/e57754
doi:10.2196/57754
PMID:

Case Report

Standardizing Corneal Transplantation Records Using openEHR: Case Study

Diana Ferreira[1], MSc; Cristiana Neto[1*], MSc; Francini Hak[1*], MSc; António Abelha[1*], PhD; Manuel Santos[1*], PhD; José Machado[1*], PhD

ALGORITMI Research Center, Intelligent Systems Associate Laboratory (LASI), School of Engineering, University of Minho, Guimarães, Portugal
[*]these authors contributed equally

Corresponding Author:
José Machado, PhD
ALGORITMI Research Center
Intelligent Systems Associate Laboratory (LASI)
School of Engineering, University of Minho
Campus of Azurém
Guimarães, 4800-058
Portugal
Phone: 351 253604438
Email: jmac@di.uminho.pt

Abstract

Background: Corneal transplantation, also known as keratoplasty, is a widely performed surgical procedure that aims to restore vision in patients with corneal damage. The success of corneal transplantation relies on the accurate and timely management of patient information, which can be enhanced using electronic health records (EHRs). However, conventional EHRs are often fragmented and lack standardization, leading to difficulties in information access and sharing, increased medical errors, and decreased patient safety. In the wake of these problems, there is a growing demand for standardized EHRs that can ensure the accuracy and consistency of patient data across health care organizations.

Objective: This paper proposes the use of openEHR structures for standardizing corneal transplantation records. The main objective of this research was to improve the quality and interoperability of EHRs in corneal transplantation, making it easier for health care providers to capture, share, and analyze clinical information.

Methods: A series of sequential steps were carried out in this study to implement standardized clinical records using openEHR specifications. These specifications furnish a methodical approach that ascertains the development of high-quality clinical records. In broad terms, the methodology followed encompasses the conduction of meetings with health care professionals and the modeling of archetypes, templates, forms, decision rules, and work plans.

Results: This research resulted in a tailored solution that streamlines health care delivery and meets the needs of medical professionals involved in the corneal transplantation process while seamlessly aligning with contemporary clinical practices. The proposed solution culminated in the successful integration within a Portuguese hospital of 3 key components of openEHR specifications: forms, Decision Logic Modules, and Work Plans. A statistical analysis of data collected from May 1, 2022, to March 31, 2023, allowed for the perception of the use of the new technologies within the corneal transplantation workflow. Despite the completion rate being only 63.9% (530/830), which can be explained by external factors such as patient health and availability of donor organs, there was an overall improvement in terms of task control and follow-up of the patients' clinical process.

Conclusions: This study shows that the adoption of openEHR structures represents a significant step forward in the standardization and optimization of corneal transplantation records. It offers a detailed demonstration of how to implement openEHR specifications and highlights the different advantages of standardizing EHRs in the field of corneal transplantation. Furthermore, it serves as a valuable reference for researchers and practitioners who are interested in advancing and improving the exploitation of EHRs in health care.

(JMIR Med Inform 2024;12:e48407) doi:10.2196/48407

KEYWORDS

electronic health record; EHR; corneal transplantation; keratoplasty; openEHR; data representation; data exchange; templates; archetypes; forms; standardization

Introduction

Background

The eye is a highly evolved and complex sensory organ possessed by a wide range of species, enabling organisms to perceive and interpret visual information from their surroundings. Vision is one of the most valuable senses for humans and plays a critical role in every facet of an individual's life [1]. The sense of vision is the result of an intricate interaction among the eyes, the brain, and the nervous system [2].

Visual impairment occurs when a pathological condition disrupts the visual system and one or more of its associated functions [1]. Blindness is a major public health issue, particularly in low- and middle-income countries where access to health care and resources is scarce [3,4]. The loss of sight can severely impact an individual's daily life, hindering their ability to perform routine tasks, interact with others, and preserve their independence [5]. On many occasions, blindness can also lead to social isolation, depression, and decreased quality of life [1,5].

In 2019, the World Health Organization estimated in the World Report on Vision that there were approximately 2.2 billion people worldwide with vision impairment or blindness [1]. The prevalence of visual disability is alarming and a source of growing global concern.

The apprehension surrounding blindness is rooted not only in the physical limitations it imposes but also in its social and economic consequences [3]. From an economic point of view, the loss of sight can cause reduced workforce participation, decreased productivity, and increased health expenses [5]. Consequently, the loss of income and the higher health care costs can drain governments with additional financial pressures, exacerbating poverty and slowing economic growth [1]. Hence, the impacts of blindness are far reaching, affecting not only the individual but also their families, communities, and society as a whole.

Without more assertive measures, the escalating demand for eye care services worldwide is projected to persist and intensify in the next few decades, posing a meaningful challenge to the health care industry and requiring innovative solutions to meet the increasing pressure for quality eye care services [6,7].

The eye is a complex organ composed of several structures that work together in the perception of the world in all its lights, colors, shapes, and movements [2]. One of the most vital structures of the visual system is the cornea, which is the clear outermost layer located at the front of the eye [8]. A transparent cornea acts as a clear window to allow light to enter the eye and reach the retina, a layer of neural tissue at the back of the eye where light is converted into electrical signals and transmitted to the brain for interpretation as visual information [2]. For an individual to have clear vision, the cornea must be transparent and free of any obstructions, such as scars or opacities, to allow light to cross the eye and access the retina [8,9].

The main causes of visual impairment are cataract, glaucoma, macular degeneration, detached retina, diabetic retinopathy, and retrolental fibroplasia [3,7,10,11]. Some of these eye conditions, such as cataract, diabetic retinopathy, and retrolental fibroplasia, can negatively impact the clarity of the cornea and lead to vision impairment [2,9]. According to the World Health Organization, corneal opacities are the fourth leading cause of blindness on a global scale [9].

In many cases, visual rehabilitation is possible with corneal transplantation [4,8]. Corneal blindness can be effectively reversed with a cornea transplant from a healthy donor. Because the cornea lacks blood vessels, the risk of graft rejection is significantly reduced, making corneal transplantation one of the most successful forms of organ transplantation in the human body [8]. Corneal transplantation, also known as keratoplasty, is a surgical procedure that replaces a damaged or diseased cornea with a healthy one to restore vision and improve the quality of life of patients [4].

The success of corneal transplantation heavily relies on the accurate and timely management of patient data, including preoperative evaluation, surgical planning, and postoperative care. Electronic health records (EHRs) have evolved as indispensable means for handling clinical information, providing a centralized repository for patient records, and facilitating communication between health care providers [12,13]. The appropriate use of EHRs can substantially bolster the positive outcomes of corneal transplant surgeries, culminating in improved patient results and a more seamless and efficient care journey. By minimizing errors and inconsistencies in patient data management, EHRs can further diminish the likelihood of unfavorable events, ultimately improving the success of corneal transplants.

Despite their widespread use, traditional EHRs are prone to lack of standardization and consistency [12]. As a consequence, inconsistency and fragmentation plague the management and documentation of corneal transplantation records, causing difficulties in accessing and sharing information, increased medical errors, and decreased patient safety. Furthermore, the lack of standardization and organization creates an obstacle for health care professionals to access complete and accurate information about patients who are due to undergo or have undergone corneal transplantation, resulting in inefficiencies and suboptimal patient care.

To address these issues, there is a growing need for standardized EHRs that can ensure the accuracy and consistency of patient data across health care organizations and locations [14,15] to ultimately improve the management and documentation processes of corneal transplantation records.

One promising approach for standardizing EHRs is the use of openEHR structures [16,17]. openEHR is an open-source

standard that provides a set of specifications and tools to support the creation of interoperable data structures and the long-term management of health data [18]. A foundational paradigm on which the openEHR framework is based is the 2-level modeling, separating domain semantics from software. Under the model-driven approach, a stable reference information model constitutes the first level of modeling, whereas formal definitions of clinical content in the form of archetypes and templates constitute the second level. Overall, the adaptability, flexibility, and scalability of openEHR's modular methodology provide a powerful solution to the challenges facing the health care industry, and it is an ideal approach for health care systems of all sizes [16-19].

Objectives

Hence, this study sought to explore the implementation of openEHR structures as a means of standardizing records in the field of corneal transplantation. In addition, this manuscript expounds upon the potential benefits that may arise from the use of openEHR specifications for standardizing corneal transplantation records, including improved data consistency and completeness and increased data accessibility and sharing, alongside the mitigation of errors and inefficiencies in data management. This paper will also delve into the challenges and limitations of implementing openEHR in the context of corneal transplantation.

Through the evaluation of the potential benefits and hurdles of using openEHR specifications, this study provides an informative resource and a valuable reference for researchers and practitioners interested in improving the use of EHRs in health care, extending beyond the field of ophthalmology and encompassing other medical disciplines as well.

One of this research's objectives was to contribute to the ongoing efforts to improve the quality and safety of patient care through the disclosure of insights into the potential of using openEHR structures for the advancement of EHRs in health care, particularly in the field of corneal transplantation.

As health care continues to evolve, the authors believe that the standardization of clinical records using openEHR structures holds great potential for ensuring that patients receive safe, effective, and high-quality care.

Methods

Overview

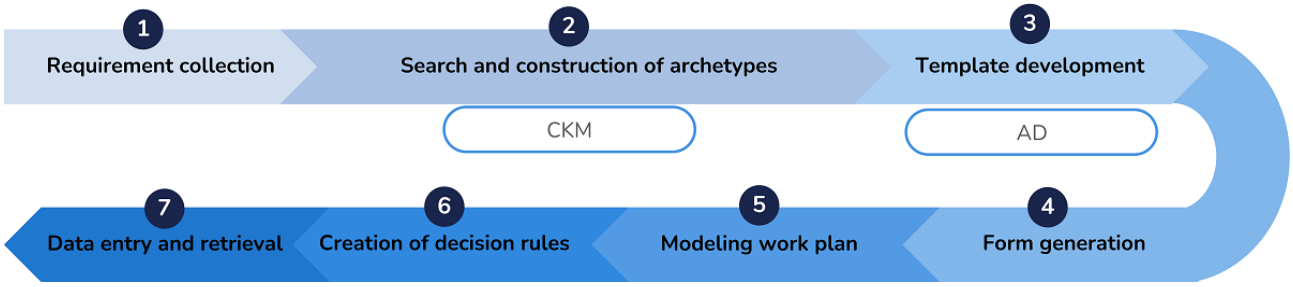
In this study, a specific methodology was defined that outlines the sequential steps involved in transforming the corneal transplantation records of a Portuguese hospital using the openEHR specifications.

In contrast to traditional approaches, openEHR's advanced modular approach separates data from applications and services, providing a level of flexibility that is unmatched in the industry [20]. This approach enables an easier adaptation to constantly evolving requirements, technological changes, health care policies, and other external factors. Moreover, by separating the data layer from the application layer, openEHR can integrate a wide range of health-related data from various sources regardless of the format in which they are stored [20].

By using archetypes and templates as a means of ensuring the consistency and accuracy of clinical data, openEHR is more supple compared to conventional approaches to clinical documentation, which often rely on free-text entries that can be arduous to comprehend and interpret across different patients and clinical settings [20]. The use of archetypes and templates allows health care institutions to minimize variability and fragmentation, which contrasts with the current uncoordinated methods. It provides a framework that allows health care professionals to customize patient-specific information in a more dynamic and structured way [19]. In addition, it ensures that clinical data remain consistently structured and semantically interoperable, leading to more precise interpretation and sharing of clinical data among different clinical systems.

In this sense, the methodology adopted in this study uses openEHR to contribute to the ongoing efforts to improve the quality and safety of patient care through the implementation of standardized EHRs. Figure 1 represents in a simplified way the different stages of the methodology carried out in this study.

Figure 1. Methodological approach followed in the standardization of corneal transplantation electronic health records using openEHR specifications. AD: Archetype Designer; CKM: Clinical Knowledge Manager.



Initially, as the objective was to develop a case study on the transformation of the corneal transplantation records of a particular hospital institution, a direct line of communication with the medical team was deemed indispensable to conduct an in-depth inquiry on the necessary requirements.

After collecting the requirements, the modeling of the openEHR structures that will support the registration and management of corneal transplantation records and the timely execution of associated tasks of corneal transplantation records began, which includes archetypes, templates, forms, decision rules, and work plans.

In the next subsections, the characterization of the work developed in each of the stages that compounds the methodology represented in Figure 1 will be described in detail.

Requirement Collection

Requirement collection is a crucial step in the development of health care systems and applications as it lays the foundation to ensure that the needs and expectations of all parties involved are understood and incorporated into the final solution.

Accordingly, the first step of the methodology started with stakeholder identification, including health care professionals and IT personnel. Overall, the process of collecting requirements for enhancing the corneal transplantation records using openEHR was a collaborative effort between medical staff, the IT personnel of the hospital's information systems department, and the developers.

After identifying a work group, a series of meetings were organized to initiate the requirement-gathering phase. This stage entailed the active participation of stakeholders, who served as the primary source of knowledge for modeling, guiding the design and development of the openEHR structures. During these collaborative meetings, a thorough analysis of the current documentation processes and data management practices was conducted to identify areas for improvement. The meetings were structured to stimulate an active feedback process from the stakeholders regarding existing workflows, specific needs and preferences, and any identified pain points and suggestions for improvement.

Over the course of these conversations, it became apparent that the information system that the hospital used for the management of corneal transplantation records was plagued by significant shortcomings, including the possibility of errors in data entry, loss of information, difficulty sharing data between health care providers, and lack of standardization.

Therefore, the main goal was to address these issues and implement a more reliable and comprehensive information management system that used openEHR structures to mitigate the problems acknowledged. Through a collaborative effort, the stakeholders were able to delimit the scope and define the requirements and use cases of the clinical domain to be modeled.

The work group identified 5 key events in the corneal transplantation process in which data-recording actions could take place. These moments were carefully analyzed to ensure that all necessary data are captured with the highest level of accuracy and consistency. Textbox 1 provides a description of each key event.

Textbox 1. Description of the 5 key events identified within the corneal transplantation process.

Key event and description

- Corneal transplantation proposal: to insert the data concerning the corneal transplantation proposal, such as type of transplant, laterality, diagnosis, priority, and motive, among others
- Contact for corneal transplantation: to record the 3 possible contact attempts for corneal transplantation, including information such as phone number, date of contact, result, reason, and date of next contact
- Schedule anesthesia consultation: to enter data regarding the scheduling of the anesthesia consultation; it contains the requesting service, the date of the appointment, the motive, and additional information
- Perform anesthesia consultation: to record the result of the anesthesia consultation, the executing service, the execution date of the appointment, and observations
- Manage suspended proposal: to register the result of the decision regarding suspended corneal transplantation proposals, either to reinstate to the list or to abandon, and the corresponding reason

Modeling and Development of Archetypes, Templates, and Forms

Archetypes, templates, and forms are interconnected concepts in openEHR that work together to ensure that clinical data are consistently collected, stored in a structured and meaningful way, and retrievable in a usable format [18,21]. By providing a standard, flexible, and scalable manner to manage clinical data, they also serve as a foundation for data sharing and exchange among different health systems and organizations, thereby promoting interoperability [22].

Archetypes are reusable, modular building blocks that describe the structure and content of clinical data elements. They define the data types, units, constraints, and other properties of specific clinical concepts, such as patient demographics, test results, and medication information [19,22].

In turn, templates are collections of archetypes that define a specific clinical record, such as a patient's progress notes, a medication prescription, or a diagnostic test result. Templates provide standardized structure and content for clinical records, ensuring that data are collected consistently in a usable and meaningful format [19].

Forms, on the other hand, provide a user-friendly interface based on templates for the input and retrieval of clinical data, supporting the entry and display of structured, semistructured, and unstructured data [23]. Forms can be customized and configured to meet the unique requirements of various clinical scenarios.

In summary, archetypes provide the building blocks for clinical data structures, templates define the standard structures for specific clinical records, and forms serve as a user-friendly interface for the input and retrieval of clinical data.

The development process of the openEHR forms in which data can be introduced in some tasks of the corneal transplantation workflow involves transforming archetypes into templates and

later into forms. For representation purposes, and to simplify the demonstration of the development process, from now on the description will focus on illustrating the development of the openEHR structures, archetypes, templates, and forms related to the corneal transplantation proposal.

As previously stated, the first step involved meetings with domain experts to define the scope of the clinical domain to be modeled and collect data requirements as specified by the archetype modeling methodology [24]. The main stakeholders involved in this process were health care professionals who were familiar with openEHR, the archetype development process, and clinical terminologies.

After determining the scope of the modeling process and identifying the clinical concepts and information elements involved, the second step entailed searching the Clinical Knowledge Manager for existing archetypes that fit the scope of the modeling scenario under consideration. The Clinical Knowledge Manager is an openEHR community pillar that enables worldwide governance of domain knowledge artifacts as well as collaborative development, management, and publishing [25]. It is an open-source library of openEHR archetypes and templates that lays the foundation for both semantic and syntactic interoperability [26].

Some archetypes were used directly, whereas others did not fully represent the data elements and had to be adapted through specialization. When no corresponding archetypes existed, new ones were created. The openEHR Reference Model defines 4 major categories of archetypes: COMPOSITION, SECTION, ENTRY, and CLUSTER. A COMPOSITION is a container class, whereas a SECTION is an organizing class, both of which contain ENTRY objects [27]. The ENTRY class is further specialized into ADMIN_ENTRY, OBSERVATION, EVALUATION, INSTRUCTION, and ACTION subclasses, of which the latter 4 are kinds of CARE_ENTRY. CLUSTERS are reusable archetypes that can be used within any ENTRY or CLUSTER [16].

For the corneal transplantation proposal, 2 archetypes were found suitable for use, namely, the service request (Figure 2), which is part of openEHR's INSTRUCTION subclass of the ENTRY class, and the anatomical location (Figure 3), which is part of openEHR's CLUSTER class.

In addition, a third archetype was created to contemplate some procedure aspects, namely, the risk, complexity, number of previous surgeries, and need for anesthesia consultation. This archetype was named "Eye Surgery Details" and belongs to the CLUSTER class. Figure 4 depicts the archetype's mind map.

At the end of this step, information concerning applicable data constraints, such as data types, cardinality, occurrences, and specific data values (eg, terminologies for coded values and ranges for numerical values), was stipulated for each archetype.

It is worth mentioning the use of Systematized Nomenclature of Medicine–Clinical Terms terminologies for mapping the coded values of the "Laterality" item belonging to the "Anatomical Location" archetype. Local terms were used for the remaining coded text items. Table 1 provides a description of the coded values used for each coded text item.

After the discovery and development of the required archetypes, the Archetype Designer tool was used to assemble and constrain the archetypes into a template that represents the requirements of the corneal transplantation proposal. A template of the COMPOSITION type was created using the "Request for Service" archetype and named "Corneal Transplantation Proposal." To begin, the "Service Request" archetype was incorporated into the content attribute. The template was then modified to remove items that were irrelevant to the clinical context being modeled, such as "Service Type," "Order Detail," and "Intent." Both the "Anatomical Location" and "Eye Surgery Details" archetypes were imported into the "Specific Details" cluster. The items "Aspect," "Anatomical Line," and "Description" of the "Anatomical Location" archetype were also excluded from the template. Subsequently, some items were assigned specific default values, the content of which can be found in Table 2.

Figure 2. Mind map view of the “Service Request” archetype. EHR: electronic health record.

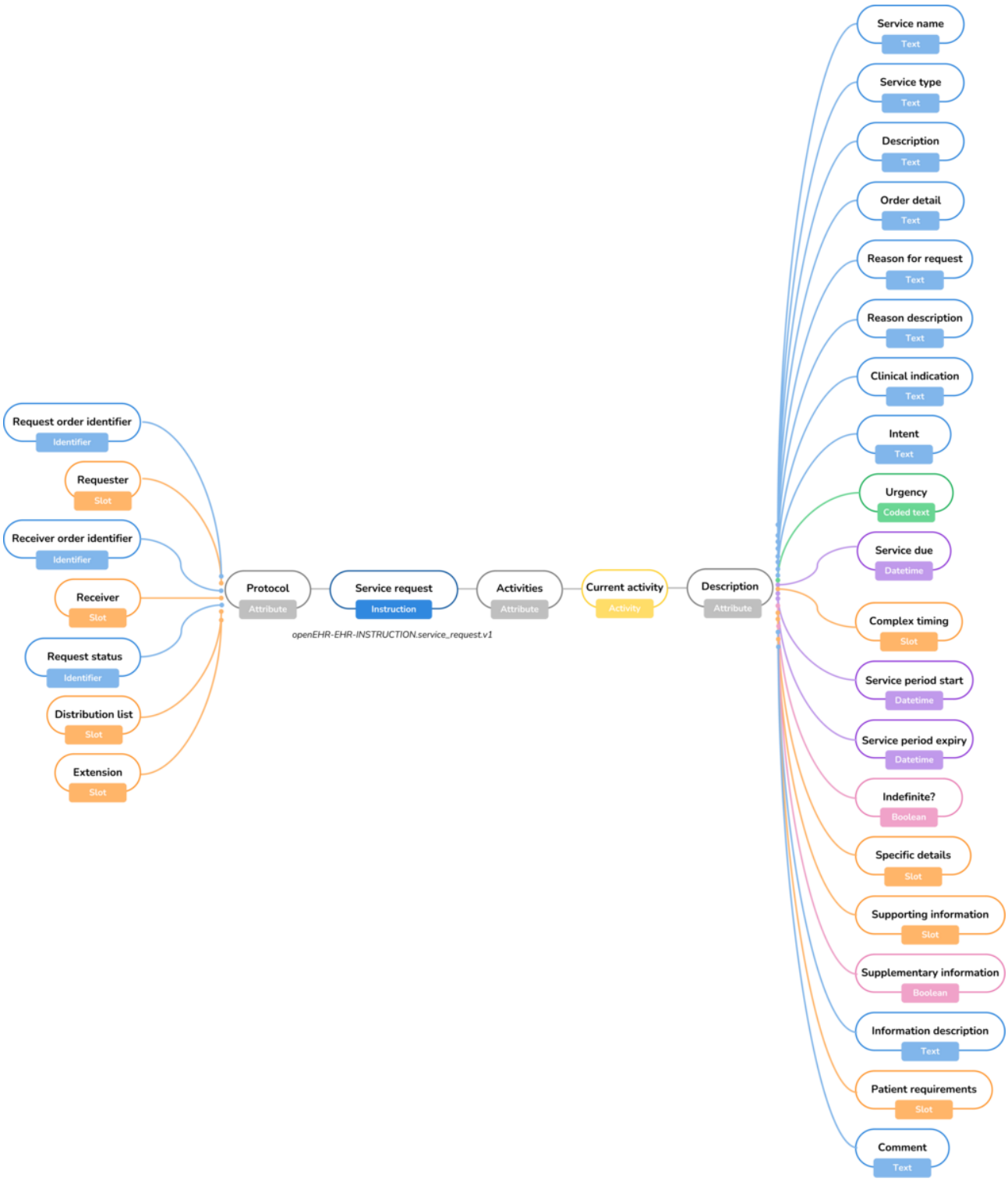


Figure 3. Mind map view of the “Anatomical Location” archetype. EHR: electronic health record.

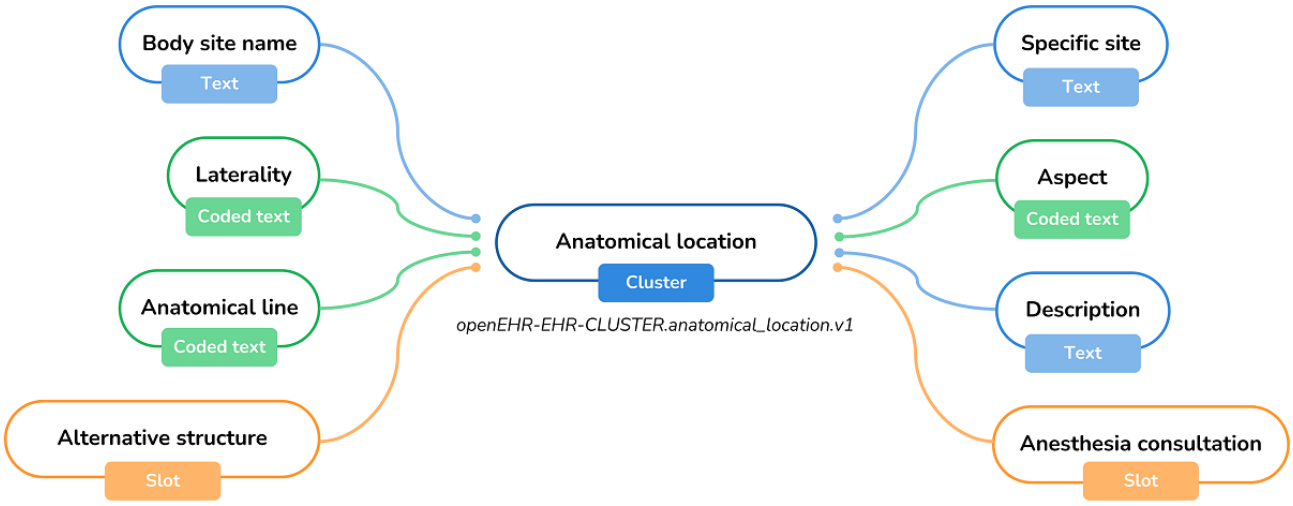


Figure 4. Mind map view of the “Eye Surgery Details” archetype. EHR: electronic health record.

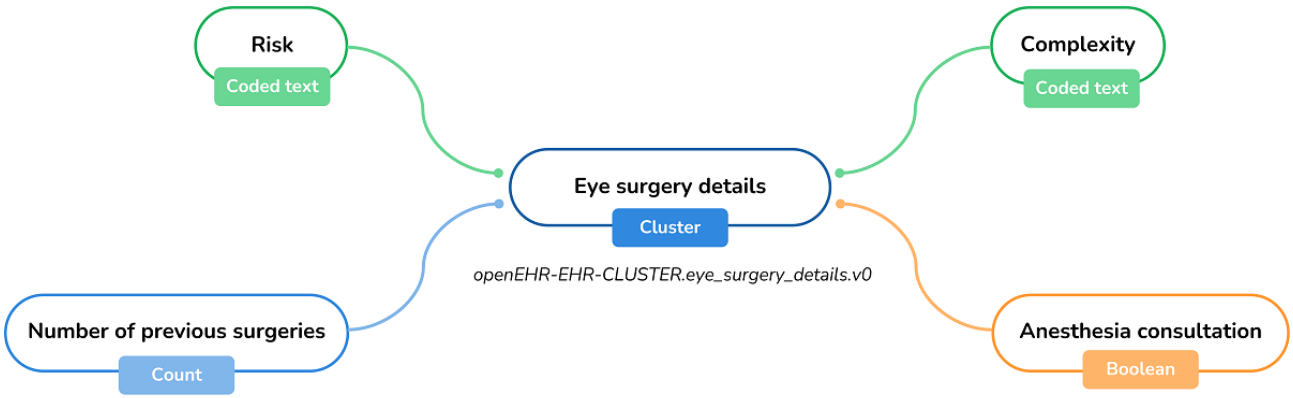


Table 1. Description of the coded values assigned to each coded text item.

Archetype	Item	Code	Value
Anatomical location	Laterality	SNOMED-CT[a]::362503005	Left eye
Anatomical location	Laterality	SNOMED-CT::362502000	Right eye
Anatomical location	Laterality	SNOMED-CT::362508001	Both eyes
Eye surgery details	Risk	at0004	High
Eye surgery details	Risk	at0005	Moderate
Eye surgery details	Risk	at0006	Low
Eye surgery details	Complexity	at0007	High
Eye surgery details	Complexity	at0008	Moderate
Eye surgery details	Complexity	at0009	Low
Service request	Diagnosis	H16319	Corneal abscess
Service request	Diagnosis	H1830	Corneal membrane alterations
Service request	Diagnosis	H52219	Irregular astigmatism
Service request	Diagnosis	H18739	Descemetocele
Service request	Diagnosis	H18519	Corneal endothelial dystrophy
Service request	Diagnosis	H18719	Corneal ectasia
Service request	Diagnosis	H1820	Corneal edema
Service request	Diagnosis	T868499	Corneal graft (complication)
Service request	Type of transplant	08R83KZ_tt_d	Total transplantation (right eye)
Service request	Type of transplant	08R83KZ_tt_e	Total transplantation (left eye)
Service request	Type of transplant	08R83KZ_dalk_d	Anterior transplantation—DALK[b] (right eye)
Service request	Type of transplant	08R83KZ_dalk_e	Anterior transplantation—DALK (left eye)
Service request	Type of transplant	08R83KZ_dsaek_d	Anterior transplantation—DSAEK[c] (right eye)
Service request	Type of transplant	08R83KZ_dsaek_e	Anterior transplantation—DSAEK (left eye)
Service request	Type of transplant	08R83KZ_dmek_d	Anterior transplantation—DMEK[d] (right eye)
Service request	Type of transplant	08R83KZ_dmek_e	Anterior transplantation—DMEK (left eye)
Service request	Urgency	at0136	Emergency
Service request	Urgency	at0137	Urgent
Service request	Urgency	at0138	Routine

[a]SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms.

[b]DALK: deep anterior lamellar keratoplasty.

[c]DSAEK: Descemet stripping endothelial keratoplasty.

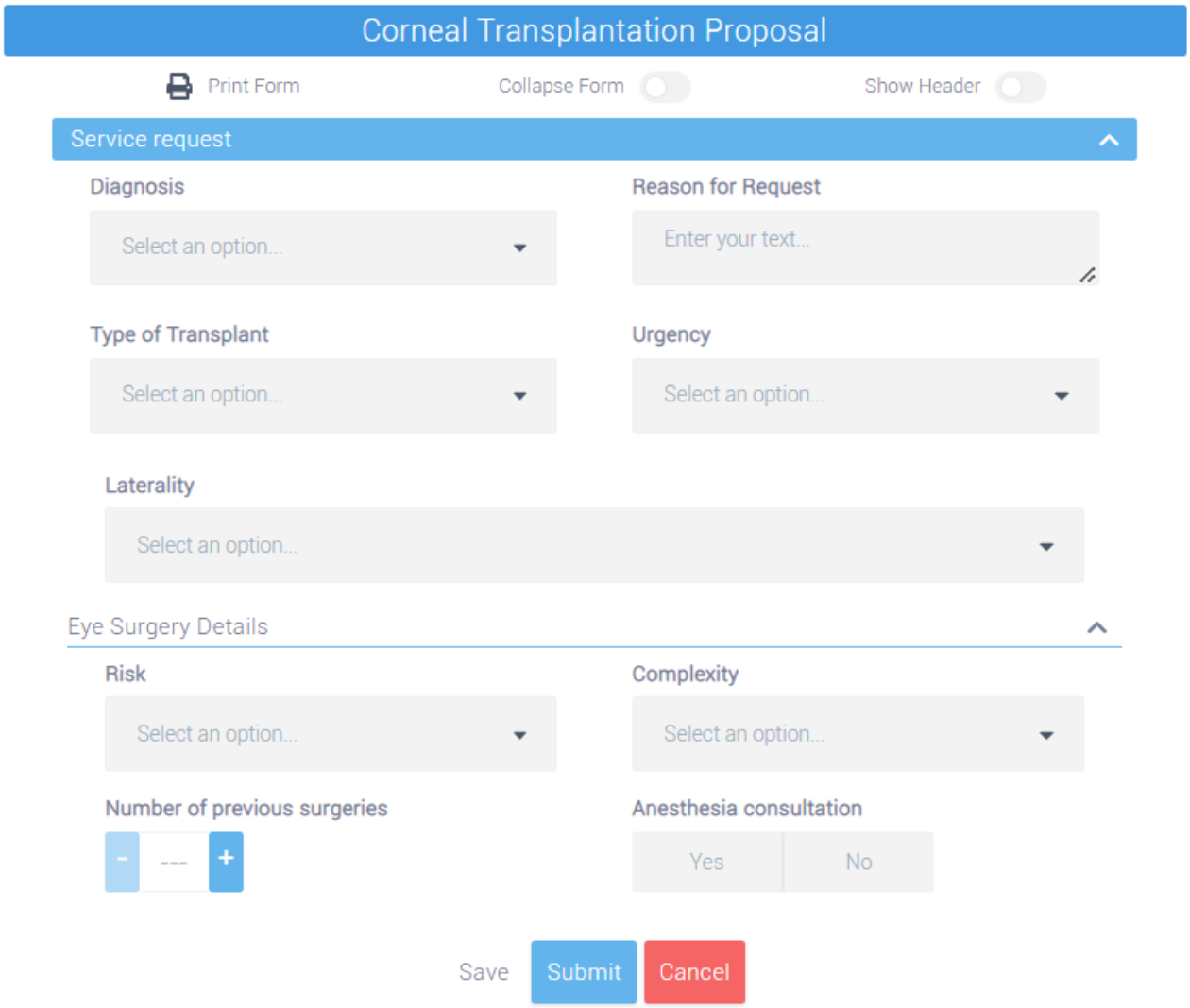
[d]DMEK: Descemet membrane endothelial keratoplasty.

Table 2. Description of the default values assigned.

Archetype	Item	Default value
Service request	Service name	Corneal transplantation
Anatomical location	Body site name	Eye
Anatomical location	Specific site	Cornea

To facilitate interpretation and manipulation, the template was exported in the Operational Template structure and later converted into the JSON Data Template structure. Finally, the JSON Data Template was injected into the Form Builder tool to format the user interface form, which can be consulted in Figure 5.

Figure 5. Graphic representation of the user interface form generated from the "Corneal Transplantation Proposal" template.



At the end of this stage, 5 forms were created. To simplify the identification of each form, Textbox 2 assigns an identification label to each form. Following that, each of these forms will be associated with specific tasks in the corneal transplantation workflow, acting as storage schemes for different patient tasks. A more detailed description of the deployment will be provided later in this paper.

Textbox 2. Forms developed for the corneal transplantation workflow.

ID and form

- F1: corneal transplantation proposal
- F2: contact for corneal transplantation
- F3: schedule anesthesia consultation
- F4: perform anesthesia consultation
- F5: manage suspended proposal

Work Plan Modeling

In recent years, a major extension has been incorporated into the openEHR specifications for addressing requirements in the area of clinical process automation, known as Task Planning [18]. Task Planning allows for the management of standardized Task Plans (TPs) and clinical workflows. The main concept of Task Planning is centered on a plan, or set of plans, that is devised to accomplish a specific goal and pertains to an active subject [28].

Within the Task Planning specification, in terms of conceptual elements, the formal concept at the highest level of hierarchy is the Work Plan (WP), which encompasses one or more TPs [29]. A WP defines a series of tasks that need to be performed in a specific order to achieve a clinical goal with respect to a

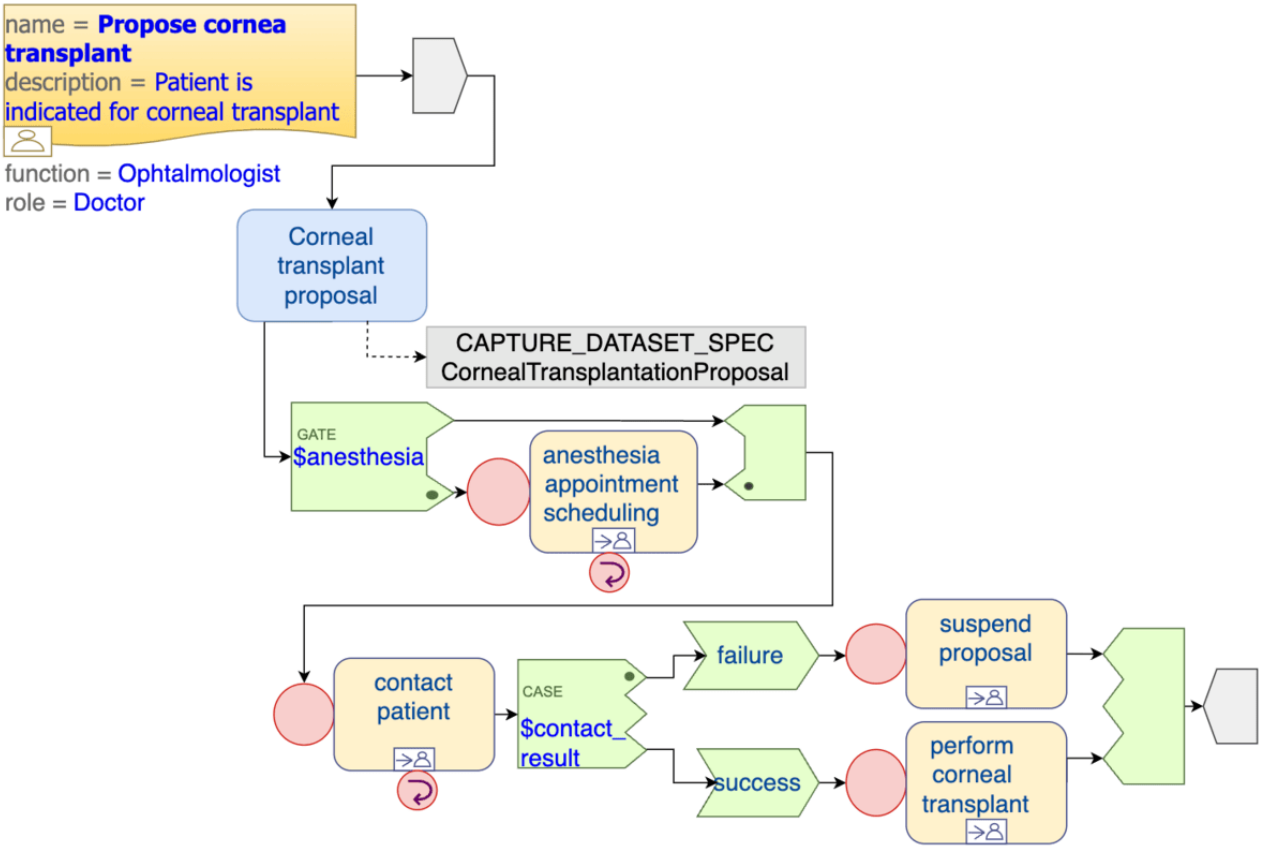
subject, human, or other subject of care [18]. It is worth noting that, as the WP is subject-centric, each subject requires a unique instance of a WP. WPs can organize and monitor the progress of clinical tasks, ensuring that all necessary steps are taken in a timely and efficient manner [29]. In turn, each TP incorporated within a WP is an explicit depiction of the work that must be performed in a particular work context by the principal performer, along with other possible participants [18]. In openEHR, the principal performer refers to the individual or entity responsible for carrying out a specific clinical action, which enables the tracking of clinical actions back to their responsible parties. The data collected through forms and the decision support provided by Decision Logic Modules (DLMs) can be used to define and update WPs in real time.

In this study, a WP regarding corneal transplantation was modeled to ensure that the implementation of the solution proceeds smoothly and is completed within the specified timeline.

The first TP defined in this WP concerns the corneal transplantation proposal. This task is available to certain ophthalmologists in specific contexts. In total, 2 TPs related to the anesthesia consultation were also included: scheduling and carrying out the consultation, which are assigned to different groups of professionals. Finally, 2 more TPs were modeled: one related to the contact for transplantation carried out by administrative staff and one related to the performance of the transplant, which is allocated to the physician who submitted the proposal in the first instance.

Figure 6 shows the first TP in a simplified way. It is interesting to note that, in this first TP, the proposal form for a corneal transplant is completed in the first performable task. After submission, the execution of the anesthesia consultation relies on the value assigned to the DV_BOOLEAN field labeled "Anesthesia Consultation" within the "Corneal Transplantation Proposal" form. If this field indicates a true value, the subsequent steps for that patient involve scheduling the anesthesia consultation; conducting the consultation; and, eventually, establishing contact for the transplantation procedure. On the contrary, if the anesthesia consultation is not necessary, the patient goes to the active waiting list represented by the cornea transplant contact tasks. If the contact is successful, the patient proceeds to undergo the transplant; if not, the proposal is suspended, and the suspension management task becomes available. All dispatchable tasks presented in this example are connected to other TPs, which mainly have a performable task with one of the associated modeled forms.

Figure 6. Representation of the top-level Task Plan.



Later in this paper, the main decisions that need to be made in the course of the WP will be explained.

Construction of Decision Rules

Overview

Clinical decision support is a key component of the openEHR architecture, responsible for automating and enhancing decision-making. The openEHR community defines decision

rules and guidelines using a specific syntax, the Decision Language (DL) [30]. DL is a formal language for representing clinical knowledge in decision-making and expressing decision support logic through rulesets.

DLMs are multisectioned modules with a specific structure for defining rules, encompassed in DL. DLMs are computerized decision-making instructions that provide a standardized and automated way to apply decision rules to patient data within the EHR in real time [30,31].

DLMs enable health care organizations to implement algorithms and rules to determine the best course of action for a given patient [32]. The output from a DLM can be used to guide clinical decision-making; provide alerts or notifications to health care providers; or drive automated actions within the EHR, such as ordering tests or medications [18].

To ensure the efficiency of the solution, decision rules were established for forms and TPs to define the logic and actions that should be taken based on specific data inputs. The rules are stored in a standardized structure that can be applied to patient data at runtime.

Overall, by automating the application of decision rules, DLMs can help reduce the risk of errors and variability in decision-making, providing more consistent care [33].

Form Decision Rules

The DLMs allow the forms to automatically adapt to the specific constraints for a given patient or scenario and to the responses of the health care professional filling in the form fields, ensuring that the information entered is always consistent with established clinical guidelines and best practices. The data collected through the forms are used as input for DLMs to support the delivery of real-time decision support.

By automating the process of adjusting forms to specific scenarios and inputs, the solution can provide decision support to health care providers, guiding them through the process of entering data and making clinical decisions.

Hence, according to the requirements gathered in the first stage of the methodology, 2 DLMs were created for guiding health care professionals in the process of filling in the forms F1—“Corneal Transplantation Proposal” and F2—“Contact for Corneal Transplantation.” Table 3 describes the association among the forms, DLMs, and rules.

Table 3. Association among forms, Decision Logic Modules (DLMs), and rules.

Form	DLM	Rule	Rule description
F1	1	1	If the patient’s age is ≥75 years, anesthesia is mandatory.
F2	2	1	If the patient intends to leave the transplant list, the motive is mandatory.
F2	2	2	If the patient needs to postpone the contact, the next contact is mandatory.

The rules associated with each DLM, including the conditions necessary to trigger a certain action, are described below.

For the “Contact for Corneal Transplantation” form, the decision logic operates as follows: If the “Result” field is set to “Cancellation,” the “Motive” field becomes mandatory. Similarly, if the “Result” field is set to “Postpone contact,” the “Next contact” field becomes mandatory.

Regarding the “Perform Anesthesia Consultation” form, the logic dictates that if the “Age” field has a value of 75 or greater, the “Anesthesia” field must be set to “Yes.”

WP Decision Rules

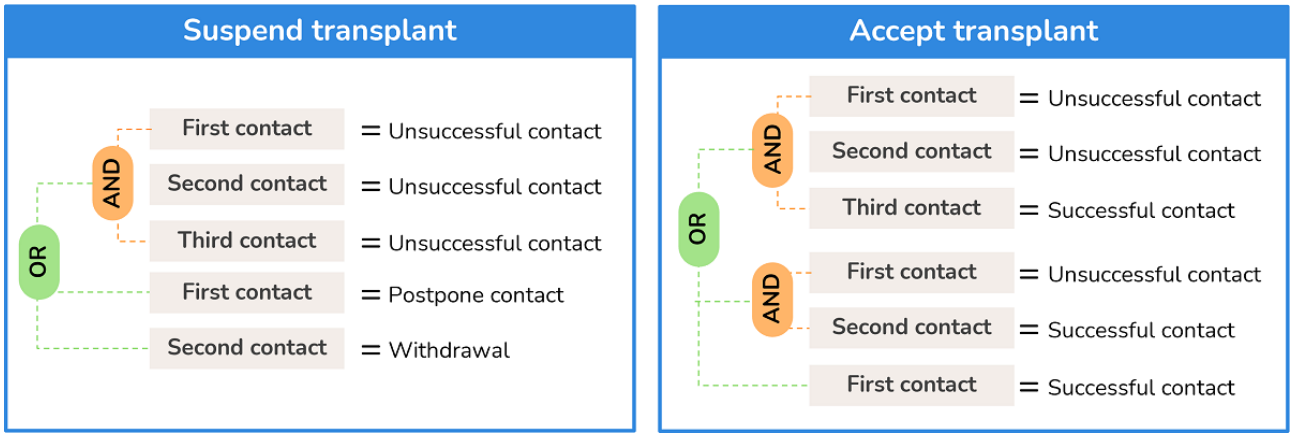
The DLM rules built to support the necessary decisions in the modeled WP were crucial for the correct functioning of the respective materializations.

The first rule to be processed concerns the decision whether the patient should proceed to an anesthesia appointment. If the priority of the proposal is urgent, the patient does not need an anesthesia appointment. If it is not urgent, the need for an anesthesia consultation is decided by the “Anesthesia” field filled in by the physician in the “Corneal Transplantation Proposal” form.

Regarding the anesthesia consultation, a rule was created to verify the success of the consultation and, thus, decide whether the patient goes to the contact for transplantation. This rule uses the data entered by the anesthesiologist in the respective form.

Then, the most complex rule was built to support the decision after the patient is called for transplant. Figure 7 represents the logic behind the decision to be made after the transplant contact, where up to 3 contacts with the patient can be registered.

Figure 7. Decision logic associated with the Work Plan.



Finally, there is still a decision that is made for suspended patients, which is based on whether they are removed from the transplant waiting list and the WP ends or the patient is reinserted into the transplant contact list, making the respective performable task available.

Ethical Considerations

In this study, all patient demographic information was anonymized through the use of the openEHR separation of the Demographic Information Model and Clinical Information Model. This approach ensures that personal identifiers are not linked to clinical data, thereby maintaining patient confidentiality and minimizing the risk of reidentification. Furthermore, the data analysis conducted in this study focused on performing statistical analyses at an administrative level. This includes examining the quantity of tasks available, the completion rate of these tasks, and referencing solely the number of patients enrolled in the study. The analysis was restricted to aggregate data, ensuring that individual patient identifiers and clinical details were not disclosed. This methodological approach allowed for a comprehensive evaluation of the operational aspects of the study without compromising patient confidentiality or violating ethical standards. The data structure of the generated forms can be found [34]. Given these protections, this study qualified for an exemption from ethical review, as per the Code of Ethical Conduct of the University of Minho.

Results

Deployment and Architecture Overview

The openEHR ecosystem provides a comprehensive solution for managing health care data by combining different tools and technologies. In openEHR, forms, DLMs, and WPs interact in a complementary manner to support the provision of effective and efficient clinical care. These modules ensure that patient data are accurate, consistent, and accessible and provide data validation, automated actions, and real-time decision support to ensure that the best course of action is taken for each individual patient. Furthermore, the integration of these modules allows for the organization and tracking of the progress of clinical tasks to ensure that all necessary clinical tasks are performed in a timely and efficient manner.

Once the modeling and validation of all openEHR structures supporting corneal transplantation EHRs were concluded, it was necessary to integrate them into an automated solution. The implementation step involved the integration of a set of tools from the openEHR ecosystem, which include Form Builder, the TP engine, and the DLM engine.

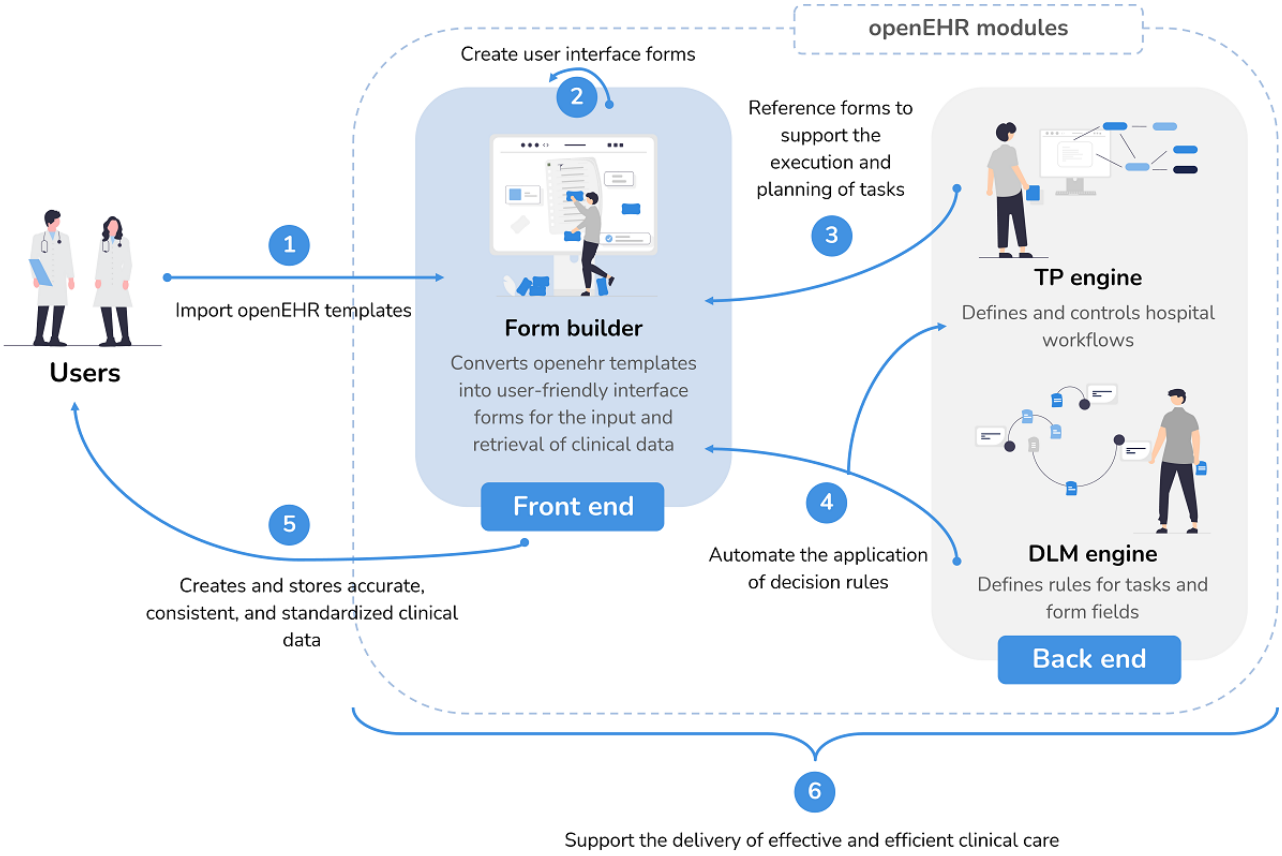
Form Builder is a web application that is designed to generate user interface forms from openEHR templates. It provides a platform for health care professionals who possess modeling knowledge to customize the user interface forms by adjusting the formatting, such as choosing colors and fonts and determining whether fields should be shown or hidden. In addition, it allows for the association of functions and refsets to form fields, which can add further functionality to the form. Using Form Builder, health care professionals can streamline the process of data entry by creating forms that are intuitive and optimized for their specific clinical workflows.

In turn, the TP engine is a tool that manages all workflows modeled by the professionals, including materialization, task status, decision management, and allocation of performers. To accomplish this, the TP module defines a formal model for processing tasks and workflows. This tool is designed to translate graphical workflow models into executable models of an organized plan that, when carried out by an engine, notifies employees of tasks. Overall, the TP engine provides an automated solution for managing and executing complex workflows in the health care setting.

The DLM engine, on the other hand, is a decision logic engine that is responsible for processing clinical or operational rules and triggering specified events based on predefined conditions. This engine plays a crucial role in supporting the logic of forms and TPs. Accordingly, it receives requests from both Form Builder and the TP engine. It ensures that decision rules are executed correctly and consistently, leading to improved patient care and outcomes.

Figure 8 illustrates the different interactions that occur among these openEHR modules.

Figure 8. Representation of openEHR modules and their interactions. DLM: Decision Logic Module; TP: Task Plan.



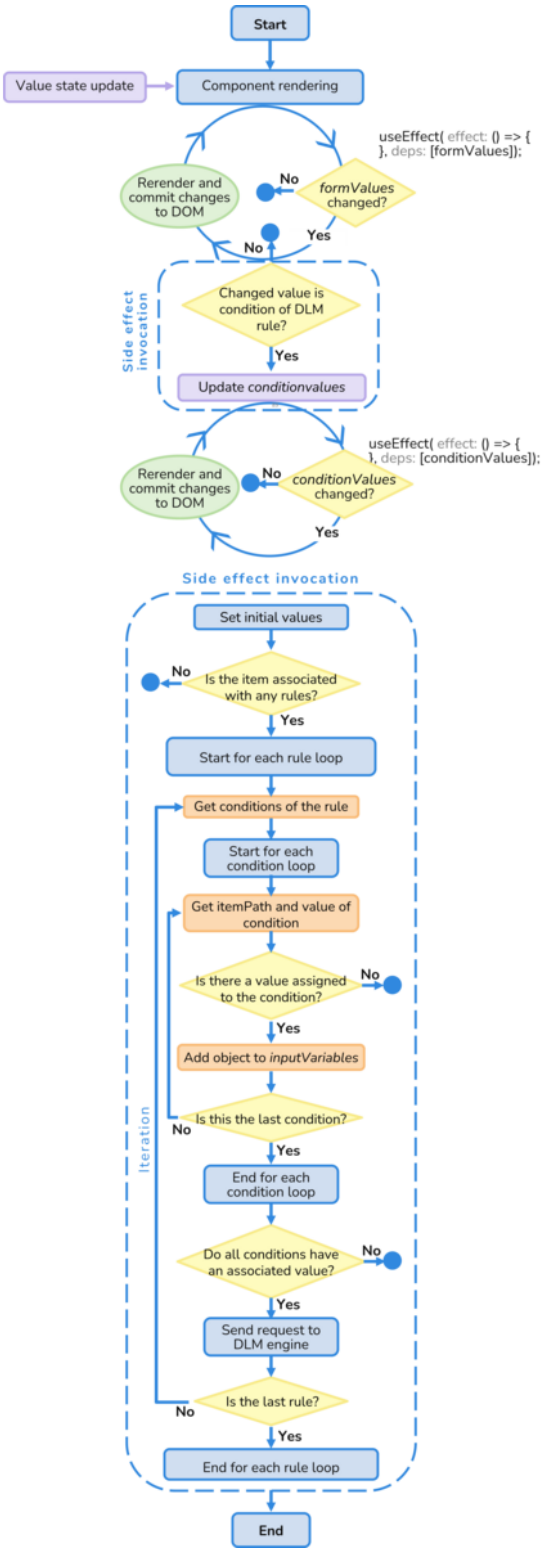
The integration of all these components was also explored in this research.

Regarding the integration of Form Builder and the DLM engine, the process starts when a health care professional updates a form field. Whenever the value of a field changes, Form Builder performs an internal processing to check whether that field is a condition of any rule within the DLM associated with the current form. If so, Form Builder identifies each associated rule and, for each of them, verifies whether there are more associated conditions that already have a value assigned to them. When all the conditions of a rule have an assigned value, Form Builder triggers a POST request to the DLM engine with the input variables. This is a well-defined sequence of computational tasks that need to be followed to properly implement the specific set of rules that have been modeled for each form. For better understanding, a graphical representation of the processing that occurs in Form Builder to verify the need to trigger any rule requests is presented in Figure 9.

Upon receiving the input variables, the DLM engine checks whether the conditions are met for executing a given rule. If the values assigned to the form fields do not fulfill a rule, an empty object is returned. On the other hand, if the values meet the conditions for executing a rule, the engine returns the path of the affected field and the type of event to execute. Form Builder then triggers the necessary actions for each field depending on the response. Figure 10 serves as an exemplification of the HTTP requests that are exchanged between Form Builder (front-end server) and the DLM engine (back-end server) in 2 distinct scenarios. In the first scenario, the health professional enters a value that triggers the execution of a rule, whereas the second scenario does not involve the activation of any rule. This diagram provides a clear representation of the data flow and communication between the 2 servers during the execution of the rule-based system.

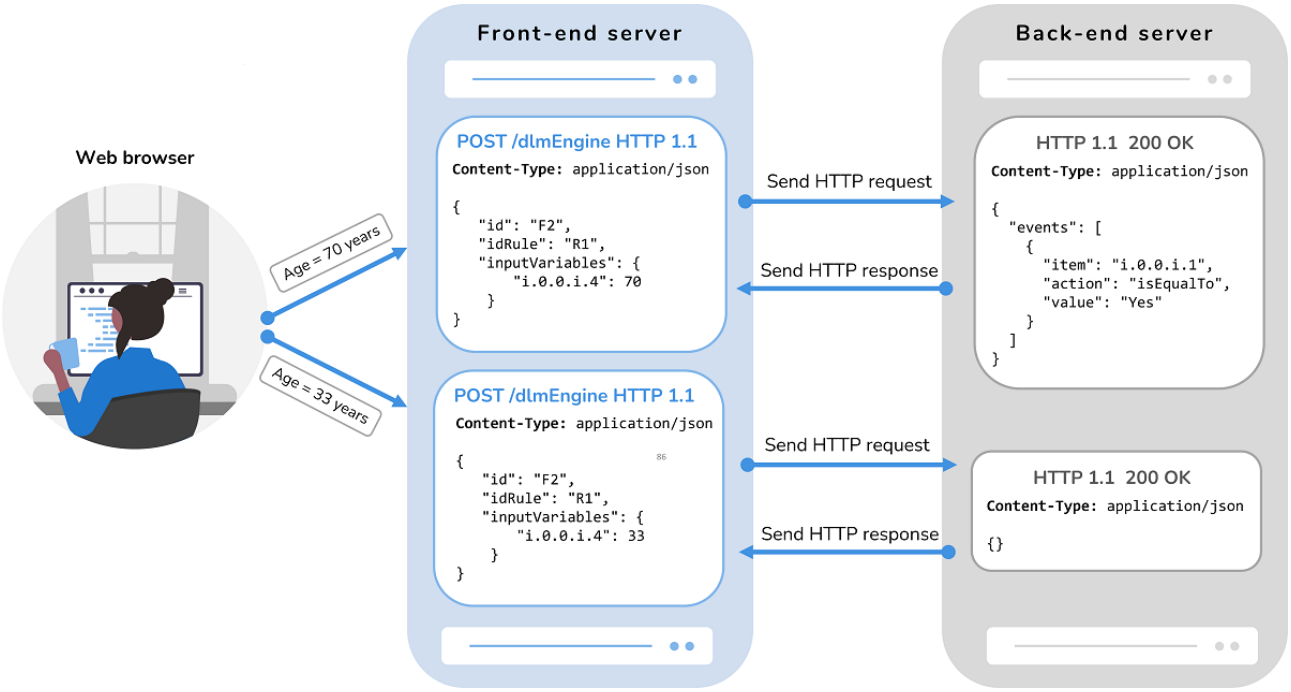
The integration of the DLM engine with the TP engine represents a remarkable achievement in the openEHR ecosystem. The successful interaction between the 2 enables the handling of all conditional structures encompassed in a WP, including condition groups, decision groups, and event groups, with the support of the DLM engine. In general, when the path of a WP materialization reaches a decision point, a request is issued to an application programming interface provided by the DLM engine with the variable associated with that point. In turn, the DLM engine processes the rules and conditions associated with the request that was made and returns a response. Through the response received, the TP engine manages to associate it with the respective branch (decision branch, condition branch, or event branch) and proceed with its execution.

Figure 9. A schematic illustration of the logic flow that occurs in the Form Builder process responsible for initiating rule requests. DLM: Decision Logic Module; DOM: Document Object Model.



In the openEHR ecosystem, the TP engine can refer openEHR forms to support the planning and execution of hospital tasks. By providing a user-friendly interface for entering and retrieving data, Form Builder makes forms available to health care professionals that can be used to collect data required for various tasks, and it ensures that the collected data are accurate and standardized. Furthermore, the responses given by health care professionals in certain fields of the form can be determinant to define which path should be triggered after submitting the form. By integrating the TP engine and Form Builder, health care providers can ensure that relevant information is captured and acted upon as part of the patient's care plan, leading to improved patient outcomes and increased efficiency in health care delivery. This integration enhances the quality of care provided to patients and enables health care professionals to make informed decisions based on accurate and timely data.

Figure 10. Diagram of HTTP request and response interactions between Form Builder, which acts as the front-end server, and the Decision Logic Module (DLM) engine, which serves as the back-end server, in 2 distinct settings.



After successfully integrating these components in a symbiotic IT environment, it was possible to install the solutions in the Portuguese hospital to implement and evaluate them in a real-world setting.

For several years, this hospital has maintained a collaborative relationship with the research center where the working group is based. Given this existing partnership, the hospital was deemed an ideal site to deploy and assess the newly developed solution.

Considering the current IT infrastructure of the hospital, the amalgamation of the proposed solution was not hampered by any intricate integration challenges and unfolded as a straightforward process. To streamline the workflow and improve efficiency, Form Builder was integrated into a web application already implemented in the hospital that presents a detailed listing of the tasks assigned to a certain user. This integration enables health care professionals to view a comprehensive list of assigned tasks and submit completed forms directly through the portal. The TP engine is the mechanism that controls and triggers the tasks made available to the medical team on the professional portal. As a result, the professional portal serves as a centralized hub for task management and data collection, enhancing the overall clinical workflow.

To ensure that only authorized health care professionals had access to corneal transplantation tasks, it was necessary to create specific members for this purpose within the demographic of professionals with their respective capabilities, roles, and functions. Once the members of each team were established, they were associated with the corresponding tasks. By establishing a clear hierarchy of roles and responsibilities, the hospital was able to ensure that the tasks related to corneal transplantation were being accessed and completed by qualified and authorized personnel.

The collaboration between the health care institution and the research center has paved the way for the exchange of knowledge and resources, allowing for a more efficient and effective implementation of the solution. Furthermore, this cooperation has fostered a culture of innovation and continuous improvement in the hospital's clinical practices, ultimately yielding beneficial outcomes for the patients.

The main challenges encountered during deployment were related to the lack of health care professionals with knowledge of openEHR and modeling skills. As a result, the team conducted several demonstrations and provided comprehensive documentation to facilitate the users' adoption of the tools. Despite these challenges, the feedback and acceptance from medical staff were generally positive as they reported ease of adaptation and expressed satisfaction with the provided tools.

Statistical Analysis

In a data-driven world, statistical analysis has become critical to gain insights and draw conclusions that may not be immediately apparent through simple visual inspections. It holds particular significance in the realm of health research, where it can be used to evaluate the effectiveness and efficiency of IT solutions such as EHRs, telemedicine platforms, and other digital health tools. In this way, health care professionals can identify areas for improvement, including the streamlining of workflows, the enhancement of usability, and the resolution of technical glitches, ultimately leading to improved patient outcomes and enhanced quality of care.

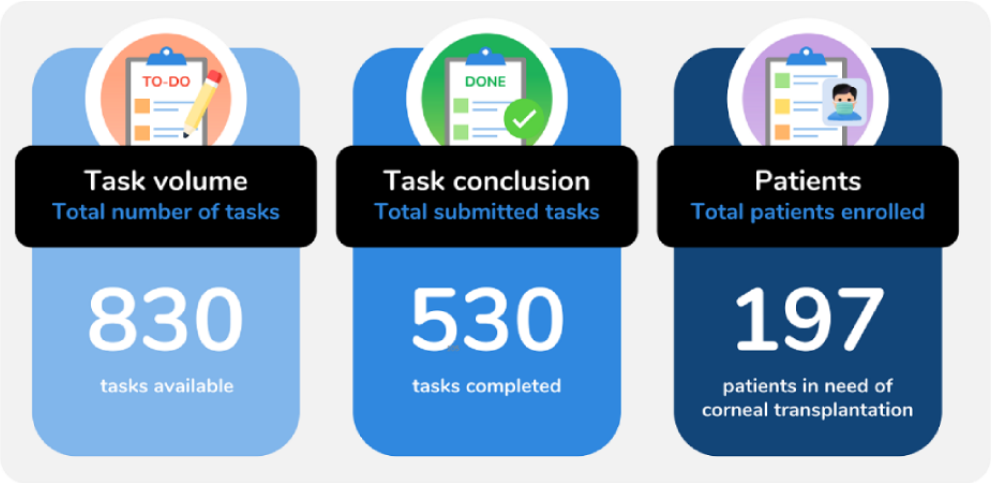
Overall, the importance of data analytics in the health care domain cannot be overstated as it has the potential to significantly impact the lives and well-being of countless individuals. Hence, to verify the efficiency and performance of the solutions offered to manage the corneal transplantation process, this section presents an analysis of the data gathered

over the period of study from May 1, 2022, to March 31, 2023. This analysis will include relevant indicators and charts. Before presenting the data analysis, a brief overview of the data collection process will be provided.

The data collection process for this study required careful planning, attention to detail, and adherence to ethical guidelines to ensure the accuracy and validity of the data. First, the databases and tables of interest were selected. Then, an anonymization process was carried out to preserve the identity and privacy of the patients, which involved removing any personal identifying information from the data. Once the data had been anonymized, the relevant SQL queries were developed to extract the data required. Finally, after the data had been extracted, they were meticulously organized into descriptive statistics in the form of indicators, charts, and graphs to ease interpretation and help convey the findings.

Figure 11 shows the graphical representation of key indicators in the corneal transplantation process, including task volume, task conclusion, and patients enrolled. The task volume represents the number of tasks available for health care professionals to fill out the corresponding forms, whereas the task conclusion corresponds to the number of tasks successfully submitted by health care professionals. On the other hand, patient enrollment indicates the number of patients registered in the corneal transplantation list.

Figure 11. Single-number indicators of the corneal transplantation process, including task volume, task conclusion, and patient enrollment.



The proportion of concluded tasks in comparison to the number of available tasks provides a clear and concise overview of the overall task completion rate and offers a general picture of how well tasks are being managed and completed. The number of available and completed tasks can help gauge the workload of health care professionals and assess the capacity of the system to handle the demands of the workflow. A higher number of completed tasks relative to available tasks indicates that the system is functioning efficiently and effectively. Meanwhile, a lower completion rate could suggest potential bottlenecks or areas for improvement in the system's design or implementation. In this study, there is 63.9% (530/830) of concluded tasks and 36.1% (300/830) of available tasks.

The total number of corneal transplantation forms submitted by health care professionals over time can be consulted in Table 4, allowing for the identification of trends and patterns in form submission and providing insights into the volume and frequency of tasks completed.

Table 4. Total number of corneal transplantation tasks submitted over time (n=530).

Month and year	Corneal transplantation tasks submitted, n (%)
March 2023	33 (6.2)
February 2023	19 (3.6)
January 2023	79 (14.9)
December 2022	53 (10)
November 2022	32 (6)
October 2022	87 (16.4)
September 2022	39 (7.4)
August 2022	49 (9.3)
July 2022	30 (5.7)
June 2022	58 (10.9)
May 2022	51 (9.6)

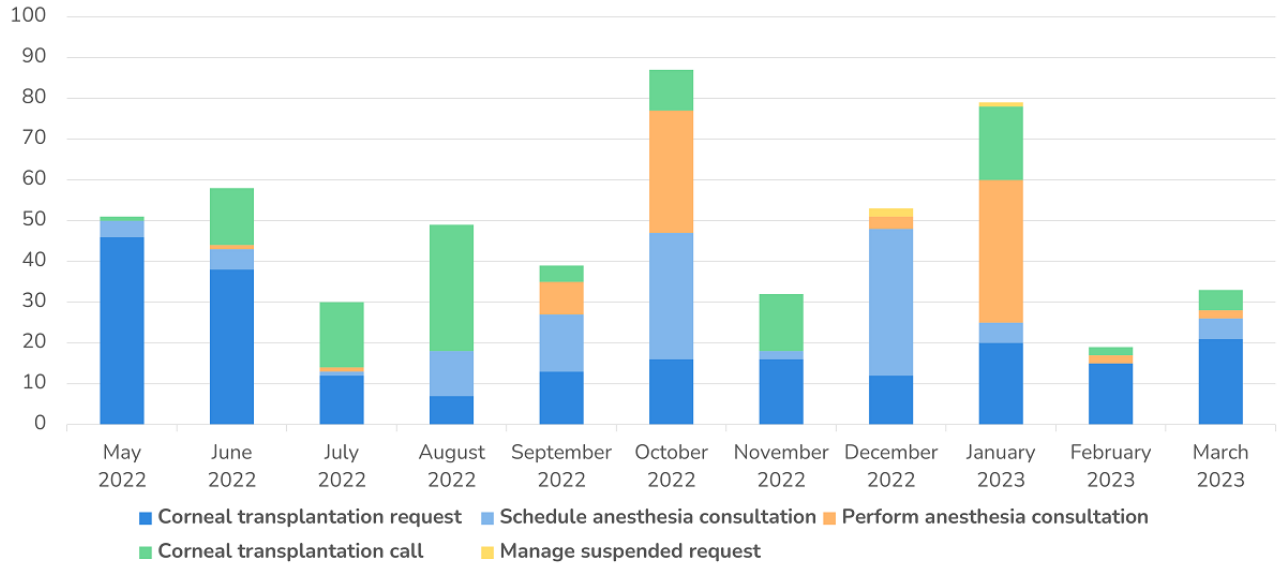
As the corneal transplantation workflow is a complex process involving the completion of a variety of tasks, Table 5 helps visualize the distribution of these tasks across the different form categories. This table displays both the numerical and percentage distribution of submitted tasks according to each form category.

Table 5. Total number of corneal transplantation tasks submitted over time (n=530).

Tasks	Corneal transplantation tasks submitted, n (%)
Corneal transplantation request	216 (40.8)
Schedule anesthesia consultation	114 (21.5)
Perform anesthesia consultation	82 (15.5)
Corneal transplantation call	115 (21.7)
Manage suspended requests	3 (0.6)

Finally, to provide a comprehensive visualization of the number of tasks submitted over time for each form category, a stacked bar chart was used in Figure 12. This chart displays the total number of tasks completed and submitted for each form category over the period under consideration, helping assess the relative contributions of each form category to the overall workflow.

Figure 12. Number of submitted tasks within the corneal transplantation workflow for each form over time.



Discussion

Principal Findings

The findings of this study offer a comprehensive and in-depth overview of the incorporation of various components of openEHR specifications, particularly the interaction of openEHR forms, WPs, and DLMs, which have not yet been fully covered in scientific literature.

To evaluate the impact of the solution, the openEHR structures were incorporated into a workflow and integrated into a health care institution. This study demonstrated the effects of the intervention on the corneal transplantation workflow, which was previously characterized by inadequate automation and an intensified risk of data loss. The use of the new technologies has the potential to substantially influence the workload of health care professionals and, consequently, affect patient care outcomes. Therefore, in addition to providing a thorough description of the implementation process, this paper also presented a statistical analysis of its effects.

The indicators presented in the *Statistical Analysis* section, which are shown in Figure 11, show the number of available tasks, completed tasks, and patients enrolled in the corneal transplantation list during the time frame of this study.

The number of available tasks was 830, which represents the total number of tasks that need to be completed. Of these tasks, 530 were completed, indicating that the completion rate of the tasks was 63.9% (530/830).

The number of available and completed tasks can help gauge the workload of health care professionals and assess the capacity of the system to handle the demands of the workflow. A higher ratio of completed tasks to available tasks indicates that the system is working efficiently and effectively. A lower completion rate, on the other hand, may suggest potential bottlenecks or areas for improvement in the system's design or implementation.

As a result, a 63.9% (530/830) completion rate falls short of expectations. This lower rate could be explained by a number of factors. First, the hospital may have limited resources to complete the tasks within the workflow, leading to delays in

completing the tasks. Furthermore, the health care professionals responsible for completing the tasks within the workflow may not have received enough training or may be inexperienced, which may imply some reluctance in adopting the new technologies.

Because most of the available tasks within the studied workflow regarded the "Contact for Corneal Transplantation," which represents the active waiting list for corneal transplantation, the workflow completion depended on external factors such as patient health and availability of donor organs. These dependencies directly affected the time required to complete the tasks and, consequently, the total workflow.

Due to the complexity of the corneal transplantation process, a multidisciplinary service team is required, namely, the anesthesia team and the ophthalmology team, which can cause communication issues between the teams and, as a result, delays in task completion.

These factors may also explain the fluctuations in the number of task submissions over time depicted in Table 4. Hence, as a result of the influence of these factors, it was not possible to establish patterns in the data over time as expected given the nonseasonal nature of the data.

Furthermore, the enrollment of 197 patients is a significant aspect of the data as it serves as a critical contextual element in assessing the workload. However, it should be noted that the number of enrolled patients, as previously mentioned, does not accurately reflect the total number of corneal transplantation proposals submitted. This is because a single patient may be registered multiple times to undergo different procedures. As a result, the number of enrolled patients may not be a direct indicator of the workload or the number of tasks that need to be completed in a corneal transplantation workflow.

In this sense, to gain a more comprehensive understanding of workload management within a hospital setting, a more specific study on the different types of tasks available was required. This analysis enables more detailed scrutiny of the specific challenges and constraints associated with each task type and allows for the identification of more targeted solutions to enhance efficiency and productivity.

A quick look into Table 5 reveals a higher rate, 40.8% (216/530), of completed tasks associated with the "Corneal Transplantation Proposal" form in comparison to the remaining tasks. This rate can be explained by the lack of resources to perform the transplants at a quicker pace. It is also worth noting that the form with the lowest submission rate was "Management of Suspended Proposal," indicating the fewer cases in which the corneal transplantation proposals were suspended.

Finally, Figure 12 depicts the number of submitted tasks according to each form over time. In the first months, it is possible to observe that the tasks associated with the "Contact for Corneal Transplantation" form increased, whereas the number of "Corneal Transplantation Proposal" tasks decreased. Since August 2022, a higher number of submitted tasks pertaining to the "Schedule Anesthesia Consultation" form can be observed. As a result, a month later, it is possible to observe an increase in the number of tasks pertaining to the "Perform Anesthesia Consultation" form.

The findings of this analysis emphasize the importance of effective workload management within a hospital setting. By monitoring and analyzing the number of tasks available and completed, hospitals can identify areas for improvement and ensure that patient care is not jeopardized. In addition, this study can help identify potential bottlenecks and areas of inefficiency in current workflows, informing the design and implementation of targeted interventions to enhance the effectiveness of health care operations.

Limitations

A paper outlining the implementation steps of clinical standards in a hospital setting is a valuable resource for identifying best practices and areas for improvement, as well as for advancing patient care. However, it is important to acknowledge its potential limitations to fully understand its impact on health care.

Resource limitations are an important concern to consider as the implementation of openEHR specifications requires staff time and training, which can pose additional challenges for health care organizations. One difficulty encountered in this study was the staff workload as the implementation of the solution presented in this paper required them to attend training sessions and meetings, as well as review new policies and procedures. Changes in work processes and team dynamics were an additional threat to the implementation of the solution as it was necessary to update team members' permissions and roles.

Training was another area in which difficulties were encountered as health care providers had to learn new skills and competencies to use the specifications effectively, which involved becoming accustomed to new technology and tools. This was particularly challenging for those who were less comfortable with technology, limiting their willingness to adopt new practices and posing resistance to change.

Finally, it was essential to consider time constraints because the findings discussed in this paper pertain to the duration of the study. The adoption of clinical standards is an ongoing process, and this paper may not be able to fully capture the long-term effects of the implementation.

Conclusions

Traditional health care is plagued by the use of disparate systems for managing patient data, leading to a fragmented view of medical records as well as inconsistencies and gaps in clinical information. Without standardized and efficient systems in place, there is a higher risk of medical errors, miscommunication, delayed or inadequate diagnoses, and suboptimal treatment decisions, which can ultimately compromise patient safety and health care quality. In addition, this issue underscores the importance of interoperability in health care.

In the case of corneal transplantation, accurate and timely management of patient information is critical for the success of the procedure and the well-being of the patient. Hence, this

study proposed the adoption of clinical standards, specifically openEHR, to address these challenges by enabling the creation of a comprehensive and shared patient record. This paper provides insights into the use of openEHR in health care and contributes to the incessant efforts to improve the quality and safety of patient care.

The implementation of openEHR specifications to standardize corneal transplantation records and streamline its workflow can yield significant benefits to patients, health care providers, and the health care system as a whole. Standardized EHRs can ensure the accuracy, consistency, and completeness of data entry and management, leading to increased patient safety and reduced medical errors. Furthermore, it serves as a centralized repository for clinical data, enabling health care providers to access information more easily and facilitating the seamless exchange of data between different health care systems. openEHR can also support clinical decision-making by providing real-time access to patient data and enabling clinicians to make more informed decisions about patient care.

In summary, the process that connects openEHR forms, WPs, and DLMs ensures that the best course of action for a given patient is taken by providing real-time decision support, data validation, and automated actions and ensuring that patient data are accurate, consistent, and accessible, as well as that all necessary clinical tasks are performed in a timely and efficient manner.

Although this study focused on the implementation of clinical standards in a specific health care setting, the principles and strategies used to implement openEHR specifications remain relevant and applicable in other health care contexts. Hence, the findings of this study hold considerable value for health care professionals, hospital administrators, and technology developers, providing critical insights into the implementation of openEHR specifications within a hospital setting and paving the way for the development of innovative solutions to optimize health care operations.

In light of these benefits, it is clear that the adoption of openEHR structures for the standardization of corneal transplantation records represents a critical step forward in the pursuit of safer, more effective, and higher-quality care. Hence, the authors believe that using openEHR specifications will become standard practice in the health care industry in the near future.

Future research could focus on the application of artificial intelligence algorithms to data extracted from standardized EHRs as training algorithms on reliable, consistent, and high-quality data leads to more robust and trustworthy results. This can enable a more efficient and effective clinical data analysis, maximizing the potential of openEHR to drive meaningful improvements in health care outcomes.

Acknowledgments

This work was supported by Foundation for Science and Technology within the R&D Units project scope (UIDB/00319/2020). DF, CN, and FH thank the Portugal Fundação para a Ciência e a Tecnologia for grants 2021.06308.BD, 2021.06507.BD, and 2021.06230.BD, respectively.

Conflicts of Interest

None declared.

References

1. World report on vision. World Health Organization. 2019. URL: https://tinyurl.com/4xer4xz3 [accessed 2024-04-29]
2. Hollins M. Understanding Blindness: An Integrative Approach. New York, NY: Routledge; 2022.
3. GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e130-e143 [FREE Full text] [doi: 10.1016/S2214-109X(20)30425-3] [Medline: 33275950]
4. Package of eye care interventions. World Health Organization. 2022. URL: https://www.who.int/publications/i/item/9789240048959 [accessed 2024-04-29]
5. Brown RL, Barrett AE. Visual impairment and quality of life among older adults: an examination of explanations for the relationship. J Gerontol B Psychol Sci Soc Sci 2011 May 14;66(3):364-373. [doi: 10.1093/geronb/gbr015] [Medline: 21402645]
6. Bourne RR, Flaxman SR, Braithwaite T, Cicinelli MV, Das A, Jonas JB, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. Lancet Glob Health 2017 Sep;5(9):e888-e897 [FREE Full text] [doi: 10.1016/S2214-109X(17)30293-0] [Medline: 28779882]
7. GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. Lancet Glob Health 2021 Feb;9(2):e144-e160 [FREE Full text] [doi: 10.1016/S2214-109X(20)30489-7] [Medline: 33275949]
8. Singh R, Gupta N, Vanathi M, Tandon R. Corneal transplantation in the modern era. Indian J Med Res 2019;150(1):7. [doi: 10.4103/ijmr.ijmr_141_19]

9. Pineda R. World corneal blindness. In: Pineda R, editor. Foundations of Corneal Disease: Past, Present and Future. Cham, Switzerland: Springer; 2020:299-305.
10. Flaxman SR, Bourne RR, Resnikoff S, Ackland P, Braithwaite T, Cicinelli MV, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. Lancet Glob Health 2017 Dec;5(12):e1221-e1234. [doi: 10.1016/S2214-109X(17)30393-5]
11. Action plan for the prevention of avoidable blindness and visual impairment, 2009-2013. World Health Organization. 2010. URL: https://www.emro.who.int/control-and-preventions-of-blindness-and-deafness/announcements/action-plan-prevention-avoidable-blindness-visual-impairment-2014-2019.html [accessed 2024-04-29]
12. Dobrow MJ, Bytautas JP, Tharmalingam S, Hagens S. Interoperable electronic health records and health information exchanges: systematic review. JMIR Med Inform 2019 Jun 06;7(2):e12607 [FREE Full text] [doi: 10.2196/12607] [Medline: 31172961]
13. Uslu A, Stausberg J. Value of the electronic medical record for hospital care: update from the literature. J Med Internet Res 2021 Dec 23;23(12):e26323 [FREE Full text] [doi: 10.2196/26323] [Medline: 34941544]
14. Ferreira D, Silva S, Abelha A, Machado J. Recommendation system using autoencoders. Appl Sci 2020 Aug 10;10(16):5510. [doi: 10.3390/app10165510]
15. Martins B, Ferreira D, Neto C, Abelha A, Machado J. Data mining for cardiovascular disease prediction. J Med Syst 2021 Jan 05;45(1):6. [doi: 10.1007/s10916-020-01682-8] [Medline: 33404894]
16. Yang L, Huang X, Li J. Discovering clinical information models online to promote interoperability of electronic health records: a feasibility study of OpenEHR. J Med Internet Res 2019 May 28;21(5):e13504 [FREE Full text] [doi: 10.2196/13504] [Medline: 31140433]
17. Min L, Atalag K, Tian Q, Chen Y, Lu X. Verifying the feasibility of implementing semantic interoperability in different countries based on the OpenEHR approach: comparative study of acute coronary syndrome registries. JMIR Med Inform 2021 Oct 19;9(10):e31288 [FREE Full text] [doi: 10.2196/31288] [Medline: 34665150]
18. openEHR specifications. openEHR. URL: https://specifications.openehr.org/ [accessed 2023-04-10]
19. Leslie H. openEHR archetype use and reuse within multilingual clinical data sets: case study. J Med Internet Res 2020 Nov 02;22(11):e23361 [FREE Full text] [doi: 10.2196/23361] [Medline: 33035176]
20. Tarenskeen D, van de Wetering R, Bakker R, Brinkkemper S. The contribution of conceptual independence to IT infrastructure flexibility: the case of openEHR. Health Policy Technol 2020 Jun;9(2):235-246. [doi: 10.1016/j.hlpt.2020.04.001]
21. Kalra D, Beale T, Heard S. The openEHR foundation. Stud Health Technol Inform 2005;115:153-173. [Medline: 16160223]
22. Oliveira D, Ferreira D, Abreu N, Leuschner P, Abelha A, Machado J. Prediction of COVID-19 diagnosis based on openEHR artefacts. Sci Rep 2022 Jul 22;12(1):12549 [FREE Full text] [doi: 10.1038/s41598-022-15968-z] [Medline: 35869091]
23. Demski H, Garde S, Hildebrand C. Open data models for smart health interconnected applications: the example of openEHR. BMC Med Inform Decis Mak 2016 Oct 22;16(1):137 [FREE Full text] [doi: 10.1186/s12911-016-0376-2] [Medline: 27770769]
24. Moner D, Maldonado JA, Robles M. Archetype modeling methodology. J Biomed Inform 2018 Mar;79:71-81 [FREE Full text] [doi: 10.1016/j.jbi.2018.02.003] [Medline: 29454107]
25. S Rubí JN, L Gondim PR. IoMT platform for pervasive healthcare data aggregation, processing, and sharing based on OneM2M and OpenEHR. Sensors (Basel) 2019 Oct 03;19(19):4283 [FREE Full text] [doi: 10.3390/s19194283] [Medline: 31623304]
26. Frade S, Beale T, Cruz-Correia RJ. OpenEHR implementation guide: towards standard low-code healthcare systems. In: Seroussi B, Ohno-Machado L, editors. MEDINFO 2021: One World, One Health - Global Partnership for Digital Innovation. Amsterdam, Netherlands: IOS Press; 2022:52-55.
27. Cardoso de Moraes JL, de Souza WL, Pires LF, do Prado AF. A methodology based on openEHR archetypes and software agents for developing e-health applications reusing legacy systems. Comput Methods Programs Biomed 2016 Oct;134:267-287 [FREE Full text] [doi: 10.1016/j.cmpb.2016.07.013] [Medline: 27480749]
28. Iglesias N, Juarez JM, Campos M. Business process model and notation and openEHR task planning for clinical pathway standards in infections: critical analysis. J Med Internet Res 2022 Sep 15;24(9):e29927 [FREE Full text] [doi: 10.2196/29927] [Medline: 36107480]
29. Iglesias N, Juarez JM, Campos M. Handling time constraints in infection clinical pathways using openEHR TP. Stud Health Technol Inform 2022 Jun 06;290:7-11. [doi: 10.3233/SHTI220021] [Medline: 35672960]
30. Beale T. Decision language specification. openEHR. URL: https://specifications-test.openehr.org/releases/PROC/latest/decision_language.html [accessed 2023-04-12]
31. Hak F, Oliveira D, Abreu N, Leuschner P, Abelha A, Santos M. An OpenEHR adoption in a Portuguese healthcare facility. Procedia Comput Sci 2020;170:1047-1052. [doi: 10.1016/j.procs.2020.03.075]
32. Beale T, Chen R. CDS, guidelines and planning overview. openEHR. URL: https://specifications-test.openehr.org/releases/PROC/latest/overview.html [accessed 2023-04-12]
33. Li M, Cai H, Nan S, Li J, Lu X, Duan H. A patient-screening tool for clinical research based on electronic health records using OpenEHR: development study. JMIR Med Inform 2021 Oct 21;9(10):e33192 [FREE Full text] [doi: 10.2196/33192] [Medline: 34673526]

34. Clinical knowledge manager. openEHR. URL: https://ckm.openehr.org/ckm/ [accessed 2024-04-29]

Abbreviations

DL: Decision Language
DLM: Decision Logic Module
EHR: electronic health record
TP: Task Plan
WP: Work Plan

Edited by C Lovis; submitted 21.04.23; peer-reviewed by H Leslie, R Marshall; comments to author 09.01.24; revised version received 03.02.24; accepted 22.07.24; published 16.09.24.

Please cite as:
Ferreira D, Neto C, Hak F, Abelha A, Santos M, Machado J
Standardizing Corneal Transplantation Records Using openEHR: Case Study
JMIR Med Inform 2024;12:e48407
URL: https://medinform.jmir.org/2024/1/e48407
doi:10.2196/48407
PMID:

Clinical Decision Support to Increase Emergency Department Naloxone Coprescribing: Implementation Report

Stuart W Sommers[1], BA; Heather J Tolle[1], MA, PhD; Katy E Trinkley[2,3], PharmD, PhD; Christine G Johnston[4], MD; Caitlin L Dietsche[5], MD; Stephanie V Eldred[5], MD; Abraham T Wick[6], PharmD; Jason A Hoppe[7], DO

1
2
3
4
5
6
7

Corresponding Author:
Stuart W Sommers, BA

Abstract

Background: Coprescribing naloxone with opioid analgesics is a Centers for Disease Control and Prevention (CDC) best practice to mitigate the risk of fatal opioid overdose, yet coprescription by emergency medicine clinicians is rare, occurring less than 5% of the time it is indicated. Clinical decision support (CDS) has been associated with increased naloxone prescribing; however, key CDS design characteristics and pragmatic outcome measures necessary to understand replicability and effectiveness have not been reported.

Objective: This study aimed to rigorously evaluate and quantify the impact of CDS designed to improve emergency department (ED) naloxone coprescribing. We hypothesized CDS would increase naloxone coprescribing and the number of naloxone prescriptions filled by patients discharged from EDs in a large health care system.

Methods: Following user-centered design principles, we designed and implemented a fully automated, interruptive, electronic health record–based CDS to nudge clinicians to coprescribe naloxone with high-risk opioid prescriptions. “High-risk” opioid prescriptions were defined as any opioid analgesic prescription ≥90 total morphine milligram equivalents per day or for patients with a prior diagnosis of opioid use disorder or opioid overdose. The Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework was used to evaluate pragmatic CDS outcomes of reach, effectiveness, adoption, implementation, and maintenance. Effectiveness was the primary outcome of interest and was assessed by (1) constructing a Bayesian structural time-series model of the number of ED visits with naloxone coprescriptions before and after CDS implementation and (2) calculating the percentage of naloxone prescriptions associated with CDS that were filled at an outpatient pharmacy. Mann-Kendall tests were used to evaluate longitudinal trends in CDS adoption. All outcomes were analyzed in R (version 4.2.2; R Core Team).

Implementation (Results): Between November 2019 and July 2023, there were 1,994,994 ED visits. CDS reached clinicians in 0.83% (16,566/1,994,994) of all visits and 15.99% (16,566/103,606) of ED visits where an opioid was prescribed at discharge. Clinicians adopted CDS, coprescribing naloxone in 34.36% (6613/19,246) of alerts. CDS was effective, increasing naloxone coprescribing from baseline by 18.1 (95% CI 17.9 - 18.3) coprescriptions per week or 2,327% (95% CI 3390 - 3490). Patients filled 43.80% (1989/4541) of naloxone coprescriptions. The CDS was implemented simultaneously at every ED and no adaptations were made to CDS postimplementation. CDS was maintained beyond the study period and maintained its effect, with adoption increasing over time (τ=0.454; P<.001).

Conclusions: Our findings advance the evidence that electronic health record–based CDS increases the number of naloxone coprescriptions and improves the distribution of naloxone. Our time series analysis controls for secular trends and strongly suggests that minimally interruptive CDS significantly improves process outcomes.

(JMIR Med Inform 2024;12:e58276) doi:10.2196/58276

KEYWORDS

clinical decision support systems; order sets; drug monitoring; opioid analgesic; opioid use; opioid prescribing; drug overdose; opioid overdose; naloxone; naloxone coprescribing; harm reduction; harm minimization

Introduction

Overdose (OD) deaths decreased in the United States from 2022 to 2023, but 81,083 people still died from opioids in 2023 [1]. Almost 10 million adults misused prescription opioids in 2019 [2], making opioids the most misused prescription drug [3]. Up to 20% of emergency department (ED) visits result in an opioid prescription, and ED opioid prescribing has been associated with increased opioid misuse, abuse, and death [4-10], underscoring the need for ED harm reduction.

Naloxone is an opioid antagonist capable of reversing opioid OD. Naloxone distribution has been associated with reductions in population-level opioid mortality [11,12]. Prescribing naloxone with opioids (naloxone coprescribing) is a Centers for Disease Control and Prevention (CDC) best practice and has been mandated in some states [13,14]. Yet, naloxone coprescribing remains rare [15-17], only occurring 2.3% of the time when >90 morphine milligram equivalents of opioids are ordered from the ED and 7.4% of the time after an ED visit for suspected opioid OD (vs epinephrine which is prescribed in 49% of ED visits for anaphylaxis) [16,17]. Stigma, workload, and time pressures may explain these gaps [18-21].

Health systems have begun implementing strategies to facilitate naloxone coprescribing [22]. Computerized clinical decision support (CDS) is a strategy to assist decision-making and improve health care quality [23,24]. When designed well, CDS have been shown to improve evidence-based prescribing [25-27], as well as opioid OD education and naloxone distribution [28-33]. CDS best practices include increasing specificity and sensitivity, triggering at the right time, making the evidence-based choice the easiest option, and tracking patient outcomes [24,34,35]. Effective CDS implementation requires attention to choice architecture, setting, and best practices to reduce bias and improve adoption [36-38].

We aimed to improve the evidence-based delivery of naloxone by developing and deploying an ED clinician-facing, electronic health record (EHR)–based CDS. We quantified the impact of CDS according to the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework [39]. By specifying the users targeted, including workflow events that triggered CDS, and describing lessons learned, we hope to encourage the deployment and testing of similar CDS beyond our health system.

Methods

Intervention

Following user-centered design (UCD) principles [40], a multidisciplinary team including 5 physicians, 2 pharmacists, and several EHR builders, with expertise in implementation science, informatics, behavioral economics, and health services research, designed a fully automated, EHR-embedded, interruptive, provider-facing CDS. The intervention was beta-tested by several ED clinicians in a practice setting for 6 months before the systemwide rollout. CDS did not interface with any technologies beyond the EHR and fired within typical workflow to recommend and facilitate the addition of a naloxone prescription before the e-signing of any high-risk opioid analgesic prescription order (Multimedia Appendix 1). High-risk criteria were adapted from the 2016 CDC guidelines for chronic pain and defined as any opioid prescription (1) resulting in >90 morphine milligram equivalents per day, (2) for a patient with an opioid use disorder (OUD) diagnosis, or (3) prior opioid OD [41]. CDS searched for Systematized Nomenclature of Medicine Clinical Terms in the "Problem List Diagnosis," "Encounter Diagnosis," and "Hospital Problem Diagnosis" lists. Alerts were suppressed if the patient had an active naloxone prescription or if the patient was discharged to hospice, given patients on end-of-life care are excluded from CDC guidelines [41]. Naloxone prescriptions stayed on the patient's medication list for 1 year.

Key design principles followed were that CDS be intuitive, trigger only when indicated, and default to a preselected naloxone order that was the least expensive option in the health care system's retail pharmacies [24,34,35]. Any provider with prescribing privileges could encounter the alert. Default selection was chosen to decrease work (clicking "Accept" added naloxone to the existing order) and because "opt-out" approaches increase the uptake of target clinical behaviors [42-45].

Accepting CDS was the path of least resistance. However, consistent with nudge theory, clinicians could bypass CDS by (1) selecting "Do Not Order" then "Accept" (2 clicks); (2) selecting prepopulated bypass options ("Doesn't meet criteria," going to "Hospice/SNF," "Already has naloxone") then "Accept" (2 clicks);" or (3) commenting (≥2 clicks) [46]. "Already has naloxone" was included to account for naloxone outside the EHR. Clinicians were returned to their prior workflow after any action.

Clinicians were educated on CDS via departmental meetings and email. Educational materials included (1) CDS rationale, (2) instructions for use, and (3) suggested patient communication. No ongoing education was provided, and no changes were made to CDS after implementation.

Study Design and Setting

This was a retrospective, observational study of ED visits in a large, not-for-profit university-affiliated, nongovernmental health care system. Located in the Rocky Mountain Region, the system has >500,000 total ED visits per year and includes 12 EDs—1 urban-academic level 1 trauma center, 2 urban community hospitals (1 a level 1 trauma center), 2 suburban community level 2 trauma centers, and 7 community free-standing EDs. The study was approved and informed consent was waived by the Colorado Multiple Institutional Review Board (COMIRB). The Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH) were followed (Multimedia Appendix 2) [47].

Data Collection, Measurements, and Outcomes

Naloxone coprescribing was defined as a clinician prescribing opioids and naloxone during the same ED visit. Deidentified patient characteristics (age, sex, race, ethnicity, preferred language, and insurance), CDS data (reasons for firing, number of firings per visit, clinician actions, and bypass reasons), and

clinical variables (whether naloxone was prescribed via CDS and the prescription was filled) were extracted monthly from the shared EHR (Epic Systems). Research data governance was linked to EHR data governance. Clinicians entered data into Epic Hyperspace and CDS responses were automatically registered in real time. Extract, transform, and load processes transferred all patient data into relational databases hosted on private virtually protected servers nightly, and a Microsoft SQL Server Management Studio query was run to further clean and filter research data into Microsoft Excel.

The RE-AIM framework was used to determine the impact of CDS [39]. More explicitly, reach was measured by examining the proportion of ED visits where CDS was triggered and whether patients' characteristics influenced opioid prescribing (and high-risk opioid prescribing, ie, CDS triggering) and naloxone coprescribing. Effectiveness (primary outcome) was assessed by evaluating the number of ED discharges with naloxone coprescriptions per week across the system before and after CDS implementation. Effectiveness was also measured by quantifying the naloxone prescription fill rate (naloxone prescription fills per naloxone orders via CDS vs other workflows) at a 24-hour ED outpatient retail pharmacy in the largest urban academic ED. This subgroup analysis was performed to determine whether increased naloxone orders translated to more naloxone reaching patients and to compare whether patients prescribed naloxone via CDS were more likely to fill their prescriptions than patients prescribed naloxone via other workflows. All prescriptions written at this ED defaulted to the ED's outpatient pharmacy—unless specifically requested by the patient—thus prescription fill data were available in the pharmacy EHR. Adoption was defined as the number of naloxone prescriptions from CDS per number of CDS firings. Due to EHR limitations, we could not measure CDS suppression. The process of implementation is described. Finally, maintenance was judged by whether CDS was maintained after the study period and by modeling changes in adoption over time.

Ethical Considerations

All data releases were cleared by a Research Services Manager who ensured the data being released were compliant with the Health Insurance Portability and Accountability Act (HIPAA) and the corresponding institutional review board exemption (#23-0458). No continuing review was required because this was secondary research and all data were deidentified. Results were shared via secure email. Individual informed patient consent was waived and no compensation was offered, given no patient participation or protected health information was shared.

Data Analysis

There is a documented need for rigorous, pragmatic evaluation when implementing new CDS [34]. Interrupted time series analyses are suggested for CDS evaluation because they control for confounding secular trends [34,48-50]. We used a Bayesian structural time-series model controlling for the number of ED visits to evaluate the impact of CDS on naloxone coprescribing (CausalImpact package; version 1.3.0; Brodersen et al) [51], Mann-Kendall tests to model longitudinal changes in CDS adoption, and chi-square tests to compare the proportions of individuals who triggered CDS and were prescribed either an opioid or opioid with naloxone across demographic categories. Equity of RE-AIM outcomes was evaluated based on patient characteristics because prior research has demonstrated an increased likelihood of opioid prescribing for White patients and increased naloxone prescribing (and coprescribing) for Black and Latine patients [52-58]. Otherwise, frequencies and percentages are reported for categorical variables. All statistical analyses were conducted in R (version 4.2.2; R Core Team) [59].

Study Sample

All ED visits with a discharge opioid prescription between March 2013 and July 2023 were included. Effectiveness was assessed by comparing weekly aggregated counts of ED visits, opioid prescriptions, and naloxone coprescriptions between the pre- (March 2013-November 2019) and postimplementation periods (November 2019-July 2023). Adoption was assessed using only post-period data. The accuracy of synthetic control models, like the Bayesian structural time-series model used, is generally improved by including more pre-period data [51]. Therefore, we extracted enough data to provide a 2:1 pre-to post-period ratio. We did not perform a prospective power calculation. Patients younger than 18 years or older than 90 years old and those who were admitted to the hospital were excluded from both periods.

Implementation (Results)

Demographics

After implementation, between November 2019 and July 2023, there were 1,994,994 eligible ED discharges. Of these, 5.19% (103,606/1,994,994) included an opioid analgesic prescription and 0.83% (16,566/1,994,994) of prescriptions met high-risk criteria. Most visits included female (n=1,083,973, 54.33%), White (n=1,357,153, 68.03%), non-Latine (n=1,519,584, 76.17%), English-speaking (n=1,866,744, 93.57%), and publicly insured (n=1,146,781, 57.48%) patients (Table 1). White, non-Latine, English-speaking, privately insured patients were prescribed opioids a greater proportion of the time compared to Black, Latine, non–English-speaking, and publicly insured patients (P<.001).

Table . Postimplementation visit demographics stratified by visit type.

	Overall: All ED[a] visits (n=1,994,994), n (%)	Subgroup 1: ED visits with an opioid prescription (n=103,606), n (%)	Subgroup 2: ED visits with a CDS[b] alert and a high-risk opioid prescription (n=16,566), n (%)	Subgroup 3: ED visits with a CDS alert and a naloxone coprescription (n=3077), n (%)
Sex				
Female	1,083,973 (54.33)	56,260 (54.30)	8670 (52.34)	1663 (54.05)
Male	905,624 (45.39)	47,094 (45.45)	7858 (47.43)	1405 (45.66)
Other	118 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)
Unknown	5279 (0.26)	252 (0.24)	38 (0.23)	9 (0.29)
Race				
White or Caucasian	1,357,153 (68.03)	78,174 (75.45)	12,724 (76.81)	2278 (74.03)
Black or African American	210,864 (10.57)	7066 (6.82)	1198 (7.23)	273 (8.87)
Other	420,055 (21.06)	18,106 (17.48)	2606 (15.73)	406 (13.19)
Unknown	6922 (0.35)	260 (0.25)	38 (0.23)	120 (3.90)
Ethnicity				
Hispanic, Latine, or Spanish origin	448,959 (22.50)	20,133 (19.43)	2637 (15.92)	565 (18.36)
Non-Hispanic, Latine, or Spanish origin	1,519,584 (76.17)	82,363 (79.50)	13,818 (83.41)	2485 (80.76)
Other	19,395 (0.97)	849 (0.82)	70 (0.42)	0 (0.00)
Unknown	7056 (0.35)	261 (0.25)	41 (0.25)	27 (0.88)
Primary language				
English	1,866,744 (93.57)	98,153 (94.74)	15,934 (96.18)	2936 (95.42)
Spanish	86,390 (4.33)	4130 (3.99)	437 (2.64)	98 (3.18)
Other	34,614 (1/74)	1031 (1.00)	154 (0.93)	33 (1.07)
Unknown	7246 (0.36)	292 (0.28)	41 (0.25)	10 (0.32)
Insurance				
Public	1,146,781 (57.48)	49,374 (47.66)	10,193 (61.53)	1937 (62.95)
Military	33,395 (1.67)	2019 (1.95)	364 (2.20)	57 (1.85)
Indigent	150,202 (7.53)	9183 (8.86)	913 (5.51)	193 (6.27)
Private or other	664,616 (33.31)	43030 (41.53)	5096 (30.76)	876 (28.47)
Unknown	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)

[a]ED: emergency department.

[b]CDS: clinical decision support.

Reach

CDS fired in 0.83% (16,566/1,994,994) of all ED visits. A total of 15.99% (16,566/103,606) of visits with a discharge opioid prescription met high-risk criteria and triggered CDS. CDS fired multiple times in 13.17% (2182/16,566) of visits (mean 1; median 1); ED clinicians interacted with CDS 19,246 times overall. Visits triggering CDS most often involved patients who were female (n=8670, 52.34%), White (n=12,724, 76.81%), non-Latine (n=13,818, 83.41%), spoke English (n=15,934, 96.18%), and had Medicaid or Medicare (n=10,193, 61.53%; Table 1). However, adjusting for the number of visits with an opioid prescription (a prerequisite for CDS triggering), CDS was more likely to trigger in visits with male, Latine, English speaking, and publicly insured patients (P<.001).

Effectiveness

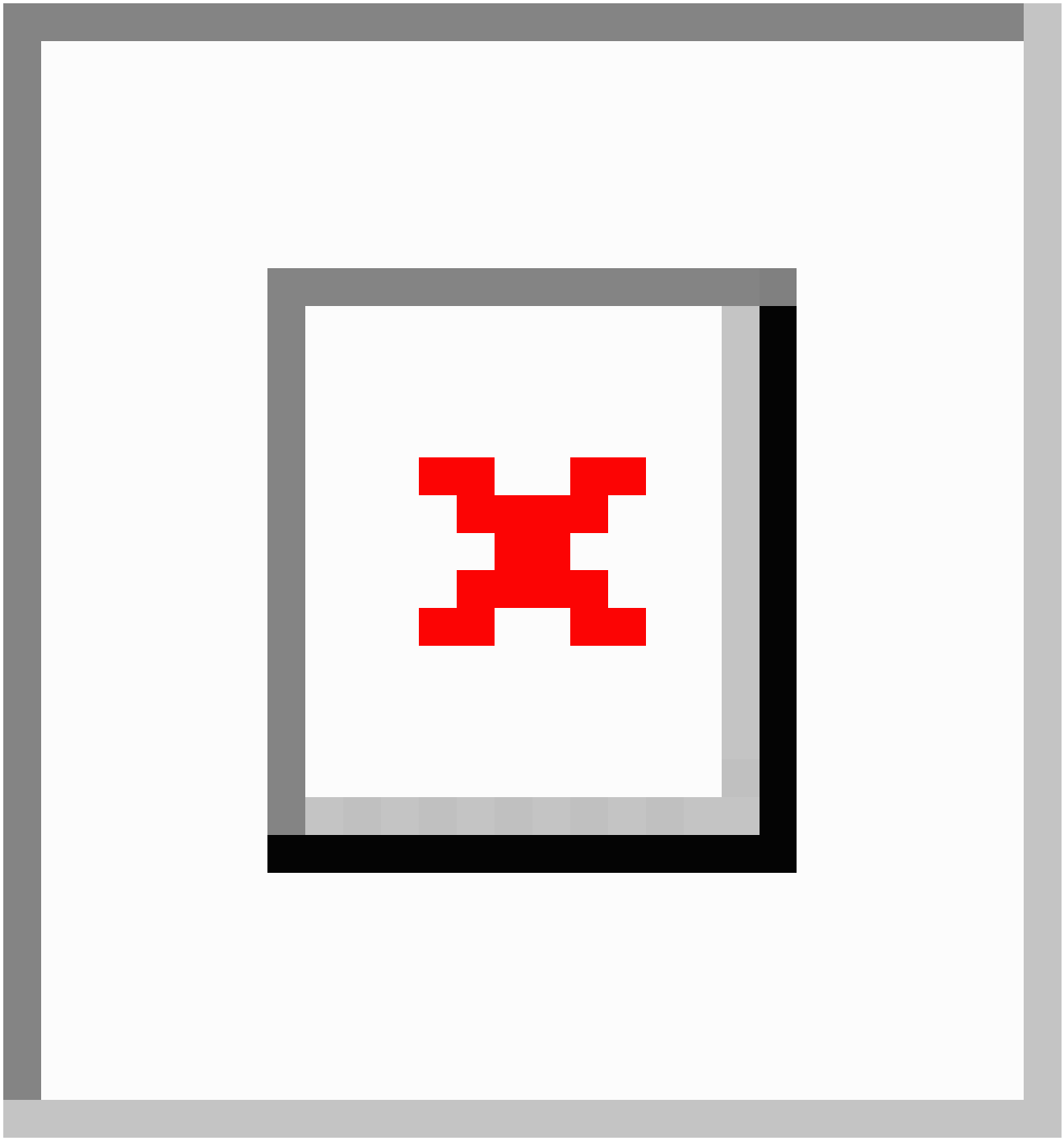
Before CDS implementation, clinicians coprescribed naloxone in 0.05% (156/318,216) of ED visits when an opioid analgesic was prescribed. After CDS implementation, ED clinicians coprescribed naloxone in 3.49% (3616/103,606) of ED visits when an opioid analgesic was prescribed. In the postimplementation period, 85.09% (3077/3616) of naloxone coprescriptions originated from CDS.

Using the number of ED visits as a covariate, the CausalImpact package predicted 0.80 (95% CI 0.55 - 1.05) ED visits with

naloxone coprescriptions per week and 150 (95% CI 100 - 200) ED visits with naloxone coprescriptions in the postimplementation period. After CDS go-live there was an immediate increase in the number of ED visits with naloxone coprescriptions each week—18.9 ED visits with naloxone coprescriptions observed on average weekly and 3616 ED visits with naloxone coprescriptions in the entire postimplementation period (Figure 1). In other words, CDS increased ED naloxone coprescribing by 18.1 (95% CI 17.85 - 18.34) naloxone coprescriptions per week or 2327% (95% CI 1702 - 3335). Black and non-Latine patients were more likely to have naloxone coprescribed when CDS triggered compared to White and Latine patients (P<.001).

During the postimplementation period, there were 4541 naloxone coprescriptions with opioid analgesics written at the urban, academic ED (mean 1.2; median 1 per visit) and 3308 (72.85%) were ordered from a CDS alert. Patients filled 49.42% (2134/4318) of their opioid prescriptions and 43.80% (1989/4541) of naloxone coprescriptions. Patients coprescribed naloxone via CDS filled their prescriptions less often than patients coprescribed naloxone via other workflows (35.64%, 1179/3308 vs 65.69%, 810/1233).

Figure 1. The CausalImpact plot of naloxone coprescribing. CDS: clinical decision support.



Adoption

ED clinicians adopted CDS, following the recommendation to coprescribe naloxone in 34.36% (6613/19,246) of alerts. Clinicians at the academic ED adopted CDS at a higher rate 61.62% (2005/3254) than at community EDs 34.70% (4608/13,280).

Implementation

This CDS was implemented simultaneously at every ED and no changes were made to CDS postimplementation. All EDs used the same EHR, and it took a CDS builder 70 hours (including meetings, communications, and build time) to design and implement CDS.

Maintenance

According to the Mann-Kendall test, CDS adoption increased over time (τ=0.454; P<.001). Because no changes were made to CDS, there were no obvious sustainability costs beyond what our health system regularly paid for EHR access. CDS is still active and currently being scaled to outpatient clinics. Sustainability decisions are made by local governance based on naloxone prescribing because it was defined as a CDC best practice. CDS are reviewed ad hoc based on technical issues and yearly otherwise.

Lessons Learned

The implementation process benefitted from the makeup of the study team, who were able to provide local context for design, identify key workflow needs, address local barriers, and serve as champions during implementation. Beta testing and CDS-specific data analytics were prioritized to identify technical and efficiency issues early. Having data analytics built and collecting data during testing was key for providing estimates on workflow interruptiveness and informing iterative improvements. For example, monitoring revealed CDS initially only searched the current visit diagnosis, failing to identify histories of OUD and OD. The trigger algorithm was changed to include any EHR documented history of OUD or OD before going live, with a significant increase in case identification. Additionally, because clinicians told champions that CDS were firing "too late," CDS were modified to trigger when clinicians entered as opposed to signed orders, facilitating clinician-patient communication before prescribing.

This project began as quality improvement, which was important for local buy-in. Also, the health system is funded, and therefore, owns the intervention. It would have been ideal to prospectively track implementation to elucidate system and per-patient costing and inform decisions about CDS maintenance. Future studies should formally evaluate patient-centered outcomes to confirm CDS as an effective and equitable implementation strategy.

Discussion

Principal Findings

A minimally interruptive CDS was readily adopted, showed a sustained effect, and significantly increased the number of ED naloxone coprescriptions. These findings support CDS as an effective implementation strategy to increase clinician uptake of naloxone best practices.

The high rate of adoption supports the need for user-centered CDS development, monitoring, and evaluation as the impact of CDS is often limited by low adoption and frequent workflow interruptions resulting in "alert fatigue" (the desensitization to important safety warnings) [24,60-64]. A Cochrane review of 122 CDS trials showed that CDS, on average, only increases the proportion of patients receiving desired care by 5.8% (95% CI 4.0% to 7.6%) [36]. The impact is variable, with the top quartile of reported improvements ranging from 10% to 62% [36]. With an adoption rate of 34.36% and a 2327% increase in the number of ED visits with naloxone coprescriptions, this CDS falls well within the top quartile of CDS improvements [36]. Interestingly, adoption increased over time. This finding differs from most other CDS literature reporting a decrease in adoption over time [65], and mirrors one other CDS study that reported a similar effect after UCD [66], perhaps suggesting that UCD improves initial and sustained adoption [66].

ED clinicians face increasingly complex workflow challenges that require validated solutions [67,68]. Previous evaluations of naloxone coprescribing CDS have not always aligned with best practices for designing, conducting, and reporting CDS interventions [28-31,34]. Prior studies have not discussed the rationale for CDS design (such as choice architecture) and have excluded key operational details (supplements and alert screenshots), making it challenging to reproduce or scale CDS [28-31,34]. The default order design of our CDS may have contributed to CDS acceptability by making choice architecture less burdensome to clinicians [34]. CDS adoption may also reflect actions in line with clinicians' and patients' positive attitudes toward naloxone prescribing and use [15,34,62].

Our Bayesian structural time series model, without a statutory mandate, offers robust evidence to support claims that CDS increases ED naloxone coprescribing. Our methods address the gap from prior studies that relied on pre-post designs and inferential statistics (logistic regression, t tests, and χ^2 tests) [30,31], which increase the risk of confounding by organizational policies, regulations, or reimbursement rules [34].

The fact that White, non-Latine, English-speaking, and privately insured patients were significantly more likely to have an opioid prescribed is concerning but consistent with prior literature [52-58]. Demographic differences in opioid and naloxone prescribing have been widely reported [52-58]. It is notable that Black and non-Hispanic patients were more likely to have naloxone coprescribed after CDS was triggered. This is the first study to report demographic differences in clinicians' responses to CDS designed to increase naloxone coprescribing. Although, Black and Latine patients are coprescribed naloxone more often at baseline. Thus, it is possible CDS increased naloxone coprescribing equally and simply failed to reduce the influence of racial and ethnic bias on opioid or naloxone prescribing [57,58]. Other CDS designers should consider these differences when implementing and evaluating CDS to ensure they do not inadvertently maintain or widen existing disparities.

Limitations

No clinical outcomes were measured, so we do not know if practice changes impacted care such as ED readmissions. No statutory mandates were implemented during this study, but we cannot be sure local educational efforts were not made to encourage naloxone coprescribing. Larger trends in opioid prescribing were not examined but are unlikely to have impacted the rate of naloxone coprescribing.

The availability of a 24-hour ED pharmacy at the academic site was another potential operational confounder in measuring naloxone fill rates. Discharged patients had to pass the pharmacy to exit the ED. This is an important consideration for sustainability since we do not compare naloxone coprescribing versus take-home naloxone (THN). THN has been reported to improve naloxone distribution by removing the need to stop at a pharmacy and may alleviate patient costs but shift medication costs to systems or public health organizations. Prior work, evaluating THN programs, has reported naloxone distribution rates as high as 87.3% [69]. However, THN programs are resource intensive [69,70], thus, might still be improved by CDS that improve the recognition of patients at risk for OD [71].

Conclusions

An EHR-based CDS encouraging ED naloxone coprescribing with opioid analgesics increased alert-based naloxone orders and overall system rates of naloxone coprescribing. The CDS had a low rate of interruption, a high rate of adoption [36], and significantly increased ED naloxone coprescribing across 12 EDs. There were no obvious sustainability costs beyond what the health system regularly paid for EHR access. These findings support claims that health care system leaders should consider CDS as an implementation strategy to address the significant gap in naloxone coprescribing [28-33].

Acknowledgments

The authors would also like to thank Natalia Truszczynski for their help editing the paper. They would like to thank Wyatt Tarter, Zhixin Lun, and the Center for Innovative Design and Analysis (CIDA) for their help in validating the statistical analyses.

Authors' Contributions

CGJ, SVE, CLD, and ATW designed and implemented the clinical decision support (CDS). SWS, HJT, and JAH conceived and operationalized the trial. SWS and HJT analyzed the data. SWS drafted the paper with contributions from JAH, HJT, CGJ, SVE, CLD, KET, and ATW. SWS takes responsibility for the paper as a whole.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Wireframe of CDS alert recommending naloxone coprescription. CDS: clinical decision support.
[JPG File, 233 KB - medinform_v12i1e58276_app1.jpg]

Multimedia Appendix 2
Completed Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH)
[PDF File, 244 KB - medinform_v12i1e58276_app2.pdf]

References

1. U.S. overdose deaths decrease in 2023, first time since 2018. Centers for Disease Control. 2024. URL: https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2024/20240515.htm#:~:text=The%20new%20data%20show%20overdose,psychostimulants%20(like%20methamphetamine)%20increased [accessed 2024-10-22]
2. Key substance use and mental health indicators in the United States: results from the 2019 national survey on drug use and health. (HHS publication no. PEP20-07-01-001, NSDUH series H-55). Substance Abuse and Mental Health Services Administration. 2020. URL: https://www.samhsa.gov/data/ [accessed 2024-10-22]
3. Driver C, Kean B, Oprescu F, Lovell GP. Knowledge, behaviors, attitudes and beliefs of physiotherapists towards the use of psychological interventions in physiotherapy practice: a systematic review. Disabil Rehabil 2017 Nov;39(22):2237-2249. [doi: 10.1080/09638288.2016.1223176] [Medline: 27635464]
4. Ali MM, Cutler E, Mutter R, et al. Opioid prescribing rates from the emergency department: down but not out. Drug Alcohol Depend 2019 Dec 1;205:107636. [doi: 10.1016/j.drugalcdep.2019.107636] [Medline: 31704377]
5. Hoppe JA, Kim H, Heard K. Association of emergency department opioid initiation with recurrent opioid use. Ann Emerg Med 2015 May;65(5):493-499. [doi: 10.1016/j.annemergmed.2014.11.015] [Medline: 25534654]
6. Bindman AB, Grumbach K, Keane D, Rauch L, Luce JM. Consequences of queuing for care at a public hospital emergency department. JAMA 1991 Aug 28;266(8):1091-1096. [Medline: 1865541]

7. Niska R, Bhuiya F, Xu J. National hospital ambulatory medical care survey: 2007 emergency department summary. Natl Health Stat Report 2010 Aug 6(26):1-31. [Medline: 20726217]
8. Chang HY, Daubresse M, Kruszewski SP, Alexander GC. Prevalence and treatment of pain in EDs in the United States, 2000 to 2010. Am J Emerg Med 2014 May;32(5):421-431. [doi: 10.1016/j.ajem.2014.01.015] [Medline: 24560834]
9. McLeod D, Nelson K. The role of the emergency department in the acute management of chronic or recurrent pain. Australas Emerg Nurs J 2013 Feb;16(1):30-36. [doi: 10.1016/j.aenj.2012.12.001] [Medline: 23622554]
10. Bohnert ASB, Valenstein M, Bair MJ, et al. Association between opioid prescribing patterns and opioid overdose-related deaths. JAMA 2011 Apr 6;305(13):1315-1321. [doi: 10.1001/jama.2011.370] [Medline: 21467284]
11. Chimbar L, Moleta Y. Naloxone effectiveness: a systematic review. J Addict Nurs 2018;29(3):167-171. [doi: 10.1097/JAN.0000000000000246]
12. Cataife G, Dong J, Davis CS. Regional and temporal effects of naloxone access laws on opioid overdose mortality. Subst Abus 2021;42(3):329-338. [doi: 10.1080/08897077.2019.1709605] [Medline: 31951788]
13. Dowell D, Ragan KR, Jones CM, Baldwin GT, Chou R. CDC clinical practice guideline for prescribing opioids for pain - United States, 2022. MMWR Recomm Rep 2022 Nov 4;71(3):1-95. [doi: 10.15585/mmwr.rr7103a1] [Medline: 36327391]
14. Green TC, Davis C, Xuan Z, Walley AY, Bratberg J. Laws mandating coprescription of naloxone and their impact on naloxone prescription in five US states, 2014-2018. Am J Public Health 2020 Jun;110(6):881-887. [doi: 10.2105/AJPH.2020.305620] [Medline: 32298179]
15. Mueller SR, Koester S, Glanz JM, Gardner EM, Binswanger IA. Attitudes toward naloxone prescribing in clinical settings: a qualitative study of patients prescribed high dose opioids for chronic non-cancer pain. J Gen Intern Med 2017 Mar;32(3):277-283. [doi: 10.1007/s11606-016-3895-8] [Medline: 27798775]
16. Chua KP, Dahlem CHY, Nguyen TD, et al. Naloxone and buprenorphine prescribing following US emergency department visits for suspected opioid overdose: August 2019 to April 2021. Ann Emerg Med 2022 Mar;79(3):225-236. [doi: 10.1016/j.annemergmed.2021.10.005] [Medline: 34802772]
17. Jones CM, Compton W, Vythilingam M, Giroir B. Naloxone co-prescribing to patients receiving prescription opioids in the medicare part D program, United States, 2016-2017. JAMA 2019 Aug 6;322(5):462-464. [doi: 10.1001/jama.2019.7988] [Medline: 31386124]
18. Eswaran V, Allen KC, Bottari DC, et al. Take-home naloxone program implementation: lessons learned from seven Chicago-area hospitals. Ann Emerg Med 2020 Sep;76(3):318-327. [doi: 10.1016/j.annemergmed.2020.02.013] [Medline: 32241746]
19. Gunn AH, Smothers ZPW, Schramm-Sapyta N, Freiermuth CE, MacEachern M, Muzyk AJ. The emergency department as an opportunity for naloxone distribution. West J Emerg Med 2018 Nov;19(6):1036-1042. [doi: 10.5811/westjem.2018.8.38829] [Medline: 30429939]
20. Holland TJ, Penm J, Dinh M, Aran S, Chaar B. Emergency department physicians' and pharmacists' perspectives on take-home naloxone. Drug Alcohol Rev 2019 Feb;38(2):169-176. [doi: 10.1111/dar.12894] [Medline: 30697852]
21. Holland TJ, Penm J, Johnson J, Sarantou M, Chaar BB. Stakeholders' perceptions of factors influencing the use of take-home-naloxone. Pharmacy (Basel) 2020 Dec 3;8(4):232. [doi: 10.3390/pharmacy8040232] [Medline: 33287294]
22. Stein BD, Smart R, Jones CM, Sheng F, Powell D, Sorbero M. Individual and community factors associated with naloxone co-prescribing among long-term opioid patients: a retrospective analysis. J Gen Intern Med 2021 Oct;36(10):2952-2957. [doi: 10.1007/s11606-020-06577-5] [Medline: 33598891]
23. Berner ES. Clinical decision support systems: state of the art. AHRQ. 2009. URL: https://www.healthit.ahrq.gov/sites/default/files/docs/page/09-0069-EF_1.pdf [accessed 2024-10-22]
24. Trinkley KE, Blakeslee WW, Matlock DD, et al. Clinician preferences for computerised clinical decision support for medications in primary care: a focus group study. BMJ Health Care Inform 2019 Apr;26(1):0. [doi: 10.1136/bmjhci-2019-000015] [Medline: 31039120]
25. White RH, Hong R, Venook AP, et al. Initiation of warfarin therapy: comparison of physician dosing with computer-assisted dosing. J Gen Intern Med 1987;2(3):141-148. [doi: 10.1007/BF02596140] [Medline: 3295148]
26. Chertow GM, Lee J, Kuperman GJ, et al. Guided medication dosing for inpatients with renal insufficiency. JAMA 2001 Dec 12;286(22):2839-2844. [doi: 10.1001/jama.286.22.2839] [Medline: 11735759]
27. Mungall DR, Anbe D, Forrester PL, et al. A prospective randomized comparison of the accuracy of computer-assisted versus GUSTO nomogram--directed heparin therapy. Clin Pharmacol Ther 1994 May;55(5):591-596. [doi: 10.1038/clpt.1994.73] [Medline: 8181203]
28. Oliva EM, Christopher MLD, Wells D, et al. Opioid overdose education and naloxone distribution: Development of the Veterans Health Administration's national program. J Am Pharm Assoc (2003) 2017 Mar;57(2):S168-S179. [doi: 10.1016/j.japh.2017.01.022]
29. Duan L, Lee MS, Adams JL, Sharp AL, Doctor JN. Opioid and naloxone prescribing following insertion of prompts in the electronic health record to encourage compliance with California State Opioid Law. JAMA Netw Open 2022 May 2;5(5):e229723. [doi: 10.1001/jamanetworkopen.2022.9723] [Medline: 35499826]
30. Srikumar JK, Daniel K, Balasanova AA. Implementation of a naloxone best practice advisory into an electronic health record. J Addict Med 2023;17(3):346-348. [doi: 10.1097/ADM.0000000000001102] [Medline: 37267187]

31. Nelson SD, McCoy AB, Rector H, et al. Assessment of a naloxone coprescribing alert for patients at risk of opioid overdose: a quality improvement project. Anesth Analg 2022 Jul 1;135(1):26-34. [doi: 10.1213/ANE.0000000000005976] [Medline: 35343932]
32. Funke M, Kaplan MC, Glover H, et al. Increasing naloxone prescribing in the emergency department through education and electronic medical record work-aids. Jt Comm J Qual Patient Saf 2021 Jun;47(6):364-375. [doi: 10.1016/j.jcjq.2021.03.002] [Medline: 33811002]
33. Marino R, Landau A, Lynch M, Callaway C, Suffoletto B. Do electronic health record prompts increase take-home naloxone administration for emergency department patients after an opioid overdose? Addiction 2019 Sep;114(9):1575-1581. [doi: 10.1111/add.14635] [Medline: 31013394]
34. Kawamoto K, McDonald CJ. Designing, conducting, and reporting clinical decision support studies: recommendations and call to action. Ann Intern Med 2020 Jun 2;172(11 Suppl):S101-S109. [doi: 10.7326/M19-0875] [Medline: 32479177]
35. Institute of Medicine (US) Committee on Quality of Health Care in America. In: Kohn LT, Corrigan JM, Donaldson MS, editors. Committee on Quality of Health Care in America: National Academies Press (US); 2000. [doi: 10.17226/9728]
36. Kwan JL, Lo L, Ferguson J, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. BMJ 2020 Sep 17;370:m3216. [doi: 10.1136/bmj.m3216] [Medline: 32943437]
37. Goddard K, Roudsari A, Wyatt JC. Automation bias - a hidden issue for clinical decision support system use. Stud Health Technol Inform 2011;164:17-22. [Medline: 21335682]
38. Gurupur V, Wan TTH. Inherent bias in artificial intelligence-based decision support systems for healthcare. Medicina (Kaunas) -> Med Kaunas 2020 Mar 20;56(3):141. [doi: 10.3390/medicina56030141] [Medline: 32244930]
39. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
40. Ray JM, Ahmed OM, Solad Y, et al. Computerized clinical decision support system for emergency department-initiated buprenorphine for opioid use disorder: user-centered design. JMIR Hum Factors 2019 Feb 27;6(1):e13121. [doi: 10.2196/13121] [Medline: 30810531]
41. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain—United States, 2016. JAMA 2016 Apr 19;315(15):1624-1645. [doi: 10.1001/jama.2016.1464] [Medline: 26977696]
42. Samuelson W, Zeckhauser R. Status quo bias in decision making. J Risk Uncertainty 1988 Mar;1(1):7-59. [doi: 10.1007/BF00055564]
43. Kahneman D, Knetsch JL, Thaler RH. Anomalies: the endowment effect, loss aversion, and status quo bias. J Econ Perspect 1991 Feb 1;5(1):193-206. [doi: 10.1257/jep.5.1.193]
44. Delgado MK, Shofer FS, Patel MS, et al. Association between electronic medical record implementation of default opioid prescription quantities and prescribing behavior in two emergency departments. J Gen Intern Med 2018 Apr;33(4):409-411. [doi: 10.1007/s11606-017-4286-5] [Medline: 29340937]
45. Halpern SD, Loewenstein G, Volpp KG, et al. Default options in advance directives influence how patients set goals for end-of-life care. Health Aff (Millwood) 2013 Feb;32(2):408-417. [doi: 10.1377/hlthaff.2012.0895] [Medline: 23381535]
46. Thaler RH, Sunstein CR. Nudge: Improving Decisions about Health, Wealth and Happiness: Penguin; 2009.
47. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694. [doi: 10.2196/46694] [Medline: 37163336]
48. Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr 2013;13(6 Suppl):S38-S44. [doi: 10.1016/j.acap.2013.08.002] [Medline: 24268083]
49. Hanbury A, Farley K, Thompson C, Wilson PM, Chambers D, Holmes H. Immediate versus sustained effects: interrupted time series analysis of a tailored intervention. Implement Sci 2013 Nov 5;8:130. [doi: 10.1186/1748-5908-8-130] [Medline: 24188718]
50. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care 2003;19(4):613-623. [doi: 10.1017/s0266462303000576] [Medline: 15095767]
51. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL. Inferring causal impact using Bayesian structural time-series models. Ann Appl Stat 2015;9(1):247-274. [doi: 10.1214/14-AOAS788]
52. Kang H, Zhang P, Lee S, Shen S, Dunham E. Racial disparities in opioid administration and prescribing in the emergency department for pain. Am J Emerg Med 2022 May;55:167-173. [doi: 10.1016/j.ajem.2022.02.043]
53. Papp J, Emerman C. Disparities in emergency department naloxone and buprenorphine initiation. West J Emerg Med 2023 Jun 30;24(4):710-716. [doi: 10.5811/westjem.58636] [Medline: 37527392]
54. Engel-Rebitzer E, Dolan AR, Aronowitz SV, et al. Patient preference and risk assessment in opioid prescribing disparities: a secondary analysis of a randomized clinical trial. JAMA Netw Open 2021 Jul 1;4(7):e2118801. [doi: 10.1001/jamanetworkopen.2021.18801] [Medline: 34323984]
55. Pletcher MJ, Kertesz SG, Kohn MA, Gonzales R. Trends in opioid prescribing by race/ethnicity for patients seeking care in US emergency departments. JAMA 2008 Jan 2;299(1):70-78. [doi: 10.1001/jama.2007.64] [Medline: 18167408]

56. Crowley AP, Sun C, Yan XS, et al. Disparities in emergency department and urgent care opioid prescribing before and after randomized clinician feedback interventions. Acad Emerg Med 2023 Aug;30(8):809-818. [doi: 10.1111/acem.14717] [Medline: 36876410]
57. Weiner SG, Carroll AD, Brisbon NM, et al. Evaluating disparities in prescribing of naloxone after emergency department treatment of opioid overdose. J Subst Abuse Treat 2022 Aug;139:108785. [doi: 10.1016/j.jsat.2022.108785] [Medline: 35537918]
58. Lin LA, Brummett CM, Waljee JF, Englesbe MJ, Gunaseelan V, Bohnert ASB. Association of opioid overdose risk factors and naloxone prescribing in US adults. J Gen Intern Med 2020 Feb;35(2):420-427. [doi: 10.1007/s11606-019-05423-7] [Medline: 31820218]
59. Team RC. R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing; 2022.
60. Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. JAMA 1998 Oct 21;280(15):1339-1346. [doi: 10.1001/jama.280.15.1339] [Medline: 9794315]
61. Hemens BJ, Holbrook A, Tonkin M, et al. Computerized clinical decision support systems for drug prescribing and management: a decision-maker-researcher partnership systematic review. Impl Sci 2011 Aug 3;6:89. [doi: 10.1186/1748-5908-6-89] [Medline: 21824383]
62. Roshanov PS, Misra S, Gerstein HC, et al. Computerized clinical decision support systems for chronic disease management: a decision-maker-researcher partnership systematic review. Impl Sci 2011 Aug 3;6:92. [doi: 10.1186/1748-5908-6-92] [Medline: 21824386]
63. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 1;18(3):327-334. [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
64. Yoo J, Lee J, Rhee PL, et al. Alert override patterns with a medication clinical decision support system in an academic emergency department: retrospective descriptive study. JMIR Med Inform 2020 Nov 4;8(11):e23351. [doi: 10.2196/23351] [Medline: 33146626]
65. McCullagh L, Mann D, Rosen L, Kannry J, McGinn T. Longitudinal adoption rates of complex decision support tools in primary care. Evid Based Med 2014 Dec;19(6):204-209. [doi: 10.1136/ebmed-2014-110054] [Medline: 25238769]
66. Trinkley KE, Wright G, Allen LA, et al. Sustained effect of clinical decision support for heart failure: a natural experiment using implementation science. Appl Clin Inform 2023 Oct;14(5):822-832. [doi: 10.1055/s-0043-1775566] [Medline: 37852249]
67. Southon FC, Sauer C, Grant CN. Information technology in complex health services: organizational impediments to successful technology transfer and diffusion. J Am Med Inform Assoc 1997;4(2):112-124. [doi: 10.1136/jamia.1997.0040112] [Medline: 9067877]
68. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Shortlife EH, Cimino JJ, editors. Biomedical Informatics: Computer Applications in Health Care and Biomedicine: Springer; 2021:795-840.
69. Moore PQ, Cheema N, Celmins LE, et al. Point-of-care naloxone distribution in the emergency department: a pilot study. Am J Health Syst Pharm 2021 Feb 8;78(4):360-366. [doi: 10.1093/ajhp/zxaa409] [Medline: 33555343]
70. Ramdin C, Chandran K, Nelson L, Mazer-Amirshahi M. Trends in naloxone prescribed at emergency department discharge: a national analysis (2012-2019). Am J Emerg Med 2023 Mar;65:162-167. [doi: 10.1016/j.ajem.2023.01.006] [Medline: 36638613]
71. Jacka BP, Ziobrowski HN, Lawrence A, et al. Implementation and maintenance of an emergency department naloxone distribution and peer recovery specialist program. Acad Emerg Med 2022 Mar;29(3):294-307. [doi: 10.1111/acem.14409] [Medline: 34738277]

Abbreviations

CDC: Centers for Disease Control and Prevention
CDS: clinical decision support
COMIRB: Colorado Multiple Institutional Review Board
ED: emergency department
HIPAA: Health Insurance Portability and Accountability Act
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
OD: overdose
OUD: opioid use disorder
RE-AIM: Reach, Effectiveness, Adoption, Implementation, and Maintenance
THN: take-home naloxone
UCD: user-centered design

Edited by C Perrin; submitted 12.03.24; peer-reviewed by E Cowan, WJ Peppard; revised version received 02.09.24; accepted 11.09.24; published 06.11.24.

Please cite as:
Sommers SW, Tolle HJ, Trinkley KE, Johnston CG, Dietsche CL, Eldred SV, Wick AT, Hoppe JA
Clinical Decision Support to Increase Emergency Department Naloxone Coprescribing: Implementation Report
JMIR Med Inform 2024;12:e58276
URL: https://medinform.jmir.org/2024/1/e58276
doi:10.2196/58276

Completion Rate and Satisfaction With Online Computer-Assisted History Taking Questionnaires in Orthopedics: Multicenter Implementation Report

Casper Craamer[1,2], MSc; Thomas Timmers[1,2], PhD; Michiel Siebelt[3], MD, PhD; Rudolf Bertijn Kool[2], MD, PhD; Carel Diekerhof[4], MD; Jan Jacob Caron[4], MD; Taco Gosens[4,5], MD, PhD; Walter van der Weegen[3], PhD

1

2

3

4

5

Corresponding Author:
Casper Craamer, MSc

Abstract

Background: Collecting the medical history during a first outpatient consultation plays an important role in making a diagnosis. However, it is a time-consuming process, and time is scarce in today's health care environment. The computer-assisted history taking (CAHT) systems allow patients to share their medical history electronically before their visit. Although multiple advantages of CAHT have been demonstrated, adoption in everyday medical practice remains low, which has been attributed to various barriers.

Objective: This study aimed to implement a CAHT questionnaire for orthopedic patients in preparation for their first outpatient consultation and analyze its completion rate and added value.

Methods: A multicenter implementation study was conducted in which all patients who were referred to the orthopedic department were invited to self-complete the CAHT questionnaire. The primary outcome of the study is the completion rate of the questionnaire. Secondary outcomes included patient and physician satisfaction. These were assessed via surveys and semistructured interviews.

Implementation (Results): In total, 5321 patients were invited, and 4932 (92.7%) fully completed the CAHT questionnaire between April 2022 and July 2022. On average, participants (n=224) rated the easiness of completing the questionnaire at 8.0 (SD 1.9; 0 - 10 scale) and the satisfaction of the consult at 8.0 (SD 1.7; 0 - 10 scale). Satisfaction with the outpatient consultation was higher in cases where the given answers were used by the orthopedic surgeon during this consultation (median 8.3, IQR 8.0 - 9.1 vs median 8.0, IQR 7.0 - 8.5; *P*<.001). Physicians (n=15) scored the average added value as 7.8 (SD 1.7; 0 - 10 scale) and unanimously recognized increased efficiency, better patient engagement, and better medical record completeness. Implementing the patient's answers into the electronic health record was deemed necessary.

Conclusions: In this study, we have shown that previously recognized barriers to implementing and adapting CAHT can now be effectively overcome. We demonstrated that almost all patients completed the CAHT questionnaire. This results in reported improvements in both the efficiency and personalization of outpatient consultations. Given the pressing need for personalized health care delivery in today's time-constrained medical environment, we recommend implementing CAHT systems in routine medical practice.

(JMIR Med Inform 2024;12:e60655) doi:10.2196/60655

KEYWORDS

computer-assisted history taking; history taking; digital medical interview; orthopedics; digital health; computer-assisted; cohort study; orthopedic; outpatient; satisfaction; patient engagement; medical record

Introduction

Background

The patient's medical history plays a crucial role in establishing an accurate diagnosis [1-3]. However, collecting the medical history during a first consultation is time-consuming, and time is scarce in today's health care environment. In addition, the first consultation can be a stressful event for a patient, resulting in anxiety and misinterpretation of the questions asked during a medical encounter [4]. Subsequently, this can potentially result in incomplete and invalid information, hindering a patient's ability to participate in shared decision-making [4].

Computer-assisted history taking (CAHT) systems, also known as digital medical interview assistant systems, are software programs that allow patients to present their medical history electronically before an outpatient consultation. For instance, this can be done remotely via a web-based portal or smartphone app prior to the scheduled consultation [5]. CAHT was first introduced in the early 1970s as an additional channel to collect highly relevant, comprehensive, and accurate patient information [6]. Multiple advantages of CAHT have been demonstrated, including saving face-to-face consultation time spent on history taking and empowering patients to be active in their own care [7]. Moreover, CAHT might enhance the comprehensiveness of patient history taking by employing standardized algorithms that expand questioning depth based on the participant's responses [8]. This approach holds the potential to uncover psychosocial and psychiatric issues potentially associated with the presenting complaint [9].

Although these findings are promising, the adoption rate of CAHT within health care remains low. This is attributed to various barriers for both health care professionals (HCPs) and patients [6]. The accessibility of health care for all comes into question while digitalizing health care. Additionally, concerns arise regarding the interoperability of data that is fragmented across multiple compartments. The ability of patients to provide accurate answers is also brought into focus when they are consulted via an online survey rather than an in-person consultation [10].

Despite these barriers in integrating CAHT into everyday medical practice, the current pressure on the health care system demands action. Since the projected growth of multiple patient populations by far exceeds the number of available HCPs in the near future, the time spent on each patient needs to be as efficient and effective as possible, without reducing (perceived) health care quality. Given the number of patients that nowadays have access to email, websites, and smart devices, and the unprecedented advances in technologies in recent years, a more successful implementation of CAHT in clinical practice could be expected. However, no research has been published about achieving this goal.

Objectives

The aims of this study were to implement an online orthopedic CAHT questionnaire that enables patients to provide their medical history prior to their first outpatient consultation and to integrate the CAHT system into the electronic health records (EHRs) of two Dutch hospitals. We subsequently analyzed the completion rate of the CAHT questionnaire, as well as HCPs' satisfaction with the collected information, its accessibility, and its added value.

Methods

Study Design and Setting

This multicenter implementation study was conducted at the orthopedic departments of the Anna Hospital (Geldrop, The Netherlands) and Elisabeth-Tweesteden Hospital (Tilburg, The Netherlands). No changes were made to the design after the study was commenced. We followed the implementation guidelines for the reporting on digital health implementations [11].

Ethical Considerations

Approval was obtained from the medical ethics committees of Anna Hospital and Elisabeth-Tweesteden Hospital. The study was exempted from the Medical Research Involving Human Subjects Act (WMO, N23.090). Patients were informed about data usage for research and publication purposes at the start of the CAHT questionnaire, with participation being voluntary.

Participant Selection

All patients aged 18 years and older who were referred to the orthopedic departments of the participating centers for their first in-hospital consultation were invited to participate in the study. Patients needed to have an email address and sufficient command of the Dutch language. Patients with a cognitive disorder and patients specifically referred for pediatric orthopedics were excluded. Inclusion criteria were assessed by hospital staff when they contacted the patients to schedule their appointment.

All physicians and orthopedic residents (n=24) working in the participating hospitals and who have had scheduled initial consultations with patients during the study period were invited to participate in the study as well [12].

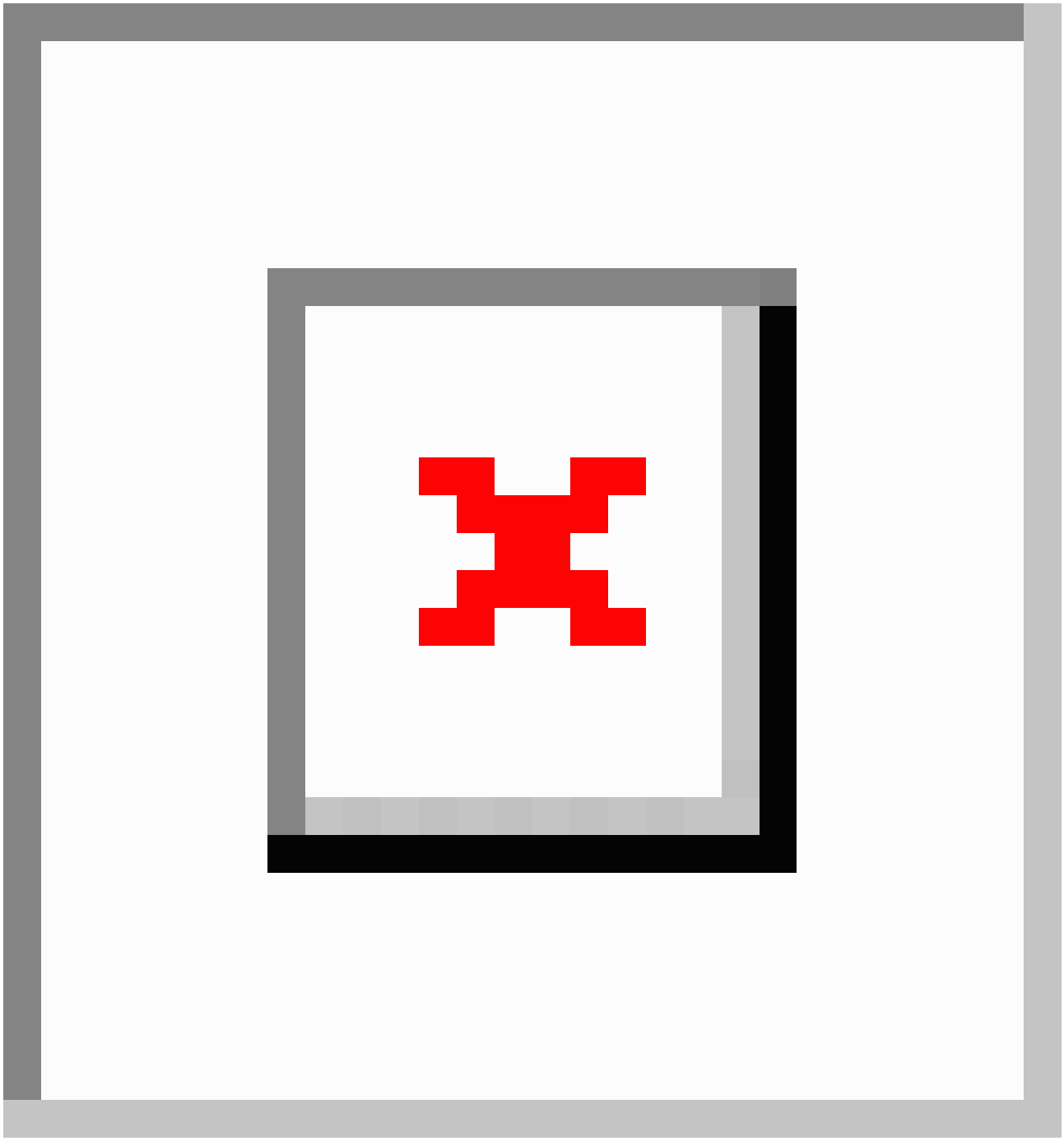
The CAHT Questionnaire

The CAHT questionnaire aimed to collect the patients' medical condition, in preparation for their first outpatient consultation. The questionnaire included several topics in the following order: affected joint, previous diagnoses or treatments, health status, personal care needs or preferences, and patient characteristics (Table 1). Some questions were generic, but most were joint-specific. Depending on the answers given by the patient on specific questions, standardized algorithms expand questioning depth using branching logic without using artificial intelligence (eg, in the case of trauma, more information was requested about the trauma origin). The online questionnaire was offered to patients in a design consistent with the hospital's branding, and some questions were supported with illustrations and instructions (Figure 1).

Table . Example of covered subjects and related questions for patients with knee complaints. Questions are translated from Dutch (the language used in the study) to English.

Topic	Question
History	• Have you ever received a diagnosis**,** by a physician or general practitioner, due to complaints in your knee and/or your lower back? • Have you ever undergone a surgical procedure for your knee and/or lower back?
Main complaint	• For which knee do you have complaints? • For how long have you experienced knee complaints? • On a scale of 0 to 10, how severe is your knee pain at rest?
Main complaint in relation to social activities	• Are you limited in playing sports or executing your hobby due to your knee complaints? • Are you limited in your job due to your knee complaints?
Personal care needs and preferences	• Your orthopedic surgeon would like to know what your main worry or question is regarding your complaints. This way the consultation can be about what's important to you. • If necessary, to what extent are you willing to undergo a surgical procedure to get rid of your knee complaints?
Effect of conservative therapies	• Have you tried muscle-strengthening physical training for a period of 4 to 6 weeks?

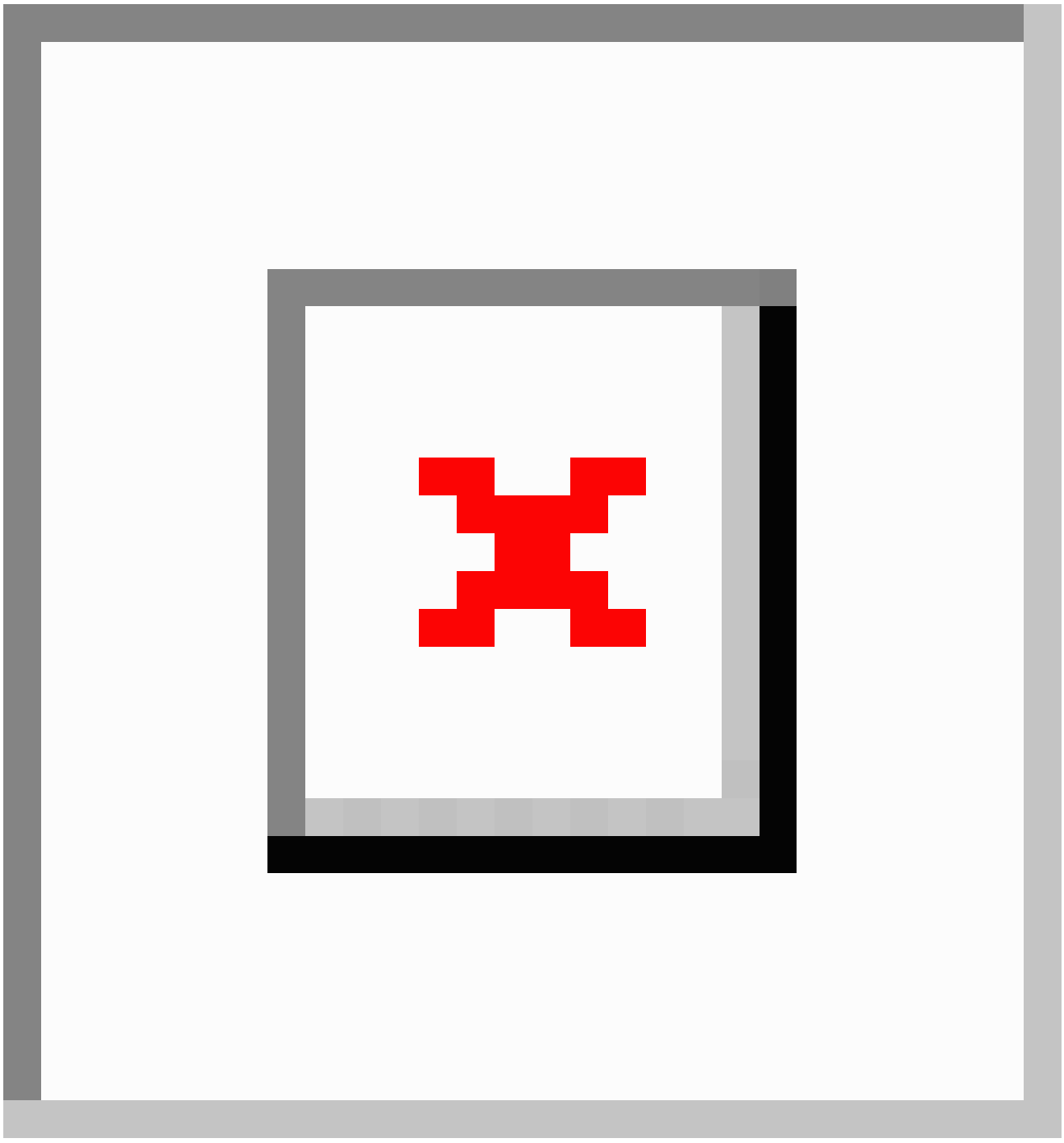
Figure 1. Example of the computer-assisted history taking questionnaire for patients with hip complaints. Patients can indicate the location of their hip pain using an illustration. Questions are translated from Dutch (the language used in the study) to English.



Development of the CAHT Questionnaire

The CAHT questionnaire and its output were developed between January 2022 and March 2022. An expert panel was created for the development of the CAHT questionnaire and its output and consisted of three experienced orthopedic surgeons from both hospitals. First, the input from the physicians at Anna Hospital was obtained, offering insights from the perspective of their professional focus area. Subsequently, the draft version was developed and internally reviewed before undergoing external assessment by three orthopedic surgeons from the Elisabeth Tweesteden Hospital. The answers to the CAHT questionnaire were automatically presented as a coherent summary in a format designed by the expert panel, without providing a differential diagnosis (Figure 2). A web link from the CAHT platform with single sign-on was integrated with the two EHRs used by the participating hospitals in March 2022. This allowed the physicians to read and interpret the data, and to alter it when deemed necessary during the consultation. All feedback on the questionnaire and its output were processed before the study commenced.

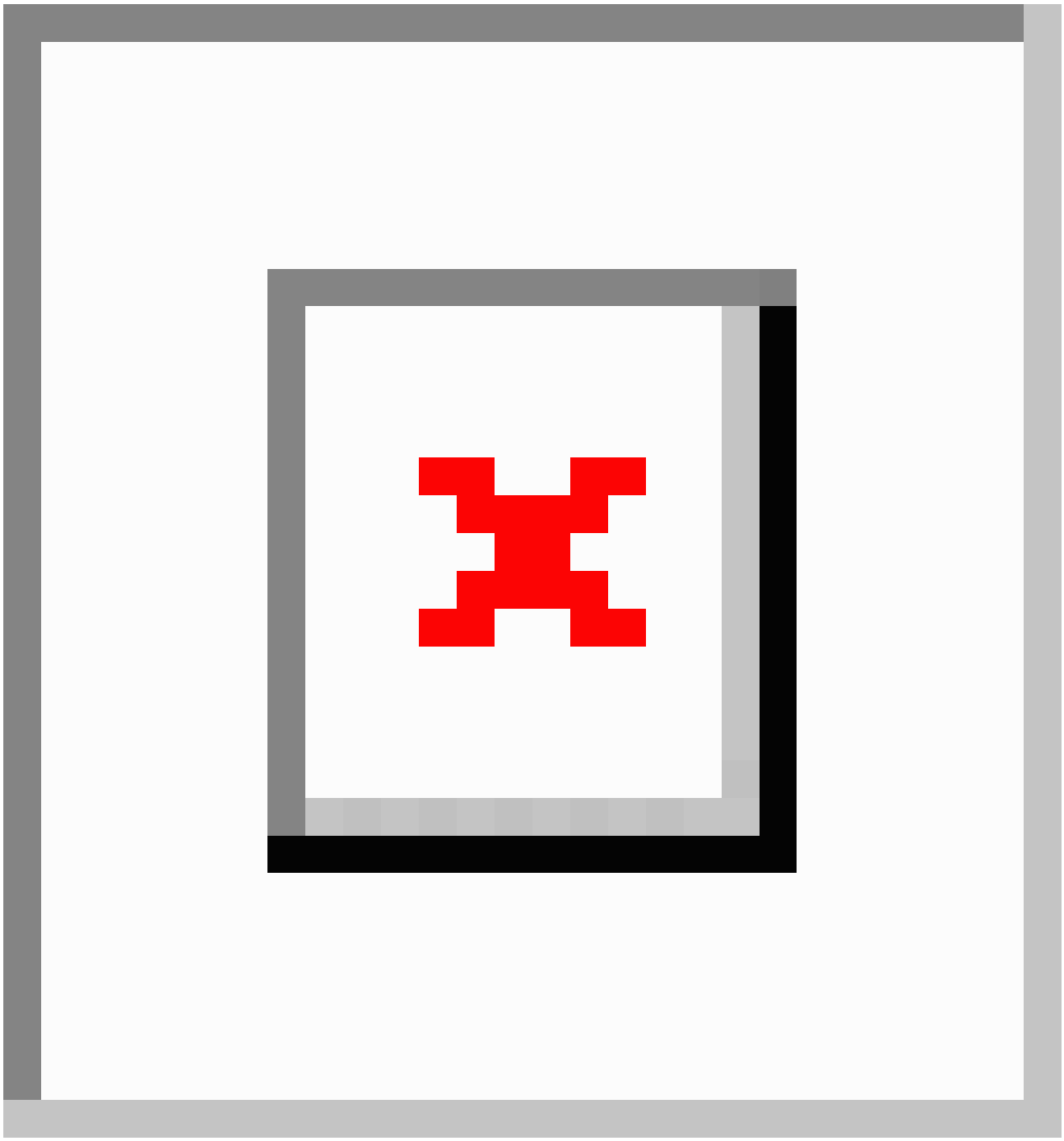
Figure 2. Example of the answers to the computer-assisted history taking questionnaire, presented as a coherent summary of a patient indicating the knee as the affected joint. The summary is translated from Dutch (the language used in the study) to English. AB: antibiotics; ACL: anterior cruciate ligament; NRS: numeric rating scale.



The CAHT Platform

OnlinePROMS (Interactive Studios, ‘s-Hertogenbosch, The Netherlands) was used as the CAHT platform. Interactive Studios has been active in the Dutch health care market for over a decade, providing sustainable solutions for remote patient monitoring. The platform meets European regulations for the privacy and security of patient-reported health data. Invitations to answer the CAHT questionnaire were automatically sent by email when a patient was added to the platform by hospital administration staff. The CAHT questionnaire became accessible to the patient after a two-factor authentication code was entered. The platform allowed hospitals to send two reminders by either email or SMS text messaging if the patient did not complete the questionnaire. Participants of Anna Hospital were reminded 1 and 2 days after the initial invitation at 9 AM via SMS text messaging or email if the phone number was unknown. The ETZ Hospital was chosen to send automated reminders 5 and 10 days after the initial invitation at 9 AM (Figure 3).

Figure 3. Flowchart of the computer-assisted history taking (CAHT) platform. After a patient was referred to the hospital and got in contact with the hospital staff, patients were included, and subsequently, an automated invitation was sent. If the patient did not respond to the invitation, a first and a second reminder were sent based on the response schedule. If the patient completed the questionnaire, a coherent summary of the answers was generated automatically from the available data and was eventually accessible from the electronic health record (EHR).



Study Outcomes

The percentage of patients who completed the CAHT questionnaire was the primary outcome. The usage statistics were collected as secondary outcomes, including the time needed to complete the CAHT questionnaire and the number of necessary reminders. Patient demographic data were collected from the CAHT questionnaire (age, sex, BMI, and affected joint). In addition, in July 2022, all patients were invited to answer a questionnaire on the usage of the CAHT questionnaire. After reaching a convenience sample (n=224), this invitation was deactivated. All physicians of the participating hospitals were asked to rate their satisfaction with the information collected with the CAHT system as well as its accessibility and added value. An overview of the study outcomes is presented in Table 2. All data were collected for the duration of the study via the OnlinePROMS platform.

Table . Overview of the study outcomes.

Outcome	Description
Completion percentage[a]	The percentage of patients who completed the CAHT[b] questionnaire in preparation for their outpatient consultation.
Duration to complete[c]	Time (in minutes) needed to complete the CAHT questionnaire.
First reminders sent[c]	The number of first reminders automatically sent by the CAHT platform in case the CAHT questionnaire was not completed yet.
Second reminders sent[c]	The number of second reminders automatically sent by the CAHT platform in case the CAHT questionnaire was not completed yet.
Patient demographics[a]	Age, sex, BMI, and affected joint.
Easiness of the CAHT questionnaire[d]	Easiness of the CAHT questionnaire on a 10-point Likert scale.
Patient satisfaction[d]	Satisfaction of the consultation on a 10-point Likert scale.
Usage by the physician[d]	The number of patients reporting that the CAHT questionnaire summary was used by the physician during consultation.
Added value[d]	The added value of a CAHT questionnaire supported medical history taking during consultation on a 10-point Likert scale.
HCP[e] data accessibility[f]	Experienced easiness for data accessibility. Satisfaction is rated on a 10-point Likert scale.
Added value for HCP[f]	The added value of a completed CAHT questionnaire during outpatient consultation on a 10-point Likert scale.

[a]Collected prior to consultation.

[b]CAHT: computer-assisted history taking.

[c]Derived from platform user statistics.

[d]Collected 1 day after consultation.

[e]HCP: health care professional.

[f]Collected at the end of the study.

Statistical Methods

Categorical variables are presented as numbers and percentages. Normally distributed continuous variables are presented as means (with the SD). Nonnormally distributed variables are presented as the median value (with the IQR). To analyze satisfaction between groups, data were considered statistically significant at $P<.05$. In case of normal distribution and variances, an independent *t* test was used. For nonnormal distributions, a nonparametric test was used. All data was analyzed using IBM SPSS Statistics for Macintosh, version 29.0 (IBM Corp).

Budget Planning

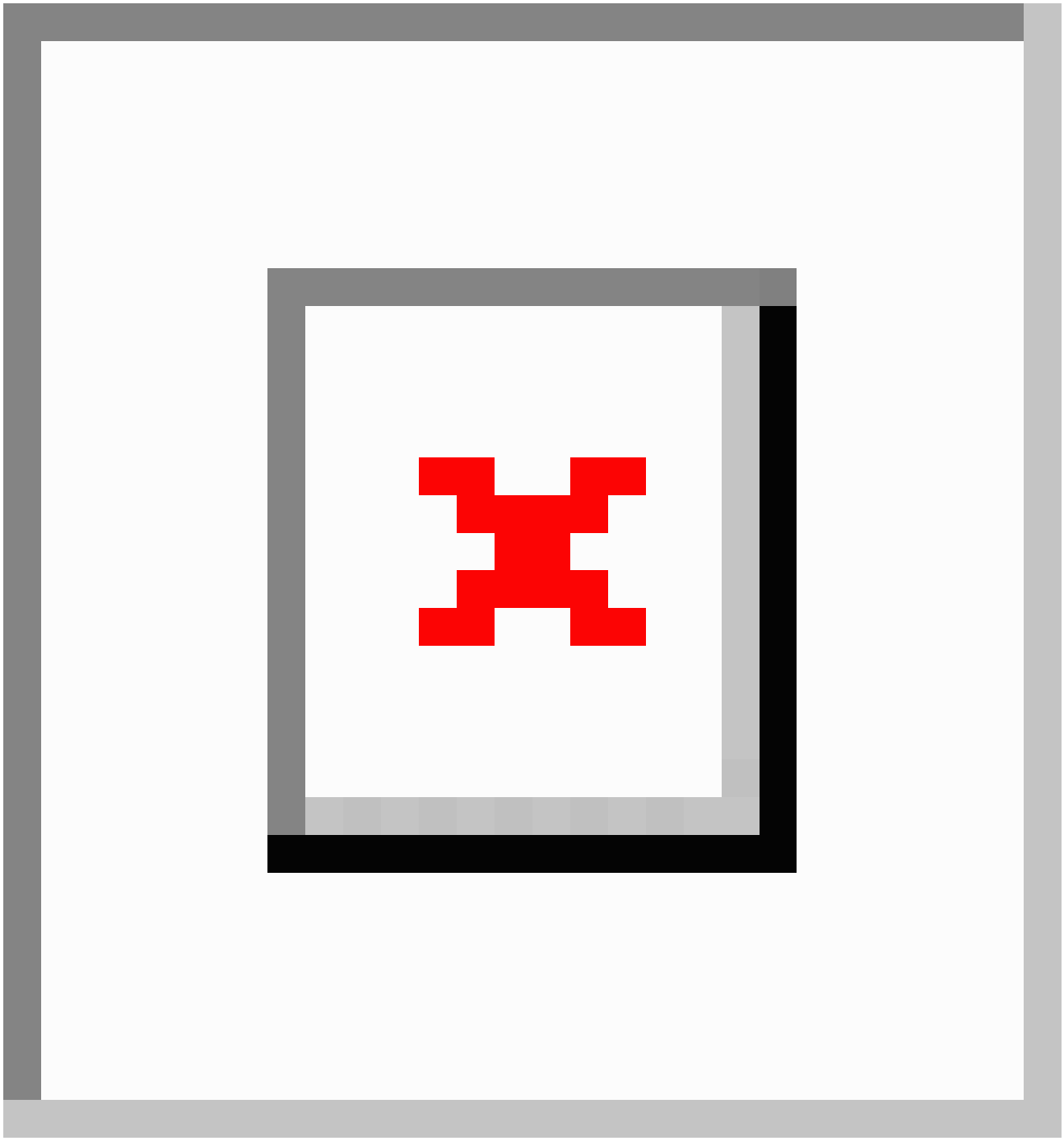
The license fee of the CAHT platform was estimated to be €1000 (US $1085) per month regardless of the number of patients included and health care providers involved. In addition, a one-time setup fee for the questionnaire was estimated to be €4000 (US $4343). Finally, a one-time setup fee for the integration with the EHR was estimated to be €4000 (US $4343).

Implementation (Results)

Study Sample

Between April 2022 and July 2022, a total of 7065 patients were scheduled to have a first consultation with an orthopedic surgeon in 1 of the 2 participating hospitals. Of these, 5321 (75.3%) met the study's inclusion criteria. In addition, a usability questionnaire was sent to 414 participants, of whom 224 (54.1%) responded (Figure 4). Semistructured interviews were performed with 15 orthopedic surgeons from the participating hospitals.

Figure 4. Flow diagram showing the inclusion and exclusion of patients. Anna: Anna Hospital; CAHT: computer-assisted history taking; ETZ: Elisabeth Tweesteden Hospital.



Completed CAHT Questionnaires

In total, 5321 patients were invited to complete the CAHT questionnaire. Out of these, 4932 (92.7%) participants fully completed the questionnaire. Anna Hospital invited 2620 patients, of whom 2516 (96.0%) completed the CAHT questionnaire. ETZ Hospital invited 2701 patients, of whom 2416 (89.4%) completed the CAHT questionnaire (Table 3).

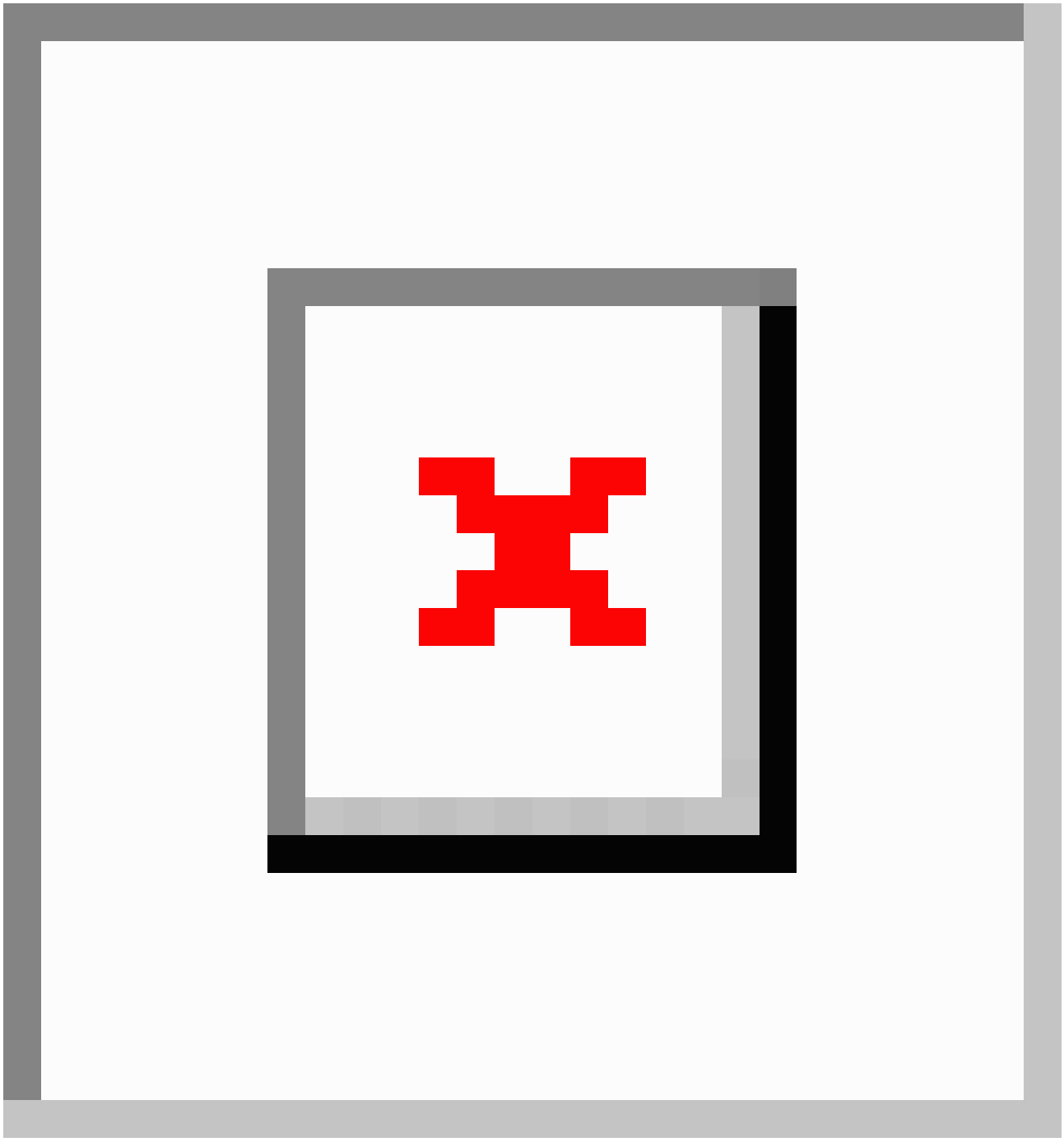
Table . Number of invitations sent and completed computer-assisted history taking questionnaires, divided by hospital.

	Invitations sent, n (%)	Completed, n (%)
Anna Hospital	2620 (49.3)	2516 (96.0)
Elisabeth Tweesteden Hospital	2701 (50.8)	2416 (89.4)
Total	5321 (100)	4932 (92.7)

Usage Statistics

The duration to complete the questionnaire for participants who started and completed the questionnaire on the same day displayed a median value of 16.2 (IQR 11.4 - 26.1) minutes. In addition to the initial invitation, Anna Hospital needed 1846 first and 707 second reminders and ETZ needed 796 first and 323 second reminders, respectively (Figure 5).

Figure 5. Flow diagram showing the response to automated sent invitations and reminders. Anna: Anna Hospital; ETZ: Elisabeth Tweesteden Hospital.



Patient Characteristics

Table 4 shows the patient characteristics. The mean participant age was 56.5 (SD 16.8) years, and 2208 out of 4932 (44.8%) participants were men with a mean BMI of 27.4 (SD 5.1) kg/m^2. The three joints most often selected as affected by the 4932 participants were the knee (n=2205, 44.7%), hip (n=962, 19.5%), and ankle/foot (n=908, 18.4%).

Table . Participant characteristics.

Characteristic	Patients from Anna Hospital (n=2516)	Patients from Elisabeth Tweesteden Hospital (n=2416)
Age (years), mean (SD)	55.4 (16.9)	57.5 (16.7)
18 - 27, n (%)	239 (9.5)	165 (6.8)
28 - 37, n (%)	206 (8.2)	190 (7.9)
38 - 47, n (%)	254 (10.1)	272 (11.3)
48 - 57, n (%)	548 (21.8)	468 (19.4)
58 - 67, n (%)	592 (23.5)	552 (22.8)
68 - 77, n (%)	511 (20.3)	529 (21.9)
78 - 87, n (%)	156 (6.2)	220 (9.1)
88 - 97, n (%)	10 (0.4)	20 (0.8)
Sex (male), n (%)	1157 (46)	1051 (43.5)
BMI (kg/m^2), mean (SD)	27.2 (16.8)	27.6 (5.2)
Affected joint[a], n (%)		
Knee	1205 (47.9)	1000 (41.4)
Hip	437 (17.4)	525 (21.7)
Ankle/foot	423 (16.8)	485 (20.1)
Shoulder	409 (16.3)	310 (12.8)
Elbow	34 (1.4)	48 (2.0)
Wrist/hand	57 (2.3)	116 (4.8)
Spine	183 (7.3)	349 (14.4)

[a]The computer-assisted history taking questionnaire allowed patients to select multiple affected joints.

Patient Satisfaction

Participants included were asked to rate the CAHT questionnaire usability (n=224); on average, they rated it an 8.0 (SD 1.9) out of 10 on easiness to complete the questionnaire and an 8.0 (SD 1.7) out of 10 considering the satisfaction of the consult. Satisfaction with the outpatient consultation was higher in cases where the given answers were used by the orthopedic surgeon during this consultation (median 8.3, IQR 8.0 - 9.1 vs median 8.0, IQR 7.0 - 8.5; *P*<.001). Out of 224 participants, 145 (64.7%) reported that the CAHT questionnaire was used by the physician during the consultation and 157 out of 224 (70.4%) participants had the feeling that their physician had a better understanding of their complaint due to the CAHT questionnaire.

Physician Satisfaction

On average, physicians (n=15) scored the added value of using the CAHT questionnaires during their consultation as 7.8 (0 - 10 scale, SD 1.7). One physician reported not using the answers at all during the consultation. Physicians' unanimously recognized benefits during outpatient consultations, such as increased efficiency, better patient engagement, and better medical record completeness. This was more the case in patients with hip or knee complaints, and less in those with foot and ankle complaints, most likely due to the foot and ankle's more complex anatomy, making it harder for patients to pinpoint the exact location of their symptoms. Physicians highlighted the questionnaire's value in eliciting pertinent information, aiding diagnosis, and providing a framework for informed decision-making. Implementing the summary generated from patients' answers in the EHR was deemed necessary to achieve this with direct access to the information from their own workspace.

Discussion

Principal Findings

This study demonstrates the feasibility of the implementation and clinical adaptation of CAHT for orthopedic patients scheduled for their first consultation in a hospital. The completion rate to answer a CAHT questionnaire before the first consultation was very high: 4932 out of 5321 (92.7%). Patients found the questionnaire easy to understand and complete. Additionally, they were more satisfied with their outpatient consultation when the summary of the CAHT questionnaire was taken into consideration by their physician. Physicians rated the CAHT questionnaire to be a useful addition to standard outpatient consultations, as insight into the personal care needs and preferences allowed them to address the main concern of the patient directly.

To achieve a high completion rate of the CAHT questionnaire, we took implementation barriers addressed in previous studies (accessibility, accuracy, and acceptability) into consideration [6]. For accessibility, linked with interoperability [13], we integrated a single sign-on web link that displayed a

comprehensive summary (and the entire questionnaire when needed) of the answers directly in the EHR. From a patient perspective, the CAHT questionnaire was easily accessible from their email inbox with a two-factor authentication code. Regarding the accuracy, there is conflicting evidence [10,14]. We aimed to obtain accurate answers by enriching some questions with illustrations or a short instruction and using easy/informal language where possible. For data acceptance [15], we created an expert panel to develop and design the CAHT questionnaire and its output. This method is underlined by research, addressing the improvement in quality and acceptance of data [16].

Examining the impact resulting from the implementation of CAHT in a multicenter study design represents a major strength of this study. By incorporating 2 hospitals (1 nonacademic teaching hospital and 1 general hospital), we were able to include a high number of patients, strengthening the generalizability and robustness of our findings. This is confirmed by the similarity of the data between hospitals in terms of demographics, completion rate, and satisfaction.

This study is not without limitations. In this study, we implemented CAHT within orthopedic departments only, limiting the generalizability to other medical departments. The study population's age and sex distribution might, however, indicate usability within other departments as well. Another limitation is the absence of nonverbal communication that occurs in face-to-face conversations. In contrast, CAHT completion allows for more time to think about the answers and fact-check them, without the stress and shortage of time experienced during outpatient visits. The reported median time of 16 minutes for completion supports this.

Lessons Learned

Our study demonstrates that implementing CAHT in the daily routine of an orthopedic department is feasible and can lead to good clinical adaptations but does require the necessary steps to be taken. Requirements are that the patients have an email address, and hospital staff must be available to invite patients to the CAHT platform. The latter would ideally be done through an automated connection with the EHR. Making the results of the questionnaire available in the EHR can be done through a single sign-on web link, offered by almost all EHR suppliers. In addition, the development of the CAHT questionnaire by an expert panel was considered important to ensure usability for HCPs. The integration of the CAHT system with the EHRs and the presentation of the CAHT questionnaire answers as a coherent summary in a format designed by the expert panel were crucial elements for its adoption as well. We hypothesize that key factors contributing to a high completion rate among patients include the predefinition of questions in a standardized order, the inclusion of smart dependencies to avoid unnecessary questions, and the presentation of the online questionnaire in a design consistent with the hospital's branding.

Future Research

Today's health care system is facing an immense burden. Time is limited, but personalized health care needs to be maintained or even improved. Increased consultation duration is associated with better health outcomes, fewer prescriptions, and better recognition of long-term and psychosocial problems [17,18], but this is simply impossible in most health care systems. Nevertheless, a physician who is supported by CAHT results might optimize consultation time in a friendly manner while improving patient-centered communication (ie, signposting, summarization, and repetition of the medical history) [19]. This may lead to more accurate diagnoses, enhanced shared decision-making, and increased patient satisfaction [20,21]. Future research should aim to study the effect of optimized face-to-face consultation time with the support of CAHT and its effect on satisfaction and cost-effectiveness.

Conclusion

Previously reported barriers to implementing and adapting CAHT in clinical practice can nowadays be resolved. In this study, we demonstrated that almost all patients completed the CAHT questionnaire before their outpatient consultation. Both patients and HCPs reported a more efficient and personalized consultation when the answers to the questionnaire were used. Given the pressing need for personalized health care delivery in today's time-constrained medical environment, we recommend implementing CAHT systems in routine medical practice.

Acknowledgments

The authors would like to thank all the health care professionals from Anna Hospital in Geldrop and Elisabeth-Tweesteden Hospital in Tilburg for their time and energy in turning this project from an idea into this manuscript. Furthermore, they would like to thank the team of Interactive Studios for creating the computer-assisted history taking platform. Finally, they would like to thank all patients who were willing to participate in the study.

Authors' Contributions

CC, TT, MS, and WVDW conceived the study and designed the trial. WVDW and TG supervised the data collection. CC managed the data. CC, TT, and WVDW provided statistical analysis. CC drafted the manuscript. All authors contributed to its revision.

Conflicts of Interest

CC and TT work at the research and development department of Interactive Studios. Interactive Studios is the company that developed the computer-assisted history taking platform platform used in this study. Interactive Studios offered the computer-assisted history taking platform used in this study free of charge. The other coauthors declare that the research was conducted in the

absence of any other commercial or financial relationships that could be construed as a potential conflict of interest. Moreover, all authors have completed the International Committee of Medical Journal Editors uniform disclosure form and declare the following: no support from any organization for the submitted work, no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years, and no other relationships or activities that could appear to have influenced the submitted work.

References

1. Lown B. The Lost Art of Healing: Practicing Compassion in Medicine: Ballantine Books; 1999.
2. Hampton JR, Harrison MJ, Mitchell JR, Prichard JS, Seymour C. Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients. Br Med J 1975 May 31;2(5969):486-489. [doi: 10.1136/bmj.2.5969.486] [Medline: 1148666]
3. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. W J Med 1992 Feb;156(2):163-165. [Medline: 1536065]
4. Fryburg DA. What's playing in your waiting room? Patient and provider stress and the impact of waiting room media. J Patient Exp 2021 Nov 29;8:23743735211049880. [doi: 10.1177/23743735211049880] [Medline: 34869835]
5. Jimenez G, Tyagi S, Osman T, et al. Improving the primary care consultation for diabetes and depression through digital medical interview assistant systems: narrative review. J Med Internet Res 2020 Aug 28;22(8):e18109. [doi: 10.2196/18109] [Medline: 32663144]
6. Spinazze P, Aardoom J, Chavannes N, Kasteleyn M. The computer will see you now: overcoming barriers to adoption of computer-assisted history taking (CAHT) in primary care. J Med Internet Res 2021 Feb 24;23(2):e19306. [doi: 10.2196/19306] [Medline: 33625360]
7. Ammenwerth E, Schnell-Inderst P, Hoerbst A. Patient empowerment by electronic health records: first results of a systematic review on the benefit of patient portals. Stud Health Technol Inform 2011;165:63-67. [Medline: 21685587]
8. Zakim D, Brandberg H, El Amrani S, et al. Computerized history-taking improves data quality for clinical decision-making-comparison of EHR and computer-acquired history data in patients with chest pain. PLoS One 2021 Sep 27;16(9):e0257677. [doi: 10.1371/journal.pone.0257677] [Medline: 34570811]
9. Pappas Y, Anandan C, Liu J, Car J, Sheikh A, Majeed A. Computer-assisted history-taking systems (CAHTS) in health care: benefits, risks and potential for further development. Inform Prim Care 2011;19(3):155-160. [doi: 10.14236/jhi.v19i3.808] [Medline: 22688224]
10. Zakim D, Braun N, Fritz P, Alscher MD. Underutilization of information and knowledge in everyday medical practice: evaluation of a computer-based solution. BMC Med Inform Decis Mak 2008 Nov 5;8:50. [doi: 10.1186/1472-6947-8-50] [Medline: 18983684]
11. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694. [doi: 10.2196/46694] [Medline: 37163336]
12. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res 2019 Oct;280:112516. [doi: 10.1016/j.psychres.2019.112516] [Medline: 31437661]
13. Reisman M. EHRs: the challenge of making electronic data usable and interoperable. P T 2017 Sep;42(9):572-575. [Medline: 28890644]
14. Chung AE, Basch EM. Incorporating the patient's voice into electronic health records through patient-reported outcomes as the "review of systems.". J Am Med Inform Assoc 2015 Jul;22(4):914-916. [doi: 10.1093/jamia/ocu007] [Medline: 25614143]
15. Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating patient-generated health data into clinical care settings or clinical decision-making: lessons learned from project HealthDesign. JMIR Hum Factors 2016 Oct 19;3(2):e26. [doi: 10.2196/humanfactors.5919] [Medline: 27760726]
16. De Vito Dabbs A, Myers BA, Mc Curry KR, et al. User-centered design and interactive health technologies for patients. Comput Inform Nurs 2009;27(3):175-183. [doi: 10.1097/NCN.0b013e31819f7c7c] [Medline: 19411947]
17. Elmore N, Burt J, Abel G, et al. Investigating the relationship between consultation length and patient experience: a cross-sectional study in primary care. Br J Gen Pract 2016 Dec;66(653):e896-e903. [doi: 10.3399/bjgp16X687733] [Medline: 27777231]
18. Howie JG, Porter AM, Heaney DJ, Hopton JL. Long to short consultation ratio: a proxy measure of quality of care for general practice. Br J Gen Pract 1991 Feb;41(343):48-54. [Medline: 2031735]
19. Scheder-Bieschin J, Blümke B, de Buijzer E, et al. Improving emergency department patient-physician conversation through an artificial intelligence symptom-taking tool: mixed methods pilot observational study. JMIR Form Res 2022 Feb 7;6(2):e28199. [doi: 10.2196/28199] [Medline: 35129452]
20. Stewart M, Brown JB, Donner A, et al. The impact of patient-centered care on outcomes. J Fam Pract 2000 Sep;49(9):796-804. [Medline: 11032203]
21. Buller MK, Buller DB. Physicians' communication style and patient satisfaction. J Health Soc Behav 1987 Dec;28(4):375-388. [Medline: 3429807]

Abbreviations

CAHT: computer-assisted history taking
EHR: electronic health record
HCP: health care professional

Edited by C Perrin; submitted 17.05.24; peer-reviewed by M Shiraishi, Z Chen; revised version received 02.10.24; accepted 13.10.24; published 13.11.24.

Please cite as:
Craamer C, Timmers T, Siebelt M, Kool RB, Diekerhof C, Caron JJ, Gosens T, van der Weegen W
Completion Rate and Satisfaction With Online Computer-Assisted History Taking Questionnaires in Orthopedics: Multicenter Implementation Report
JMIR Med Inform 2024;12:e60655
URL: https://medinform.jmir.org/2024/1/e60655
doi:10.2196/60655

Implementing a Biomedical Data Warehouse From Blueprint to Bedside in a Regional French University Hospital Setting: Unveiling Processes, Overcoming Challenges, and Extracting Clinical Insight

Matilde Karakachoff[1], MSc; Thomas Goronflot[1], MSc; Sandrine Coudol[1], MSc; Delphine Toublant[1,2], MSc; Adrien Bazoge[1,3], MSc, PhD; Pacôme Constant Dit Beaufils[1,4], MD; Emilie Varey[1,5], MSc; Christophe Leux[6], MD, PhD; Nicolas Mauduit[6], MD; Matthieu Wargny[1], MD, PhD; Pierre-Antoine Gourraud[1,7], MPH, PhD

1

2

3

4

5

6

7

Corresponding Author:
Pierre-Antoine Gourraud, MPH, PhD

Abstract

Background: Biomedical data warehouses (BDWs) have become an essential tool to facilitate the reuse of health data for both research and decisional applications. Beyond technical issues, the implementation of BDWs requires strong institutional data governance and operational knowledge of the European and national legal framework for the management of research data access and use.

Objective: In this paper, we describe the compound process of implementation and the contents of a regional university hospital BDW.

Methods: We present the actions and challenges regarding organizational changes, technical architecture, and shared governance that took place to develop the Nantes BDW. We describe the process to access clinical contents, give details about patient data protection, and use examples to illustrate merging clinical insights.

Implementation (Results): More than 68 million textual documents and 543 million pieces of coded information concerning approximately 1.5 million patients admitted to CHUN between 2002 and 2022 can be queried and transformed to be made available to investigators. Since its creation in 2018, 269 projects have benefited from the Nantes BDW. Access to data is organized according to data use and regulatory requirements.

Conclusions: Data use is entirely determined by the scientific question posed. It is the vector of legitimacy of data access for secondary use. Enabling access to a BDW is a game changer for research and all operational situations in need of data. Finally, data governance must prevail over technical issues in institution data strategy vis-à-vis care professionals and patients alike.

(JMIR Med Inform 2024;12:e50194) doi:10.2196/50194

KEYWORDS

data warehouse; biomedical data warehouse; clinical data repository; electronic health records; data reuse; secondary use; clinical routine data; real-world data; implementation report

Introduction

The increasing use of electronic health records in research settings presents physicians with the systematic yet secondary use of data collected from multiple sources [1-3]. Indeed, hospital information systems (HISs) face a technical challenge to harmonize and integrate application systems and clinical databases that are highly heterogeneous, are based on editor-specific software formats, and use nonstandardized terminologies [3,4].

Moreover, institutions must face the legal and ethical challenge of granting secondary access to data due to national and international laws vis-à-vis patient privacy [5-7]. The reuse of data produced during the care process implies operational knowledge of ethics and legacy concepts that must be solved through well-defined data governance and access policies [7].

Repositories must ensure not only the technical aspects to data access but also the decision criteria granting access [6]. Indeed, institutional data governance is also pivotal when considering legal and ethical principles such as patient informed consent and privacy data protection [2,8,9]. Last but not least, following the line of traditional epidemiology, clinical data reuse requires treatment within a validated and standardized methodological framework to ensure a qualitative result from a scientific and clinical point of view [9].

Despite these difficulties, the rise of biomedical data warehouses (BDWs) is transforming research processes for epidemiology and clinical studies [5,10-12]. Patient data constitute well-defined profiles that can be used to facilitate the enrichment of cohorts [13,14], patient selection and follow-up for clinical trials [15], phenotypic detection, and detailed descriptions of symptoms. BDWs can facilitate the development and performance of personalized and precision medicine, including through the use of big data and artificial intelligence methods.

The implementation of BDWs is conducted at various geographical levels in France. To our knowledge, 24 active hospital BDWs were set up between 2008 and 2023 [16-21]. Regional coordination often occurs within specialized networks of BDWs such as the "Ouest Data Hub," which is specifically designed for university hospitals in Western France [22]. This can take place in thematic networks and is well advanced in cancer data [23]. National or European Union–wide initiatives also propose a development and coordination framework to deal with the different challenges of implementation. In particular, France initiated in 2019 a national project called "Health Data Hub" [24,25] that promotes centralized coordination and increases the visibility of data sources on a nationwide level.

In this paper we report our 5-year experience regarding organizational changes, technical architecture, and governance, supporting the implementation of the Nantes University Hospital Biomedical Data Warehouse (NBDW). We describe the process to access clinical content. We also give figures concerning sources and contents included in the repository, and provide some insight into the projects to which the NBDW contributed. Finally, we propose three indicators to measure the effectiveness of the setting up operation.

Methods

Overview

In 2018, Centre Hospitalo-Universitaire de Nantes (CHUN; University Hospital of Nantes) implemented a BDW to facilitate secondary use of personal health data originally collected in the context of patient care for research and to offer single secure access to up-to-date data from different sources within the CHUN HIS, accommodating a wide range of data types, including demographic and clinical information, consultations, billing codes, diagnoses, laboratory results, medical notes, and drug administration in a unified view.

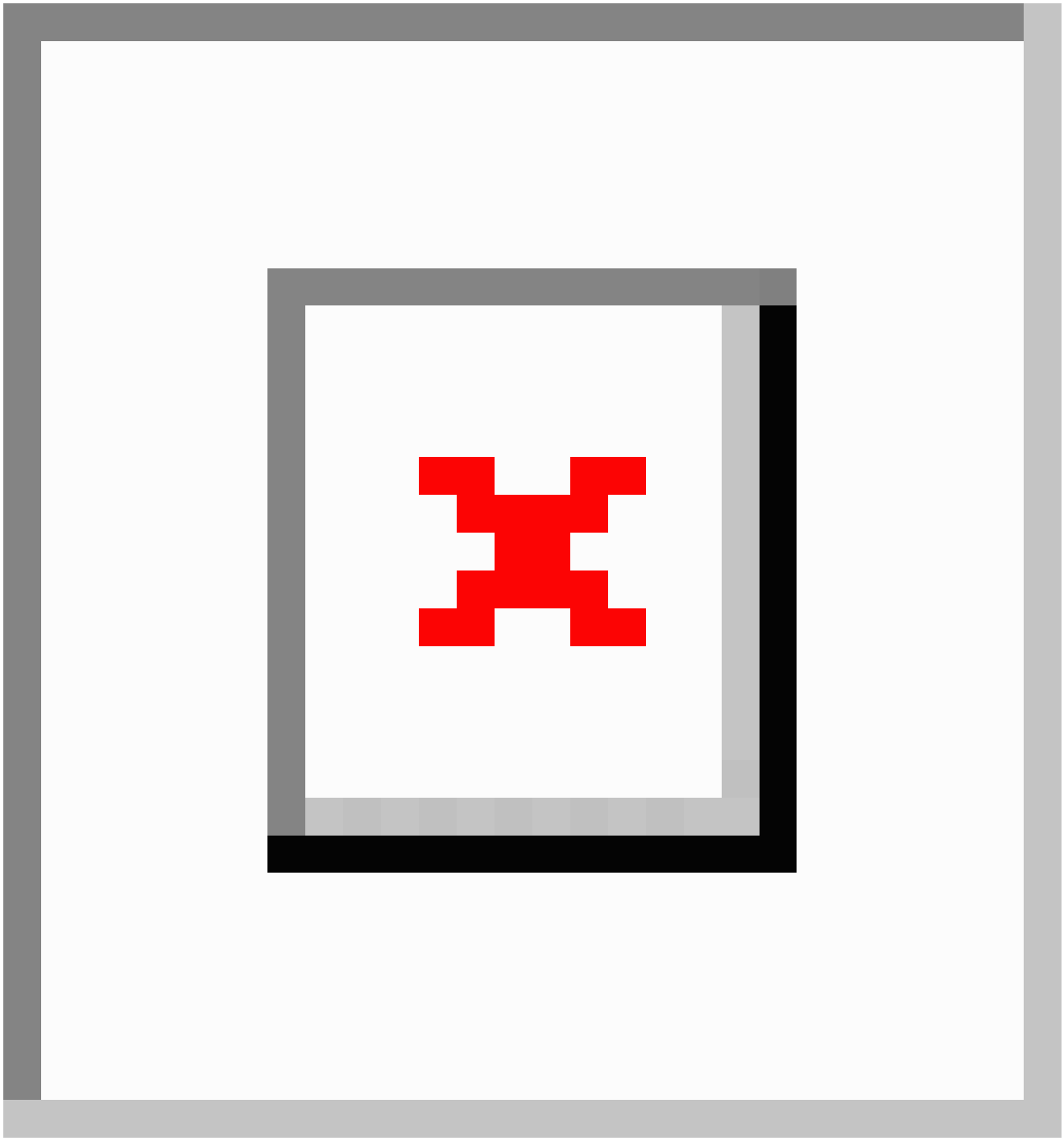
The focal point of the intervention geared toward implementing the NBDW entailed orchestrating the reorganization across the involved hospital research entities. This restructuring initiative encompassed not only the resolution of technical hurdles but also the delineation of governance structures, regulatory frameworks, and parameters governing data access.

This implementation report adheres to the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) reporting guidelines [26].

From Blueprint to the Functional Organization of the NBDW

Implementation of the NBDW required the de novo organization of a functional research network within a 3-fold structure (Figure 1). First, enforcement of the governance policy and NBDW legal framework was achieved by the Research Administration Department (RAD), which assumed legal and ethical responsibility. Second, the creation of a clinical data center (CDC) took charge of the scientific question and the potential need for NBDW data to coassume scientific responsibility with project investigators. The last layer of technical responsibility was established by the hospital's Information Technology Department (ITD). IT priorities are defined by legal and governance constraints, and address the data needs of each project. The new organization allows the distribution of tasks and responsibilities in the implementation and management of the NBDW.

Figure 1. A functional framework dedicated to the CHUN NBDW. A 3-fold governance structure constitutes the functional research framework accompanying the development of the NBDW. The ITD ensures the IT infrastructure's quality, transparency, and data deidentification. Through an ETL process, data flows are extracted and loaded to the NBDW. The RAD is an administrative department that assumes legal and ethical responsibility, and lays the regulatory framework that will define data use and access policies. The CDC assumes scientific data stewardship, is in charge of scientific responsibility, supports the methodology of the study, and grants data access. Those three structures are organized and structured to ensure research quality, ethics, and technical transparency. White arrows represent the data request and access process undertaken by investigators. CDC: clinical data center; CHUN: Centre Hospitalo-Universitaire de Nantes; ETL: extract, transform, and load; ITD: Information Technology Department; NBDW: Nantes Biomedical Data Warehouse; NLP: natural language processing; RAD: Research Administration Department.



Legal Functional Framework

On behalf of the hospital, the RAD defines the framework of data use and access policies, security, and patient consent and privacy, and is in charge of the construction, organization, and enforcement of the NBDW regulatory framework. It works in conjunction with the hospital data protection officer, represented by an officer specifically dedicated to research data. Furthermore, the RAD structure requires documented monitoring to ensure a constant adaptation of the NBDW regulatory framework to answer to changes in legal standards. All projects are publicly described on the web [27] for all potential patients contributing data.

IT Stewardship and Interoperability

The ITD is in charge of the collection, storage aggregation, quality, and integrity of clinical data. This structure carries out a continuous process to conform to technical standards. Setting up the addition of new hospital data sources is an ongoing process that started in July 2018 and is still relevant.

The NBDW is set up on virtual machines. An Oracle database enables massive data storage, integrating multiple hospital data sources into a unique set of tables containing data on patients, consultations, diagnoses, laboratory results, medical notes, and inpatient drug administration. The NBDW uses eHOP [22,28] software to organize and query the database. eHOP is a platform developed by Rennes University using a public-private partnership on the enterprise application integration Enovacom Suite Version 2 (ESV2) of Enovacom (Orange Business Service-Santé). eHOP carries out the acquisition, transformation, and integration of the data coming from various HISs with different formats and standards (Health Level 7 [HL7]; Harmoniser et promouvoir l'informatique médicale [HPRIM]; PN13; Logical Observation Identifiers Names and Codes [LOINC]; and Word documents, PDF, CSV, and text). ESV2 provides an automatic, scheduled (daily, weekly, or monthly), and monitored data supply from the NBDW.

Scientific Data Stewardship and Mediated Access to BDW

Beginning in 2018, the CDC has had a specific team dedicated to the reuse of health data for research, relying on the expertise of public health doctors, data and computer scientists, epidemiologists, biostatisticians, and project managers promoting epidemiology analyses and providing support to clinical investigators and researchers in data access. In conjunction with the ITD, the CDC defines the standards necessary for data integration and data use practices, and ensures data control and the scientific quality of the analyses.

Target and Data Access

The end users of the NBDW are investigators (internal or external to CHUN) aiming to advance the institution's research. A per-project access is created for CHUN investigators through a standardized approval process consisting of 4 main steps (Figure 1). First, investigators register their request and submit a research protocol through a portal hosted by the CHUN intranet. Second, the CDC validation board, which meets once a week, analyzes three dimensions of the submitted research projects: compliance with ethical principles, scientific relevance, and feasibility. On completion of the approval process, the project is registered on an internal database for the completion of legal requirements. Third, the CDC processes all the necessary queries on the NBDW to select a group of patients relevant to the scientific question. During this step, ongoing collaboration with the investigator is necessary, in particular, to precisely define the scientific purpose, eligible population, and data of interest. Fourth, data are made available internally through a data mart, hosted by the CHUN intranet. The data mart system provides regulated, parsimonious (only the required data regarding the targeted population), and time-limited access to investigators. Data are completely deidentified. Moreover, investigators can make simple queries and seek data on patients contained in the data marts through eHOP, which is enriched with a set of tools for simple textual and structured data queries.

Projects requiring data management and extraction to integrate a research database are declared to the public registry of CHUN projects as a guarantee of transparency and to allow patient opposition. At this step, more complex methods for the extraction of information through natural language processing (NLP) [29], regular expression tools, or other structured data [30] may be applied. Finally, data extraction is constrained to strictly necessary data, following the parsimony principle, and only if access to data can be done in a secure environment.

In the case of a project supported by an external project leader from CHUN (academic or private partner), the same process as described above takes place with the exception of the following differences: the project might be supported by a clinical team that submits the research protocol through the portal; a partnership agreement must be signed between the hospital and the partner, and the data mart is only available through a specific virtual working space (data are still internally hosted).

Data Protection and Patient Consent

To comply with national and international privacy regulations, data integration is subjected to a deidentification algorithm. Data are stored in two independent and separate Oracle schemas to separate pseudonymized data from nominative or other directly reidentifying information to which access is strongly limited. Data separation is supplemented by access management and traceability of the actions carried out (ie, AuditLog). Most notably, the platform includes a functionality for collecting and applying patient consent to the use of personal data, ensuring compliance with French law and European General Data Protection Regulation requirements [31].

Regulatory Approval

In alignment with the French Data Protection Act (Loi Informatique et Libertés, 1978), the use of personal data for health research and evaluation requires compliance with a reference methodology, representing good practices. Without such compliance, personal data use must be authorized by the Commission Nationale de l'Informatique et des Libertés (CNIL; French National Commission for Information Technology and Civil Liberties). At the launch of the NBDW, no research methodology existed for data warehouses in the field. Therefore, approval from the CNIL was mandatory to initiate implementation. Submission to the CNIL covered legal responsibilities, data processing details, access, governance, and more. Comprehensive data access details were provided, extending to researchers whether affiliated with CHUN or not. Private entities are permitted to engage in research projects based on the NBDW, ensuring adherence to this resolution and French regulations. The Data Protection Impact Assessment for NBDW was an integral part of the submission to the CNIL, serving as a mandatory document. The authorization to set up and use the NBDW was granted on July 19, 2018, by the CNIL (resolution 2018 - 295).

Budget Planning and Sustainability

Estimating the costs associated with implementing and maintaining a data warehouse is challenging owing to several factors. First, the NBDW is part of an institutional strategy, making it difficult to consider it as a stand-alone entity. Second, implementing it involves the collaboration and coordination of multiple structures and experts, complicating the estimation of resource use. Third, hidden costs are difficult to anticipate and consider, including system failures and delays, unplanned license renewals and upgrades, adjustments to regulatory and legal requirements, unplanned changes in HISs, and infrastructure upgrades. We made an estimation by considering three budget lines—infrastructure, license, and human resources—and two different periods—completion in 5 years (2018 - 2022) and maintenance in 2 years (2022 and 2023)—for a total of €2.6 million (US $2.8 million).

In terms of sustainability, an annual operational budget is allocated for maintenance and updates. Specific needs for the integration of new hospital data sources are financed through project-based funding. Moreover, an economic model is currently being defined to incorporate additional charges for infrastructure costs in the case of external research projects.

Ethical Considerations

An ethics statement is included in the regulatory approval granted by CNIL with resolution number 2018 - 295 [32].

Implementation (Results)

Description of the Sources, Concepts, and Contents

The NBDW integrates multiple hospital data sources into a unique and structured set of tables containing data on patient demographic and administrative records, inpatient drug administration, inpatient constants and anthropometric scores and metrics, anatomic pathology notes, inpatient and outpatient medical laboratory results, narrative medical notes (including admission/discharge summaries, inpatient anesthesia notes, outpatient consultation notes, nurse notes), *International Classification of Diseases, 10th Revision* (*ICD-10*) and French Classification Commune des Actes Médicaux (CCAM; Common Classification of Medical Procedures) codes for inpatient diagnoses and procedures, and medical imaging reports. Table 1 shows principal concepts and contents according to different HIS data sources or software integrated up to now. Some data sources contain only narrative notes, and some sources contain both unstructured and structured data.

Table . Nantes University Hospital Biomedical Data Warehouse principal sources and concepts. Data extracted May 10, 2023.

Concepts	Software	Period	Patients, n	Documents, n	Documents in 2022, n	Structured data, n
Inpatient drug prescriptions	MILLENNIUM	2015-today	318,456	39,681,513	5,673,000	248,289,146
Cardiology narrative notes	CARDIOREPORT	2015-today	9278	38,041	6644	—[a]
Consultation clinical narrative notes	GAM-CLINICOM	2002-today	1,053,386	6,507,860	22,982	—
Constants and anthropometric data	MILLENNIUM	2015-today	440,250	3,306,589	638,969	61,142,757
Anatomic pathology notes	DIAMIC	2015-today	131,812	229,081	27,244	1246
Biology laboratory results	DXLAB	2012-today	701,804	10,752,384	1,672,218	155,825,060
Clinical narrative notes	MILLENNIUM	2015-today	569,114	3,967,073	907,573	5,395,233
ICD-10[b] and clinical procedure codes	CLINICOM	2006-today	725,802	5,318,712	401,607	105,246,620
Radiology reports	QDOC	2015-today	284,937	904,625	122,900	—
Nurse transmissions	TRANSMISSIONS	2017-today	131,508	1,546,114	320,293	1,546,114

[a]Not applicable.

[b]*ICD-10*: *International Classification of Diseases, 10th Revision.*

NBDW Figures and Populations

CHUN, a tertiary care hospital ranked seventh in France in terms of activity [33], provides care over a population catchment area of 1.4 million inhabitants. It provides follow-up and long-term health care for both in- and outpatients. It has 2993 hospital beds, delivers 4380 babies, and conducts more than 1 million consultations and external medical procedures per year [34]. It also carries out practical teaching for 1200 medical students, 800 medical residents, and over 2000 non–medical students.

The NDBW includes information on approximately 1.5 million patients admitted between 2003 and 2022 (Table 2). More than 1.2 million hospitalizations are associated with approximately

12.3 million *ICD-10*–coded diagnoses and 7.3 million clinical procedure codes. Together with more than 6.3 million external consultations, the NBDW contains more than 11 million textual documents. These narrative notes integrated as free-text documents can be interrogated and turned into structured data for research.

The yearly number of patients, hospitalizations, consultations, and narrative notes has increased over time (Multimedia Appendix 1) with growth rates between 2003 and 2019 ranging from 109% (hospitalizations) to 430% (outpatient consultations).

Table . Nantes University Hospital Biomedical Data Warehouse figures and contents, 2003 - 2022.

Contents	Since 2003, n	2022 only, n
Patients[a]	1,597,498	300,804
Hospitalizations[b]	2,635,809	183,361
Outpatient consultations	6,358,271	524,948
Clinical narrative notes[c]	11,634,761	826,670
Diagnoses[d]	12,251,148	956,688
Clinical procedures[e]	7,272,346	504,571

[a]Patients with ≥1 clinical narrative note or a structured document, including inpatient and patients admitted for outward consultations.

[b]Inpatient hospitalizations in medical, surgical, and obstetric services, including complete hospitalizations, day-hospital admissions, and recurring visits.

[c]Clinical narrative notes (with the exclusion of vital signs and anthropometric data; *International Classification of Diseases, 10th Revision* [*ICD-10*] and clinical procedure codes; laboratory results; inpatient drug administrations; and nurse transmissions).

[d]Medical diagnoses following the *ICD-10* for medical, surgical, and obstetric hospitalizations.

[e]Classification Commune des Actes Médicaux for medical, surgical, and obstetric hospitalizations.

Projects and Effectiveness Outcomes

The availability of NBDW data makes it possible to provide data in response to a wide array of scientific questions and the need for data in the analysis and management of care and organization. Prior to the creation of NBDW data, researchers were limited to the interrogation of medico-administrative–structured information out of the scope of data reuse consent. It only covered structured data such as *ICD-10* diagnoses associated with hospitalizations; medical procedures through CCAM codes; and, to a lesser extent owing to availability issues, laboratory results. Obtaining results from different queries performed independently on different HISs and data manager services was time-consuming if not impossible for both legal and technical reasons.

The NBDW currently facilitates queries of clinical concepts in both structured and unstructured free-text notes in an integrated environment and with respect to data protection policies and laws. BDW data requests may be divided into three different types of research projects according to their purpose: (1) optimize patient screening in both clinical trials and observational studies, (2) enrich case reports or electronic case report forms for disease surveillance, and (3) evaluate and improve clinical practices and resource management. To illustrate different types of studies, three concrete examples of data use are described in Multimedia Appendix 2 [35-37].

The first outcome to measure the effectiveness of the NBDW was defined as the number of studies supported through the project tracking portal (Figure 1). Since 2018, 577 requests have been made and treated by the CDC. Among them, 269 projects involved patients included thanks to NBDW queries and research tools (second outcome), and for 115 of them, data marts were created to give investigators access to data (third outcome).

Discussion

Lessons Learned

The development of clinical data warehouses has provided unprecedented access to a large amount of diverse data from clinical care. However, it requires a dedicated effort in terms of governance, data access rules with respect to patient consent and data protection, and technical challenges. The reorganization of structures within a functional research framework is the first factor in the success of the NBDW. Collaboration between departments has not only facilitated seamless communication but also engendered the innovation necessary to deal with the complexities of health care data management. It was an opportunity to test new IT technologies such as distributed infrastructure and anonymization techniques such as deidentification. The interaction between structures has also been a fundamental element in the process of obtaining BDW authorization from the CNIL. Indeed, the obtention of regulatory approval is the result of a long negotiation that required expertise in addressing legal, technical, and scientific research requirements (see Regulatory Approval section for more details).

The creation of the CDC, the result of multidisciplinary teamwork composed of computer scientists, NLP engineers, statisticians, physicians, epidemiologists, and project managers, has probably been the second factor of success. The weekly CDC validation boards, supported by the RAD and ITD structures, verify project compliance vis-à-vis three aspects: ethics, scientific relevance, and practical feasibility, giving support and access to the NBDW in a secure context. However, it is also a leading driver in the shift toward data-driven governance in hospitals.

Implementation of the NBDW has required a continuous and still relevant process to conform to new regulatory and technical standards, and to add new hospital sources and ongoing improvements. An important lesson learned was that each of the 269 NBDW projects in the past 5 years has been an opportunity to revisit the contents of the BDW, further the quality control process, and lead the data transformation process, creating data use value.

Establishing networks and working together is probably the best lesson learned. In France, some experiences have led to changes in health data foresight [28] and promoted the implementation of interregional [22] and national hubs. The NBDW has benefited from and contributed to the "Ouest Data Hub," a network of Western France university hospital data warehouses. The aim is both to facilitate the reproducibility of data analyses, share resources and best querying practices, and promote adapted and standardized terminologies and nomenclatures between centers. Moreover, BDWs use the same software for both integration and querying, allowing them to be interrogated using consistent queries and rules. This approach ensures a high level of interoperability and accessibility, facilitating seamless interaction and adherence to the Findable, Accessible, Interoperable, Reusable (FAIR) principles.

In hindsight, if this process were to be revisited, there are certain facets that we would address differently. Specifically, in the initial stages of implementing the NBDW, primary emphasis was placed on deploying the software and IT infrastructure recommended by the regional network, ODH, aimed at facilitating the establishment of the repository and its querying system. While acknowledging the benefits inherent in this strategy, a more thorough examination of alternative IT solutions is warranted to mitigate reliance on a singular approach and to streamline potential transitions to alternative and variated options.

Conclusions

In conclusion, conducting health studies using electronic health records requires careful attention to ensure accurate results owing to a lack of a systematic quality process. The data quality control procedure is a long and necessary process [17,38,39]. The future challenge will be the setting up of standardized and shared quality control pipelines to ensure quality results, not only at the local level but also in a regional and national context of future data sharing. By extending long-term investments in IT and data in care institutions, the development of NLP and text-mining tools will further accelerate the use of BDW, facilitating data-driven decision-making discussions from top management to patients.

Any research project and analysis of care-based research performed in health management institutions could benefit from the deployment of organized data access. The collaborative nature of data production and the information and privacy protection for patients require mediated and expert access to the BDW. It demonstrates that technical solutions are partial answers to better data-driven practices and must lead to a clear governance strategy

Acknowledgments

The authors wish to thank P Boistard, M Lebigre, C Cartau, A Magnan, P Sudreau, P Lecerf, Dr S Sacher-Huvelin, Dr V Guardiolle, Dr J Esbelin, M Lazarevic, A Royer, and AC de Reboul for their assistance.

This work was financially supported, in part, by the Agence Nationale de la Recherche AIBy4 under contract ANR-20-THIA-0011 and the cluster DELPHI - NExT under contract ANR-16-IDEX-0007, and integrated into the France 2030 plan by Région Pays de la Loire and Nantes Métropoles.

Conflicts of Interest

PAG is the founder of Methodomics (2008) and the cofounder of Big data Santé (2018). He consults for major pharmaceutical companies and start-ups, all of which are handled through academic pipelines (AstraZeneca, Biogen, Boston Scientific, Cook, Docaposte, Edimark, Ellipses, Elsevier, Methodomics, Merck, Mérieux, Octopize, and Sanofi-Genzyme). PAG is a volunteer board member at the AXA not-for-profit mutual insurance company (2021). He has no prescription activity with either drugs or devices. The other authors declare no potential conflicts of interest to disclose.

Multimedia Appendix 1
Nantes University Hospital Biomedical Data Warehouse yearly data volume by type of data.
[DOCX File, 131 KB - medinform_v12i1e50194_app1.docx]

Multimedia Appendix 2
Three projects as an example of case experiences based on the Nantes University Hospital Biomedical Data Warehouse.
[DOCX File, 232 KB - medinform_v12i1e50194_app2.docx]

Checklist 1
iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations).
[DOCX File, 29 KB - medinform_v12i1e50194_app3.docx]

References

1. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017 Aug;26(1):38-52. [doi: 10.15265/IY-2017-007] [Medline: 28480475]
2. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14(1):1-9. [doi: 10.1197/jamia.M2273] [Medline: 17077452]
3. Haarbrandt B, Tute E, Marschollek M. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. J Biomed Inform 2016 Oct;63:277-294. [doi: 10.1016/j.jbi.2016.08.007] [Medline: 27507090]
4. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. JMIR Form Res 2020 Aug 27;4(8):e17687. [doi: 10.2196/17687] [Medline: 32852280]
5. Chute CG, Beck SA, Fisk TB, Mohr DN. The enterprise data trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 2010;17(2):131-135. [doi: 10.1136/jamia.2009.002691] [Medline: 20190054]
6. Pavlenko E, Strech D, Langhof H. Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. BMC Med Inform Decis Mak 2020 Jul 11;20(1):157. [doi: 10.1186/s12911-020-01177-z] [Medline: 32652989]
7. Holmes JH, Elliott TE, Brown JS, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. J Am Med Inform Assoc 2014;21(4):730-736. [doi: 10.1136/amiajnl-2013-002370] [Medline: 24682495]
8. Bloomrosen M, Detmer D. Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association. J Am Med Inform Assoc 2008;15(6):715-722. [doi: 10.1197/jamia.M2905] [Medline: 18755988]
9. Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. Health Serv Res 2010 Oct;45(5 Pt 2):1442-1455. [doi: 10.1111/j.1475-6773.2010.01140.x] [Medline: 21054365]
10. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 2;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
11. Fennelly O, Cunningham C, Grogan L, et al. Successfully implementing a national electronic health record: a rapid umbrella review. Int J Med Inform 2020 Dec;144:104281. [doi: 10.1016/j.ijmedinf.2020.104281] [Medline: 33017724]
12. Constant Dit Beaufils P, Karakachoff M, Gourraud PA, Bourcier R. Management of unruptured intracranial aneurysms: how real-world evidence can help to lift off barriers. J Neuroradiol 2023 Mar;50(2):206-208. [doi: 10.1016/j.neurad.2023.01.156] [Medline: 36724868]
13. Kurian AW, Mitani A, Desai M, et al. Breast cancer treatment across health care systems: linking electronic medical records and state registry data to enable outcomes research. Cancer 2014 Jan 1;120(1):103-111. [doi: 10.1002/cncr.28395] [Medline: 24101577]
14. Greenberg AE, Hays H, Castel AD, et al. Development of a large urban longitudinal HIV clinical cohort using a web-based platform to merge electronically and manually abstracted data from disparate medical record systems: technical challenges and innovative solutions. J Am Med Inform Assoc 2016 May;23(3):635-643. [doi: 10.1093/jamia/ocv176] [Medline: 26721732]
15. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inform 2019 Sep;129:13-19. [doi: 10.1016/j.ijmedinf.2019.05.018] [Medline: 31445247]
16. Artemova S, Madiot PE, Caporossi A, PREDIMED group, Mossuz P, Moreau-Gaudry A. PREDIMED: clinical data warehouse of Grenoble Alpes University Hospital. Stud Health Technol Inform 2019 Aug 21;264:1421-1422. [doi: 10.3233/SHTI190464] [Medline: 31438161]
17. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: a 8-years follow-up experience. Int J Med Inform 2017 Jun;102:21-28. [doi: 10.1016/j.ijmedinf.2017.02.006] [Medline: 28495345]
18. Wack M. Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy: déploiement technique, intégration et gouvernance des données [Doctoral thesis].: Université de Lorraine; 2017 Oct 7 URL: http://docnum.univ-lorraine.fr/public/BUMED_T_2017_WACK_MAXIME.pdf [accessed 2024-06-17]
19. Pressat-Laffouilhère T, Balayé P, Dahamna B, et al. Evaluation of Doc'EDS: a French semantic search tool to query health documents from a clinical data warehouse. BMC Med Inform Decis Mak 2022 Feb 8;22(1):34. [doi: 10.1186/s12911-022-01762-4] [Medline: 35135538]
20. Entrepôts de données de santé hospitaliers en France. Haute Autorité de Santé. 2022 Nov 17. URL: https://www.has-sante.fr/jcms/p_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france [accessed 2023-01-03]
21. Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: a case study in France. PLOS Digit Health 2023 Jul 6;2(7):e0000298. [doi: 10.1371/journal.pdig.0000298] [Medline: 37410797]

22. Madec J, Bouzillé G, Riou C, et al. eHOP clinical data warehouse: from a prototype to the creation of an inter-regional clinical data centers network. Stud Health Technol Inform 2019 Aug 21;264:1536-1537. [doi: 10.3233/SHTI190522] [Medline: 31438219]
23. Bocquet F, Raimbourg J, Bigot F, Simmet V, Campone M, Frenel JS. Opportunities and obstacles to the development of health data warehouses in hospitals in France: the recent experience of comprehensive cancer centers. Int J Environ Res Public Health 2023 Jan 16;20(2):1645. [doi: 10.3390/ijerph20021645] [Medline: 36674399]
24. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics initiatives: two national projects to promote data sharing in healthcare. Yearb Med Inform 2019 Aug;28(1):195-202. [doi: 10.1055/s-0039-1677917] [Medline: 31419832]
25. Goldberg M, Zins M. Health data hub: why and how? Med Sci (Paris) 2021 Mar;37(3):271-276. [doi: 10.1051/medsci/2021016] [Medline: 33739275]
26. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25(1):e46694. [doi: 10.2196/46694] [Medline: 37163336]
27. Liste de recherches sur données et/ou échantillons menées au CHU de Nantes, notamment à partir de l'entrepôt. Centre Hospitalier Universitaire de Nantes. URL: https://www.chu-nantes.fr/liste-des-etudes-menees-au-chu-de-nantes-utilisant-des-donnees-de-l-entrepot-de-recherche [accessed 2024-01-26]
28. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. Stud Health Technol Inform 2015;210:702-706. [Medline: 25991243]
29. Labrak Y, Bazoge A, Dufour R, et al. DrBERT: a robust pre-trained model in French for biomedical and clinical domains. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): Association for Computational Linguistics; 2023:16207-16221. [doi: 10.18653/v1/2023.acl-long.896]
30. Guardiolle V, Bazoge A, Morin E, et al. Linking biomedical data warehouse records with the national mortality database in France: large-scale matching algorithm. JMIR Med Inform 2022 Nov 1;10(11):e36711. [doi: 10.2196/36711] [Medline: 36318244]
31. General Data Protection Regulation (GDPR). URL: https://gdpr-info.eu/ [accessed 2023-06-06]
32. Délibération 2018-295 Du 19 Juillet 2018. Légifrance. 2018 Oct 23. URL: https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000037509951 [accessed 2024-06-11]
33. Bases statistiques SAE. Données statistiques publiques en santé et social. 2015 Jan 6. URL: https://data.drees.solidarites-sante.gouv.fr/explore/dataset/708_bases-statistiques-sae/ [accessed 2021-12-28]
34. En chiffres. Centre Hospitalier Universitaire de Nantes. URL: https://www.chu-nantes.fr/activite-et-chiffres-cles [accessed 2023-01-03]
35. Bourcier R, Chatel S, Bourcereau E, et al. Understanding the pathophysiology of intracranial aneurysm: the ICAN Project. Neurosurgery 2017 Apr 1;80(4):621-626. [doi: 10.1093/neuros/nyw135] [Medline: 28362927]
36. Cotton F, Kremer S, Hannoun S, Vukusic S, Dousset V, Imaging Working Group of the Observatoire Français de la Sclérose en Plaques. OFSEP, a nationwide cohort of people with multiple sclerosis: consensus minimal MRI protocol. J Neuroradiol 2015 Jun;42(3):133-140. [doi: 10.1016/j.neurad.2014.12.001] [Medline: 25660217]
37. Lucas DN, Yentis SM, Kinsella SM, et al. Urgency of caesarean section: a new classification. J R Soc Med 2000 Jul;93(7):346-350. [doi: 10.1177/014107680009300703] [Medline: 10928020]
38. Jantzen R, Rance B, Katsahian S, Burgun A, Looten V. The need of an open data quality policy: the case of the "transparency - health" database in the prevention of conflict of interest. Stud Health Technol Inform 2018;247:611-615. [Medline: 29678033]
39. Looten V, Kong Win Chang L, Neuraz A, et al. What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse. Comput Methods Programs Biomed 2019 Nov;181:104825. [doi: 10.1016/j.cmpb.2018.12.030] [Medline: 30612785]

Abbreviations

BDW: biomedical data warehouse
CCAM: Classification Commune des Actes Médicaux (English: Common Classification of Medical Procedures)
CDC: clinical data center
CHUN: Centre Hospitalo-Universitaire de Nantes (English: University Hospital of Nantes)
CNIL: Commission Nationale de l'Informatique et des Libertés (English: French National Commission for Information Technology and Civil Liberties)
ESV2: Enovacom Suite Version 2
FAIR: Findable, Accessible, Interoperable, Reusable
HIS: hospital information system
HL7: Health Level 7
HPRIM: Harmoniser et promouvoir l'informatique médicale

***ICD-10*:** *International Classification of Diseases, 10th Revision*
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
ITD: Information Technology Department
LOINC: Logical Observation Identifiers Names and Codes
NBDW: Nantes University Hospital Biomedical Data Warehouse
NLP: natural language processing
RAD: Research Administration Department

Edited by C Perrin; submitted 22.06.23; peer-reviewed by D Reuter, K Gierend, O Steichen; revised version received 08.04.24; accepted 17.04.24; published 24.06.24.

Please cite as:
Karakachoff M, Goronflot T, Coudol S, Toublant D, Bazoge A, Constant Dit Beaufils P, Varey E, Leux C, Mauduit N, Wargny M, Gourraud PA
Implementing a Biomedical Data Warehouse From Blueprint to Bedside in a Regional French University Hospital Setting: Unveiling Processes, Overcoming Challenges, and Extracting Clinical Insight
JMIR Med Inform 2024;12:e50194
URL: https://medinform.jmir.org/2024/1/e50194
doi:10.2196/50194

Implementation Report

Design and Implementation of an Inpatient Fall Risk Management Information System

Ying Wang[1,2], MSM; Mengyao Jiang[2], MSN; Mei He[2], MSN; Meijie Du[2], MSN

[1]School of Management, Wuhan University of Technology, Wuhan, China

[2]Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Corresponding Author:
Ying Wang, MSM
School of Management
Wuhan University of Technology
122 Luoshi Road
Hongshan District
Wuhan, 430070
China
Phone: 86 027 83662317
Email: wangying_tjh@hotmail.com

Abstract

Background: Falls had been identified as one of the nursing-sensitive indicators for nursing care in hospitals. With technological progress, health information systems make it possible for health care professionals to manage patient care better. However, there is a dearth of research on health information systems used to manage inpatient falls.

Objective: This study aimed to design and implement a novel hospital-based fall risk management information system (FRMIS) to prevent inpatient falls and improve nursing quality.

Methods: This implementation was conducted at a large academic medical center in central China. We established a nurse-led multidisciplinary fall prevention team in January 2016. The hospital's fall risk management problems were summarized by interviewing fall-related stakeholders, observing fall prevention workflow and post–fall care process, and investigating patients' satisfaction. The FRMIS was developed using an iterative design process, involving collaboration among health care professionals, software developers, and system architects. We used process indicators and outcome indicators to evaluate the implementation effect.

Results: The FRMIS includes a fall risk assessment platform, a fall risk warning platform, a fall preventive strategies platform, fall incident reporting, and a tracking improvement platform. Since the implementation of the FRMIS, the inpatient fall rate was significantly lower than that before implementation (P<.05). In addition, the percentage of major fall-related injuries was significantly lower than that before implementation. The implementation rate of fall-related process indicators and the reporting rate of high risk of falls were significantly different before and after system implementation (P<.05).

Conclusions: The FRMIS provides support to nursing staff in preventing falls among hospitalized patients while facilitating process control for nursing managers.

(JMIR Med Inform 2024;12:e46501) doi:10.2196/46501

KEYWORDS

fall; hospital information system; patient safety; quality improvement; management; implementation

Introduction

Context

Falls are one of the nursing-sensitive indicators for nursing care [1], which are a leading cause of fatal and nonfatal health loss globally [2,3]. Reducing and preventing falls has become an international health priority. Falls—common adverse events reported in hospitals—have been identified as a nursing-sensitive quality indicator of patient care.

Given the growing technological progress, health IT may help enhance the quality and safety of provided care, facilitating the effectiveness and efficiency of the clinical workflow, and supporting the provision of integrated multidisciplinary care [4-11]. The hospital information system (HIS) is a promising

approach to improve care quality and safety in the complex hospital environment. Despite extensive literature on fall risk factors and preventive strategies [12-18], few studies have focused on health information systems for managing inpatient falls.

Problem Statement

To address these issues, we formed a nurse-led multidisciplinary fall prevention team in January 2016, including the hospital administrative staff, quality management specialists, physicians, nurses, pharmacists, and informatics staff. This team retrospectively analyzed 19 inpatient fall cases that occurred in 2015 (fall rate 0.015‰), ranking first among all in-hospital nursing adverse events. Among the fall cases, 30%-40% of patients had grade ≥3 injuries, which significantly exceeded the 3.978% proportion seen in similar hospitals during this period. Falls caused severe harm and financial burden to inpatients, with 3 patients experiencing severe head injuries and 2 having hip fractures. The longest hospital stay resulting from falls reached 36 days.

The hospital's fall risk management problems were summarized by interviewing fall-related stakeholders, observing fall prevention workflow and postfall care process, and investigating patients' satisfaction; these included (1) nonachievement of real-time fall risk assessment, real-time uploading, and information sharing; (2) absence of fall risk warning management; (3) complicated fall risk management workflow; (4) absence of process control in fall prevention (such as process control for different fall risk levels, process control for different time nodes, etc); and (5) lack of standardized pathways for inpatient fall incident reporting and improvement tracking.

Similar Interventions

Several studies have highlighted the benefits of using health information systems for patient fall management. For example, Giles et al [19] reported that data collected from nursing information systems can be used to identify high-fall-risk patients. Mei et al [20] designed an electronic patient fall reporting system in a US long-term residential care facility, which could improve the fall reporting process and subsequent quality improvement efforts. Katsulis et al [21] combined the Fall TIPS (Tailoring Interventions for Patient Safety) [22] with a clinical decision support system, which increased its ease of use over the paper version. Jacobsohn et al [23] developed an automated clinical decision support system for identifying and referring older adult emergency department patients at the risk of future falls. Mlaver et al [24] at the Brigham and Women's hospital developed a valuable electronic health record–embedded dashboard that collected inpatient fall risk data. However, the abovementioned fall information system only focused on a specific domain of fall management. So far, there is still no report about an HIS for overall fall risk management.

Aims and Objectives

This implementation aims to design and implement a fall risk management information system (FRMIS) to reduce falls among inpatients and improve nursing quality. Our goal is to create a culture of safety and reduce the incidence of falls hospital-wide, ensuring the well-being and security of all patients.

Methods

This study adhered to the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) checklist [25].

Blueprint Summary

This FRMIS consists of 4 major functional platforms to facilitate comprehensive fall prevention pathway management as shown in Textbox 1.

Textbox 1. The 4 major functional platforms of the fall risk management information system.

A fall risk assessment platform (Multimedia Appendix 1):

The assigned nurse uses a personal digital assistant to conduct fall risk assessments within 4 hours of patient admission. Upon completion, the personal digital assistant automatically compiles the Morse Fall Risk Score [26,27] and risk level, marking it in the electronic nursing record. All patients' Morse Fall Risk Scores are collected and shared in real time through the information platform. Simultaneously, nurses receive nursing guidelines specific to different fall risk levels. They implement corresponding fall prevention measures such as hanging “Fall Prevention” warning signs near high-risk patients' beds, distributing “Fall Prevention Information Sheets” to guide patients and their families on preventive measures, and documenting and passing on relevant information during shift changes. The head nurse conducts daily inspections and guidance on the accuracy of Morse fall assessments and the implementation of fall prevention measures, upon completion of departmental reviews.

A fall risk warning platform (Multimedia Appendix 1):

Patients at different fall risk levels are color-coded for easy identification: high-risk (Morse Fall Risk Score≥45), red; moderate (score approximately 25-44), yellow; and low (score approximately 0-24), green. The fall risk warning module comprehensively displays the daily number of high-risk falls, department distribution and ranking, percentage of the population at the risk of falls, specific bed locations, medical diagnoses, Morse Fall Risk Scores, and assessment times through charts and color-coded indicators. This provides nursing managers real-time insights into the key populations, departments, and information related to fall risk management, enabling proactive fall risk prevention and providing precise information support for effective fall prevention process control.

A fall preventive strategies platform (see Multimedia Appendix 1):

Evidence-based fall prevention strategies are developed, incorporating fall event analysis and expert discussions to extract key process monitoring indicators. An electronic fall prevention bundle strategies quality tracking checklist was established for accurate assessment of fall risk, increased awareness of preventive measures, enhanced handover process for high-risk patients, environmental safety, implementation of fall prevention knowledge training, and guidance on proper use of assistive devices. Nursing department and ward-level managers can use mobile devices (iPads) to conduct targeted goal management and quality inspections of fall prevention strategies. Real-time monitoring is conducted on key fall process indicators such as accuracy of Morse fall risk assessments, implementation of health education, adherence to handover procedures, and compliance with environmental safety measures.

A fall incident reporting and tracking platform (see Multimedia Appendix 1):

The platform regulates the reporting process for inpatient fall events. After a fall incident occurs, the ward head nurse promptly logs into the fall incident reporting platform to proactively report the incident. They provide details such as the time and location of the fall, the sequence of events, whether the patient was injured, the extent of the injury, and the emergency treatment process. Once the information platform receives the ward's report, it immediately sends text messages to the chief nurse and members of the nursing department's safety management team. On the platform, safety management team members can quickly trace the Morse Fall Risk Score, risk level, appropriateness of fall prevention interventions, timeliness of assessments, and any dynamic evaluations associated with that patient. After gaining a comprehensive understanding of the patient's relevant information, they visit the ward in a timely manner to conduct on-site inspections and tracking. They provide guidance to the department by applying root cause analysis to thoroughly analyze the fall event, identify the underlying causes, and propose areas of improvement directly on the web-based platform. Ward head nurses and the chief nurse can access expert guidance instantly on the platform and make necessary improvements based on the advice provided.

Technical Design

The FRMIS was developed using an iterative design process, involving collaboration among health care professionals, software developers, and system architects. The design aimed to create a user-friendly interface, incorporate data integration capabilities, and enable real-time reporting functionalities. In order to meet the usage needs of both PC and mobile devices, the development language selected for this system includes C#, jQuery, and Java; the development tools used were Visual Studio (Microsoft Corp) and Eclipse (The Eclipse Foundation), and the development platforms used were Windows and Android.

Target

The FRMIS was designed to assist nursing staff in preventing inpatient falls through IT, facilitating process control for nursing managers and ensuring patient safety.

Data

Our hospital has a dedicated computer center, which serves as the technical support department for network security. It is responsible for the construction and operation of hospital network security protection measures. The collection of various data in the FRMIS complies with relevant national laws and regulations. The data collection scope follows the principle of “minimum necessary” and adopts measures such as data desensitization, data encryption, and link encryption to prevent data leakage during the data collection process.

Interoperability

To maximize the effectiveness of the FRMIS, standardization of data elements and the development of interface systems to allow seamless data exchange between our HISs were necessary. The FRMIS used Health Level Seven Fast Healthcare Interoperability Resources (HL7FHIR) to enable seamless data exchange and streamline workflows.

Participating Entities

The FRMIS project has obtained the approval and support of hospital management, who have provided strong guarantees in terms of personnel, resources, funding, and working hours required for the implementation of the research plan. Our hospital is an advanced information management hospital with state-of-the-art scientific technologies. The computer center has rich experience in developing information management platforms; they have independently developed and implemented 19 hospital operational management systems. The FRMIS’s development was initiated by the nursing department, with the

assistance of the computer center to fulfill the corresponding requirements.

Budget Planning

The FRMIS development process lasted about 4 months, and the total development cost was approximately 500,000 Renminbi (approximately US $68,300). The subsequent maintenance costs were estimated to be 8% of the total development cost annually. Funding for the FRMIS's development and maintenance was provided by our hospital. The ownership of the FRMIS belongs to Tongji Hospital.

Sustainability

The FRMIS's implementation was carried out through the issuance of relevant policy documents by the nursing department, ensuring its clinical adoption. All risk assessment and incident reports concerning the inpatient falls were conducted through this information system thus far, replacing the previous paper-based forms. Over the past few years of using this system, our hospital's computer center staff has been maintaining and fixing occasional bugs that occur during clinical implementation of this system. The computer center staff also made necessary modifications and improvements to certain details as needed to enhance system functionality, optimize workflows, and adapt to evolving health care practices.

Statistical Analysis

Statistical comparisons were made on the fall incidence rate among inpatients and the reporting rate of high-fall-risk patients before and after FRMIS implementation. Data entry and statistical analysis were performed using SPSS (version 17.0; IBM Corp). The chi-square test was used to compare the differences in the fall incidence rate among inpatients, the rate of high-fall-risk patients, and the implementation rate of preventive fall quality bundle strategy indicators before and after FRMIS implementation. A value of $P<.05$ was considered statistically significant.

Ethics Approval

The study was approved by the institutional review board of Tongji hospital (protocol TJ-IRB20191209).

Implementation (Results)

Coverage

Our hospital is a large academic medical center in central China. In 2016, the hospital had a total of 4000 open beds, 106 nursing wards, and 53 specialized nursing units. The average daily admission rate ranges from 4500 to 5000 patients, with a total of 193,709 admitted patients throughout the year. The cumulative number of bed-days reached 1,756,946, of which 277,365 (15.79%) were for critical patients.

Outcomes

We carried out the process and outcome evaluation with regard to the FRMIS's implementation. The process evaluation indicators include (1) the accuracy rate of the Morse fall risk assessment: number of accurate Morse fall risk assessments / total number of Morse fall risk assessments inspected; (2) implementation rate of fall prevention health education: number of implemented health education check items / number of patients inspected × total number of fall prevention health education check items; (3) implementation rate of shift handoff: number of implemented shift handoff check items / number of patients inspected × total number of fall prevention shift handoff check items; (4) implementation rate of environment safety: number of implemented environment safety check items / the number of patients inspected × the total number of environment safety check items.

The staff of the quality control office in the nursing department reviewed the FRMIS on a daily basis to identify the clinical departments where high-fall-risk patients were distributed across the hospital. For departments with more than 5 high-fall-risk patients and a proportion exceeding 20% of the total patients, we assigned 2 supervisory staff from the quality control team. They used the electronic form "Fall Prevention Bundle Strategy Quality Tracking Form" (see Multimedia Appendix 1) on an iPad to conduct quality inspections on the nursing units for the high-fall-risk patient population, randomly checking the implementation rate of fall prevention bundle strategy indicators (fall risk assessment, fall-related health education, fall-related shift handoff, and environment safety). Before implementing the FRMIS, a total of 1250 patients were randomly sampled for inspection. After implementing FRMIS, a total of 1806 patients were randomly sampled for inspection. Additionally, a comparative analysis was performed on the hospitalization period between February and October 2017 (after FRMIS implementation, the total bed days occupied by inpatients was 1,323,667) and between February and October 2015 (before FRMIS implementation, the total bed days occupied by inpatients was 1,303,094) to evaluate the hospital-wide reporting rate of high-fall-risk cases, incidence rate of patient falls, and severity of fall-related injuries.

The results showed that since the FRMIS's implementation, the inpatient falls rate was significantly lower than that before implementation ($P<.001$), as shown in Table 1. In addition, the percentage of major fall-related injuries was significantly lower than that before implementation, as shown in Table 2. The implementation rate of fall-related process indicators and the reporting rate of high risk of falls were significantly different before and after system implementation ($P<.001$), as shown in Table 3.

Table 1. Comparison of fall-related outcome indicators.

	Before implementation (total bed days=1,303,094), n (%)	After implementation (total bed days=1,323,667), n (%)	Chi-square (*df*)	*P* value
High-fall-risk patients' reports	1036 (0.8)	3007 (2.3)	931.7 (1)	<.001
Fall incident reports	23 (0.02)	11 (0.01)	4.4 (1)	<.001

Table 2. Results of fall-related injuries.

	Cases of fall-related injury, n			
	No injury	Minor	Moderate	Major
Before implementation	15	28	12	2
After implementation	20	13	9	0

Table 3. Comparison of fall-related process indicators.

	Before implementation (n=1250), n (%)	After implementation (n=1806), n (%)	Chi-square (*df*)	*P* value
Fall risk assessment	1056 (84.48)	1709 (95.73)	88 (1)	<.001
Fall-related health education	1107 (88.56)	1769 (97.95)	117.5 (1)	<.001
Fall-related shift handoff	1114 (89.12)	1767 (97.84)	104 (1)	<.001
Environment safety	1127 (90.16)	1796 (99.45)	153 (1)	<.001

Lessons Learned

The FRMIS's development and implementation followed a structured process, starting with needs assessment and culminating in ongoing monitoring and improvement. With this multidisciplinary team and comprehensive approach, we were able to provide a more robust and effective fall risk management system for the entire hospital. The FRMIS addressed the shortcomings of paper-based reporting, such as untimely fall assessments, delayed reporting, information transmission delays, loss of assessment forms, and incomplete tracking information. The FRMIS achieved a holistic fall prevention strategy that spanned from risk assessment to postfall intervention, which brought several benefits to both patients and health care providers. The FRMIS alerted nursing staff about high-risk patients, enabling timely interventions and reducing fall occurrences. It also standardized the reporting process for fall events, allowing for efficient tracking and analysis of incidents.

Discussion

Principal Findings

This study has designed and implemented an FMRIS at the hospital level. The novel system provided a simple, intuitive, and highly operational prevention management model, encompassing fall risk assessment, high fall risk screening, forecasting, and monitoring. It significantly improved the procedural and standardized levels of fall management for hospitalized patients, having prompted nurses to proactively implement fall preventive interventions, conducted timely fall risk assessments, reduced underreporting of high-fall-risk patients, and increased the forecast rate of high-fall-risk patients.

Unlike previous studies that focus on a specific stage of fall management (such as risk identification [19] or fall incident reporting [28]) or patients in a specific department [23], our system catered to the entire process of fall risk management for all inpatients. The FRMIS showed promise in enhancing patient safety, reducing fall incidents, and improving overall care quality.

To facilitate the successful implementation of the FRMIS in clinical practice, we first developed the Standardized Management Guidelines for Preventing Inpatient Falls at the hospital-wide level. This policy document comprehensively revises and improves clinical fall prevention efforts, which include patient fall risk assessment, health education, fall preventive interventions, fall management workflow, fall incident reporting, and system record-keeping. The policy document was distributed in hard copy by the nursing department to all departments and also uploaded electronically on the hospital's Office Automation platform. It mandated each clinical department to conduct fall prevention training based on the guidelines, requiring all nurses' participation and proficiency. This document served as a supporting tool, providing nurses with guidance on how to use the FRMIS effectively in their clinical practice to prevent inpatient falls.

In addition, we conducted standardized nurse training through a web-based platform. Three main implementation strategies were used. First, we conducted diverse forms of training, including ward-, department-, and hospital-level fall prevention training, as well as case-based warning education, bedside simulation assessment, experience sharing sessions, and special lectures, to comprehensively implement the content of the Standardized Management Guidelines for Preventing Falls. Second, we performed objective evaluation. We incorporated simulated case examinations for patient fall prevention into the clinical skills evaluation of nurses, head nurse position evaluation, and their performance appraisal to comprehensively assess the level of knowledge of fall management guidelines and the emergency handling capabilities for patient fall incidents. Third, we achieved full participation among all nurses. The training rate and assessment results of nurses in the wards were included in the performance management projects of ward head nurses, achieving the participation of all nurses and comprehensive evaluation of standardized fall prevention training. Based on the strategies mentioned above, the FRMIS's implementation in clinical practice has been relatively successful.

Limitations

This study still has certain limitations that should be acknowledged. First, the FRMIS was specifically designed and implemented by our hospital's computer center. It is currently applicable to 3 different hospital campuses within our institution

but has not been widely disseminated to other hospitals or integrated with diverse HISs. Therefore, its applicability and effectiveness in different hospital contexts remains uncertain. Second, the FRMIS heavily relied on the voluntary reporting by clinical nurses. The accuracy of these fall risk reports needed to be individually verified by staff members in the quality control office of the nursing department. This process is currently manual and lacks automation, which may introduce delays and potential inconsistencies. In the future, further improvements could be made by integrating artificial intelligence (AI) technologies. By automatically extracting fall risk factors from patients' electronic medical records, the system could achieve automated risk stratification and reduce dependence on manual reporting.

Despite these limitations, it is important to note that this study represents a significant step toward enhancing inpatient fall risk management through the FRMIS implementation. Future research and development efforts could focus on expanding the system's applicability to other hospitals, integrating AI capabilities for automated risk assessment, and improving data accuracy and automation processes. These advancements would contribute to more comprehensive and intelligent fall risk management practices for inpatients.

Conclusions

The design and implementation of an FRMIS significantly contributed to the prevention and management of falls among inpatients. The FRMIS enhanced patient safety through IT, providing comprehensive support for fall prevention and ensuring efficient management of fall events in health care settings.

Acknowledgments

We sincerely thank the nursing management and the participating nurses of the Tongji hospital for their support and participation in this study. This study was partly funded by the Huazhong University of Science and Technology Independent Innovation Fund (2013YQ008, 2018KFYYXJJ016), Chinese Nursing Association Research Project (ZHKY202204), and China Nursing Management Research Fund (CNM-2020-03).

Data Availability

The data sets used or analyzed during this study available from the corresponding author on reasonable request.

Authors' Contributions

WY designed the study. JMY and HM collected the data. HM and DMJ analyzed the data. JMY wrote the original draft of the manuscript. WY and HM reviewed and edited the manuscript. WY applied for funding. All authors have read and agreed to the version of the manuscript intended for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The presentation of the fall risk management information system.
[DOCX File , 766 KB - medinform_v12i1e46501_app1.docx]

References

1. Oner B, Zengul FD, Oner N, Ivankova NV, Karadag A, Patrician PA. Nursing-sensitive indicators for nursing care: A systematic review (1997-2017). Nurs Open 2021 May;8(3):1005-1022 [FREE Full text] [doi: 10.1002/nop2.654] [Medline: 34482649]
2. James S, Lucchesi L, Bisignano C, Castle C, Dingels Z, Fox J, et al. The global burden of falls: global, regional and national estimates of morbidity and mortality from the Global Burden of Disease Study 2017. Inj Prev 2020 Oct;26(Supp 1):i3-i11 [FREE Full text] [doi: 10.1136/injuryprev-2019-043286] [Medline: 31941758]
3. Burns ER, Stevens JA, Lee R. The direct costs of fatal and non-fatal falls among older adults - United States. J Safety Res 2016 Sep;58:99-103 [FREE Full text] [doi: 10.1016/j.jsr.2016.05.001] [Medline: 27620939]
4. Eslami Andargoli A, Scheepers H, Rajendran D, Sohal A. Health information systems evaluation frameworks: a systematic review. Int J Med Inform 2017 Jan;97:195-209. [doi: 10.1016/j.ijmedinf.2016.10.008] [Medline: 27919378]
5. Gaspar AGM, Lapão LV. eHealth for addressing balance disorders in the elderly: systematic review. J Med Internet Res 2021 Apr 28;23(4):e22215 [FREE Full text] [doi: 10.2196/22215] [Medline: 33908890]
6. Field M, Fong K, Shade C. Use of electronic visibility boards to improve patient care quality, safety, and flow on inpatient pediatric acute care units. J Pediatr Nurs 2018 Jul;41:69-76. [doi: 10.1016/j.pedn.2018.01.015] [Medline: 29395791]

7. Woodward M, De Pennington N, Grandidge C, McCulloch P, Morgan L. Development and evaluation of an electronic hospital referral system: a human factors approach. Ergonomics 2020 Jun;63(6):710-723. [doi: 10.1080/00140139.2020.1748232] [Medline: 32220218]
8. Chow CB, Leung M, Lai A, Chow YH, Chung J, Tong KM, et al. Development of an electronic emergency department-based geo-information injury surveillance system in Hong Kong. Injury 2012 Jun;43(6):739-748. [doi: 10.1016/j.injury.2011.08.008] [Medline: 21924722]
9. Carrillo I, Mira JJ, Vicente MA, Fernandez C, Guilabert M, Ferrús L, et al. Design and testing of BACRA, a web-based tool for middle managers at health care facilities to lead the search for solutions to patient safety incidents. J Med Internet Res 2016 Sep 27;18(9):e257 [FREE Full text] [doi: 10.2196/jmir.5942] [Medline: 27678308]
10. Balaguera HU, Wise D, Ng CY, Tso H, Chiang W, Hutchinson AM, et al. Using a medical intranet of things system to prevent bed falls in an acute care hospital: a pilot study. J Med Internet Res 2017 May 04;19(5):e150 [FREE Full text] [doi: 10.2196/jmir.7131] [Medline: 28473306]
11. Dalal AK, Fuller T, Garabedian P, Ergai A, Balint C, Bates DW, et al. Systems engineering and human factors support of a system of novel EHR-integrated tools to prevent harm in the hospital. J Am Med Inform Assoc 2019 Jun 01;26(6):553-560 [FREE Full text] [doi: 10.1093/jamia/ocz002] [Medline: 30903660]
12. Stockwell-Smith G, Adeleye A, Chaboyer W, Cooke M, Phelan M, Todd J, et al. Interventions to prevent in-hospital falls in older people with cognitive impairment for further research: a mixed studies review. J Clin Nurs 2020 Sep;29(17-18):3445-3460. [doi: 10.1111/jocn.15383] [Medline: 32578913]
13. Tricco AC, Thomas SM, Veroniki AA, Hamid JS, Cogo E, Strifler L, et al. Quality improvement strategies to prevent falls in older adults: a systematic review and network meta-analysis. Age Ageing 2019 May 01;48(3):337-346 [FREE Full text] [doi: 10.1093/ageing/afy219] [Medline: 30721919]
14. LeLaurin JH, Shorr RI. Preventing falls in hospitalized patients: state of the science. Clin Geriatr Med 2019 May;35(2):273-283 [FREE Full text] [doi: 10.1016/j.cger.2019.01.007] [Medline: 30929888]
15. Cameron ID, Dyer SM, Panagoda CE, Murray GR, Hill KD, Cumming RG, et al. Interventions for preventing falls in older people in care facilities and hospitals. Cochrane Database Syst Rev 2018 Sep 07;9(9):CD005465 [FREE Full text] [doi: 10.1002/14651858.CD005465.pub4] [Medline: 30191554]
16. Ambrens M, Tiedemann A, Delbaere K, Alley S, Vandelanotte C. The effect of eHealth-based falls prevention programmes on balance in people aged 65 years and over living in the community: protocol for a systematic review of randomised controlled trials. BMJ Open 2020 Jan 15;10(1):e031200 [FREE Full text] [doi: 10.1136/bmjopen-2019-031200] [Medline: 31948985]
17. Turner K, Staggs V, Potter C, Cramer E, Shorr R, Mion LC. Fall prevention implementation strategies in use at 60 United States hospitals: a descriptive study. BMJ Qual Saf 2020 Dec;29(12):1000-1007 [FREE Full text] [doi: 10.1136/bmjqs-2019-010642] [Medline: 32188712]
18. Morgan L, Flynn L, Robertson E, New S, Forde-Johnston C, McCulloch P. Intentional Rounding: a staff-led quality improvement intervention in the prevention of patient falls. J Clin Nurs 2017 Jan;26(1-2):115-124. [doi: 10.1111/jocn.13401] [Medline: 27219073]
19. Giles LC, Whitehead CH, Jeffers L, McErlean B, Thompson D, Crotty M. Falls in hospitalized patients: can nursing information systems data predict falls? Comput Inform Nurs 2006;24(3):167-172. [doi: 10.1097/00024665-200605000-00014] [Medline: 16707948]
20. Mei YY, Marquard J, Jacelon C, DeFeo AL. Designing and evaluating an electronic patient falls reporting system: perspectives for the implementation of health information technology in long-term residential care facilities. Int J Med Inform 2013 Nov;82(11):e294-e306. [doi: 10.1016/j.ijmedinf.2011.03.008] [Medline: 21482183]
21. Katsulis Z, Ergai A, Leung WY, Schenkel L, Rai A, Adelman J, et al. Iterative user centered design for development of a patient-centered fall prevention toolkit. Appl Ergon 2016 Sep;56:117-126. [doi: 10.1016/j.apergo.2016.03.011] [Medline: 27184319]
22. Dykes PC, Duckworth M, Cunningham S, Dubois S, Driscoll M, Feliciano Z, et al. Pilot Testing Fall TIPS (Tailoring Interventions for Patient Safety): a Patient-Centered Fall Prevention Toolkit. Jt Comm J Qual Patient Saf 2017 Aug;43(8):403-413. [doi: 10.1016/j.jcjq.2017.05.002] [Medline: 28738986]
23. Jacobsohn GC, Leaf M, Liao F, Maru AP, Engstrom CJ, Salwei ME, et al. Collaborative design and implementation of a clinical decision support system for automated fall-risk identification and referrals in emergency departments. Healthc (Amst) 2022 Mar;10(1):100598 [FREE Full text] [doi: 10.1016/j.hjdsi.2021.100598] [Medline: 34923354]
24. Mlaver E, Schnipper JL, Boxer RB, Breuer DJ, Gershanik EF, Dykes PC, et al. User-centered collaborative design and development of an inpatient safety dashboard. Jt Comm J Qual Patient Saf 2017 Dec;43(12):676-685. [doi: 10.1016/j.jcjq.2017.05.010] [Medline: 29173289]
25. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
26. Morse J. Preventing Patient Falls. Thousand Oaks, CA: Sage Publications; 1997.

27. Schwendimann R, De Geest S, Milisen K. Evaluation of the Morse Fall Scale in hospitalised patients. Age Ageing 2006 May;35(3):311-313. [doi: 10.1093/ageing/afj066] [Medline: 16527829]
28. Gardner LA, Bray PJ, Finley E, Sterner C, Ignudo TL, Stauffer CL, et al. Standardizing falls reporting: using data from adverse event reporting to drive quality improvement. J Patient Saf 2019 Jun;15(2):135-142. [doi: 10.1097/PTS.0000000000000204] [Medline: 26332598]

Abbreviations

AI: artificial intelligence
FRMIS: fall risk management information system
HIS: hospital information system
HL7FHIR: Health Level Seven Fast Healthcare Interoperability Resources
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
TIPS: Tailoring Interventions for Patient Safety

Edited by C Perrin; submitted 14.02.23; peer-reviewed by M Binandeh, R Hu, A Krupp; comments to author 06.04.23; revised version received 15.08.23; accepted 29.11.23; published 02.01.24.

Please cite as:
Wang Y, Jiang M, He M, Du M
Design and Implementation of an Inpatient Fall Risk Management Information System
JMIR Med Inform 2024;12:e46501
URL: https://medinform.jmir.org/2024/1/e46501
doi:10.2196/46501
PMID:38165733

Implementation Report

A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report

Mustafa Mahir Ulgu[1], MD, PhD; Gokce Banu Laleci Erturkmen[2], PhD; Mustafa Yuksel[2], PhD; Tuncay Namli[2], PhD; Şenan Postacı[2], MSc; Mert Gencturk[2], PhD; Yildiray Kabak[2], PhD; A Anil Sinaci[2], PhD; Suat Gonul[2], PhD; Asuman Dogac[2], PhD; Zübeyde Özkan Altunay[1], MD; Banu Ekinci[1], MD; Sahin Aydin[1], MSc; Suayip Birinci[1], MD

[1]Ministry of Health Turkey, Ankara, Turkey

[2]Software Research Development and Consultancy Corporation, Ankara, Turkey

Corresponding Author:
Gokce Banu Laleci Erturkmen, PhD
Software Research Development and Consultancy Corporation
Orta Dogu Teknik Universitesi Teknokent Silikon Blok Kat 1 No 16
Ankara, 06800
Turkey
Phone: 90 3122101763
Email: gokce@srdc.com.tr

Abstract

Background: The increasing population of older adults has led to a rise in the demand for health care services, with chronic diseases being a major burden. Person-centered integrated care is required to address these challenges; hence, the Turkish Ministry of Health has initiated strategies to implement an integrated health care model for chronic disease management. We aim to present the design, development, nationwide implementation, and initial performance results of the national Disease Management Platform (DMP).

Objective: This paper's objective is to present the design decisions taken and technical solutions provided to ensure successful nationwide implementation by addressing several challenges, including interoperability with existing IT systems, integration with clinical workflow, enabling transition of care, ease of use by health care professionals, scalability, high performance, and adaptability.

Methods: The DMP is implemented as an integrated care solution that heavily uses clinical decision support services to coordinate effective screening and management of chronic diseases in adherence to evidence-based clinical guidelines and, hence, to increase the quality of health care delivery. The DMP is designed and implemented to be easily integrated with the existing regional and national health IT systems via conformance to international health IT standards, such as Health Level Seven Fast Healthcare Interoperability Resources. A repeatable cocreation strategy has been used to design and develop new disease modules to ensure extensibility while ensuring ease of use and seamless integration into the regular clinical workflow during patient encounters. The DMP is horizontally scalable in case of high load to ensure high performance.

Results: As of September 2023, the DMP has been used by 25,568 health professionals to perform 73,715,269 encounters for 16,058,904 unique citizens. It has been used to screen and monitor chronic diseases such as obesity, cardiovascular risk, diabetes, and hypertension, resulting in the diagnosis of 3,545,573 patients with obesity, 534,423 patients with high cardiovascular risk, 490,346 patients with diabetes, and 144,768 patients with hypertension.

Conclusions: It has been demonstrated that the platform can scale horizontally and efficiently provides services to thousands of family medicine practitioners without performance problems. The system seamlessly interoperates with existing health IT solutions and runs as a part of the clinical workflow of physicians at the point of care. By automatically accessing and processing patient data from various sources to provide personalized care plan guidance, it maximizes the effect of evidence-based decision support services by seamless integration with point-of-care electronic health record systems. As the system is built on international code systems and standards, adaptation and deployment to additional regional and national settings become easily possible. The nationwide DMP as an integrated care solution has been operational since January 2020, coordinating effective screening and management of chronic diseases in adherence to evidence-based clinical guidelines.

(JMIR Med Inform 2024;12:e49986) doi:10.2196/49986

KEYWORDS

chronic disease management; clinical decision support services; integrated care; interoperability; evidence-based medicine; medicine; disease management; management; implementation; decision support; clinical decision; support; chronic disease; physician-centered; risk assessment; tracking; diagnosis

Introduction

As in the rest of the world, the aging population is increasing rapidly in Turkey. A recent TurkStat report predicts that by 2030, the older adult population will be 12.9%, rising to 22.6% in 2060 and 25.6% in 2080 [1]. Noncommunicable diseases are the leading cause of death and disability in Turkey, posing a significant burden [2]. The elevated health costs for older adults strain Turkey's health care system. To address this, the Turkish Ministry of Health (MOH) has implemented a national strategy emphasizing multidisciplinary teams, led by family physicians. The goal is to enhance early detection and manage complications of noncommunicable diseases through systematic screening programs under the national Disease Management Platform (DMP) project launched in late 2018.

The growing use of digital health solutions such as electronic health records (EHRs) presents an opportunity to enhance chronic disease management. Clinical decision support services (CDSSs) can assist in making patient-centered and evidence-based decisions [3,4]. Digital tools and systems that collect and use patient information to provide decision support for health care professionals (HCPs), including patient-specific assessments and recommendations, can promote adherence to national guidelines, ultimately resulting in enhanced quality of care [5-9]. Research demonstrated that computerized decision support tailored to the patient successfully improved decision-making [10,11]. Such tools enhanced the decision-making abilities of HCPs in various domains, including effective prescription decisions [12,13], adherence to guidelines for cardiac rehabilitation [14], management of hypertension and diabetes [15-21], cancer screening [22,23], and computerized order decisions [24,25].

Building on these results, the national DMP is designed as an integrated care platform for chronic disease management in Turkey in a family physician–centered manner. It aims to effectively implement clinical treatment protocols, ensuring easy adherence with decision support services. These services focus on early diagnosis, followed by structured treatment recommendations during routine follow-ups. The DMP enhances standardization of care, improving health care efficiency and quality. It also facilitates seamless transitions between primary care and specialist services, reducing costs, minimizing risks, eliminating redundant tests, and easing the burden on patients.

To ensure successful implementation of a DMP aimed at achieving these strategic objectives, several technical challenges need to be addressed. Our design decisions consider the crucial factor of integrating CDSSs seamlessly into clinicians' daily workflow [26,27]. Despite the potential of CDSSs for evidence-based medicine, significant effort is needed to realize these benefits [28]. The DMP must smoothly integrate with physicians' workflows, necessitating interoperability with existing health IT systems. CDSS guidance should be user-friendly, ensuring a natural flow for clinical protocol implementation. With a target audience of over 26,000 practitioners in Turkey, serving a population of over 85 million, the platform must ensure high performance and scalability. It should easily expand to address additional diseases within a reasonable timeframe and prioritize reusability and compliance with international health IT standards for versatile deployment.

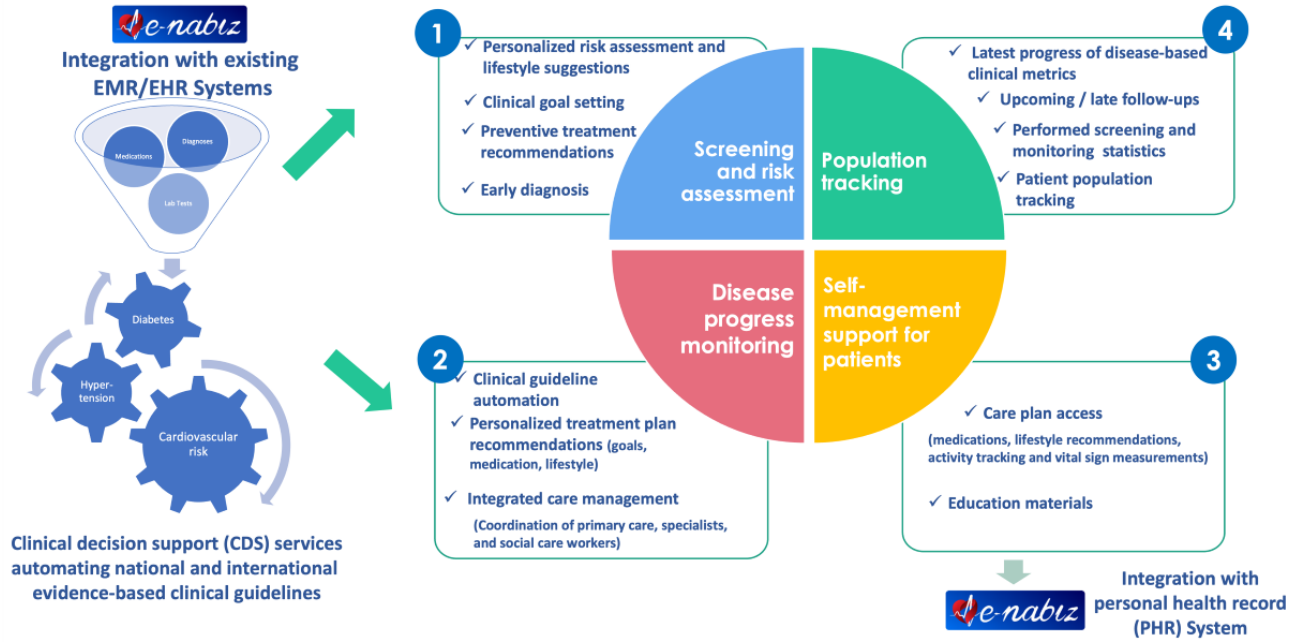
This paper outlines the design, development, nationwide implementation, and initial performance results of the national DMP in Turkey. The DMP can be categorized as a "2.3-Healthcare Provider Decision Support System" in terms of World Health Organization "Classification of digital health interventions" [29]. This implementation report will focus on the results of the deployment and implementation of the DMP in Turkey serving to more than 26,000 family medicine practitioners (FMPs) in the country. The objective is to share our experiences in building the DMP, as an implementation report in line with *iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations* [30]. We detail the design decisions and technical solutions aimed at ensuring interoperability with existing IT systems, integration with clinical workflow, enabling smooth transition of care, user-friendliness for HCPs, scalability, and adaptability in the *Methods* section. The *Implementation (Results)* section presents the outcomes of the nationwide implementation (number of users, number of screening and monitoring encounters, number of patients covered via these encounters, number of patients diagnosed as a result of screening encounters, and treatment goal achievements [such as blood pressure targets, hemoglobin A_{1c} {HbA_{1c}}, and cholesterol targets]), demonstrating how these objectives were achieved. Additionally, we outline current limitations and identify areas for future work to further enhance the clinical impact.

Methods

Overall System Architecture and Design Decisions

The DMP has been designed and implemented to enable the following 4 high-level features as summarized in Figure 1:

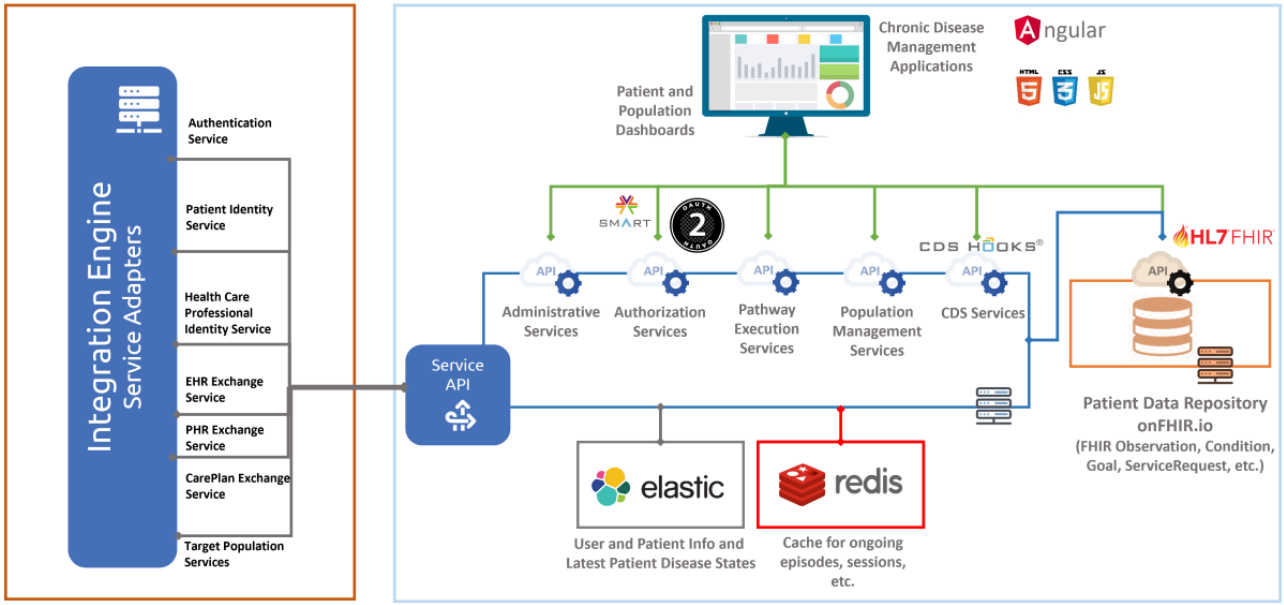
Figure 1. Overall aims of the disease management platform architecture. EHR: electronic health record; EMR: electronic medical record.



- Screening and risk assessment for healthy population: a web-based platform for FMPs facilitates screening for the healthy population. For instance, diabetes screening is required every 3 years for citizens aged over 40 years without a diabetes diagnosis. The full eligibility criteria for both screening and monitoring are presented in Multimedia Appendix 1. The system offers personalized risk assessments, early diagnosis, individualized goals, preventive treatment, and lifestyle suggestions aligned with national care pathways. Diagnosed patients enter the disease progress monitoring program, whereas undiagnosed individuals receive intensified screening based on risk and lifestyle recommendations.
- Disease progress monitoring: for diagnosed patients, the platform facilitates creating and updating personalized care plans during regular follow-up encounters, aligning with evidence-based national care pathways. It assesses laboratory results, conducts risk assessments, recommends personalized treatment goals and medications, suggests follow-up appointments, and refers to specialists when necessary for consultations and complication management. Patients in the monitoring program are categorized based on their control of clinical parameters, symptoms, and goal achievement status, guiding decisions on follow-up frequency, secondary care referrals, and medication plan updates.
- Self-management support for patients: a care plan with instructions for FMPs, specialists, and patients is shared with Turkey's e-Nabiz platform, the national EHR and personal health record (PHR) system. Patients can then access details about care plan activities, including medications, educational materials, self-measurement activities, and lifestyle recommendations.
- Population tracking: each FMP manages 2000 to 4000 patients based on their region's population. The population tracking module allows them to filter and manage patients for upcoming or overdue screening and monitoring encounters, access statistics on the screened population, send SMS invitations to patients, and monitor goal achievement for clinical parameters such as fasting plasma glucose, HbA_{1c}, and blood pressure.

The overall system architecture of the DMP is depicted in Figure 2.

Figure 2. High-level system architecture of the disease management platform. API: application programming interface; CDS: clinical decision support; EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven; PHR: personal health record.



Seamless Integration and Interoperability With Existing Systems

The DMP is designed and implemented for seamless integration with existing regional and national health IT systems. To achieve this, we have designed the core data model and data processing architecture of the DMP based on Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) Release 4 [31]. FHIR has gained widespread adoption in the health care industry [32-36] and endorsed by country-wide implementations in the United States [34], United Kingdom [37], and Germany [38].

The DMP core data model conforms to HL7 FHIR Release 4 to encompass basic EHR components as well as resources for representing a patient's care plan. An open-source HL7 FHIR Repository, namely onFHIR.io [39], serves as the main component of the data management layer (Figure 2). onFHIR.io uses MongoDB as a database and provides real-time data subscription with the help of Apache Kafka. The DMP web application directly accesses patient and care plan data through RESTful interfaces provided by onFHIR.io, enabling fine-grained access control over all FHIR resources in compliance with the SMART on FHIR authorization guidelines and scopes [40].

In Turkey, the MOH operates e-Nabiz, a central national EHR infrastructure [41]. This system collects patient records as encounter summaries from nationwide health care providers, with patients also inputting vital signs and activity data. e-Nabiz codes data using international and national medical terminology, such as *International Classification of Diseases, Tenth Revision* (ICD-10). It is a document-based repository accessed through a Representational State Transfer application programming interface [42], and interoperability adapters in the DMP project (EHR exchange and PHR exchange services in Figure 2) communicate with it to retrieve patient data. These adapters transform proprietary XML formats to HL7 FHIR-based data models and store them in the Patient Data Repository. This transformation includes both structural and semantic mapping, incorporating a strategy of incremental synchronization. On initial DMP access, the patient's longitudinal EHR is mapped to FHIR, and subsequent encounters retrieve and transform only new, unsynchronized data.

To secure patient data access, clinicians authenticate to the DMP through the MOH's central authentication and authorization services using the OpenID Connect protocol. The DMP uses a role-based access control mechanism, catering to different disease management roles. Before data access and synchronization, a check ensures that the user has the required access rights via the MOH's central authentication service. If authorized, the DMP generates a patient-specific JavaScript Object Notation Web Token with corresponding permissions, serving as an OAuth2.0 bearer token for all interactions within the DMP.

In the DMP, FMPs perform screening and monitoring encounters based on predefined eligibility criteria. For instance, hypertension monitoring is required every 3 months for patients with a hypertension diagnosis and on antihypertensive medications. These criteria are executed in the e-Nabiz data warehouse, and both DMP and family medicine information systems retrieve target population lists through target population services (Figure 2). FMPs can easily identify if a visiting patient is on the screening or monitoring list via family medicine information systems, initiating a DMP encounter directly with a single sign-on integration.

The care plan created with the help of the DMP is stored as an HL7 FHIR *CarePlan* resource in the Patient Data Repository. It is shared with the e-Nabiz system via the Care Plan Exchange Service (Figure 2), enabling it to be accessible to the patient via e-Nabiz interfaces.

The DMP uses Elasticsearch technology for storing user information, basic patient attributes, and their current screening and monitoring statuses for each disease. Elasticsearch also serves as a system log repository. We have developed a Kibana

interface for monitoring system performance and geographical statistics. Redis is used as a caching system to temporally store information about ongoing encounters and user authorization access tokens.

Automation of National Care Pathways as a Clinical Workflow for FMPs

The interfaces of the DMP have been designed with ease of use in mind to allow for seamless integration into the regular clinical workflow. It is implemented as a cocreation activity with the involvement of system analysts, software engineers, and a clinical reference group set up by the MOH Department of Chronic Diseases and Elderly Health including multidisciplinary HCPs.

The national evidence-based care pathways have been collaboratively analyzed, leading to the identification of common steps, such as physical examination, medical history review, risk assessment, medication review, lab results review, diagnosis, clinical goal setting, pharmacological treatment planning, and nonpharmacological treatment planning. Each care pathway is designed modularly within the DMP as a series of pages corresponding to these common steps. These are organized as a flow of pages that is followed automatically based on patient parameters.

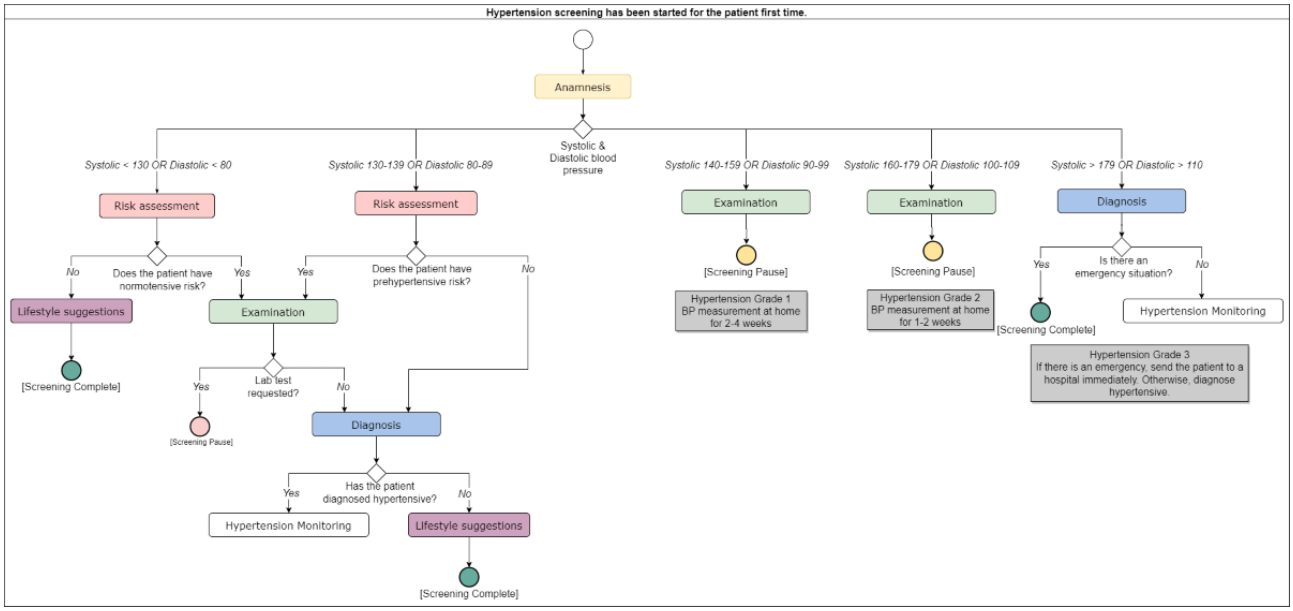
Each page is meticulously designed, specifying patient parameters for assessment. Most data come from the national EHR system, enabling clinicians to review prefilled pages with the latest parameters and make adjustments as needed. Validity periods for each parameter are identified, emphasizing recency, and they are enforced by the system and reminded to FMPs. Additionally, scaled assessments (eg, Mini Nutritional Assessment), risk assessments (eg, cardiovascular risk), and associated algorithms (eg, SCORE-Turkey) are also identified. Business rules within the pages are designed for personalized suggestions aligned with evidence-based care pathways.

All of these are thoroughly documented after discussions in cocreation workshops. Mock-up screens are designed, and flow diagrams are created to identify transition criteria between pages. These materials undergo further review and finalization in subsequent cocreation workshops. As an illustrative example, Figure 3 depicts a sample flow for hypertension screening.

After cocreation, each step's design becomes a web-based interface in the DMP application, developed with the Angular framework. A "Pathway Execution Service" state manager automates the flow diagram for disease screening or monitoring, adapting to patient parameters. This allows FMPs to use a wizard-like interface for encounters, facilitating adherence to national clinical pathways.

Transitions between disease modules are also modeled and implemented. For example, in hypertension screening, if a patient's fasting plasma glucose exceeds 110 mg/dL, the system prompts FMPs to consider a diabetes screening if not already monitored for diabetes. In response, the patient's diabetes screening schedule is automatically updated.

Figure 3. Hypertension screening flow. BP: blood pressure.



CDSS Implementation

CDSSs are a core component of DMP to enable patient-tailored recommendations. On the basis of the documented business rules from the design phase, we have designed CDSSs as automated processes. These processes link patient-specific data with evidence-based knowledge from national care pathways. We can categorize the CDSS implemented based on their functionality as follows:

- Risk assessment via scored algorithms (eg, SCORE-Turkey): FMPs are provided with explanatory guidance about scoring, referencing validated scoring assessment algorithms (see Figure 4).
- Diagnosis recommendations based on the patient's current condition and risk assessment: in screening operations, the CDSS recommends diagnoses to FMPs using predefined ICD-10 codes.

- Guidance for lab test ordering and interpretation: a personalized list of required lab tests is determined based on the patient's disease state, risks, and other comorbidities. The CDSS also provides notifications for when these lab tests should be renewed on expiration.
- Diagnosis and referral suggestions are recommended based on patient parameters such as lab results. For example, referral to a nephrologist is recommended when the estimated glomerular filtration rate result is below 60 mL/min/1.73 m^2.
- Treatment goals (eg, low-density lipoprotein cholesterol) are recommended based on the patient's risk, disease stage, and comorbidities. In Figure 5, an example screen for goal planning is presented. The physician can always manually update these targets based on their assessments.
- Medication suggestions are recommended for treatment planning based on disease stage, response to previous medications, existing medications, and comorbidities. Certain medications are marked as contraindications based on the existing comorbidities of the patient.
- Referral suggestions for preventive consultation visits are recommended, especially for complication management. For instance, a yearly retinopathy check with an ophthalmologist is advised during diabetes monitoring encounters.
- Follow-up visits are recommended based on the current status of the patient. For instance, screening in each 2 years is suggested for patients with low cardiovascular disease risk, whereas once a year screening is suggested for high-risk patients.
- Automated care pathway transitions for patients with multiple morbidities are personalized based on specific disease criteria. For instance, if a patient aged over 40 years has not had their cardiovascular risk score calculated, the DMP guides FMPs to continue with the cardiovascular risk module during hypertension or diabetes monitoring.

In the DMP, all CDSS implementations adhere to the CDS Hooks specifications [43]. As a standard published by HL7, it provides an API specification for CDS calls. Both input parameters and output suggestions are defined in reference to HL7 FHIR resources, facilitating plug-and-play interoperability with platforms that support HL7 FHIR. The CDS Hooks–compliant approach allows easy expansion with CDSSs created by external entities and to simplify deployment in different settings already adhering to HL7 FHIR.

Figure 4. An example screenshot from the Cardiovascular Risk Screening Module presenting individualized risk calculation. (The system is implemented in a multilingual manner supporting Turkish and English by default.) CVD: cardiovascular disease.

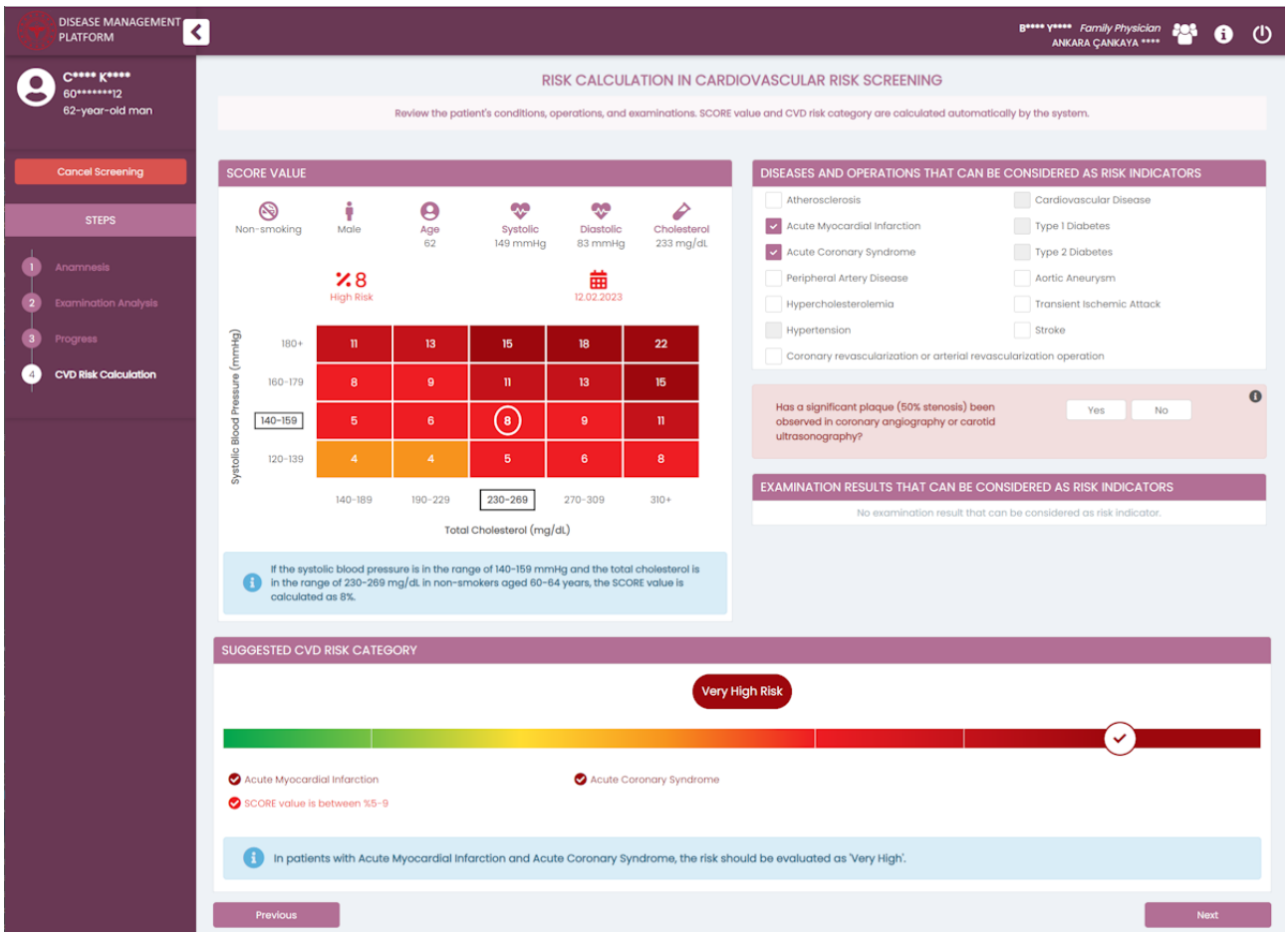
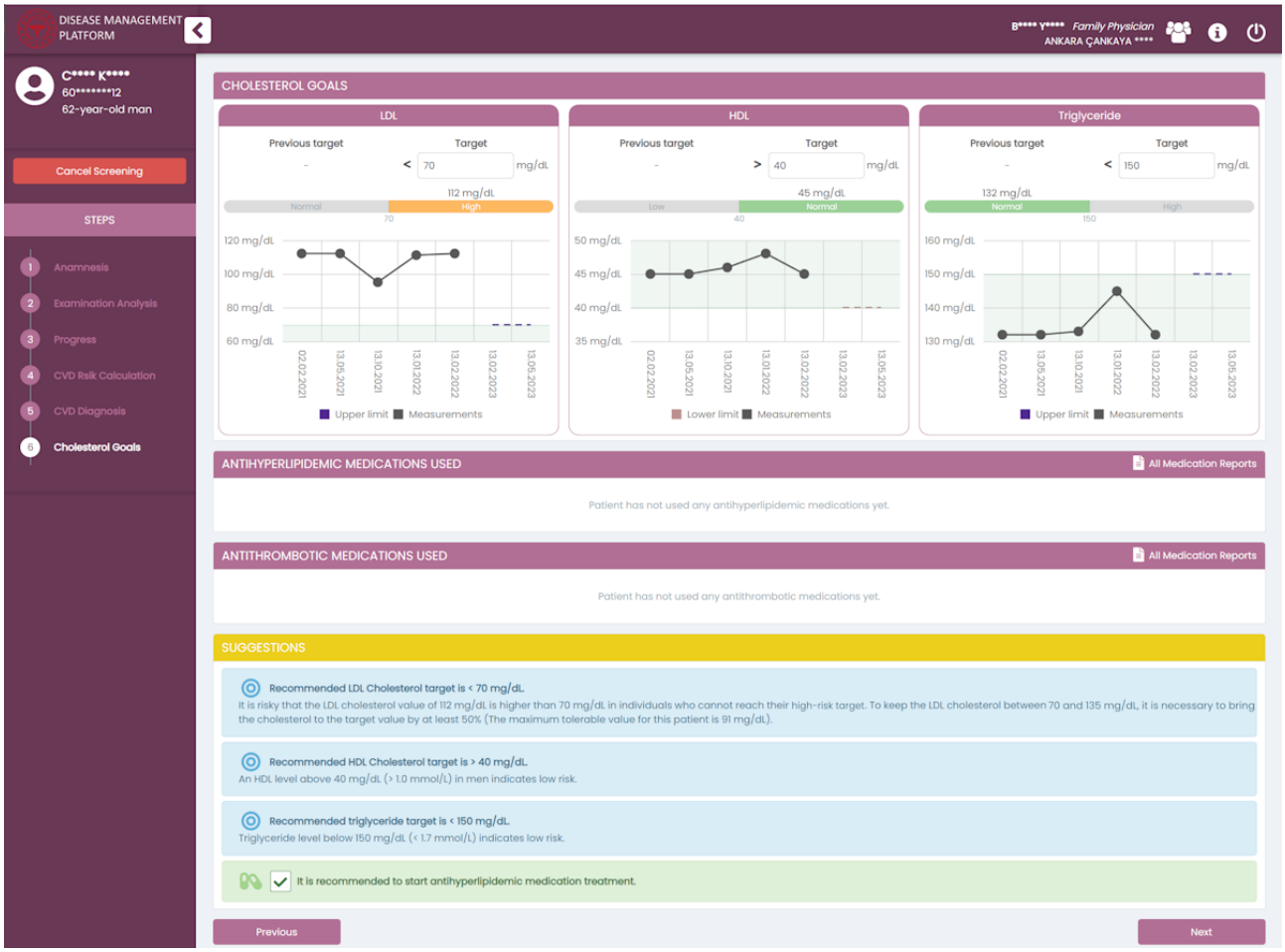


Figure 5. An example screenshot from the disease management platform presenting personalized lipid goals for the patient. HDL: high-density lipoprotein; LDL: low-density lipoprotein.



Ensuring Performance of the System

The DMP is designed for high horizontal scalability, using 2 servers for the web application and 3 for the Patient Data Repository, forming an onFHIR.io server cluster. Nginx acts as both a reverse proxy and a load balancer to distribute traffic across these backend servers. onFHIR.io servers connect to a horizontally scalable MongoDB cluster for data distribution and replication. Elasticsearch log and data store operate on a cluster hosted on 6 servers.

Testing, Piloting, and Deployment of Disease Modules

The system is developed by SRDC on behalf of the Turkish MOH with the support of Türksat and Innova. The final product is owned by the Turkish MOH. The initial version of the DMP, including modules for screening and monitoring of type 2 diabetes, hypertension, and cardiovascular risk management, was extensively tested by the clinical reference group. It underwent a 3-month pilot phase in 4 cities in late 2019. The pilot phase involved 14,351 encounters conducted by 219 FMPs for 5521 patients. Two more modules for obesity screening and monitoring, as well as older adult monitoring, were added to the system during this period. After the feedback is addressed and the system is retested, the system has been operationalized in whole Turkey by January 2020. On June 30, 2021, the MOH has published a directive incentivizing FMPs to conduct screening and monitoring for diabetes, hypertension, cardiovascular disease risk management, obesity, and older adult monitoring via the DMP. The system with incentivization calculations has been operational in whole Turkey since July 1, 2021.

Beyond existing modules, the system now includes monitoring modules for coronary artery disease, chronic kidney disease, stroke, chronic obstructive pulmonary disease, and asthma. The cocreation process, covering requirement analysis, mock-up design, implementation, and testing, took 3 months for each module, showcasing the process's repeatability for swift module additions. These new disease modules are not yet public in the operational system.

Implementation (Results)

The system is being used extensively throughout the whole country. As of September 18, 2023, a total of 73,715,269 screening and monitoring encounters have been performed by 25,568 users (24,627 FMPs and 941 FMP nurses) for 16,058,904 unique citizens. Among these citizens, 56.2% (n=9,025,104) are female and 43.8% (n=7,033,800) are male. The average number of DMP encounters per patient is 4.59. The distribution of encounters per DMP module and the breakdown between screening and monitoring is provided in Table 1.

In Turkey, there are 26,600 FMP units, with each unit using 1 FMP at a time. As of September 18, 2023, FMPs working at

26,210 (98.5%) unique FMP units have logged into the DMP at least once, and 22,982 (86.4%) FMP units have performed at least 1 encounter.

Table 2 details the nationwide coverage rates per disease module and encounter type as of September 18, 2023. It includes the cumulative target population size and the unique number of patients screened or monitored at least once. During this period, DMP screenings led to new diagnoses: 144,768 for hypertension, 490,346 for diabetes, 534,423 for high cardiovascular risk, and 3,545,573 for obesity. These individuals were diagnosed with these chronic diseases for the first time, following evidence-based clinical guidelines.

Age histograms of DMP patients who have been screened or monitored at least once are provided per sex in Figure 6.

Piloting studies occurred from October to December 2019, and the system has been fully operational nationwide since January 2020. Use notably increased with FMP salary incentivization calculations on July 1, 2021 (Figure 7), showing monthly encounter numbers by module from the start of 2021. Since then, encounters have steadily risen, with minor drops during summer holidays, and the distribution among DMP modules has remained consistent.

Figure 8 displays the distribution of total DMP encounters per city in Turkey, with colors intensifying as encounter numbers rise. Although higher numbers generally align with city populations, outliers exist, as seen in the top 10 performing cities outlined in Table 3. Despite Istanbul having Turkey's largest population, it only slightly surpasses Ankara in DMP encounters. This is mainly due to the high patient load per FMP in Istanbul. FMPs overseeing over 4000 citizens are exempt from DMP use due to their heavy workload. Table 3 also provides patient average age and encounter duration information.

The performance of the FMPs is assessed monthly. The cumulative targets and realized achievement rates for January 2023 are provided in Table 4. An achievement rate of 23.1% (4,508,841/19,546,041) for the entire population represents significant advancement compared with the 3.9% (511,198/13,117,900) achievement rate in July 2021.

The DMP system recommends personalized treatment goals such as systolic blood pressure, low-density lipoprotein cholesterol, and weight based on clinical guidelines. After a treatment goal is set, the DMP also assesses progress toward the goal in subsequent encounters. As of September 18, 2023, approximately 12.4 million of these treatment goals have been assessed, and the achievement rates are presented in Table 5. These assessments provide valuable information for FMPs caring for their patients.

At present, the performance of the FMPs is quantitatively calculated based on the number of performed encounters. However, the MOH envisions transitioning to a qualitative performance evaluation in midterm, where treatment goals and their achievement rates will play a significant role.

The system is highly performant and scalable. On a selected working day, February 14, 2023, the onFHIR.io HL7 FHIR Repository handled a total of 105.7 million FHIR interactions with an average response time of 31.3 milliseconds. During peak times of the day, the system can effortlessly manage up to 5000 FHIR interactions per second. Multimedia Appendix 2 illustrates the distribution and average response time of FHIR interactions on this day.

Among all FHIR requests, 57.4% (60.7 million) are search interactions, which are extensively used by the DMP web app to find, display, and forward specific clinical concept values to CDSS. Following search interactions, update interactions make up 33.2% (35.1 million) of the requests and are also used for resource creation when a provided resource ID is available. The average response times for read and search interactions are only 3.9 and 6.4 milliseconds, respectively. In the case of transactions and batch interactions, the average response times are even lower than update interaction alone, thanks to the parallelization of contained requests within onFHIR.io. As of September 18, 2023, onFHIR.io maintains a repository of 16.3 billion FHIR resources, totaling 22.4 terabytes in size, including care planning data by DMP and EHR/PHR data synchronized from e-Nabiz.

Table 1. Total screening and monitoring encounters per module.

Module	Screening (n=45,166,536), n (%)	Monitoring (n=28,548,733), n (%)	Total (n=73,715,269), n (%)
Hypertension	13,857,594 (30.7)	12,046,449 (42.2)	25,904,043 (35.1)
Obesity	18,029,994 (39.9)	800,480 (2.8)	18,830,474 (25.5)
Diabetes	8,914,193 (19.7)	5,071,646 (17.8)[a]	13,985,839 (19.0)
CVD[b] risk	4,364,755 (9.7)	9,182,814 (32.2)	13,547,569 (18.4)
Older adult	N/A[c]	1,447,344 (5.1)	1,447,344 (2.0)

[a]Only the patients monitored in primary care are listed; advanced obesity cases (a BMI over 40 kg/m^2 or a BMI between 30 and 40 kg/m^2 supported with additional comorbidities) are monitored in secondary and tertiary care.

[b]CVD: cardiovascular disease.

[c]N/A: not applicable.

Table 2. Coverage rate of citizens in target population lists.

Module and encounter type	All citizens in target population, n	Screened and monitored patients, n	Coverage rate (%)
Hypertension			
Screening	48,443,467	10,820,774	22.3
Monitoring	14,943,378	4,083,057	27.3
Obesity			
Screening	59,956,288	14,640,013	24.4
Monitoring	769,654[a]	383,920	49.9
Diabetes			
Screening	27,450,172	6,486,947	23.6
Monitoring	7,588,543	2,472,585	32.6
CVD[b] risk			
Screening	17,276,617	3,319,070	19.2
Monitoring	17,759,500	5,078,665	28.6
Older adult			
Monitoring	8,770,474	1,056,766	12.0

[a]Only those in the primary care obesity monitoring list, as explained in Table 1.

[b]CVD: cardiovascular disease.

Figure 6. Age histograms of disease management platform patients: female on the left and male on the right.

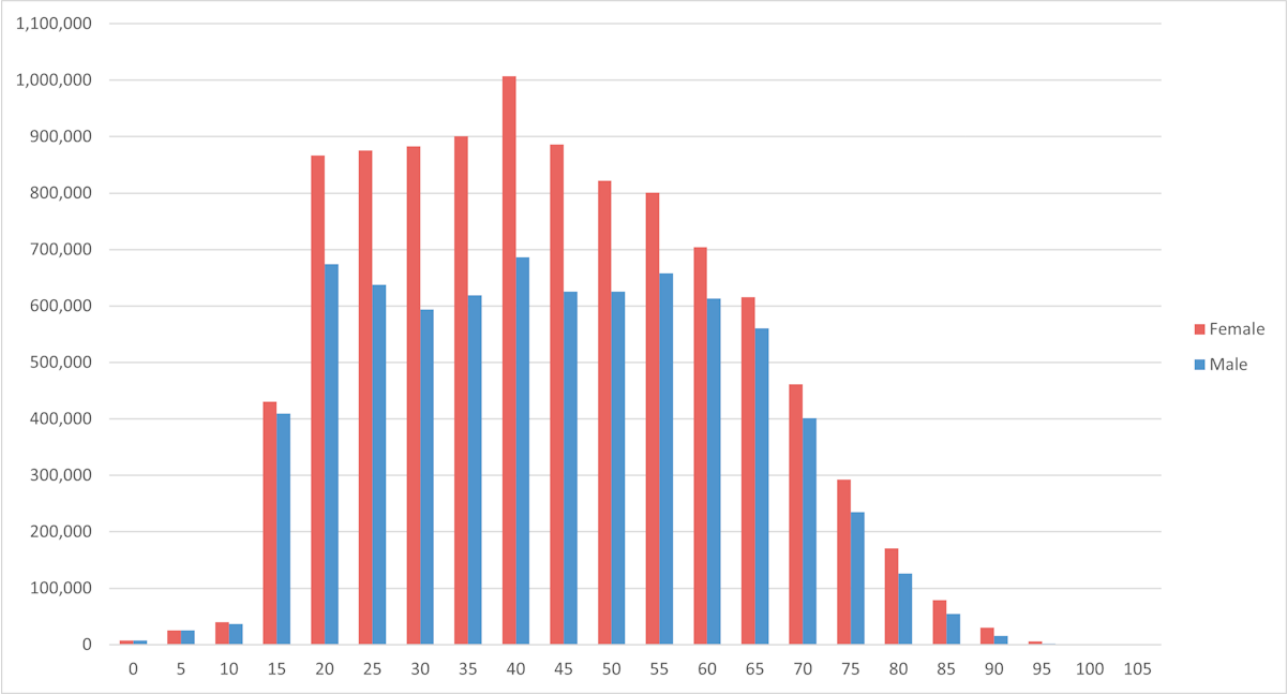


Figure 7. Disease management platform encounters per month by module. CVD: cardiovascular disease.

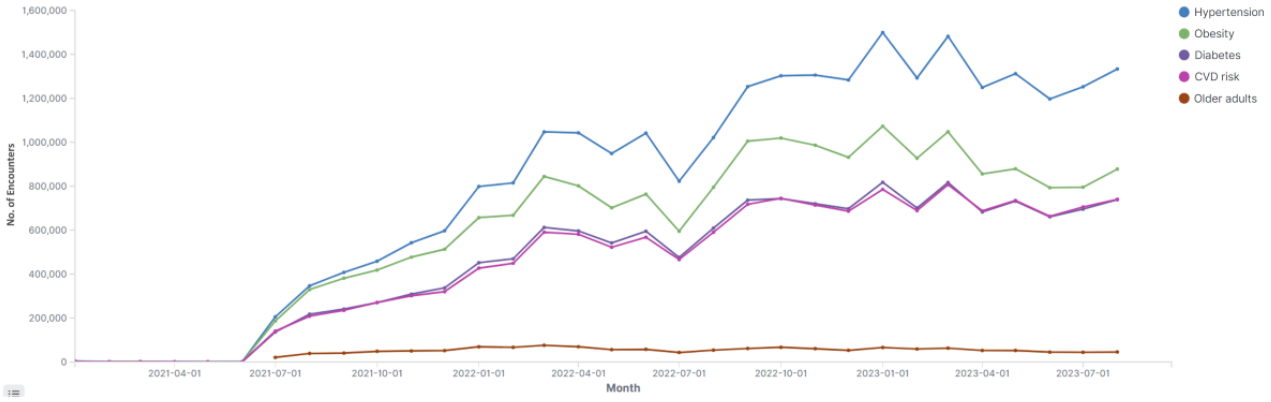


Figure 8. Encounters by city on a map.

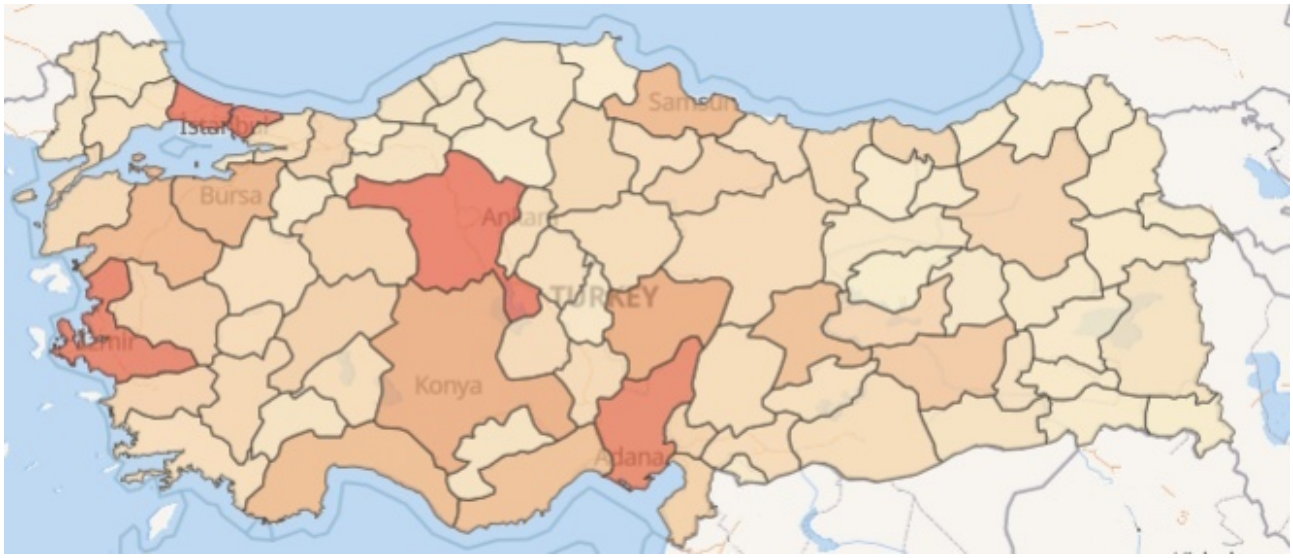


Table 3. Top 10 performing cities.

City	Total population[a]	Rank[b]	Number of encounters	Number of patients	Average age of patients (years)	Average duration (minutes)
Istanbul	15,907,951	1	4,974,972	1,286,640	50.5	1.21
Ankara	5,782,285	2	4,918,745	1,084,797	51.7	1.13
Izmir	4,462,056	3	4,395,653	937,607	53.4	1.18
Adana	2,274,106	7	3,531,441	709,075	51.3	0.99
Kayseri	1,441,523	15	2,633,349	479,849	51.7	1.06
Antalya	2,688,004	5	2,617,274	636,535	51.4	1.08
Konya	2,296,347	6	2,607,459	579,927	51.2	1.15
Bursa	3,194,720	4	2,408,817	525,016	52.3	1.18
Balikesir	1,257,590	17	2,322,111	425,171	55.7	1.13
Samsun	1,368,488	16	2,237,385	422,438	54.2	1.06

[a]2022 census data by the Turkish Statistical Institute (TurkStat).

[b]The rank of cities in Turkey by total population count.

Table 4. Screening and monitoring encounters per module in January 2023.

Module and encounter type	Monthly target	Number of encounters	Achievement rate (%)
Hypertension			
Screening	3,868,662	789,699	20.4
Monitoring	4,795,117	709,004	14.8
Obesity			
Screening	4,849,306	1,083,414	22.3
Monitoring	53,770	51,512	95.8
Diabetes			
Screening	898,665	670,502	74.6
Monitoring	2,128,955	279,071	13.1
CVD[a] risk			
Screening	742,634	262,419	35.3
Monitoring	1,488,004	589,523	39.6
Older adult			
Monitoring	720,928	73,697	10.2
Total	19,546,041	4,508,841	23.1

[a]CVD: cardiovascular disease.

Table 5. Achievement rates of treatment goals.

Treatment goal	Achievement rate (%)
Systolic BP[a]	88.8
Diastolic BP	94.2
Fasting glucose	52.0
HbA_{1c}[b]	61.5
LDL[c] cholesterol	14.8
HDL[d] cholesterol	63.2
Triglyceride	52.6
Weight	5.6
BMI	6.3
Waist circumference	2.9

[a]BP: blood pressure.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]LDL: low-density lipoprotein.

[d]HDL: high-density lipoprotein.

Discussion

Principle Findings and Lessons Learned

We have demonstrated that as of September 18, 2023, the DMP has been used by more than 25,000 users to conduct over 73 million screening and monitoring encounters for more than 16 million individuals. The national directive incentivizing FMPs to conduct screening and monitoring for chronic diseases is one of the contributing factors to this success.

We demonstrated the platform's efficient horizontal scalability, serving thousands of HCPs daily without performance issues. DMP screenings identified approximately 150,000 new hypertension cases, over 490,000 diabetes cases, more than 500,000 high cardiovascular risk cases, and over 3.5 million obesity cases. This allowed timely treatment in line with evidence-based guidelines.

We have shown that the system seamlessly interoperates with existing national EHR via HL7 FHIR. It enables accessing and processing patient data from various sources to provide personalized care plan guidance, maximizing the effectiveness

of evidence-based decision support services. The DMP has achieved all 5 levels of the 5S Model as proposed by Haynes [44] for the successful implementation of information services for evidence-based health care decisions. Continuous cocreation activity involving members of the Turkish MOH has contributed this success, along with the interoperability architecture based on international standards. On the other hand, we have collected feedback from FMPs to encourage us to also enable seamless integration with the national e-Prescription and national appointment system. FMPs need to manually input prescription and appointment recommendations into the other systems. Future plans include integrating these national systems directly to the DMP as well.

Although we have demonstrated that, through a repeatable and well-defined cocreation methodology, the system can be easily extended to address additional diseases, it still requires implementation effort from developers. We plan to extend the DMP system with administrative interfaces. This will enable subject matter experts from the MOH to create new disease screening and monitoring modules using form-based design interfaces.

Finally, although FMPs conduct screening and monitoring, specialists can view patient dashboards but cannot perform encounters; this can be easily enabled with the DMP's role-based access control mechanism, pending organizational decisions for national-scale implementation.

The system is operated as a part of national health IT ecosystem funded by the budget of the Turkish MOH. Open-source technologies have been used; hence, additional licensing fee has not incurred. Approximately 80% of the budget is spent for software development, 15% for project management, and 5% for training costs. Initial development phase has lasted 2 years. In the last 2.5 years, the system is under maintenance, and new disease modules have been developed.

Prospective Benefits and Impact

The system paves the way forward value-based care, where patient outcomes are monitored, and providers are incentivized for improving health. Currently, the DMP sets individual clinical goals (eg, HbA_{1c} and BMI) based on evidence-based guidelines. It monitors FMP performance in achieving these targets through close screening and monitoring. FMPs are presently incentivized based on screening and monitoring visits, but the system is ready to adopt value-based care by monitoring clinical targets.

DMP implementation opens opportunities to collect real-time research data, measuring the effectiveness of nationwide disease management protocols. Continuously gathering information about patients' disease status and recording outcomes from screening and monitoring visits, the generated data provide valuable insights into disease management.

Conclusions

This paper introduces a nationwide DMP designed for effective chronic disease screening and management, aligning with evidence-based clinical guidelines to enhance health care quality. With its user-friendly interfaces, it guides FMPs through personalized care planning with checklists for medication orders, referrals, lab tests, and risk screening. The system has been operational nationwide since January 2020. We have demonstrated seamless EHR integration, scalability, performance, and effectiveness in early diagnosis and meeting clinical targets. Future work includes a comprehensive study to analyze the direct clinical and cost-saving effects of the DMP on chronic disease management in Turkey.

Acknowledgments

The authors wish to acknowledge administrative and technical support by the following departments of the Turkish Ministry of Health: Department of Public Health Informatics, Department of Chronic Diseases and Elderly Health, Department of Healthy Nutrition and Active Life, Department of Data Management, and Department of Standards and Accreditation; Türksat, and Innova.

Data Availability

The data sets generated and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

GBLE and MMU carried out conceptualization of the paper; GBLE, MY, TN, SP, MG, YK, AAS, SG, AD, ZOA, and BE established the methodology of the study; GBLE, MY, TN, SP, MG, and YK developed the software; ZOA, BE, SA, MMU, and SB contributed validation studies; GBLE, MY, and TN carried out formal analysis; MY, TN, SP, MG, YK, AAS, and SG contributed to data curation; GBLE, MY, TN, SP, MG, and MMU wrote the manuscript; AAS, YK, SG, ZOA, BE, and SA reviewed and edited the manuscript; GBLE, MY, TN, SP, and MG created the visualizations in the manuscript; AD and SB supervised the study; GBLE, MY, BE, and MMU coordinated project administration. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Eligibility criteria for target populations for screening and monitoring encounters.
[DOCX File , 25 KB - medinform_v12i1e49986_app1.docx]

Multimedia Appendix 2
FHIR interactions in a day: request count on the left and average response time on the right. FHIR: Fast Healthcare Interoperability Resources.
[PNG File , 98 KB - medinform_v12i1e49986_app2.png]

References

1. Life tables, 2017-2019. Turkish Statistical Institute. 2020. URL: https://data.tuik.gov.tr/Bulten/Index?p=Hayat-Tablolari-2017-2019-33711 [accessed 2023-03-22]
2. Non-communicable diseases and risk factors cohort study for Turkey. Republic of Turkey Ministry of Health. URL: https://hsgm.saglik.gov.tr/depo/birimler/kronik-hastaliklar-ve-yasli-sagligi-db/Dokumanlar/Raporlar/v9s_NCDkohort_.pdf [accessed 2023-11-02]
3. Yucesan M, Gul M, Mete S, Celik E. A forecasting model for patient arrivals of an emergency department in healthcare management systems. In: Bouchemal N, editor. Intelligent Systems for Healthcare Management and Delivery. Hershey, Pennsylvania: IGI Global; 2019:266-284.
4. Omid P, Bilal A, Despotou G, Keung SNLC, Mohamad Y, Gappa H, et al. CAREPATH methodology for development of computer interpretable, integrated clinical guidelines. 2023 Presented at: DSAI 2022: 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion; August 31-September 2, 2022; Lisbon, Portugal. [doi: 10.1145/3563137.3563155]
5. Recommendations on digital interventions for health system strengthening–research considerations. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/WHO-RHR-19.9 [accessed 2023-11-02]
6. Agarwal S, Glenton C, Tamrat T, Henschke N, Maayan N, Fønhus MS, et al. Decision-support tools via mobile devices to improve quality of care in primary healthcare settings. Cochrane Database Syst Rev 2021;7(7):CD012944 [FREE Full text] [doi: 10.1002/14651858.CD012944.pub2] [Medline: 34314020]
7. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
8. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
9. Fortmann J, Lutz M, Spreckelsen C. System for context-specific visualization of clinical practice guidelines (GuLiNav): concept and software implementation. JMIR Form Res 2022;6(6):e28013 [FREE Full text] [doi: 10.2196/28013] [Medline: 35731571]
10. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293(10):1223-1238 [FREE Full text] [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
11. Shiffman RN, Liaw Y, Brandt CA, Corb GJ. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. J Am Med Inform Assoc 1999;6(2):104-114 [FREE Full text] [doi: 10.1136/jamia.1999.0060104] [Medline: 10094063]
12. Poller L, Shiach CR, MacCallum PK, Johansen AM, Münster AM, Magalhães A, et al. Multicentre randomised study of computerised anticoagulant dosage. European concerted action on anticoagulation. Lancet 1998;352(9139):1505-1509 [FREE Full text] [doi: 10.1016/s0140-6736(98)04147-6] [Medline: 9820298]
13. Samore MH, Bateman K, Alder SC, Hannah E, Donnelly S, Stoddard GJ, et al. Clinical decision support and appropriateness of antimicrobial prescribing: a randomized trial. JAMA 2005;294(18):2305-2314 [FREE Full text] [doi: 10.1001/jama.294.18.2305] [Medline: 16278358]
14. Goud R, de Keizer NF, ter Riet G, Wyatt JC, Hasman A, Hellemans IM, et al. Effect of guideline based computerised decision support on decision making of multidisciplinary teams: cluster randomised trial in cardiac rehabilitation. BMJ 2009;338:b1440 [FREE Full text] [doi: 10.1136/bmj.b1440] [Medline: 19398471]
15. Filippi A, Sabatini A, Badioli L, Samani F, Mazzaglia G, Catapano A, et al. Effects of an automated electronic reminder in changing the antiplatelet drug-prescribing behavior among Italian general practitioners in diabetic patients: an intervention trial. Diabetes Care 2003;26(5):1497-1500 [FREE Full text] [doi: 10.2337/diacare.26.5.1497] [Medline: 12716811]
16. Erturkmen GBL, Yuksel M, Sarigul B, Arvanitis TN, Lindman P, Chen R, et al. A collaborative platform for management of chronic diseases via guideline-driven individualized care plans. Comput Struct Biotechnol J 2019;17:869-885 [FREE Full text] [doi: 10.1016/j.csbj.2019.06.003] [Medline: 31333814]

17. von Tottleben M, Grinyer K, Arfa A, Traore L, Verdoy D, Keung SNLC, et al. An integrated care platform system (C3-Cloud) for care planning, decision support, and empowerment of patients with multimorbidity: protocol for a technology trial. JMIR Res Protoc 2022;11(7):e21994 [FREE Full text] [doi: 10.2196/21994] [Medline: 35830239]
18. Lobach DF, Hammond WE. Computerized decision support based on a clinical practice guideline improves compliance with care standards. Am J Med 1997;102(1):89-98 [FREE Full text] [doi: 10.1016/s0002-9343(96)00382-8] [Medline: 9209205]
19. Dexter PR, Perkins S, Overhage JM, Maharry K, Kohler RB, McDonald CJ. A computerized reminder system to increase the use of preventive care for hospitalized patients. N Engl J Med 2001;345(13):965-970 [FREE Full text] [doi: 10.1056/NEJMsa010181] [Medline: 11575289]
20. Wang Z, An J, Lin H, Zhou J, Liu F, Chen J, et al. Pathway-driven coordinated telehealth system for management of patients with single or multiple chronic diseases in China: system development and retrospective study. JMIR Med Inform 2021;9(5):e27228 [FREE Full text] [doi: 10.2196/27228] [Medline: 33998999]
21. Ramirez M, Chen K, Follett RW, Mangione CM, Moreno G, Bell DS. Impact of a "chart closure" hard stop alert on prescribing for elevated blood pressures among patients with diabetes: quasi-experimental study. JMIR Med Inform 2020;8(4):e16421 [FREE Full text] [doi: 10.2196/16421] [Medline: 32301741]
22. Burack RC, Gimotty PA, Simon M, Moncrease A, Dews P. The effect of adding Pap smear information to a mammography reminder system in an HMO: results of randomized controlled trial. Prev Med 2003;36(5):547-554 [FREE Full text] [doi: 10.1016/s0091-7435(02)00062-2] [Medline: 12689799]
23. McPhee SJ, Bird JA, Fordham D, Rodnick JE, Osborn EH. Promoting cancer prevention activities by primary care physicians. Results of a randomized, controlled trial. JAMA 1991;266(4):538-544. [Medline: 2061981]
24. Bates DW, Kuperman GJ, Rittenberg E, Teich JM, Fiskio J, Ma'luf N, et al. A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. Am J Med 1999;106(2):144-150. [doi: 10.1016/s0002-9343(98)00410-0] [Medline: 10230742]
25. Overhage JM, Tierney WM, Zhou XH, McDonald CJ. A randomized trial of "corollary orders" to prevent errors of omission. J Am Med Inform Assoc 1997;4(5):364-375 [FREE Full text] [doi: 10.1136/jamia.1997.0040364] [Medline: 9292842]
26. Wasylewicz ATM, Scheepers-Hoeks AMJW. Clinical decision support systems. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science. Cham (CH): Springer International Publishing; 2019.
27. Stagg BC, Stein JD, Medeiros FA, Wirostko B, Crandall A, Hartnett ME, et al. Special commentary: using clinical decision support systems to bring predictive models to the glaucoma clinic. Ophthalmol Glaucoma 2021;4(1):5-9 [FREE Full text] [doi: 10.1016/j.ogla.2020.08.006] [Medline: 32810611]
28. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc 2001;8(6):527-534 [FREE Full text] [doi: 10.1136/jamia.2001.0080527] [Medline: 11687560]
29. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://apps.who.int/iris/handle/10665/260480[accessed [accessed 2023-11-02]
30. Franck CP, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
31. Welcome to FHIR. HL7 FHIR Release 4. URL: https://hl7.org/fhir/R4/ [accessed 2023-12-22]
32. HL7 FHIR Accelerator™ Program. HL7 International. URL: https://www.hl7.org/about/fhir-accelerator/ [accessed 2023-11-02]
33. What is FHIR? The Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/sites/default/files/2019-08/ONCFHIRFSWhatIsFHIR.pdf [accessed 2023-11-02]
34. Heat wave: the U.S. is poised to catch FHIR in 2019. HealthITBuzz. 2018. URL: https://www.healthit.gov/buzz-blog/interoperability/heat-wave-the-u-s-is-poised-to-catch-fhir-in-2019[accessed [accessed 2023-11-02]
35. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
36. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
37. FHIR (Fast Healthcare Interoperability Resources). NHS Digital. 2022. URL: https://digital.nhs.uk/services/fhir-apis [accessed 2023-11-02]
38. Medical Informatics Initiative demonstrates ability to standardise healthcare data across Germany according to FHIR. Medical Informatics Initiative Germany. 2019. URL: https://www.medizininformatik-initiative.de/en/medizininformatik-initiative-bundesweit-einheitliche-auswertbarkeit-von-versorgungsdaten [accessed 2023-11-02]
39. HL7 FHIR® based secure data repository. onFHIR.io. URL: https://onfhir.io/ [accessed 2023-11-02]
40. SMART App launch: scopes and launch context. HL7 International. URL: http://hl7.org/fhir/smart-app-launch/1.0.0/scopes-and-launch-context/index.html [accessed 2023-11-02]

41. Birinci S. National Healthcare Technology Initiative. In: Kacır MF, Seker M, Dogrul M, editors. National Technology Initiative. Ankara: Turkish Academy of Sciences Publication; 2022:305-328.
42. USS Services. URL: https://usskurumsal.saglik.gov.tr/kurumsalservisler/#HYP [accessed 2023-11-02]
43. HL7 CDS hooks. HL7 International. URL: https://cds-hooks.hl7.org/ [accessed 2023-11-02]
44. Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evid Based Med 2006;11(6):162-164 [FREE Full text] [doi: 10.1136/ebm.11.6.162-a] [Medline: 17213159]

Abbreviations

CDSS: clinical decision support service
DMP: Disease Management Platform
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
FMP: family medicine practitioner
HbA_{1c}: hemoglobin A_{1c}
HCP: health care professional
HL7: Health Level Seven
ICD-10: International Classification of Diseases, Tenth Revision
MOH: Ministry of Health
PHR: personal health record

Edited by C Perrin; submitted 16.06.23; peer-reviewed by J Galvez-Olortegui, J Pevnick; comments to author 13.09.23; revised version received 21.09.23; accepted 29.11.23; published 19.01.24.

Please cite as:
Ulgu MM, Laleci Erturkmen GB, Yuksel M, Namli T, Postacı Ş, Gencturk M, Kabak Y, Sinaci AA, Gonul S, Dogac A, Özkan Altunay Z, Ekinci B, Aydin S, Birinci S
A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report
JMIR Med Inform 2024;12:e49986
URL: https://medinform.jmir.org/2024/1/e49986
doi:10.2196/49986
PMID:38241077

Implementation Report

Ten Years of Experience With a Telemedicine Platform Dedicated to Health Care Personnel: Implementation Report

Claudio Azzolini[1,2,3], MD; Elias Premi[3,4], MD; Simone Donati[3,5], MD; Andrea Falco[2,6], MEng; Aldo Torreggiani[7], MEng; Francesco Sicurello[3,8], BSc; Andreina Baj[5], MD; Lorenzo Azzi[5], MD; Alessandro Orro[2,3,8], MEng; Giovanni Porta[5], MD; Giovanna Azzolini[2], BSc; Marco Sorrentino[2,9], JD; Paolo Melillo[10], MEng; Francesco Testa[10], MD; Francesca Simonelli[10], MD; Gianfranco Giardina[11], MEng; Umberto Paolucci[12], MEng

[1]Advisory Council of e-Health and Telemedicine, University of Insubria of Varese-Como, Varese, Italy
[2]TM95 Srl, Milan, Italy
[3]Italian Association of Telemedicine and Medical Informatics, Milan, Italy
[4]Department of Life Sciences and Biotechnologies, University of Insubria, Varese-Como, Italy
[5]Department of Medicine and Surgery, University of Insubria, Varese-Como, Italy
[6]Alfa Design Studio, Milan, Italy
[7]T&C Srl, Milan, Italy
[8]Institute of Biomedical Technologies, National Research Council, Milan, Italy
[9]Bms Farm Law Firm, Milan, Italy
[10]Multidisciplinary Department of Medical, Surgical and Dental Sciences, University of Campania Luigi Vanvitelli, Naples, Italy
[11]DDay.it, Milan, Italy
[12]Up Invest Srl, Milan, Italy

Corresponding Author:
Claudio Azzolini, MD
Advisory Council of e-Health and Telemedicine
University of Insubria of Varese-Como
Via Guicciardini 9
Varese, 21100
Italy
Phone: 39 0332393603
Email: claudio.azzolini@uninsubria.it

Abstract

Background: Telemedicine, a term that encompasses several applications and tasks, generally involves the remote management and treatment of patients by physicians. It is known as transversal telemedicine when practiced among health care professionals (HCPs).

Objective: We describe the experience of implementing our telemedicine Eumeda platform for HCPs over the last 10 years.

Methods: A web-based informatics platform was developed that had continuously updated hypertext created using advanced technology and the following features: security, data insertion, dedicated software for image analysis, and the ability to export data for statistical surveys. Customizable files called “modules” were designed and built for different fields of medicine, mainly in the ophthalmology subspecialty. Each module was used by HCPs with different authorization profiles.

Implementation (Results): Twelve representative modules for different projects are presented in this manuscript. These modules evolved over time, with varying degrees of interconnectivity, including the participation of a number of centers in 19 cities across Italy. The number of HCP operators involved in each single module ranged from 6 to 114 (average 21.8, SD 28.5). Data related to 2574 participants were inserted across all the modules. The average percentage of completed text/image fields in the 12 modules was 65.7%. All modules were evaluated in terms of access, acceptability, and medical efficacy. In their final evaluation, the participants judged the modules to be useful and efficient for clinical use.

Conclusions: Our results demonstrate the usefulness of the telemedicine platform for HCPs in terms of improved knowledge in medicine, patient care, scientific research, teaching, and the choice of therapies. It would be useful to start similar projects across various health care fields, considering that in the near future medicine as we know it will completely change.

(JMIR Med Inform 2024;12:e42847) doi:10.2196/42847

KEYWORDS

telemedicine; ophthalmology; eHealth; informatics platform; health care professional; patient care; information technology; data warehouse

Introduction

Context

Medicine has typically involved physicians engaging face to face with patients. However, many teleconsultation projects have now been developed, particularly during the COVID-19 pandemic era, which has boosted teleconsultations in all medical specialties [1-4].

Alongside telemedicine between physicians and patients, there is also transversal telemedicine, which is conducted between health care professionals (HCPs). Our experience with this topic started in 1996 and has demonstrated the feasibility of training young ophthalmic vitreoretinal surgeons working in nonoptimal environments (postwar Bosnia), using telemedicine (via a satellite link) in Milan and Sarajevo [5,6]. Input from the above experiences [7,8] constituted the basis for our understanding of the needs of HCPs and the developmental direction of the dedicated telemedicine platform, giving users access, with appropriate personal authorization, from anywhere and at any time.

Problem Statement

The problem to be solved is the difficulty of sharing patients' clinical data and images among health personnel for efficient evaluation. This process should be multidisciplinary, involving actors such as physicians from different specialties, nurses, technicians, orthoptists, geneticists, residents, and tutors who need access to a common database holding key patient information.

Similar Interventions

Our scientific literature analysis identified a number of publications about implementation projects involving telemedicine platforms. These projects were mainly based on COVID-19 management and aimed to support different systems to provide health care in emergency conditions [9-11]. The purpose of these initiatives is to foster telecare and telemonitoring and to reduce the need for patients to visit hospitals or medical centers [12-17]. Our program is oriented in a different direction: the Eumeda web-based medical platform was developed for sharing patients' medical data among physicians. The platform has expanded its services to many HCPs. This paper describes how database modules for the clinical databank and trials, as well as second opinion services, were created and have now been implemented.

Methods

Aims and Objectives

The aim of this implementation program was to broaden the applications of our telemedicine platform with a transversal approach targeted at health care personnel. This process took place over the last 10 years with the creation of different projects aimed at clinical data collection, teleconsultation, and gathering second opinions. Various modules have been built for the platform (Textbox 1) for use by HCPs at different times. Twelve representative modules for different clinical projects are described in Table 1 [18-23].

We identified outcome measures and evaluated overall parameters for access, acceptability, and medical efficacy of the platform (Textbox 2).

Textbox 1. Building a module in 8 steps. The time required for the final release varies between modules (from 1 to 3 months for more complex ones). The original source code for the modules created belongs to the medical platform.

1. Initial agreement between the entity applying the module (university, company, institution, or representative association) and the manager of the medical platform (MP)

2. Signing of detailed operational form (with project requirements, such as the type of project, number of health care professionals and structures involved, and the importance of images) by the main users of the module and the scientific coordinator (who has knowledge of medicine planning and the potentiality and limits of medical informatics) of the MP

3. "Shoulder-to-shoulder" work by the scientific coordinator of the MP and main team programmer of the MP

4. Development of alpha software (not yet stable and still incomplete) to be shown to the entity that will use the module for changes and additions

5. Development of beta software with almost all functionalities

6. Massive data entry by the MP programmer to find bugs or software incompatibilities

7. Completion of beta software with automatic control functionalities (eg, alert icons to prevent inappropriate data from being entered, numerical limitations, and priorities to be respected in data entry or blocking of inappropriate saving) for users to check and identify any small changes required

8. Release of final version in a meeting with users, with explanatory text embedded in the module

Table 1. The left-hand column lists the 12 representative projects for which many modules have been built for the medical platform. The modules designed have been managed by health care personnel over the last 10 years in different locations in Italy.

Module	Description	Module type	Purpose	Timeframe of project activity	Holder	Sponsor
1	Teleconsultation in retinal diseases [18]	Second opinions[a]	Feasibility of second opinions among physicians	1 Month (during 2011)	Insubria University, Varese-Como	Comed Research nonprofit association, Milan
2	Age-related maculopathy [19]	Group[b] (10 locations)	Acceleration of anti–vascular endothelial growth factor therapy	19 Months (2011-2012)	T&C Srl, Milan, Italy	Novartis Pharma SpA, Origgio, Italy
3	Retinal pathology samples and correlated genes [20]	Data[c]	Collection of data on gene expression	4 Months (2012-2013)	Insubria University, Varese-Como	Insubria University, Varese-Como
4	Epiretinal macular membrane [21]	Data	Collection of data on disease morphology and functionality	10 Months (2015-2016)	Insubria University, Varese-Como	Insubria University, Varese-Como
5	Inherited eye diseases	Data	Collection of data on genetic eye diseases	2017-present	Ophthalmological Unit II, University of Naples	Ophthalmological Unit II, University of Naples; Rome Foundation
6	Retinal dystrophy due to rare *RPE65* gene mutation [22]	Group (9 locations)	Collection of data on disease	16 Months (2018-2020)	Ophthalmological Unit II, University of Naples	Retina Italia nonprofit association, Milan
7	Second opinions among resident physicians	Second opinions	Feasibility of second opinions in didactics	4 Months (during 2019)	Comed Research nonprofit association, Milan	Bayer Italy SpA, Milan
8	Instrumental data in multiple sclerosis	Group (2 locations)	Collection of multidisciplinary data on disease	2019-present	Neurological Unit, Insubria, University Varese-Como	Insubria University, Varese-Como
9	Epidemiological data on COVID-19 in workers	Group (2 locations)	Search for COVID-19 in throat, saliva, and tears	3 Months (during 2020)	SEA Company, Milan Linate-Malpensa Airports	SEA Company, Milan Linate-Milan Malpensa Airports
10	SARS-CoV-2 on throat and ocular surfaces [23]	Group (2 locations)	Search for SARS-CoV-2 in throat and tears in COVID-19 patients	1 Month (during 2020)	T&C Srl, Milan. Italy	Insubria University, Varese-Como
11	Potential malignant oral lesions	Group (2 locations)	Collection of data on disease	2021-present	Orthodontics Unit, Insubria University, Varese-Como	Insubria University, Varese-Como
12	Maculopathies and anti-aging medicine	Data	Collection of data on diseases and follow-up	2022-present	Claude Boscher, MD	Claude Boscher, MD

[a]Second opinions: second opinions from health care professionals at the same or a different institution.

[b]Group: shared database used by health care professionals at more than one institution.

[c]Data: shared database used by health care professionals at a single institution.

Textbox 2. Result options for the questionnaire for each health care professional, with relative scores. The final score is given by the sum of the partial scores (maximum 9, minimum 3). Scores equal to or higher than 6 are considered to indicate approval.

Access to the network by computer or mobile devices

- Poor: score of 1
- Good: score of 2
- Very good: score of 3

Acceptability of the procedures

- Poor: score of 1
- Good: score of 2
- Very good: score of 3

Medical efficacy

- Poor: score of 1
- Good: score of 2
- Very good: score of 3

Blueprint Summary

Design of Key Features and Roadmap

The design of the implementation program was oriented to develop three types of operational modules, integrated with one another where necessary: (1) a databank of diseases for clinical or scientific studies, (2) a database for groups of HCPs in different locations, giving them access to shared data from trial studies, and (3) a functionality enabling physicians to seek second opinions. The key points of the implemented modules were easy accessibility, complete acceptability for HCPs, data reliability, and overall medical efficacy considering all health specialties. The roadmap followed these principles and several new projects involving HCPs produced specific modules, which were created for the platform and take advantage of its benefits as a whole.

Technological Design and Infrastructure

Since 2010, the Eumeda platform has used continuously updated versions of PHP, an HTML-embedded web scripting language built to a high standard using advanced technology [24], which has the advantage of speed, flexibility, low use of resources, and compatibility with all web servers. PHP does not require a high level of machine resources to run and is therefore very fast and lends itself to applications with external integration.

Main Features of the Platform

Information technology services can be accessed via monitors or mobile devices and include current advanced technologies, such as the following: 24-7, 365-day-a-year access, easy data image insertion in electronic medical records, image comparison and overlapping, and SMS and email notification, when necessary, for fast interactivity.

Customizable Modules

The platform includes customizable files called “modules” that are designed and built for each project according to its needs in collaboration with professionals from different knowledge areas (Textbox 1). Each module functions to support the features and advantages of the entire platform. No data are sent directly to or from HCPs’ hardware. HCPs are able to see data in the central database, accessing this information remotely. All HCPs have a personal access code depending on their authorization level, enabling them to view, insert, or modify data in specific fields, close the electronic medical record (EMR) data temporarily or permanently, and export data for statistical surveys. The platform allows for individual and group interaction among HCPs at different sites. A remote “prompt assistance” service is provided for each module when necessary.

Module Functionality

Several functions can be activated, with open pop-ups showing the rationale of each study, its population, the provenance of resources, and the operating HCPs with various authorization profiles. The data entry procedure is quick, intuitive (Figure 1), and guided by many system alerts in the case of errors. When necessary, warning notifications are sent to users via SMS or email. Special software can be created, if requested, to support HCPs’ data evaluation and clinical decisions [25-27] (Figure 2). Data extraction for statistical surveys is immediate (Figure 1). At the end of each study period, the HCPs evaluated the project using a 3-point scoring system (Textbox 2).

Figure 1. Example of the main tasks and procedures for a module (module 6 in Table 1) on the medical platform: (1) entry to the system by the health care professional with their personal access key, after which they select the modules that they are qualified and authorized to use; (2) access to a list of operative centers with their own lists of patients and respective electronic medical records relating to the first and follow-up visits; (3) individual patient electronic medical record folder, which allows for the easy and quick insertion of data and multiple images at any time, as well as access to successive masks (a repository of images is available that allows image overlapping and comparison; Figure 3); (4) quick data extraction for statistical purposes.

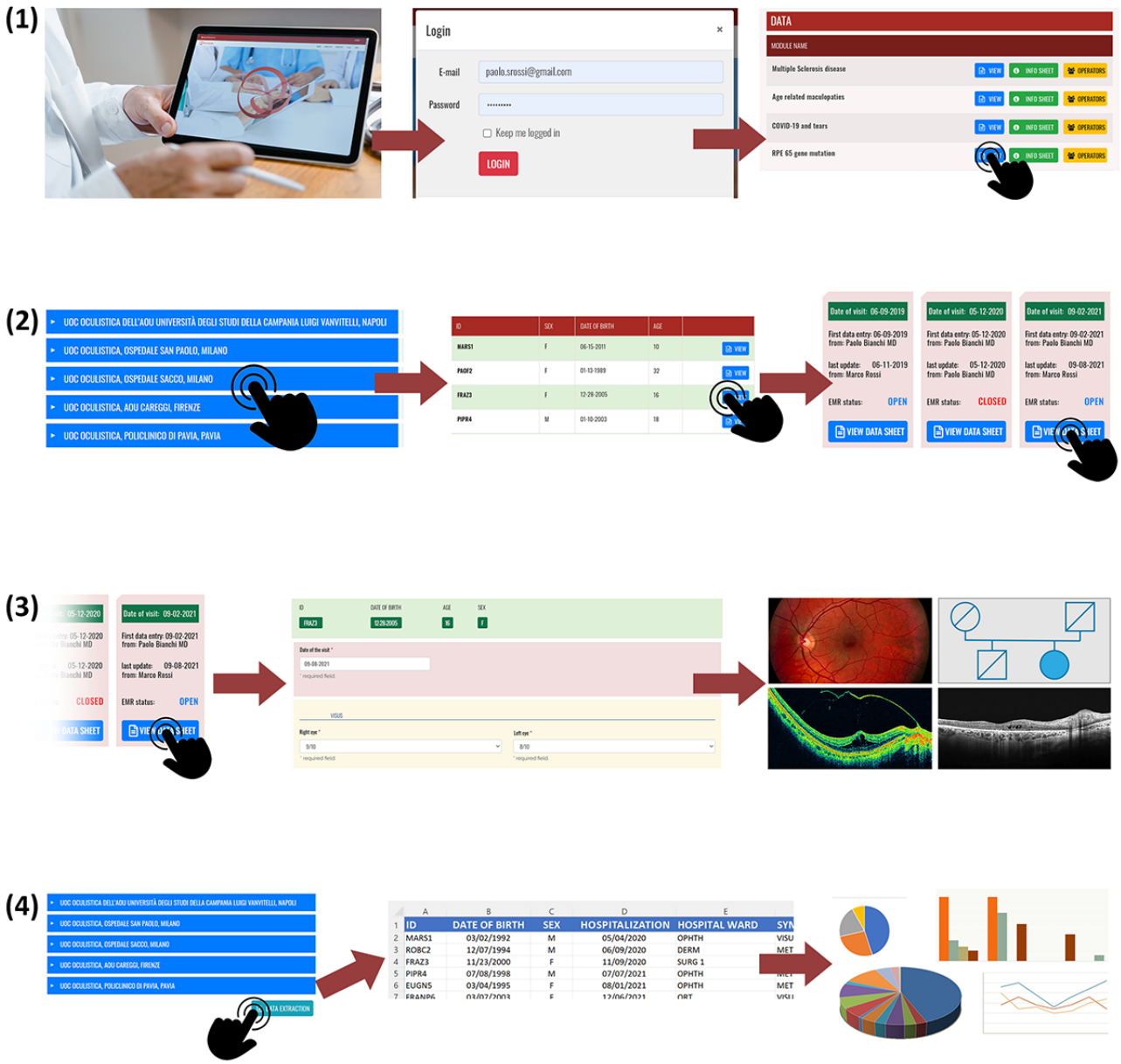
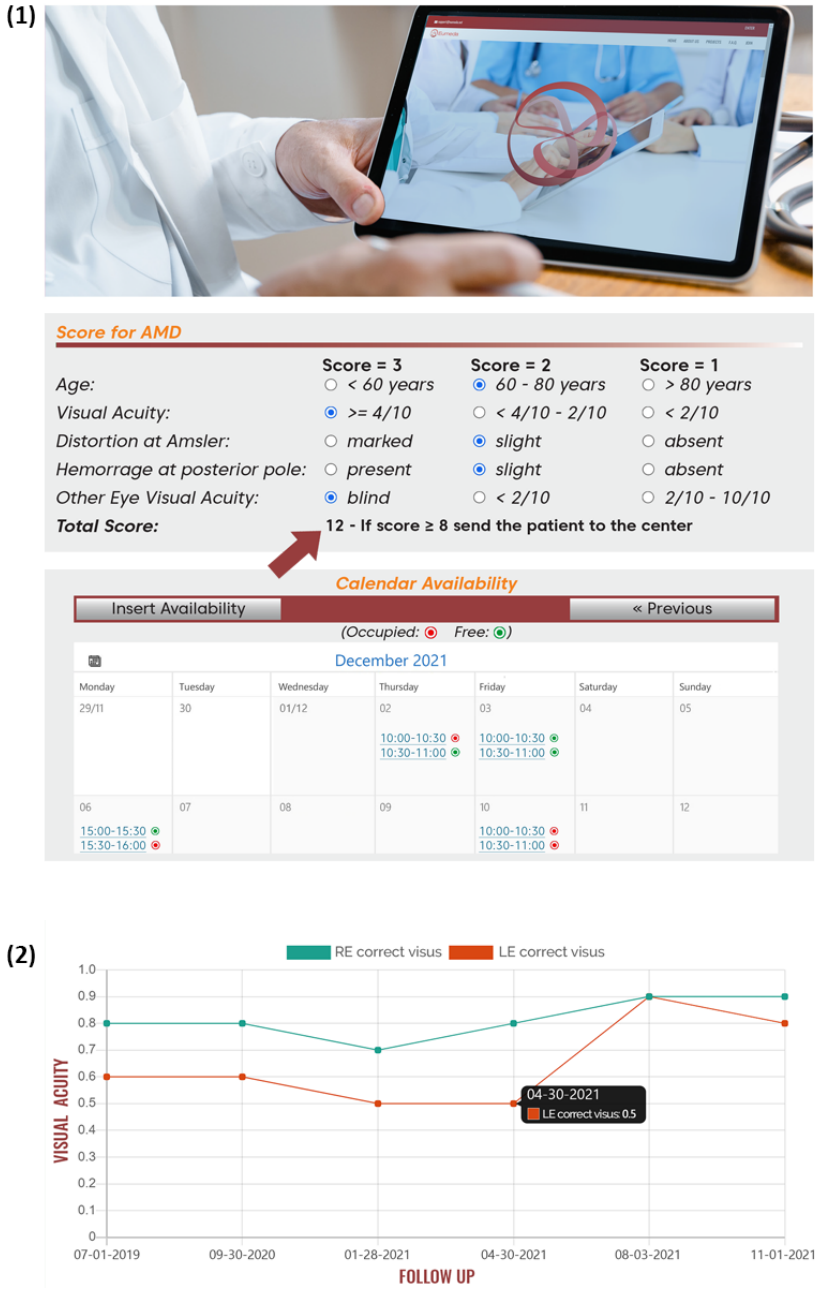


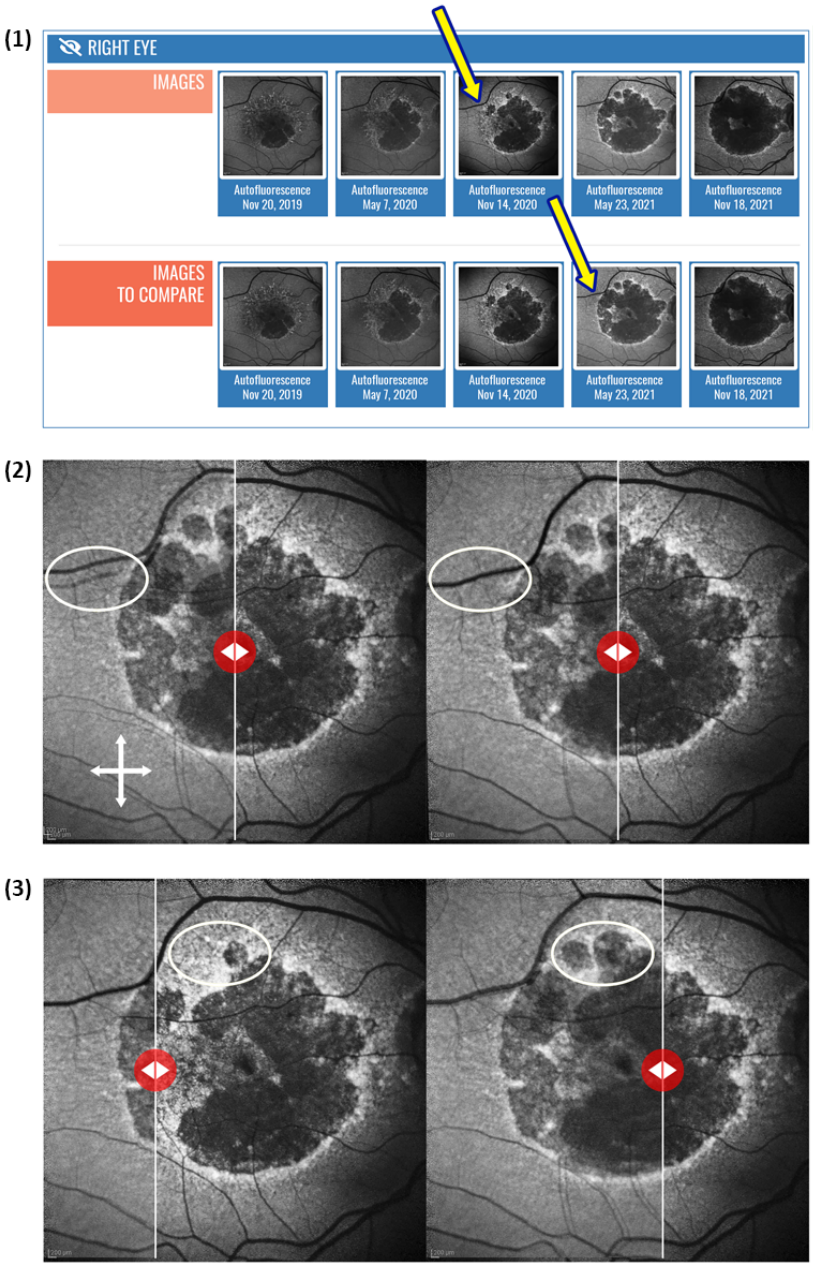
Figure 2. Examples of 2 special software programs designed to support health care professional activity. (1) For medical care decisions, each diagnostic variable of a disease is given a numeric value, and the software automatically provides a total score (shown by the arrow). If the value exceeds a defined score, the software advises general physicians to send the patient to an appropriate center at the next available appointment (module 2 in Table 1). (2) For tracking patients' clinical course, visual acuity data (or any other numerical data) are inserted into a patient's electronic medical record and the graph is updated in real time. The health care professional can see at a glance the functional course of the disease. RE: right eye; LE: left eye.



Images

Dedicated software allows the uploading of even high-resolution images and videos in a few seconds. A shared whiteboard for all images is available for each module. Image magnification and comparison software enables morphological changes to be observed over time in detail (Figure 3).

Figure 3. Example of image comparison: (1) selection of 2 images from a patient's electronic medical record (in this case, the patient had degenerative retinal maculopathy) uploaded at the 6-month follow-up (shown by the arrows); (2) creation of an overlap image that can be adjusted by clicking and moving the white cross; a special transparency application allows the images to be accurately superimposed on each other (shown by the white rings); (3) evaluation of morphological changes in the disease over time (within the white rings) by moving the overlap line back and forth (shown by the white line) using the red button.



Type of Technology

The Eumeda software platform is closed-source and owned by a private company that grants access through contracts. The platform was developed with the Wappler (Wappler.io) integrated development environment.

Targets

The target user base includes physicians of different specialties, nurses, technicians, orthoptists, geneticists, residents, and tutors, all of whom rely on access to a common database holding key patient information.

The target sites are hospitals, private offices, and medical hub centers working with spoke-peripheral centers. We involved medical structures equipped with technology and staff prepared to use hardware and software (Figure 4).

Figure 4. Locations of health care professionals (round dots) in Italy who have used the platform over the years (the map shows the Italian names of the cities). The Supervisory Center (SC) in Milan has responsibility for the data warehouse and help desk as well as a general coordination role.

Data

Data Location

All data were uploaded and stored in a data warehouse in Milan, Italy (Datasys Srl from 2001 to 2012; then Aruba Business Srl, provided by IRQ10 Srl, from 2013 until the present), to ensure data security and uninterrupted availability. Automated daily backups are a security measure guarding against the loss of data.

Data Entry Policy

HCPs must agree to the liability agreement, ownership agreement, and a code of conduct before using the platform. In all modules, data entry is performed in accordance with the guidelines of the Declaration of Helsinki and its subsequent revisions [28]. Informed consent forms are collected by the health facilities. In cases where data analysis included a therapeutic choice, approval from the relevant ethics committee or a qualified local committee was obtained, as in modules 6 and 11 in Table 1.

Data Security and Privacy

Data security and privacy are guaranteed by the latest generation of servers with secure backups. Data are protected on several levels: (1) individual HCPs receive access keys generated by the system; (2) subjects' personal data are encrypted and stored in a separate table in the cloud; (3) the system binds clinical data to personal data only when accessing hardware with a special algorithm; and (4) if necessary, a ready-to-use informed consent form can be downloaded for signature.

Responsibility for and Ownership of Data

According to European Union (EU) and Italian rules, liability for entered medical data, including cloud storage, lies with the HCP entering such data (acting as the "controller," as clearly defined in EU Regulation 679/16) and the manager of the medical platform and server farm (acting as "processors," as clearly defined in EU Regulation 679/16). Ownership of the data, including the purpose and methods for processing the data, belongs to the entity applying the module, who acts as the "controller."

Interoperability

The Eumeda software platform is not accessible from specific application programming interface clients by users, so it does not use data standards such as Health Level Seven. However, it implements the International Classification of Diseases, 10th

revision, system internally to classify pathologies. Clinical imaging is managed with current standard protocols.

Participating Entities

A nonprofit organization (Comed Research) initially implemented (from 2001) the projects. Subsequently, a joint venture between 2 for-profit companies (T&C Srl and TM95 Srl) managed the platform. These partners supply hardware, create software, or participate as web designers, hosting companies, or law firms. The funders of the implementation projects are public universities, hospitals and foundations, nonprofit medical associations, private companies, and physicians working in private offices (Table 1).

The society that created the main platform is the owner of all the implementations. The entities that applied the modules hold the ownership and the intellectual property.

Budget Planning

The total budget covered different phases according to implementation progression and project type for a period of 1 to 3 years. The costs included preliminary planning and the final draft (up to 30%), programming for final front view on the computer screen (30%), and user training (10%), as well as the pilot phase (5%), operation (10%), initial service (10%), and ongoing reports (5%). Selected projects could be conducted for free based on their importance or visibility for the platform.

Sustainability

The projects were initially funded (from 2001) by a nonprofit organization (Comed Research), which relies on donations from companies or nonprofit medical associations. Since January 1, 2017, the platform has been managed by a joint venture between 2 for-profit companies. The business model is based on the type and duration of the projects developed during the implementation phase, financed by different entities. The end of the project foresees the dissemination of the results with potential permanent effects.

Implementation (Results)

Coverage

The projects developed during the implementation phase have national coverage, encompassing a large number of Italian regions and their referent hospitals. The developed modules evolved over time, with varying degrees of interconnectivity, in different centers in 19 cities across Italy (Figure 4). In 2 modules (modules 1 and 2 in Table 1), HCPs from the referring regional areas were closely involved. The number of HCPs (at different levels) using individual modules ranged from 6 to 114 (average 21.8, SD 28.5).

Outcomes

Implementing the telemedicine platform allowed us to build several modules that could be used by HCPs at different times. The characteristics of 12 representative modules used over the last 10 years for different clinical projects are shown in Table 1.

Over time, our experience has led us to concentrate on three types of operational modules, integrated with one another if necessary: (1) a databank of diseases for clinical or scientific studies (eg, module 4; Table 1), (2) a database for groups of HCPs in different locations, giving them access to shared data (eg, module 6; Table 1), and (3) a functionality enabling physicians to seek second opinions (eg, module 7; Table 1).

The overall outcomes are reported in Table 2. Up to now, more than 250 HCPs have used the platform for several effective and operational projects. The total number of participants inserted in the modules is 2574. The percentage of data entered in the text or image fields for each module ranged from 20% to 95% (with an average of 65.7%). The evaluation score for each module was calculated as the sum of 3 partial scores (Textbox 2): out of all the modules, the first (module 1, the first to be created) was the one with the lowest evaluation score (Table 2). The average number of requests for technological support varied from 5 per month (in the case of simpler modules) to 9 (for more complex ones).

Table 2. Results pertaining to the designed modules shown in Table 1.

Module	Description	Centers involved, n	HCPs[a] involved, n	Participants whose data were inserted, n	Text/image fields for each EMR[b], n (fields that were filled in, %)	Beneficial effects	Evaluation score[c] (minimum positive score)
1	Teleconsultation in retinal diseases [18]	1 Retina center, 17 territorial offices	18	52	30 (60)	Useful teleconsultation among doctors	109 (108)
2	Age-related maculopathy [19]	11 Retina centers	114	678	65 (85)	Improvements in patients' functional final outcomes	803 (684)
3	Retinal pathology samples and correlated genes [20]	1 Ophthalmological center, 1 genetic center	11	12	65 (80)	Better understanding of molecular mechanisms	Not acquired
4	Epiretinal macular membrane [21]	2 Ophthalmologic centers, 1 human anatomy center	11	28	25 (65)	Identification of ultramicroscopic features of membranes	80 (66)
5	Inherited eye diseases[d]	1 Ophthalmological center, 1 genetic center	14	1145[e]	480[e] (20)	Increased knowledge of genetic eye diseases	In progress
6	Retinal dystrophy due to rare *RPE65* gene mutation [22]	9 Retinal-genetic centers	28	60	260 (65)	Identification of suitable patients for therapy	200 (168)
7	Second opinions among resident physicians[d]	4 University ophthalmological departments	19	110	12 (85)	Resident physicians' learning accelerated	140 (114)
8	Instrumental data in multiple sclerosis[d]	2 Neurological centers, 2 ophthalmological centers	6	58[e]	450[e] (18)	Recognition of the disease in the subclinical stage	In progress
9	Epidemiological data on COVID-19 in workers[d]	2 Care offices at 2 airports	9	298	30 (90)	Collection of useful diagnostic data on COVID-19 and how the disease is transmitted	75 (54)
10	SARS-CoV-2 on throat and ocular surfaces [23]	14 Medical units	20	108	34 (95)	Increased knowledge of COVID-19	165 (120)
11	Potential malignant oral lesions	4 Medical units	6	15[e]	50[e] (68)	Better prevention and therapy	In progress
12	Maculopathies and anti-aging medicine	1 Retina center	6	10[e]	110[e] (58)	Significantly better care	In progress

[a]HCP: health care professional (physicians from different specialties, nurses, technicians, orthoptists, geneticists, residents, tutors [employees were excluded]).

[b]EMR: electronic medical record (for each patient, considering first visit and all follow-ups).

[c]Sum of 3 partial scores for access, acceptability, and medical efficacy at end of the active working period (described in Textbox 2).

[d]Unpublished data.

[e]At the time of writing this paper.

Clinical fallout can be identified more easily with the use of this telemedicine platform because of the visibility of a database shared by HCPs (modules 3, 4, 5, 8, and 9; Table 2). Furthermore, data on rare diseases (collected from a large number of centers) can be used to identify patients who would benefit from expensive new therapies (module 6; Table 2). By sharing medical data, physicians and residents can learn better and faster (modules 1 and 7; Table 2), and the possibility of having a databank helps them to discover potential, as yet unknown disease complications (module 10; Table 2). Patient follow-up with dedicated software helps HCPs to locate better treatment options, identify preventive interventions (modules 2 , 11 and 12; Table 2) and track patients and their outcomes in real time.

Lessons Learned

Our program has multiple success factors that may be considered in future implementations or in the creation of similar telemedicine platforms and modules. First, the technological infrastructure of the platform is modern, highly versatile, and continuously updated by technical staff. The use of the latest generation of servers with secure, daily backups guarantees that no data loss occurs, while data security and privacy are protected on several levels, as specified in the Methods section. Second, data entry and retrieval in each module are immediate. Each module has different blocks of information that are well separated, including an explanation of the rationale of the study and practical guidance on how to insert data, as well as different HCP access profiles, patient IDs, EMRs, images, and statistical surveys. Third, no images are transmitted among HCPs. All images are stored on the main server and are viewed remotely without any deterioration. A dedicated procedure even allows the insertion of high-resolution images (through common connection links) immediately or very quickly. Rapid viewing is greatly appreciated by users, in addition to the possibility of enlarging, comparing, and superimposing, as well as being able to see in detail the morphological changes, even minimal ones, of a pathology over time (Figure 3). Lastly, different authorization profiles are given to HCPs, which enables them to access modules on the central server once they have agreed to abide by the terms of the liability and ownership agreements and the code of conduct, using personal passwords to view, change, or modify data and images. Module coordinators usually have total control of their respective modules and can compile statistical surveys using all the data, while other HCPs may only be able to enter data and images in accordance with their remit and authorized access level. A great amount of work has been undertaken to ensure that the user-friendly platform is up and running. A remote service is available by mail or telephone. All modules are visible both from monitors and mobile devices. In particular, the second opinion module may be suitable for use with mobile health (mHealth).

We consider the following points more as challenges than limits to implementation. The construction phase of each module is of critical importance, and a single medical interlocutor must be the voice of all HCPs (Textbox 1). The main mandatory factors involved in building a module include a scientific coordinator as the central figure and the participation of someone with both medical and IT skills. Finally, older HCPs tended to struggle with working on the platform, while the younger operators adapted quickly, were not disconcerted by the technology, and showed interest and satisfaction with the projects they carried out [29-34].

The presumed budgets of each project, divided into direct (eg, coordinators, IT programmers, law firms) and indirect (eg, travel, equipment, insurance) costs have been considered in the final balance.

The following recommendations may assist in overcoming many barriers to telemedicine practice among HCPs. First, the amount of preparatory work needed (Textbox 1) tends to be underestimated. Second, it is difficult to create systems for sharing text and images with appropriate levels of usability. Third, bureaucracy is often an obstacle, and self-regulation codes in telemedicine need official authorization. Fourth, a suitable “network culture” is still lacking in medicine, due to multiple technical and human factors. The success of telemedicine among HCPs requires participation, responsibility, and a desire for effective collaboration to develop knowledge for the benefit of professionals and patients.

Discussion

Principal Findings

The technological infrastructure of telemedicine intended strictly for HCPs is specific to this field, is not easy to implement, and must be customized for each individual project. Key persons such as scientific coordinators (with specific knowledge of medicine and IT) and program managers must be well chosen for projects to succeed. The results of our projects have shown a range of benefits, including increased medical efficacy and clinical knowledge, improved patient care, enhanced teaching and integration among hospitals, and a more effective choice of therapies. It is necessary for work to be carried out on organizational, bureaucratic, and network culture issues where these are not yet fully accepted and on sustainable business plans.

We identified some difficulties and limitations in our implementation project that may also be considered useful for future or similar telemedicine projects. Building a module in the absence of straightforward ideas forced us to make major changes during construction, meaning that the preliminary work completed had to be discarded and redone. All the software involved in the platform modules must be customized according to the needs of the HCPs, which requires time and hard work. Too many text or image fields to fill out and include in EMRs make the system difficult to use and produce a very large final database that is not fully used (as happened with module 5).

Conclusion

In conclusion, our experience was that both physicians and patients were always satisfied to be part of this “community of health” supported by groups of HCPs working for their benefit and making them feel cared for. The detailed description of our implementation program may be useful to shorten the learning curve for others seeking to implement similar projects in many fields of medicine, which must be able to adapt to the continuously changing nature of medicine now and in the future.

Acknowledgments

We would like to thank the hundreds of health care professionals who have used and contributed over time to the development of the platform. Special thanks go to all engineers, programmers, and other personnel who in their various capacities have worked to improve the platform over the last 10 years, especially Francesco Oggionni (IT professional) and Valerio Tartaglia (IT professional). We are grateful to Ferruccio Fazio, MD, former Italian minister of health, and Gianfranco Ferla, MD, for their support and advice. We would also like to thank Roberta Romagnolo (Lexikon) for editing the manuscript.

Data Availability

The data on which this manuscript is based are available upon request to the corresponding author.

Conflicts of Interest

None declared.

References

1. Ebbert JO, Ramar P, Tulledge-Scheitel SM, Njeru JW, Rosedahl JK, Roellinger D, et al. Patient preferences for telehealth services in a large multispecialty practice. J Telemed Telecare 2023 May;29(4):298-303. [doi: 10.1177/1357633X20980302] [Medline: 33461397]
2. Schulz T, Long K, Kanhutu K, Bayrak I, Johnson D, Fazio T. Telehealth during the coronavirus disease 2019 pandemic: Rapid expansion of telehealth outpatient use during a pandemic is possible if the programme is previously established. J Telemed Telecare 2022 Jul;28(6):445-451 [FREE Full text] [doi: 10.1177/1357633X20942045] [Medline: 32686556]
3. Reitzle L, Schmidt C, Färber F, Huebl L, Wieler LH, Ziese T, et al. Perceived access to health care services and relevance of telemedicine during the COVID-19 pandemic in Germany. Int J Environ Res Public Health 2021 Jul 19;18(14):7661 [FREE Full text] [doi: 10.3390/ijerph18147661] [Medline: 34300110]
4. Capusan KY, Fenster T. Patient satisfaction with telehealth during the COVID-19 pandemic in a pediatric pulmonary clinic. J Pediatr Health Care 2021;35(6):587-591 [FREE Full text] [doi: 10.1016/j.pedhc.2021.07.014] [Medline: 34417077]
5. Azzolini C, Fontanella G, Mason A. A pilot study to train vitreoretinal surgeons by telemedicine. 1997 Presented at: Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting; May 11-16; Fort Lauderdale, FL p. 397.
6. Karcić S, Azzolini C, Alikadić-Husović A. [Telemedicine in vitreoretinal surgery]. Med Arh 1999;53(3 Suppl 3):73-75. [Medline: 10870633]
7. Contini F, Prati M, Donati S, Azzolini C. Idiopathic macular hole: Multicentric clinical trial. 2006 Presented at: Association for Research in Vision and Ophthalmology Meeting; May; Fort Lauderdale, FL URL: https://iovs.arvojournals.org/article.aspx?articleid=2391305&resultClick=1
8. Mason A, Feliciani F, Morelli P. The Italian telemedicine SHARED project. 1998 Presented at: American Telemedicine Association Third Annual Conference; Orlando, FL p. 5.
9. Li S, Wang C, Lu W, Lin Y, Yen DC. Design and implementation of a telecare information platform. J Med Syst 2012 Jun;36(3):1629-1650. [doi: 10.1007/s10916-010-9625-6] [Medline: 21120592]
10. Clin L, Leitritz MA, Dietter J, Dynowski M, Burgert O, Ueffing M, et al. Design, implementation and operation of a reading center platform for clinical studies. Stud Health Technol Inform 2017;235:33-37. [doi: 10.3233/978-1-61499-753-5-33] [Medline: 28423750]
11. Lopez E, Berlin M, Stein R, Cozzi E, Bermudez A, Mandirola Brieux H, et al. Results of the use of the teleconsultation platform after 2 months of implementation. Stud Health Technol Inform 2020 Jun 16;270:1377-1378. [doi: 10.3233/SHTI200450] [Medline: 32570667]
12. Hasson SP, Waissengrin B, Shachar E, Hodruj M, Fayngor R, Brezis M, et al. Rapid implementation of telemedicine during the COVID-19 pandemic: Perspectives and preferences of patients with cancer. Oncologist 2021 Apr;26(4):e679-e685 [FREE Full text] [doi: 10.1002/onco.13676] [Medline: 33453121]
13. Beltrán V, von Marttens A, Acuña-Mardones P, Sanzana-Luengo C, Rueda-Velásquez SJ, Alvarado E, et al. Implementation of a teledentistry platform for dental emergencies for the elderly in the context of the COVID-19 pandemic in Chile. Biomed Res Int ;2022:6889285 [FREE Full text] [doi: 10.1155/2022/6889285] [Medline: 35330690]
14. Espay AJ, Hausdorff JM, Sánchez-Ferro Á, Klucken J, Merola A, Bonato P, et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019 May;34(5):657-663 [FREE Full text] [doi: 10.1002/mds.27671] [Medline: 30901495]
15. Nadar M, Jouvet P, Tucci M, Toledano B, Cyr M, Sicotte C. The implementation of a synchronous telemedicine platform linking off-site pediatric intensivists and on-site fellows in a pediatric intensive care unit: A feasibility study. Int J Med Inform 2019 Sep;129:219-225. [doi: 10.1016/j.ijmedinf.2019.06.009] [Medline: 31445259]

16. Kern C, Fu DJ, Kortuem K, Huemer J, Barker D, Davis A, et al. Implementation of a cloud-based referral platform in ophthalmology: making telemedicine services a reality in eye care. Br J Ophthalmol 2020 Mar;104(3):312-317 [FREE Full text] [doi: 10.1136/bjophthalmol-2019-314161] [Medline: 31320383]
17. Brenner B, Brancolini S, Eshraghi Y, Guirguis M, Durbhakula S, Provenzano D, et al. Telemedicine implementation in pain medicine: A survey evaluation of pain medicine practices in spring 2020. Pain Physician 2022 Aug;25(5):387-390 [FREE Full text] [Medline: 35901479]
18. Azzolini C. A pilot teleconsultation network for retinal diseases in ophthalmology. J Telemed Telecare 2011;17(1):20-24. [doi: 10.1258/jtt.2010.100305] [Medline: 21097561]
19. Azzolini C, Torreggiani A, Eandi C, Donati S, Oum MA, Vinciguerra R, et al. A teleconsultation network improves the efficacy of anti-VEGF therapy in retinal diseases. J Telemed Telecare 2013 Dec;19(8):437-442. [doi: 10.1177/1357633X13501760] [Medline: 24162839]
20. Azzolini C, Pagani IS, Pirrone C, Borroni D, Donati S, Al Oum M, et al. Expression of VEGF-A, Otx homeobox and p53 family genes in proliferative vitreoretinopathy. Mediators Inflamm 2013:857380 [FREE Full text] [doi: 10.1155/2013/857380] [Medline: 24227910]
21. Azzolini C, Congiu T, Donati S, Passi A, Basso P, Piantanida E, et al. Multilayer microstructure of idiopathic epiretinal macular membranes. Eur J Ophthalmol 2017 Nov 08;27(6):762-768. [doi: 10.5301/ejo.5000982] [Medline: 28525683]
22. Testa F, Murro V, Signorini S, Colombo L, Iarossi G, Parmeggiani F, et al. RPE65-associated retinopathies in the Italian population: a longitudinal natural history study. Invest Ophthalmol Vis Sci 2022 Feb 01;63(2):13 [FREE Full text] [doi: 10.1167/iovs.63.2.13] [Medline: 35129589]
23. Azzolini C, Donati S, Premi E, Baj A, Siracusa C, Genoni A, et al. SARS-CoV-2 on ocular surfaces in a cohort of patients with COVID-19 from the Lombardy Region, Italy. JAMA Ophthalmol 2021 Sep 01;139(9):956-963 [FREE Full text] [doi: 10.1001/jamaophthalmol.2020.5464] [Medline: 33662099]
24. PHP Manual. The PHP Group. 1997. URL: https://www.php.net/download-docs.php [accessed 2023-12-20]
25. Campbell JP, Lee AY, Abràmoff M, Keane PA, Ting DS, Lum F, et al. Reporting guidelines for artificial intelligence in medical research. Ophthalmology 2020 Dec;127(12):1596-1599 [FREE Full text] [doi: 10.1016/j.ophtha.2020.09.009] [Medline: 32920029]
26. Azzolini C, Donati S, Falco A. The digital citizen: duties and rights to build a fairer future society. 2022 Presented at: XXII Infopoverty World Conference; December 1; New York, NY.
27. Azzolini C, Donati S. The digital era: new horizons in medicine and rehabilitation. 2023 Presented at: XXIII National Congress of the Italian Association of Telemedicine and Medical Informatics; November 24-25; Rome, Italy.
28. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
29. Pit SW, Velovski S, Cockrell K, Bailey J. A qualitative exploration of medical students' placement experiences with telehealth during COVID-19 and recommendations to prepare our future medical workforce. BMC Med Educ 2021 Aug 16;21(1):431 [FREE Full text] [doi: 10.1186/s12909-021-02719-3] [Medline: 34399758]
30. Mahabamunuge J, Farmer L, Pessolano J, Lakhi N. Implementation and assessment of a novel telehealth education curriculum for undergraduate medical students. J Adv Med Educ Prof 2021 Jul;9(3):127-135 [FREE Full text] [doi: 10.30476/jamp.2021.89447.1375] [Medline: 34277843]
31. Curioso WH, Peña-Ayudante WR, Oscuvilca-Tapia E. COVID-19 reveals the urgent need to strengthen nursing informatics competencies: a view from Peru. Inform Health Soc Care 2021 Sep 02;46(3):229-233. [doi: 10.1080/17538157.2021.1941974] [Medline: 34292802]
32. Bell KA, Porter C, Woods AD, Akkurt ZM, Feldman SR. Impact of the COVID-19 pandemic on dermatology departments' support of medical students: A survey study. Dermatol Online J 2021 Jul 15;27(7):14 [FREE Full text] [doi: 10.5070/D327754376] [Medline: 34391338]
33. Coe TM, McBroom TJ, Brownlee SA, Regan K, Bartels S, Saillant N, et al. Medical students and patients benefit from virtual non-medical interactions due to COVID-19. J Med Educ Curric Dev 2021;8:23821205211028343 [FREE Full text] [doi: 10.1177/23821205211028343] [Medline: 34368454]
34. Frankl S, Joshi A, Onorato S, Jawahir GL, Pelletier SR, Dalrymple JL, et al. Preparing future doctors for telemedicine: an asynchronous curriculum for medical students implemented during the COVID-19 pandemic. Acad Med 2021 Dec 01;96(12):1696-1701 [FREE Full text] [doi: 10.1097/ACM.0000000000004260] [Medline: 34323861]

Abbreviations

EU: European Union
EMR: electronic medical record
HCP: health care professional
mHealth: mobile health

Edited by C Perrin; submitted 13.10.22; peer-reviewed by J Gurp, van, M Venturini, S Pesälä, V Ramos; comments to author 27.01.23; revised version received 13.07.23; accepted 29.11.23; published 26.01.24.

Please cite as:
Azzolini C, Premi E, Donati S, Falco A, Torreggiani A, Sicurello F, Baj A, Azzi L, Orro A, Porta G, Azzolini G, Sorrentino M, Melillo P, Testa F, Simonelli F, Giardina G, Paolucci U
Ten Years of Experience With a Telemedicine Platform Dedicated to Health Care Personnel: Implementation Report
JMIR Med Inform 2024;12:e42847
URL: https://medinform.jmir.org/2024/1/e42847
doi:10.2196/42847
PMID:38277199

Implementation Report

Learnings From Implementation of Technology-Enabled Mental Health Interventions in India: Implementation Report

Sudha Kallakuri[1], MSc; Sridevi Gara[1], BE; Mahesh Godi[1], BE; Sandhya Kanaka Yatirajula[1], PhD; Srilatha Paslawar[1], MPH; Mercian Daniel[1], PhD; David Peiris[2,3], MBBS, MIPH, PhD; Pallab Kumar Maulik[1,3,4,5,6], MSc, MD, PhD

[1]George Institute for Global Health, New Delhi, India

[2]George Institute for Global Health, Sydney, Australia

[3]Faculty of Medicine, University of New South Wales, Sydney, Australia

[4]Department of Brain Sciences, Imperial College London, London, United Kingdom

[5]Prasanna School of Public Health, Manipal Academy of Higher Education, Manipal, India

[6]George Institute for Global Health, London, United Kingdom

Corresponding Author:
Sudha Kallakuri, MSc
George Institute for Global Health
308 Elegance Tower, Third Floor
Plot No 8, Jasola District Centre
New Delhi, 110025
India
Phone: 91 11 4158 8091
Email: skallakuri1@georgeinstitute.org.in

Abstract

Background: Recent years have witnessed an increase in the use of technology-enabled interventions for delivering mental health care in different settings. Technological solutions have been advocated to increase access to care, especially in primary health care settings in low- and middle-income countries, to facilitate task-sharing given the lack of trained mental health professionals.

Objective: This report describes the experiences and challenges faced during the development and implementation of technology-enabled interventions for mental health among adults and adolescents in rural and urban settings of India.

Methods: A detailed overview of the technological frameworks used in various studies, including the Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health pilot study, SMART Mental Health cluster randomized controlled trial, and Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums (ARTEMIS) study, is provided. This includes the mobile apps that were used to collect data and the use of the database to store the data that were collected. Based on the experiences faced, the technological enhancements and adaptations made at the mobile app and database levels are described in detail.

Implementation (Results): Development of descriptive analytics at the database level; enabling offline and online data storage modalities; customizing the Open Medical Record System platform to suit the study requirements; modifying the encryption settings, thereby making the system more secure; and merging different apps for simultaneous data collection were some of the enhancements made across different projects.

Conclusions: Technology-enabled interventions prove to be a useful solution to cater to large populations in low-resource settings. The development of mobile apps is subject to the context and the area where they would be implemented. This paper outlines the need for careful testing using an iterative process that may support future research using similar technology.

Trial Registration: SMART Mental Health trial: Clinical Trial Registry India CTRI/2018/08/015355; https://ctri.nic.in/Clinicaltrials/pmaindet2.php?EncHid=MjMyNTQ=&Enc=&userName=CTRI/2018/08/015355. ARTEMIS trial: Clinical Trial Registry India CTRI/2022/02/040307; https://ctri.nic.in/Clinicaltrials/pmaindet2.php?EncHid=NDcxMTE=&Enc=&userName=CTRI/2022/02/040307

(JMIR Med Inform 2024;12:e47504) doi:10.2196/47504

KEYWORDS

mental health; technological interventions; digital health; community intervention; implementation; eHealth; India; Asia; development; health technology

Introduction

The burden of mental disorders [1] and the treatment gap due to untreated mental disorders in low- and middle-income countries (LMICs) such as India is estimated to range between 75% and 85% [2], with 1 in every 27 individuals being treated for depression [3]. Technological solutions have been advocated to increase access to care, especially in primary health care settings in LMICs, to facilitate task-sharing, given the lack of trained mental health professionals. Research has indicated the effectiveness of employing technologies for addressing complex health concerns among people with mental illnesses. However, the cost-effectiveness of technology-enabled interventions compared to in-person interventions has not yet been established [4].

Technology-enabled service delivery models have increased access to care and facilitated service monitoring, with mobile health (mHealth) being one such strategy. The World Health Organization (WHO) defines mHealth as "a medical and public health practice that is supported by mobile devices, such as mobile phones, patient monitoring devices, and other wireless devices" [5]. mHealth in the form of electronic decision support systems (EDSSs) has been widely adopted by service users and providers for monitoring health status and for diagnosing and managing a range of health conditions, including mental disorders and substance use [6]. mHealth use has increased with increasing penetration of mobile network connectivity [7].

This paper highlights the processes involved in the development and implementation of technology-enabled interventions employed in three projects across rural and urban settings in India among adults and adolescents: the Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health (SMH) Pilot project [8], and two cluster randomized controlled trials (cRCTs), SMH trial [9] and the Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums (ARTEMIS) trial [10].

All three projects used EDSSs to facilitate the identification, diagnosis, and management of common mental disorders (CMDs), including depression, anxiety, psychological distress, and increased suicide risk. A task-sharing approach was used where nonphysician health workers known as Accredited Social Health Activists (ASHAs) and primary care doctors worked together to support people at high risk of CMDs [8-10].

This implementation report describes the experiences of using technology in implementing these three mental health projects, following implementation reporting guidelines [11].

Methods

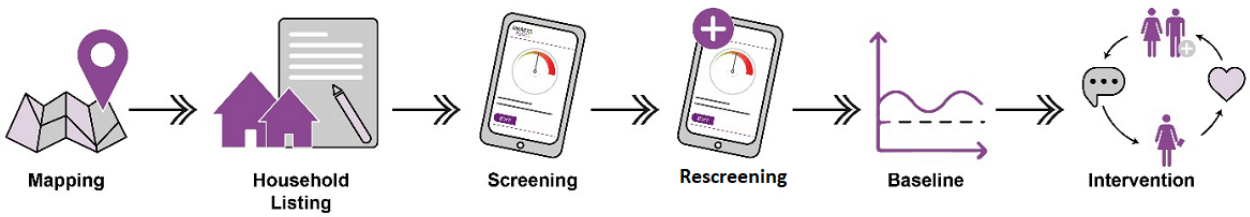
Aims and Objectives

This paper highlights the processes involved in the development and implementation of technology-enabled interventions employed in three projects across rural and urban settings among adults and adolescents in India.

Blueprint Summary

The overall technological framework of the SMH pilot study has two main components: a mobile app and a database. Different mobile apps were developed to collect data at divergent phases of the study (Figure 1). All apps were installed on 7-inch Android tablets for use by ASHAs/community women volunteers (CWVs), or primary health center (PHC) doctors. ASHAs are local women trained from the community with 8th-10th–grade education levels to support the implementation of health programs. While ASHAs work contractually, they are incentivized for their involvement in other projects. CWVs are women who reside in the same community where the study is being done. These CWVs were chosen from the slums and would have similar education level as ASHAs. They were trained on basic knowledge about mental health, along with the stigma and care of individuals with stress, depression, and increased suicide risk.

Figure 1. Different phases of the study.



The three studies underwent a formative phase, testing study tools and mobile apps while gauging user acceptance [12,13]. Iterations were made based on user feedback before the intervention phase. The technical team assessed the app and released a test version for research team testing. Once confirmed, a definitive version was used for data collection.

In the preintervention phase, geographical mapping and demarcation of the village boundaries were performed, followed by house listing to obtain accurate census data. Custom apps were developed for each step, including population screening for identifying individuals at risk of CMDs, which involved data collectors and ASHAs using specific screening tools. After screening, baseline data on various variables were collected before the intervention was implemented (Figure 1).

Technical Framework Design

The key components of the EDSS included the ASHA app, doctors app, and priority listing app (Table 1). Each ASHA had a finite set of individuals who lived in the geographical location covered by her. The tablets had encrypted, password-protected individual logins unique for every ASHA. Individuals screening positive were referred to primary care doctors for clinical diagnosis and treatment based on predetermined cut-off scores. The doctors used the WHO mhGAP-IG tool (version 1.0) [14] for diagnosing and treating people with CMDs, offering algorithm-based diagnoses and evidence-based treatment recommendations, including comorbidities. Doctors followed these recommendations, entering the type of care provided (pharmacological, psychological, referral, or combinations thereof) into their app. Doctors input the data to generate a traffic light–coded priorities list for ASHAs, indicating the status of screen-positive individuals in their area. Using color coding due to the low education levels of ASHAs, the list included pertinent questions on treatment adherence, social support, and stressors for each color category. The list was dynamic, changing based on doctors' updates during patient follow-up visits.

Table 1. Details of the apps used for the three studies and the target of the intervention.

App	Phase of the study	Users
SMART MH[a] (Pilot) and SMART MH trial focused on rural adults		
Listing app	Listing (household census data collection)	Data collectors
Screening app	Household screening for common mental disorders	ASHAs[b]
Baseline data collection app	Baseline: collected data on different variables and stressors triggering anxiety/depression	Data collectors
Intervention (ASHA app)	Intervention: for regular follow up of adults at high risk of CMDs[c] who sought care from the doctor or have yet to seek care	ASHAs
Intervention (doctor app)	Intervention: diagnosis and treatment for CMDs among adults	Primary care doctors
3M, 6M, and 12M app	Assessments at 3, 6, and 12 months of the intervention	Data collectors
ARTEMIS[d] trial focused on adolescents		
Listing app	Listing (household census data collection)	Data collectors
Screening app	Household screening for common mental disorders	Data collectors
Baseline data collection app	Baseline: collected data on different variables and stressors triggering anxiety/depression	Data collectors
Intervention (ASHA app)	Intervention: for regular follow up of adolescents who are at high risk of CMDs who sought care from the doctor or have yet to seek care	ASHAs
Intervention (doctor app)	Intervention: diagnosis and treatment for CMDs	Primary care doctors
3M, 6M, and 12 M app	Assessments at 3, 6, and 12 months of the intervention	Data collectors

[a]SMART MH: Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health.

[b]ASHA: Accredited Social Health Activist.

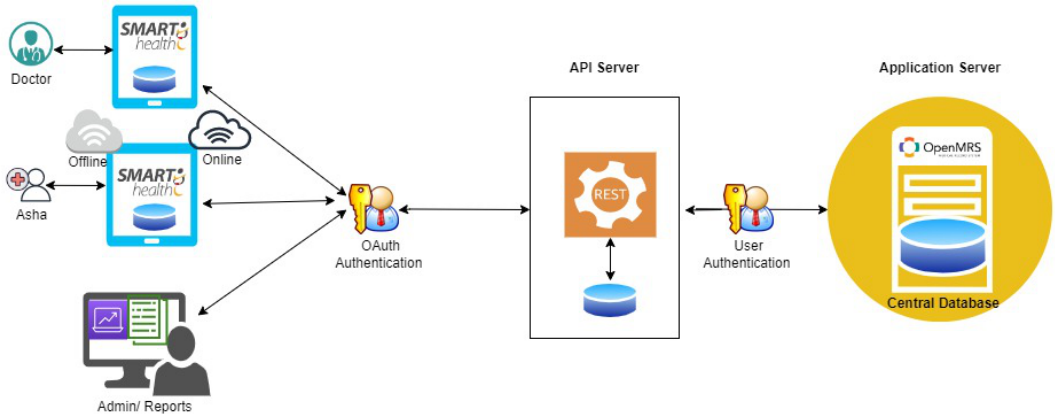
[c]CMD: common mental disorder.

[d]ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums.

Identifying an Electronic Medical Record System

All three projects utilized apps based on the Open Medical Record System (OpenMRS) [15], a community-driven open-source software for medical record storage and processing. OpenMRS is robust, scalable for large interventions, and customizable to study workflows and data collection needs. OpenMRS was chosen for these projects as it is freely available. Based on our earlier experience, the functionalities were suitable for our mental health projects [16]. Data collected on tablets underwent authentication and were transferred to the application programming interface (API) server, which were then sent to the application server housing the central OpenMRS database (Figure 2).

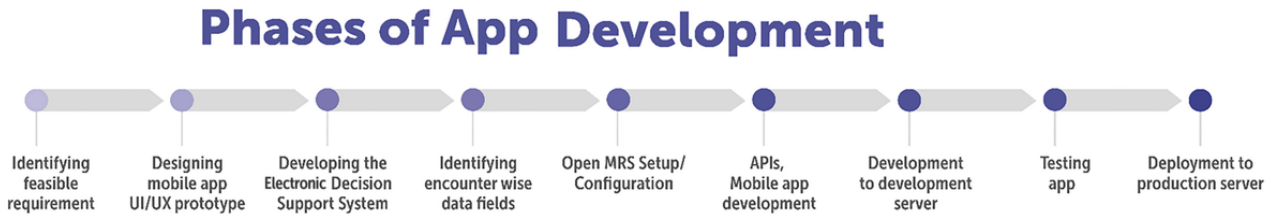
Figure 2. Workflow of data. API: application programming interface; Asha: Accredited Social Health Activist.



Phases of App Development

The apps went through different phases of development, enhancement, and adaptations across the three projects to suit the specific requirements of each project (Figure 3).

Figure 3. Phases of app development. API: application programming interface; MRS: Medical Record System; UI: user interface; UX: user experience.



Based on the scope of work (ie, a detailed document or description that outlines the specific tasks, activities, objectives, and deliverables associated with a particular project) received, the technological team assessed requirements and checked the feasibility of incorporating them.

The next step involved the design of the mobile app user interface (UI)/user experience (UX) prototype, which was an interactive mock-up of the mobile app. The prototype contained key UIs, screens, and simulated functions without any working code or final design elements. This provided a better understanding of the real-time UI and UX before production.

Subsequently, the EDSSs were designed according to standard existing diagnosis and management guidelines, which were programmed to develop the most appropriate apps. To identify encounter data fields, individual interactions by ASHAs/doctors were recorded as separate encounters in OpenMRS. Different study phases had distinct data points, necessitating a logical flow of questions. Specific roles were assigned, tailoring the data collection tools to individual responsibilities. For instance, the follow-ups for ASHAs used priority-listing questions, whereas the doctors app incorporated mhGAP tool queries. This ensured targeted and relevant data capture for each study participant.

The next step involved configuring project-specific technical details such as concepts, encounter types, visit frequencies, user roles, and API settings within OpenMRS. Additionally, custom tables were created to facilitate real-time reporting and analytics, ensuring efficient data management and analysis for the project.

The final step was the development of the mobile app and APIs, which was carried out as a multistage process. The set up followed the sequence of development, test stage, and production environments. The final prototype for the mobile app involved integrating the EDSS into the app. The SMH apps supported online/offline features. Standard security integrations were enabled while developing the mobile app in the local database in the three different environments. In the test environment, the integrated feature was assessed with test data to evaluate the impact of the load of data and the performance of the app. In the stage environment, this phase included an exact replica of a production environment for testing. In the production environment, the software or products were made live for use. Once the development of the app was complete and certified by the quality assurance team, it was deployed for the production environment. Screenshots of the app are provided in Figures 4-6.

Continuous modifications and maintenance of the app were applied across the projects' lifetimes.

Figure 4. App screenshot 1.

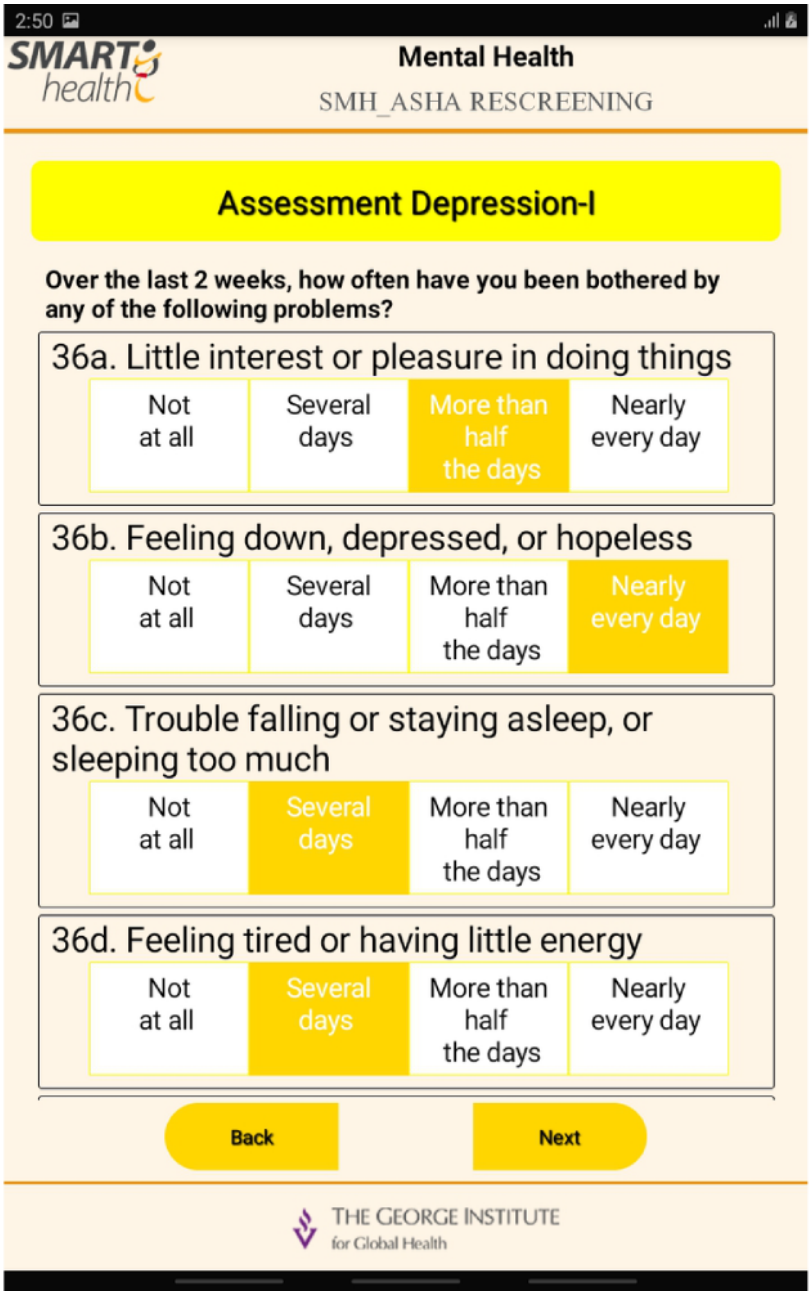


Figure 5. App screenshot 2.

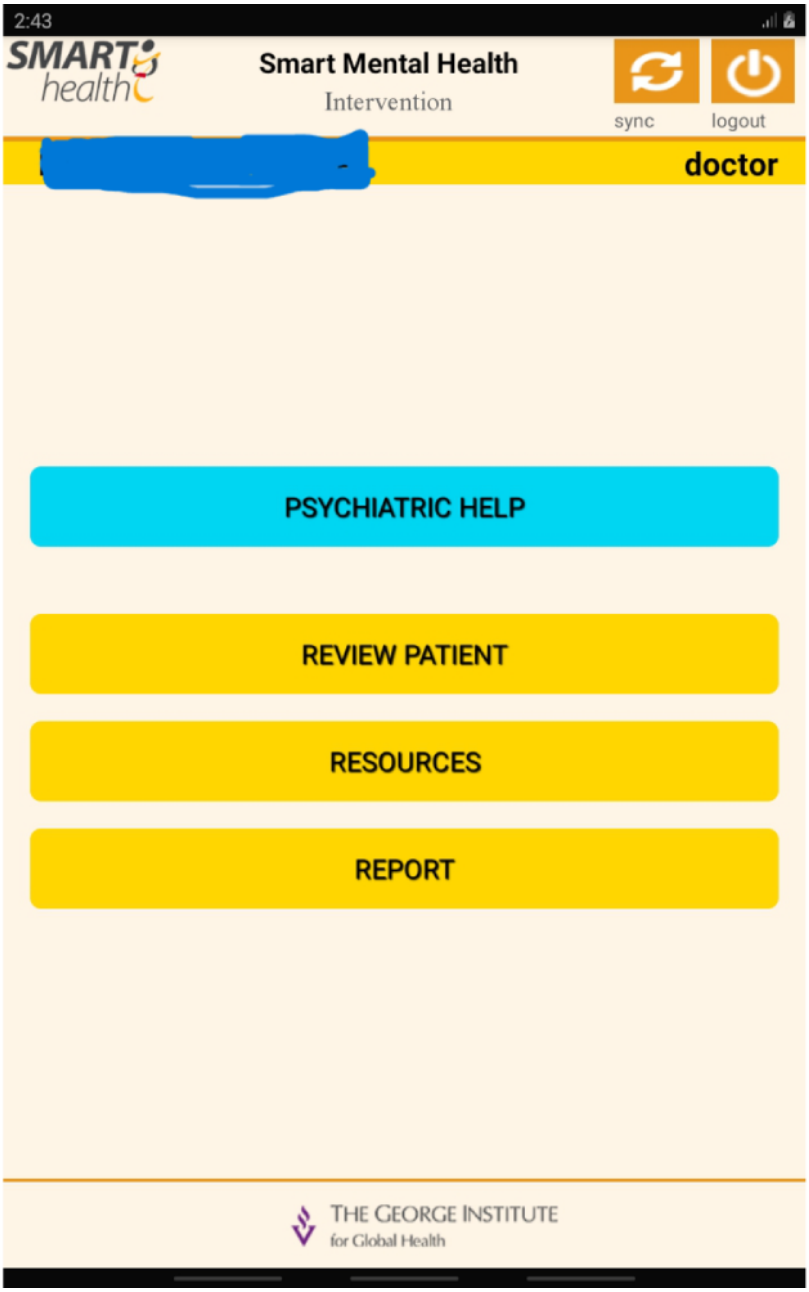
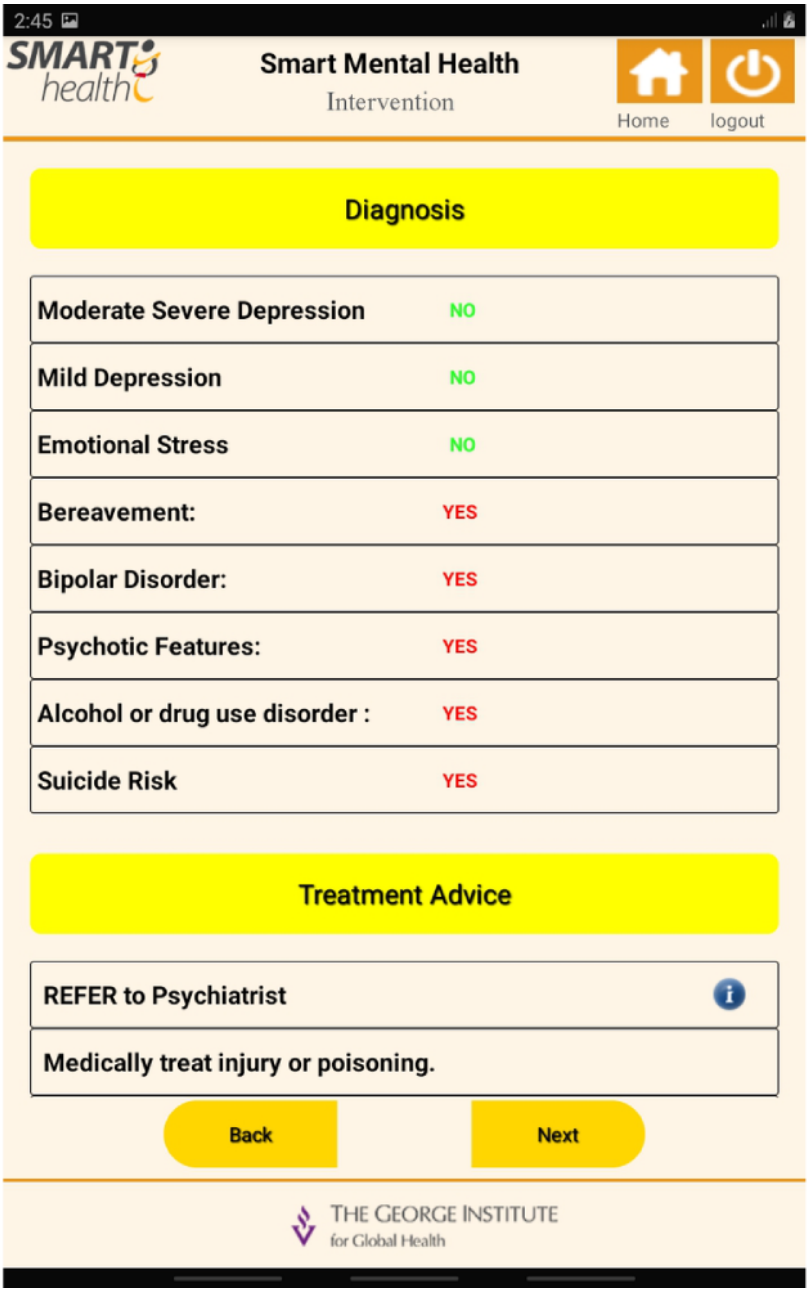


Figure 6. App screenshot 3.



Target

The SMH Pilot was implemented in 42 villages across rural and tribal areas of Andhra Pradesh [8,17] with the goal of understanding the feasibility and acceptability of using mobile technology and task-sharing approaches to address CMDs. This project covered approximately 50,000 adults and informed the subsequent SMH Trial, which took place in villages across Haryana and Andhra Pradesh, screening 165,000 adults in 133 villages and 44 PHCs. Currently, ARTEMIS is being implemented among 70,000 adolescents (10-19 years old) in 60 urban slum clusters in Vijayawada (Andhra Pradesh) and New Delhi.

Ethical Considerations

All collected data are securely stored on central servers in Hyderabad, with restricted access limited to the project team. Participants provided written consent and received detailed information about data collection at various time points. The SMART Mental Health pilot study was approved by the Independent Ethics Review Committee of the Centre for Chronic Disease Control (IRB00006330) for studies CCDC_IEC_03_2014 and CCDC_IEC_02_2014 on October 1, 2014; the SMART Mental Health cluster randomized controlled trial was approved by the George Institute Ethics Committee (009/2018) on April 27, 2018; and the ARTEMIS trial was approved by the George Institute Ethics Committee (17/2020) on September 4, 2020. The study tools were approved by The George Institute Ethics Committee, and each participant was assigned a unique identification number at the study's outset. Data were consistently deidentified before any sharing, and only research staff and the study's implementation and statistical teams had access to the data, ensuring that confidentiality and ethical standards were maintained throughout the research process.

Participating Entities

The studies have received funding from various international organizations such as Wellcome Trust/Department of Biotechnology (India Alliance), National Health and Medical Research Council Australia, and the UK Medical Research Council. Importantly, these funders are not involved in data collection or analysis and do not have access to the data. Government agencies, although collaborators, also do not manage or analyze the data. The SMH app is under intellectual property rights of the developer, The George Institute India. Local government consultation occurred for support, but they have no role in data governance.

Budget Planning

A predefined budget was allocated to the development and implementation of the technological interventions. The main costs incurred included the cost of the server (INR 500,000=US $6862) and the time cost of an Android developer and a technical lead (INR 200,000=US $2868/month for the initial 6 months for development and then a 25% time cost for maintenance). The other costs included the procurement of tablets for data collection.

Interoperability

The apps used in the three studies followed the Health Level 7 (HL7)/Fast Healthcare Interoperability Resources (FHIR) standards for exchanging patient information between a server and mobile app in JavaScript Object Notation (JSON) format. HL7 has also developed other standards, including the HL7 Clinical Document Architecture. We used FHIR in our apps as it was designed to facilitate interoperability of health care systems, allowing different health care apps and devices to easily exchange and share data. As the FHIR standard is based on modern web technologies such as Representational State Transfer principles, JSON, and Extensible Markup Language, it provides a flexible and scalable approach to health care data exchange, making it easier for developers to build interoperable apps.

Sustainability

The study was developed and implemented in collaboration with the Andhra Pradesh and Haryana governments. The tool has been previously utilized in two studies with adults while undergoing several phases of enhancements and is currently being used in the ARTEMIS study with adolescents. Poststudy, the tool will be shared with government and other nongovernmental organizations interested in using it.

Implementation (Results)

Coverage

The overall coverage of the number of study participants, ASHAs/CWVs, and doctors reached in the three studies is detailed in Table 2.

Table 2. Coverage of participants across the three projects.

Project	Study participants reached, n	ASHAs[a]/CWVs[b] included, n	Doctors included, n
SMART MH[c] Pilot (2014 to 2019)	50,000 adults	40	14
SMART MH Trial (2018 to 2022)	165,000 adults	175	50
ARTEMIS[d] (2020-2024)	69,600 adolescents (10-19 years old)	104	27

[a]ASHA: Accredited Social Health Activist.

[b]CWV: community woman volunteer.

[c]SMART MH: Systematic Medical Appraisal and Referral Treatment Mental Health.

[d]ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums.

Outcomes and Technical Amendments Made to the Data Repository

OpenMRS indicated the overall number of instances that data were collected for each individual participant at different time points in the study. As OpenMRS has a report generation model, it was difficult to compare different data points for the same person or between participants across the same time point. Hence, an intermediary database was developed in house to facilitate the process of running customized reports, which enabled comparison of data at different time points. This process evolved following considerable testing at the backend to obtain the desired output in terms of data visualization. Some customizations were made to OpenMRS to suit study requirements (Table 3).

Table 3. Steps of configurations made to the Open Medical Record System (OpenMRS).

Configuration of OpenMRS modules	Features for the study
Creation of a concept dictionary	Every data point to be used for the study was created as a concept and given a short name
Role management	The roles of each user were fixed and were restricted based on the type of activity they were expected to do; for example, the project manager was only given access to user data management and downloading reports
User management	As per our project flow, the different users were allocated to each role, such as ASHAs[a], doctors, field staff/data collector, project manager, and administrator
Encounter management	Each entry into the tab for a specific user (ie, ASHA, doctor, data collector) was recorded as an encounter with a unique encounter ID, which helped to differentiate the number of encounters that had taken place for each study participant
Managing encounter types	Based on the different phases of the study, each phase was also considered as a separate encounter, such as the screening, rescreening, ASHA follow-up, and doctor follow-up phases
Manage observations	Each data point was considered as a separate observation
Managing persons	Demographic data for every app user (ASHA, doctor) or participant were stored as person details
Managing patients	In this feature, any additional personal identifiers/demographic details identified could be modified/configured
Cohort management	Specific cohorts were created for every phase of the project, matched to the user. This enabled the users to access data of people who were in their own cohort. This helped them to identify and follow up the individuals easily. This was done both for ASHAs and doctors, with each doctor in a particular PHC[b] having a defined set of ASHAs, who in turn had a defined set of high-risk individuals
Multilocation data management	This was a custom development made to the system to ensure the data of one location (state) were not merged with data from another location. This was relevant to the SMH[c] and ARTEMIS[d] trials, which involved two different geographical locations.

[a]ASHA: Accredited Social Health Activist.

[b]PHC: primary health center.

[c]SMH: Systematic Medical Appraisal and Referral Treatment Mental Health.

[d]ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums.

In India, internet connectivity varies, particularly in rural regions. To address this, the quality of connectivity was assessed in each study area and hotspots were identified. Offline and online data storage methods were implemented, allowing local storage on tablets in areas with poor connectivity. Data could later be uploaded to the central server once connectivity was restored. Additionally, networks at certain PHCs were improved, increasing the bandwidth to enable ASHAs and doctors to upload data when in proximity to these PHCs.

Lessons Learned

There were several lessons learned while designing and implementing these interventions, which resulted in several enhancements to the systems for improving UX and achieve the study outcomes. Following the SMH Pilot, issues were identified in the EDSS that needed to be corrected for the SMH and ARTEMIS trials (Table 4). During the project, unforeseen challenges arose due to the COVID-19 pandemic. Face-to-face training for health workers was impossible, leading to the preparation of training materials delivered with the assistance of field staff. Additionally, some tablets used by health workers broke down, necessitating replacements and revealing bugs in the app. The SMH cRCT project faced difficulties because of COVID-19, and different mitigation strategies were adopted to ensure implementation of the different stages of the project. However, due to the rapidly changing situation, those also had to be modified quickly. Considering all the issues encountered earlier, we tried to mitigate all these challenges encountered during the SMH pilot study and cRCT, leading to enhancement of the apps developed for the ARTEMIS project. To have a smooth transition from the test environment to the live environment, the technical team performed additional checks by testing the apps by the field staff and creating data that were uploaded to the server to confirm whether all the fields are being populated correctly. This helped in reducing the errors while data were being captured in live scenarios.

Table 4. Enhancements made to the electronic decision support system.

Issues that needed amendments	Solutions for the problems/issues
Daily monitoring of data at the field level and comparison of data across sites, localities, and users was very difficult. Monitoring of clinical data of patients was also difficult	Development of descriptive analytics at the database level while implementing the SMH[a] trial was done to ease monitoring of data. There were many enhancements made at that level, in terms of representing real-time data from different aspects of the study. This included identification of mental health service use, the burden of different mental health conditions, and comparison of different conditions, among other factors. These analytics could be viewed by comparisons made across regions, gender, and age groups. These were represented through pictorial modes such as graphs and pie charts (see Figure 7 for examples)
Monitoring an individual's mental health status over time was not possible	Analytics were developed to track the PHQ9[b] and GAD7[c] scores of an individual in the different phases of the study. Data captured periodically during monitoring could be viewed as graphs and charts based on the longitudinal data at the backend using analytics.
The performance of ASHAs[d] could not be tracked well	There were enhancements made to the ASHA app, which tracked the performance of each ASHA and provided data about the numbers of screenings and follow-ups performed, including the time taken for each. Random audio recordings of their interactions were also captured to ensure quality checks.
As the database is encrypted and stored in a password-protected, secure location, it is hard to gain access to data by reverse engineering or decoding	The app is protected with multifactor authentication using a password and lock pin as an enhancement to the existing setup.
User interface and functioning of the app were not clear	Several changes were made to the user interface, including a change of font size, color, and creating different section headers using attractive symbols/pictures, for better user experience
Enabling online training during COVID-19	Some of the training materials were embedded in the mobile apps to enable easier access for trainees using virtual modes during COVID-19.
Real-time monitoring of the activities of field staff was required to ensure increased data quality	Random audio recording of interaction of field staff with study participants or high-risk individuals was enabled. The time taken for each screening was also made available at the database level for these audio recordings. This helped the implementation team to monitor data collection and quality.
Merging of two apps, namely household listing and participant screening, into one app	This merger made it possible for simultaneous data collection for both listing and screening, which saved time for both the participant and field staff and reduced multiple visits to the same household for data collection.

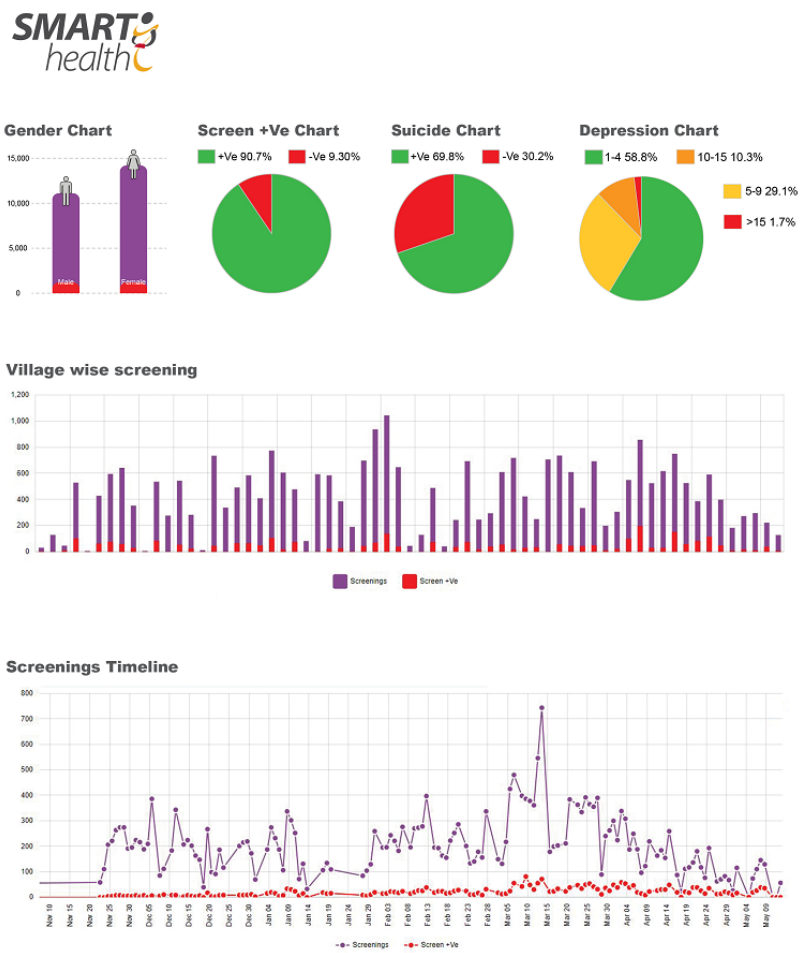
[a]SMH: Systematic Medical Appraisal and Referral Treatment Mental Health.

[b]PHQ9: Patient Health Questionnaire-9.

[c]GAD7: Generalized Anxiety Disorder-7.

[d]ASHA: Accredited Social Health Activist.

Figure 7. Snapshot of the analytics developed to monitor the progress of the study outcomes. +Ve: positive; -Ve: negative.



Discussion

Principal Findings

This paper outlines the experience of employing technology in mental health service delivery across rural and urban India in three projects. We have highlighted the implementation challenges and app adaptations based on user feedback, offering insights valuable for technology-based mental health projects in resource-limited settings. Technology-enabled interventions have shown effectiveness in diagnosing, treating, and following up on various health conditions [18,19]. Most mHealth interventions used in India have been disease-specific and do not involve a health systems approach. One example of a more health system–focused app is the Government to Government web-based monitoring information system that has been set up by the Ministry of Health & Family Welfare, Government of India, to monitor the National Health Mission and other health programs. To increase effectiveness, these innovations should focus on creating new avenues to integrate tools that have encouraging and sustainable outcomes related to access, equity, quality, and responsiveness. The SMH app can be integrated with government systems after specific modifications. The use of electronic medical record systems and telemedicine are examples of some of the interventions implemented and found to be beneficial for health care delivery for large populations, especially in LMICs [16,19,20]. However, there is a need to understand the local context and setting while developing or enhancing any existing app, as some of the original features may not be relevant to the local context, making further adaptations critical.

One way to enhance the functional capabilities of apps such as SMH is to link the app with telemedicine facilities that amplify the ability to connect to remotely located consumers with specialists located in larger cities [13]. For example, machine learning has been applied for suicide prediction, matching patients to appropriate treatment, improving the efficacy of mental health care by clinicians, and monitoring patients for treatment adherence with the help of smartphones and sensors [21].

Another way to leverage technology in mental health is by using artificial intelligence. A recent systematic review recommended the use of artificial intelligence technologies as accurate and effective strategies in the diagnosis and treatment of mental health conditions [22]. Virtual reality technology has proven to be a useful and powerful tool in addiction research [23]. The user interacts with the virtual reality environment, offering an environment close to real life that is dynamic in nature and requires active participation. These environments can be used to develop psychotherapeutic interventions by adding a personal

touch, having predictable conditions with additional features such as embodiment, eye tracking, and other biological factors [24].

There is still substantial work to be done in terms of scaling up these interventions and understanding their feasibility and acceptability across different settings and populations. Use of novel strategies such as videogaming can be explored to implement mental health interventions that can be customized to specific populations [25]. Such techniques should be considered in future iterations of the technology platform [26,27].

Limitations

There were a few limitations in our apps. First, the mobile apps developed were limited to stress, depression, anxiety, and increased suicide risk; however, the principles of including other mental health conditions would be quite similar. Second, although the projects had a system of referring participants requiring specialist care to mental health professionals, it was beyond the scope of the projects to track the care provided by the mental health professionals through our app. This was because our app was developed through primary health system–focused application for use in low-resource settings and was not linked to any central electronic health record system as is possible in more developed health systems with more robust data capture and record-sharing capabilities, such as the National Health Service in the United Kingdom or health systems in Australia. Third, the current apps are compatible on Android platforms and could not be expanded to other operating systems. Finally, the apps developed were specifically created following consultations with local stakeholders; hence, their generalizability across other settings will need to be assessed after adaptation is complete.

Future Recommendations

Given our experiences, we have compiled a set of suggested recommendations for technology-based interventions in similar settings, which are presented in Textbox 1.

Textbox 1. Recommendations for technology-based interventions.

- Inclusion of the technical team from the outset when study protocols are being developed.
- All study-related tools and database designing should be finalized in consultation with relevant experts.
- A protocol that details the process of server support in terms of setup/maintenance needs should be developed and followed.
- The server needs to factor in the size of the data set and latest versions of operating systems in reducing any issues faced.
- App user interface/user experience should be designed and assessed for acceptability by targeted populations. The use of reports or data analytics for the study must be discussed and finalized as per study needs.
- Develop systems that can be used across any kind of device, are compatible for software or version upgrades, and are web-based and easily programmable.
- A technical guide with frequently asked questions outlining the various aspects of technology, such as navigation, problem-solving, and reporting of issues, should be developed to facilitate staff training.
- The infrastructure and the architecture of the app should be flexible for making modifications or scaling up the app. The scalability is measured by the number of requests an app can manage and support the app effectively. A decision needs to be taken in terms of adding resources to the computing system for scaling either horizontally (adding more machines to the existing pool) or vertically (adding more power to the existing machines). Both types of scaling are similar as they add computing resources to the infrastructure; however, there are distinct differences between the two in terms of implementation and performance.

Conclusion

In conclusion, the development of any health-related app is subject to the context and the area where it would be implemented. There is a need for careful testing using iterative processes, allocate human and budgetary resources that are adequate, and integrate apps with larger electronic health record systems that inform health systems.

Acknowledgments

We would like to acknowledge the entire Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health team at the Andhra Pradesh, Haryana site and the staff of Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums (ARTEMIS) at the Vijayawada and Delhi sites who provided input while assessing the apps, which greatly helped in making specific enhancements to the apps. This work was supported by ARTEMIS (to SG, MG, and SK), Australian National Health and Medical Research Council (NHMRC) Global Alliances for Chronic Disease (GACD) (SMART for Common Mental Disorders in India; grant APP1143911), and UK Research and Innovation (UKRI)/Medical Research Council (MRC) grant (ARTEMIS; MR/S023224/1 to PKM). DP is partially or wholly supported through the SMART Mental Health NHMRC/GACD grant. PKM is the principal investigator on the ARTEMIS Project and coprincipal investigator on the SMART Mental Health Project and is partially supported by both projects. DP is supported by fellowships from the NHMRC of Australia (1,143,904) and the Heart Foundation of Australia (101,890). SKY is supported by the ARTEMIS Project (UKRI/MRC grant MR/S023224/1), SP is supported by the ARTEMIS Project and another project titled MentAL Health Risk Factors among Older AdolesceNts living in Urban Slums: An InTervention to Improve ResIlience (ANUMATI) funded by the Indian Council of Medical Research

(grant 2019-0531). MD is supported by SMART Mental Health funded by NHMRC Australia (grant APP1143911) and the International Study of Discrimination and Stigma Outcomes (INDIGO) Partnership Research Programme funded by the UK MRC (MR/R023697/1). The funding bodies played no role in the design of the study and in the conceptualization and writing of the manuscript.

Authors' Contributions

SK, SG, PKM: conceptualization. SK and SG: writing of first draft. MG, SK, SKY, SP, MD, DP, and PM: review & editing. All authors read and approved the final manuscript.

Conflicts of Interest

All authors are employees of The George Institute, which has a part-owned social enterprise, George Health Enterprises, with commercial relationships involving digital health innovations.

References

1. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016 Feb;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
2. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bull World Health Organ 2004 Nov;82(11):858-866 [FREE Full text] [Medline: 15640922]
3. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 2017 Feb 02;210(2):119-124 [FREE Full text] [doi: 10.1192/bjp.bp.116.188078] [Medline: 27908899]
4. Naslund JA, Marsch LA, McHugo GJ, Bartels SJ. Emerging mHealth and eHealth interventions for serious mental illness: a review of the literature. J Ment Health 2015 May 28;24(5):321-332 [FREE Full text] [doi: 10.3109/09638237.2015.1019054] [Medline: 26017625]
5. World Health Organization. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. Geneva: World Health Organization; 2011.
6. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
7. Agrawal N. Telephone network and internet penetration in India: a pragmatic study using data analytics. Global J Enterprise Inf Syst 2021;13(1):42-48 [FREE Full text]
8. Maulik PK, Devarapalli S, Kallakuri S, Bhattacharya A, Peiris D, Patel A. The systematic medical appraisal referral and treatment mental health project: quasi-experimental study to evaluate a technology-enabled mental health services delivery model implemented in rural India. J Med Internet Res 2020 Feb 27;22(2):e15553 [FREE Full text] [doi: 10.2196/15553] [Medline: 32130125]
9. Daniel M, Maulik PK, Kallakuri S, Kaur A, Devarapalli S, Mukherjee A, et al. An integrated community and primary healthcare worker intervention to reduce stigma and improve management of common mental disorders in rural India: protocol for the SMART Mental Health programme. Trials 2021 Mar 02;22(1):179 [FREE Full text] [doi: 10.1186/s13063-021-05136-5] [Medline: 33653406]
10. Yatirajula SK, Kallakuri S, Paslawar S, Mukherjee A, Bhattacharya A, Chatterjee S, et al. An intervention to reduce stigma and improve management of depression, risk of suicide/self-harm and other significant emotional or medically unexplained complaints among adolescents living in urban slums: protocol for the ARTEMIS project. Trials 2022 Jul 29;23(1):612 [FREE Full text] [doi: 10.1186/s13063-022-06539-8] [Medline: 35906663]
11. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
12. Maulik PK, Tewari A, Devarapalli S, Kallakuri S, Patel A. The Systematic Medical Appraisal, Referral and Treatment (SMART) Mental Health Project: development and testing of electronic decision support system and formative research to understand perceptions about mental health in rural India. PLoS One 2016 Oct 12;11(10):e0164404 [FREE Full text] [doi: 10.1371/journal.pone.0164404] [Medline: 27732652]
13. Daniel M, Kaur A, Mukherjee A, Bhattacharya A, Tewari A, Sagar R, et al. The systematic medical appraisal, referral and treatment (SMART) mental health programme: formative research informing a cluster randomized controlled trial. SSM Mental Health 2023 Dec;3:100223. [doi: 10.1016/j.ssmmh.2023.100223]
14. Scaling up care for mental, neurological and substance use disorders: mhGAP. World Health Organization. 2008. URL: https://www.who.int/activities/scaling-up-mental-health-care [accessed 2024-01-14]
15. OpenMRS. URL: https://openmrs.org/ [accessed 2024-01-22]
16. Peiris D, Praveen D, Mogulluru K, Ameer MA, Raghu A, Li Q, et al. SMARThealth India: a stepped-wedge, cluster randomised controlled trial of a community health worker managed mobile health intervention for people assessed at high

cardiovascular disease risk in rural India. PLoS One 2019 Mar 26;14(3):e0213708 [FREE Full text] [doi: 10.1371/journal.pone.0213708] [Medline: 30913216]
17. Maulik PK, Kallakuri S, Devarapalli S, Vadlamani VK, Jha V, Patel A. Increasing use of mental health services in remote areas using mobile technology: a pre-post evaluation of the SMART Mental Health project in rural India. J Glob Health 2017 Jun;7(1):010408 [FREE Full text] [doi: 10.7189/jogh.07.010408] [Medline: 28400954]
18. Bassi A, John O, Praveen D, Maulik PK, Panda R, Jha V. Current status and future directions of mHealth interventions for health system strengthening in India: systematic review. JMIR Mhealth Uhealth 2018 Oct 26;6(10):e11440 [FREE Full text] [doi: 10.2196/11440] [Medline: 30368435]
19. Koppar AR, Sridhar V. A workflow solution for electronic health records to improve healthcare delivery efficiency in rural India. 2009 Presented at: 2009 International Conference on eHealth, Telemedicine, and Social Medicine; February 1-7, 2009; Cancun, Mexico. [doi: 10.1109/etelemed.2009.30]
20. Acharya R, Rai J. Evaluation of patient and doctor perception toward the use of telemedicine in Apollo Tele Health Services, India. J Family Med Prim Care 2016;5(4):798-803 [FREE Full text] [doi: 10.4103/2249-4863.201174] [Medline: 28348994]
21. Haggerty E. Healthcare and digital transformation. Network Security 2017 Aug;2017(8):7-11. [doi: 10.1016/s1353-4858(17)30081-8]
22. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim H, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 07;21(11):116 [FREE Full text] [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
23. Segawa T, Baudry T, Bourla A, Blanc J, Peretti C, Mouchabac S, et al. Virtual reality (VR) in assessment and treatment of addictive disorders: a systematic review. Front Neurosci 2019 Jan 10;13:1409 [FREE Full text] [doi: 10.3389/fnins.2019.01409] [Medline: 31998066]
24. Cavalcante Passos I, Mwangi B, Kapczinski F. Personalized psychiatry: Big data analytics in mental health. New York, NY: SpringerLink; 2019.
25. Hamari J, Keronen L. Why do people play games? A meta-analysis. Int J Inf Manag 2017 Jun;37(3):125-141. [doi: 10.1016/j.ijinfomgt.2017.01.006]
26. Wilkinson N, Ang RP, Goh DH. Online video game therapy for mental health concerns: a review. Int J Soc Psychiatry 2008 Jul 01;54(4):370-382. [doi: 10.1177/0020764008091659] [Medline: 18720897]
27. Li J, Theng Y, Foo S. Game-based digital interventions for depression therapy: a systematic review and meta-analysis. Cyberpsychol Behav Soc Netw 2014 Aug;17(8):519-527 [FREE Full text] [doi: 10.1089/cyber.2013.0481] [Medline: 24810933]

Abbreviations

API: application programming interface
ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums
ASHA: Accredited Social Health Activist
CMD: common mental disorder
CWV: community woman volunteer
cRCT: cluster randomized controlled trial
EDSS: electronic decision support system
FHIR: Fast Healthcare Interoperability Resource
HL7: Health Level 7
JSON: JavaScript Object Notation
LMIC: low- and middle-income country
mHealth: mobile health
OpenMRS: Open Medical Record System
PHC: primary health center
SMH: Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health
UI: user interface
UX: user experience
WHO: World Health Organization

Edited by C Perrin; submitted 22.03.23; peer-reviewed by N Mungoli, E Korshakova; comments to author 24.04.23; revised version received 21.05.23; accepted 29.11.23; published 15.02.24.

Please cite as:
Kallakuri S, Gara S, Godi M, Yatirajula SK, Paslawar S, Daniel M, Peiris D, Maulik PK
Learnings From Implementation of Technology-Enabled Mental Health Interventions in India: Implementation Report
JMIR Med Inform 2024;12:e47504
URL: https://medinform.jmir.org/2024/1/e47504
doi:10.2196/47504
PMID:38358790

Implementation Report

A Mobile App (Concerto) to Empower Hospitalized Patients in a Swiss University Hospital: Development, Design, and Implementation Report

Damien Dietrich[1,2], MD; Helena Bornet dit Vorgeat[3], BSc, MBA; Caroline Perrin Franck[1], PhD; Quentin Ligier[3], MSc

[1]Geneva Hub for Global Digital Health, Faculty of Medicine, University of Geneva, Geneva, Switzerland

[2]Kheops Technologies SA, Plan-Les-Ouates, Switzerland

[3]Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Damien Dietrich, MD
Geneva Hub for Global Digital Health
Faculty of Medicine
University of Geneva
Campus Biotech
9 Chemin des Mines
Geneva, 1202
Switzerland
Phone: 41 227714730
Email: damien.dietrich@gmail.com

Abstract

Background: Patient empowerment can be associated with better health outcomes, especially in the management of chronic diseases. Digital health has the potential to promote patient empowerment.

Objective: Concerto is a mobile app designed to promote patient empowerment in an in-patient setting. This implementation report focuses on the lessons learned during its implementation.

Methods: The app was conceptualized and prototyped during a hackathon. Concerto uses hospital information system (HIS) data to offer the following key functionalities: a care schedule, targeted medical information, practical information, information about the on-duty care team, and a medical round preparation module. Funding was obtained following a feasibility study, and the app was developed and implemented in four pilot divisions of a Swiss University Hospital using institution-owned tablets.

Implementation (Results): The project lasted for 2 years with effective implementation in the four pilot divisions and was maintained within budget. The induced workload on caregivers impaired project sustainability and warranted a change in our implementation strategy. The presence of a killer function would have facilitated the deployment. Furthermore, our experience is in line with the well-accepted need for both high-quality user training and a suitable selection of superusers. Finally, by presenting HIS data directly to the patient, Concerto highlighted the data that are not fit for purpose and triggered data curation and standardization initiatives.

Conclusions: This implementation report presents a real-world example of designing, developing, and implementing a patient-empowering mobile app in a university hospital in-patient setting with a particular focus on the lessons learned. One limitation of the study is the lack of definition of a “key success” indicator.

(JMIR Med Inform 2024;12:e47914) doi:10.2196/47914

KEYWORDS

patient empowerment; mobile apps; digital health; mobile health; implementation science; health care system; hospital information system; health promotion

Introduction

Context

During recent decades, medicine has been moving from a focus on paternalistic approaches toward a paradigm of patient-centeredness, highlighting patient partnership and participation. Patient empowerment refers to a metaconcept with no unique definition [1]. However, it is commonly accepted that empowered patients possess key capacities and resources to be able to (1) participate in shared decision-making, (2) manage their own health, and (3) self-empower themselves [1].

Patient Empowerment and Clinical Outcomes

Some studies have demonstrated a positive association between patient empowerment and improved clinical outcomes or their proxy. This is best documented in the context of chronic diseases, especially diabetes. Wong et al [2] compared serum glycated hemoglobin (HbA_{1C}) and low-density lipoprotein cholesterol (LDL-C) levels in a group following implementation of a patient empowerment program (PEP) or the standard of care, resulting in decreased LDL-C levels in the PEP group. Similarly, Lian et al [3] found a lower incidence of all-cause mortality, cardiovascular events, and diabetes mellitus complications following participation in a PEP. In a review of randomized controlled trials, a decrease in HbA_{1C} and blood pressure levels was associated with empowerment interventions for patients with diabetes in sub-Saharan Africa [4]. In a meta-analysis, Baldoni et al [5] reported an improvement in HbA_{1C} levels following collective empowerment strategies. In a systematic review, Shnaigat et al [6] identified patient activation, a concept related to empowerment, as a valid strategy to improve outcomes of patients with chronic obstructive pulmonary disease.

However, it is important to highlight that several studies also reported no beneficial effects of empowerment programs. A 2017 meta-analysis found no statistically significant positive effect of empowerment on HbA_{1C} levels, despite five included studies reporting positive results [7]. Santoso et al [8] reported a lack of evidence to demonstrate a positive association between women's empowerment and outcomes of child nutrition.

The lack of clear definitions and measures for empowerment may explain these controversial findings. Differences in program design could also contribute to this variability; therefore, further research identifying determinants for a successful intervention is needed. Indeed, the authors of the cited studies often reported the poor availability of high-quality research.

Digital Health and Patient Empowerment

With the variety of solutions that could be envisioned, digital health is seen as a promising tool to promote patient empowerment and, indirectly, outcomes. However, mixed results are seen in the related literature.

In a systematic review, Johansson et al [9] showed that online communities support patient empowerment. Sosa et al [10] reported that a text messaging–based empowering intervention following head and neck surgery was both highly appreciated by patients and feasible. Conversely, Ammenwerth et al [11] reported no clinically relevant effect of patient portals on patient empowerment or health-related outcomes in a systematic review. Vitger et al [12] failed to demonstrate a positive effect of digital interventions to support shared decision-making, which was likely due to the small number of high-quality studies available. Verweel et al [13] found limited evidence demonstrating a positive effect of a digital intervention for health literacy. Finally, Thomas et al [14] reported that the quality and adequacy of the content of patient-empowering mobile apps varied greatly, urging for a more rigorous design and further testing before implementation. To our knowledge, no study has directly shown a link between mobile health app–induced empowerment and direct health outcomes.

Overall, few high-quality studies assessing the effect of digital health interventions on patient empowerment are available. Research is needed to confirm or deny the high perceived potential of digital tools.

Concerto: A Mobile App Designed to Promote Patient Empowerment

Concerto is a mobile app designed to promote the empowerment of hospitalized patients. The app was initially designed during a hackathon in 2015 by a multidisciplinary team including health care and IT professionals as well as one patient. Building on the hackathon prototype and after a feasibility study, the Geneva University Hospitals (HUG) launched a project aiming at developing and implementing a fully functional mobile app delivered on institution-owned tablets in four pilot divisions (oncology, neurorehabilitation, orthopedics, and pediatrics) and assessing its effectiveness. Following this pilot study, the mobile app was further refined and deployed institution-wide based on a bring-your-own-device (BYOD) approach. This implementation report focuses on the pilot study only, with the objective to highlight the lessons learned. The report is structured following the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) guidelines [15].

Methods

Design and Agile Development

Building on the prototype developed during the hackathon, the foreseen functionalities of Concerto were first compared with patients' expectations using focus groups and a semiquantitative questionnaire. A feasibility study was then performed to assess the availability and quality of the necessary data in the hospital health information system (HIS), which has been developed mainly in-house during the last 30 years.

Based on the patients' insights, further described in Dietrich et al [16], version 1.0 of Concerto was specified and developed using an agile methodology with frequent user testing among hospitalized patients. The main functionalities of this version of Concerto included:

1. An up-to-date calendar on which patients can visualize their care schedule and better understand their daily planning with the aim to reduce the impact of these events and be better prepared for them.

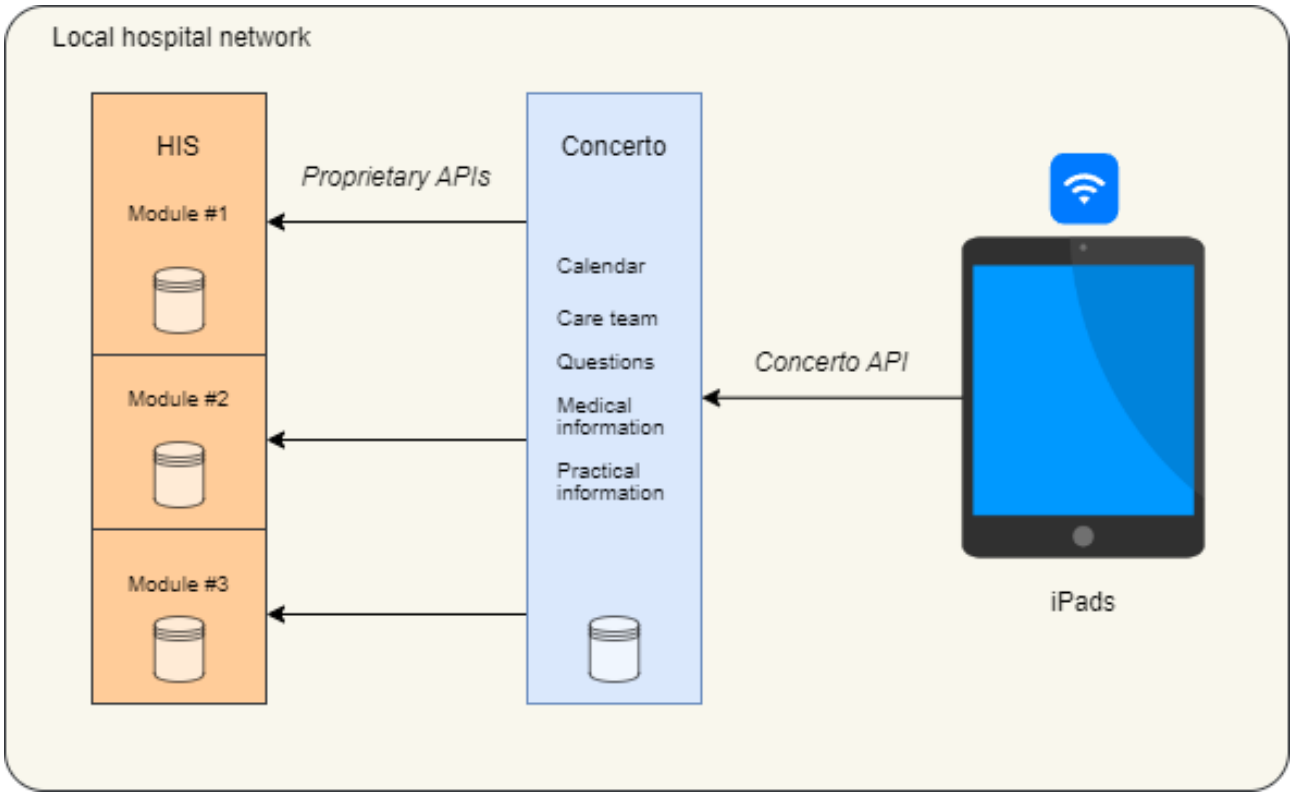
2. A care team module on which the patient can obtain information about their on-duty care team, including names and photographs, to better know the professionals they will meet during their stay and facilitate communication.
3. A “questions” module on which patients can prepare their questions for health care professionals during the medical round and thus elevate the level of communication.
4. An “information” module on which patients have access to targeted medical information. Expected benefits of this module were to achieve better situation awareness, better treatment adherence, and early detection of complications.
5. A “practical information” module on which patients can find useful information about their stay at the hospital to improve their overall experience.
6. A social network module on which patients can interact with HUG accounts.

In subsequent versions, a new module was added, allowing the patient to choose their meal directly on the app rather than via a form completed by the nurse. This module was designed to simultaneously improve the patient experience while decreasing the nurse workload.

The app was developed in web languages (an Angular project), encapsulated as an iOS app (with Apache Cordova), and deployed on institution-owned tablets using a mobile device management solution. The key arguments for internal development over acquisition of a commercial solution were that (1) a significant part of the development work was about interfacing with the HIS, and (2) to our knowledge, no adequate and mature commercial solution was available at the time, although such solutions have emerged since then.

The initial version of the app connected directly to the custom-made HUG HIS using its proprietary interfaces for the sake of development simplicity and to alleviate time constraints. Further versions of the app have used industry standards such as Health Level 7 Fast Healthcare Interoperability Resources, with the vision to enable Concerto to connect more easily to other HISs in the future. This update has required new developments on the HIS side and was not achievable during the pilot phase described in this report. Figure 1 presents the simplified architecture of Concerto.

Figure 1. Technical overview of Concerto. Through the hospital's private Wi-Fi, the patient's tablet connects to the Concerto server, which contains some patient-generated data and the business logic to provide the app data. The server connects to different modules of the hospital information system (HIS) to retrieve other data using diverse application programming interfaces (APIs).



Implementation

The definition of the logistics necessary to deliver Concerto on institution-managed tablets was an important part of the project. The following process was repeated for each patient: (1) setting up the tablet, including defining a personal passcode; (2) two-factor authentication in the Concerto app using the patient ID, scanned from the identification bracelet, and an SMS text messaging challenge; (3) on-demand charging; (4) disinfecting the tablet after the patient’s discharge; and (5) reinitializing and erasing the tablet. Tablets were charged and stored under key-secured storage in the nurse office. Each tablet was protected using individual cases. Hygiene procedures were validated by the Infection Prevention and Control Division of the HUG.

Once version 1.0 became available, caregivers of the different divisions were trained for 30 minutes during hands-on sessions

in which (1) the project and app were presented; (2) the logistics of the tablets were explained; and (3) most importantly, they had the opportunity to familiarize themselves with the tool. At least one caregiver was defined as a "superuser" on a voluntary basis and was implicated from the beginning of the project. The specific responsibilities of superusers included (1) acquiring deep knowledge of the app, (2) being the focal point for exchange with the project team, and (3) acting as the referent for day-to-day questions of caregivers. A typical division included 20 beds and comprised a pool of over 50 caregivers that were trained during different sessions. Importantly, as in many hospital projects, caregivers did not have dedicated time for the project. Therefore, they had to manage making themselves available during a normal day of work.

One unit was scheduled for launch every 2 weeks, with constant presence of one member of the Concerto team during the first few days. Only patients able to interact with a mobile app, as assessed by their caregivers, were offered to use the app. To this end, caregivers used a communication flyer describing the functionalities of the app, the modalities of its use, as well as data and privacy considerations.

Bugs, feedback, and general satisfaction were systematically consigned to fuel the improvement-and-fix backlog.

Data Considerations

At the stage presented during preparation of this report, Concerto worked mainly in "read-only" mode for personal health data available in the HIS and for insensitive, impersonal information. The information patients accessed from the HIS was part of their medical records. According to Swiss law, every patient owns the data contained in their medical record, except for personal notes of health care professionals, which were out of the scope of Concerto. Accordingly, Concerto facilitated access to data already owned by the patients.

The access to this sensitive personal information required a secure log-in based on the patient's ID number and a second-factor authentication with an SMS text challenge. The use of institution-owned devices allowed Concerto to access data in the hospital's local network, preventing unwanted access from the rest of the world.

The only personal information entered in Concerto included any questions patients may have had before interacting with their caregivers. This information was stored in the HIS and deleted after the hospital stay. Tablets were erased and reinitialized between patients, ensuring that no information leakage was possible between patients using the same tablet.

To summarize, Concerto facilitated the access to personal health information owned by the patient without the possibility to modify information from the app, and further allowed the patient to enter personal health information stored in the HIS that is inaccessible to others with all information systematically erased after the patient's hospital stay.

Overall, the project was compliant with the Swiss Law for Data Protection [17].

Funding and Budget Planning

The feasibility study and initial concept were self-funded by the eHealth and Telemedicine Division of the HUG, with the budget including salaries for a junior developer and a senior project manager.

The pilot project was then funded by private foundations based in Geneva, which included the salaries as well as necessary materials (tablets, covers, and software licenses).

Overall, the order of magnitude of the project costs ranged between US $150,000 and US $200,000, from which 25% was used for materials.

Ethical Considerations

This study is based on an internal project of a Swiss University Hospital, aiming for quality improvement. As such, no patient or participant was included specifically for this study. Moreover, no patient data of any kind were collected. Accordingly, this study does not qualify for a review by the Geneva Canton Ethics Board (Commission Cantonale d'Éthique de la Recherche sur l'Être Humain [18]). As there were no participants involved in the research, no consent, compensation scheme, or privacy and confidentiality considerations applies.

Implementation (Results)

Project Summary

Concerto was implemented in four pilot divisions; a typical division includes 20 beds and comprises a pool of over 50 caregivers.

The timeline of the various stages of the project is provided in Figure 2. From the initial hackathon to acquiring the funding, approximately 1 year was necessary to refine the concept with patients and assess the feasibility of the app. Following funding acquisition, 6 months of development were needed, followed by 6 months of piloting in the four selected hospital divisions. Overall, the project took 2 years.

The budget was respected. However, additional funding would have been welcome to help free the caregivers from their clinical duties to enable better implementation (see below for further discussion of this point).

The development team considered the agile development phase to be efficient and productive.

Critical to the development process, the organization of focus groups and one-on-one interviews with patients were facilitated owing to the clinical background of the project manager. The development team reported that early contacts with the IT division during the feasibility study helped to improve communication and hence efficiency. Finally, dedicated support of the management unlocked political stalling.

During the pilot phase, the app was proposed to all eligible patients (see the Implementation subsection of the Methods). The percentage of eligible patients was unfortunately not systematically monitored but varied according to the profile of hospitalized patients across the different divisions and over

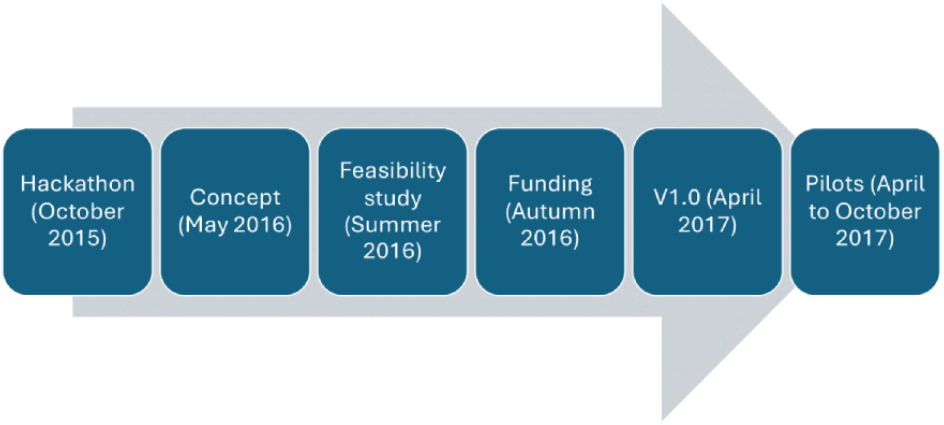
time. An eligibility rate below 50% of all hospitalized patients was common.

The percentage of acceptance was also not systematically collected. However, the project team recalled acceptance to be relatively less variable than the percentage of eligible patients and consistently high (over 80%).

The dropout rate should have also been monitored carefully to identify the reasons for dropping out.

Most importantly, navigating the logistics of the tablet emerged as a particular challenge for caregivers. Despite the support of the project team, this impaired the inclusion of patients and consequently use of the app. More precisely, caregivers reported difficulties in assisting patients with the log-in and reinitialization procedures, and all logistics steps were reported as being too time-consuming. Based on this finding, it was decided to stop the pilot phase and transition to a BYOD approach.

Figure 2. Timeline of the main phases of the project.



Lessons Learned and Determinants of Success and Failure

As often reported in the field, most challenges were encountered during the implementation (pilot) phase.

Our strategy to use institution-owned tablets added an important workload on the care teams because they were in charge of managing the tablet fleets in their divisions. This strategy was based on a rationale for cybersecurity and development; however, we underestimated the additional work it would generate for already overwhelmed caregivers. With such a strategy, it is our experience that protected and dedicated time for training caregivers is mandatory, at least for superusers. It is well recognized that the quality of training represents a key success factor for the implementation of electronic medical records (EMRs) [19]. Our impression is that this also applies to our project. Accordingly, our two first reported determinants of success are (1) having an implementation strategy that minimizes the impact on already overworked health care professionals and (2) including quality training time protected from the daily routine.

Similarly, we noticed that implementation was easier in care units in which the superuser was both convinced of the project's benefits and was an influential figure among their peers. Accordingly, our third observed determinant of success is that the presence of a "killer function," which on its own brings tremendous value, would have increased adoption by stakeholders. Even though such a function was not identified during patient focus groups, it was revealed during the implementation as the possibility for the patient to choose and order their own meals. Indeed, this function had the potential to both empower the patient and free up time for the caregivers. We consider that having such a functionality will be particularly relevant before the full-scale implementation.

The communication with the project's stakeholders was considered to be a key factor to maintain motivation and trust in the project. In particular, reactivity in fixing identified bugs or transparency about delays was appreciated.

Finally, we realized that the quality of the HIS medical information fueling Concerto was not always appropriate for display in a patient mobile app. This was either because the information was not timely or was incorrect in some cases, but most importantly because its label was too technical. This issue was associated with disadvantages such as a lack of confidence in the project as well as advantages such as a welcome transparency about HIS data, triggering continuous improvements. For example, specific agenda labels designed for patients were created in the HIS owing to the implementation of Concerto.

Discussion

This implementation report presents a real-world example of designing, developing, and implementing a patient-empowering mobile app in an in-patient setting of a Swiss public university hospital. The lessons learned, as presented in the Implementation (Results) section, are summarized in Table 1.

As described in the Introduction section, patient empowerment is a metaconcept. Hence, it is difficult to monitor with a single indicator. For this reason, a key success indicator was not defined at the beginning of this project, which has complicated its evaluation. This represents a limitation of this report, as an objective metric would have been important for complete evaluation. Simple monitoring metrics (eg, eligibility, number of users, and dropout and acceptance ratios) should have also

been collected and are planned for the next app version. A randomized controlled trial assessing the effectiveness of the Concerto mobile app on a patient situation awareness score has been designed and should be conducted in the near future. This trial will allow for better evaluation of the cost-effectiveness of such a project. Overall, data on the effectiveness of eHealth projects are often lacking, and the creation of a dedicated "Implementation Report" article type in *JMIR Medical Informatics* is helping to fill this gap.

The generalizability of our study is another limitation. Indeed, the innovation ecosystem and the EMR landscape at the HUG are very specific and different constraints may be experienced in other settings. However, we believe the reported lessons learned remain relevant in various environments.

In response to one of the main lessons learned with the pilot implementation of Concerto, a BYOD version of the app was developed. With this version, every patient was able to use the app on their personal devices, including computers, tablets, or smartphones. This decision was made to limit the workload on caregivers and improve the adoption rate. New functionalities such as the possibility for patients to choose their meal were also developed to answer unmet needs for both end users and stakeholders impacted by implementation of the app (ie, caregivers). Important challenges in terms of cybersecurity, interoperability, and compatibility had to be met with development of the BYOD version. These will be further described in a forthcoming implementation report focusing on this project phase.

Table 1. Main lessons learned and associated perceived relevance.

Lessons learned	Perceived relevance[a]
Minimize the workload of caregivers or, if possible, decrease it	5/5
Plan protected time for training end users	4/5
Select a convinced and influential superuser	3/5
Wait for a killer function before implementing the app	5/5
Maintain trust through reactivity and transparent communication	4/5

[a]Based on perceived experience, lessons learned were identified by the authors and their relevance was assessed by consensus using a score ranging from 1 (minimally important) to 5 (maximally important).

Acknowledgments

The Fondation Privée des Hôpitaux Universitaires de Genève was the main sponsor for the development and implementation of Concerto as described in this paper.

Authors' Contributions

DD was the project manager for Concerto during the project phases described in this report and wrote the manuscript. HBdV has been the project manager for Concerto after the project phases described in this implementation report and reviewed the manuscript. CPF has reviewed the manuscript. QL has been the lead developer of Concerto during the project phases described in this report and reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Bravo P, Edwards A, Barr PJ, Scholl I, Elwyn G, McAllister M, Cochrane Healthcare Quality Research Group, Cardiff University. Conceptualising patient empowerment: a mixed methods study. BMC Health Serv Res 2015 Jul 01;15:252 [FREE Full text] [doi: 10.1186/s12913-015-0907-z] [Medline: 26126998]
2. Wong CKH, Wong WCW, Lam CLK, Wan YF, Wong WHT, Chung KL, et al. Effects of patient empowerment programme (PEP) on clinical outcomes and health service utilization in type 2 diabetes mellitus in primary care: an observational matched cohort study. PLoS One 2014 May 1;9(5):e95328 [FREE Full text] [doi: 10.1371/journal.pone.0095328] [Medline: 24788804]
3. Lian J, McGhee SM, So C, Chau J, Wong CKH, Wong WCW, et al. Five-year cost-effectiveness of the Patient Empowerment Programme (PEP) for type 2 diabetes mellitus in primary care. Diabetes Obes Metab 2017 Sep 05;19(9):1312-1316. [doi: 10.1111/dom.12919] [Medline: 28230312]
4. Mogueo A, Oga-Omenka C, Hatem M, Kuate Defo B. Effectiveness of interventions based on patient empowerment in the control of type 2 diabetes in sub-Saharan Africa: A review of randomized controlled trials. Endocrinol Diabetes Metab 2021 Jan 25;4(1):e00174 [FREE Full text] [doi: 10.1002/edm2.174] [Medline: 33532614]

5. Baldoni NR, Aquino JA, Sanches-Giraud C, Di Lorenzo Oliveira C, de Figueiredo RC, Cardoso CS, et al. Collective empowerment strategies for patients with diabetes mellitus: a systematic review and meta-analysis. Prim Care Diabetes 2017 Apr;11(2):201-211. [doi: 10.1016/j.pcd.2016.09.006] [Medline: 27780683]
6. Shnaigat M, Downie S, Hosseinzadeh H. Effectiveness of patient activation interventions on chronic obstructive pulmonary disease self-management outcomes: a systematic review. Aust J Rural Health 2022 Feb 16;30(1):8-21. [doi: 10.1111/ajr.12828] [Medline: 35034409]
7. Aquino JA, Baldoni NR, Flôr CR, Sanches C, Di Lorenzo Oliveira C, Alves GCS, et al. Effectiveness of individual strategies for the empowerment of patients with diabetes mellitus: a systematic review with meta-analysis. Prim Care Diabetes 2018 Apr;12(2):97-110. [doi: 10.1016/j.pcd.2017.10.004] [Medline: 29162491]
8. Santoso MV, Kerr RB, Hoddinott J, Garigipati P, Olmos S, Young SL. Role of women's empowerment in child nutrition outcomes: a systematic review. Adv Nutr 2019 Nov 01;10(6):1138-1151 [FREE Full text] [doi: 10.1093/advances/nmz056] [Medline: 31298299]
9. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021 Feb 09;23(2):e19910 [FREE Full text] [doi: 10.2196/19910] [Medline: 33560233]
10. Sosa A, Heineman N, Thomas K, Tang K, Feinstein M, Martin MY, et al. Improving patient health engagement with mobile texting: a pilot study in the head and neck postoperative setting. Head Neck 2017 May 06;39(5):988-995 [FREE Full text] [doi: 10.1002/hed.24718] [Medline: 28263468]
11. Ammenwerth E, Hoerbst A, Lannig S, Mueller G, Siebert U, Schnell-Inderst P. Effects of adult patient portals on patient empowerment and health-related outcomes: a systematic review. Stud Health Technol Inform 2019 Aug 21;264:1106-1110. [doi: 10.3233/SHTI190397] [Medline: 31438096]
12. Vitger T, Korsbek L, Austin SF, Petersen L, Nordentoft M, Hjorthøj C. Digital shared decision-making interventions in mental healthcare: a systematic review and meta-analysis. Front Psychiatry 2021 Sep 6;12:691251 [FREE Full text] [doi: 10.3389/fpsyt.2021.691251] [Medline: 34552514]
13. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
14. Thomas TH, Go K, Go K, McKinley NJ, Dougherty KR, You K, et al. Empowerment through technology: a systematic evaluation of the content and quality of mobile applications to empower individuals with cancer. Int J Med Inform 2022 Jul;163:104782 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104782] [Medline: 35525126]
15. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
16. Dietrich D, Vorgeat H, Mazouri S, Ligier Q, Geissbuhler A. Engager les patients dans leurs soins à travers une application mobile: le projet Concerto. Rev Med Suisse 2018;617:1543-1547. [doi: 10.53738/revmed.2018.14.617.1543]
17. Federal Act on Data Protection. Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2022/491/en [accessed 2024-03-18]
18. CCER - obtain authorization for medical research on humans. Geneva Canton Ethics Board. URL: https://www.ge.ch/ccer-obtenir-autorisation-recherche-medicale-etre-humain [accessed 2024-03-18]
19. Pantaleoni J, Stevens L, Mailes E, Goad B, Longhurst C. Successful physician training program for large scale EMR implementation. Appl Clin Inform 2015 Dec 19;6(1):80-95 [FREE Full text] [doi: 10.4338/ACI-2014-09-CR-0076] [Medline: 25848415]

Abbreviations

BYOD: bring your own device
EMR: electronic medical record
HbA_{1c}: glycated hemoglobin
HIS: health information system
HUG: Hôpitaux Universitaires de Genève (University of Geneva Hospitals)
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
LDL-C: low-density lipoprotein cholesterol
PEP: patient empowerment program

Edited by A Benis; submitted 05.04.23; peer-reviewed by S Delaigue, KM Kuo; comments to author 10.06.23; revised version received 31.07.23; accepted 06.09.23; published 28.03.24.

Please cite as:
Dietrich D, Bornet dit Vorgeat H, Perrin Franck C, Ligier Q
A Mobile App (Concerto) to Empower Hospitalized Patients in a Swiss University Hospital: Development, Design, and Implementation Report
JMIR Med Inform 2024;12:e47914
URL: https://medinform.jmir.org/2024/1/e47914
doi:10.2196/47914
PMID:38546728

Implementation Report

Implementation of the Observational Medical Outcomes Partnership Model in Electronic Medical Record Systems: Evaluation Study Using Factor Analysis and Decision-Making Trial and Evaluation Laboratory-Best-Worst Methods

Ming Luo[1], BS; Yu Gu[1], BS; Feilong Zhou[2], BS; Shaohong Chen[2], BS

[1]Meizhou People's Hospital, Meizhou, China

[2]Shenzhen Luohu District People's Hospital, Shenzhen, China

Corresponding Author:
Shaohong Chen, BS
Shenzhen Luohu District People's Hospital
Number 47, Youyi Road
Luohu District
Shenzhen, 518000
China
Phone: 86 13631629007
Email: shaohong2023@163.com

Abstract

Background: Electronic medical record (EMR) systems are essential in health care for collecting and storing patient medical data. They provide critical information to doctors and caregivers, facilitating improved decision-making and patient care. Despite their significance, optimizing EMR systems is crucial for enhancing health care quality. Implementing the Observational Medical Outcomes Partnership (OMOP) shared data model represents a promising approach to improve EMR performance and overall health care outcomes.

Objective: This study aims to evaluate the effects of implementing the OMOP shared data model in EMR systems and to assess its impact on enhancing health care quality.

Methods: In this study, 3 distinct methodologies are used to explore various aspects of health care information systems. First, factor analysis is utilized to investigate the correlations between EMR systems and attitudes toward OMOP. Second, the best-worst method (BWM) is applied to determine the weights of criteria and subcriteria. Lastly, the decision-making trial and evaluation laboratory technique is used to illustrate the interactions and interdependencies among the identified criteria.

Results: In this research, we evaluated the AliHealth EMR system by surveying 98 users and practitioners to assess its effectiveness and user satisfaction. The study reveals that among all components, “EMR resolution” holds the highest importance with a weight of 0.31007783, highlighting its significant role in the evaluation. Conversely, “EMR ease of use” has the lowest weight of 0.1860467, indicating that stakeholders prioritize the resolution aspect over ease of use in their assessment of EMR systems.

Conclusions: The findings highlight that stakeholders prioritize certain aspects of EMR systems, with “EMR resolution” being the most valued component.

(JMIR Med Inform 2024;12:e58498) doi:10.2196/58498

KEYWORDS

electronic medical record; Technology Acceptance Model; external factors; perception; attitude; behavioral inclination; OMOP

Introduction

Background

Today, electronic files are a standard tool for enhancing the efficiency and effectiveness of care services, thanks to their speed, accuracy, intelligent systems, reminders, and decision support. One of the most commonly used systems is the electronic medical record (EMR), a computerized system used by care providers, including hospitals and doctors’ offices [1,2]. It is designed for entering, storing, displaying, retrieving, and printing patients’ medical records. The system offers several

benefits, such as enhancing the quality of care provided to patients, better organizing information, and improving the timeliness, accuracy, and completeness of documentation [3]. Several factors drive health care providers and organizations to adopt EMR systems now or in the near future [4-6]. These include the ability to create and access patient records electronically, allowing patients to access their own files, preventing medication errors and allergic reactions, reducing medical errors, and providing immediate access to information across different locations. Additionally, EMR systems offer decision-support technology, streamline workflows, and enhance the overall efficiency of clinical processes. They also improve the quality of treatment and facilitate information sharing between general practitioners and specialists. Reducing medical errors and improving clinical data collection are key benefits of EMR systems. Health care providers and organizations increasingly recognize the need to adopt EMR systems to deliver more effective and efficient services [7]. However, the implementation of EMRs faces resistance, particularly from health care personnel, including doctors. To address this challenge, it is essential to create the necessary conditions for the acceptance of this system [8].

The acceptance of an EMR system is based on 2 major factors: perceived usefulness and perceived ease of use [9]. An individual's perception of the usefulness of information technology refers to the belief that using a particular technology will improve job performance or facilitate more efficient task execution within the organization [10]. This support may be reflected in reduced task completion time or the provision of timely information [11,12]. Perceived ease of use reflects the collective belief within an organization that a specific system is straightforward and requires minimal effort to operate [13-16]. In essence, tasks that are perceived as simpler are more likely to be embraced by users. The primary aim of this research is to explore the factors influencing the acceptance and use of the EMR system and the Observational Medical Outcomes Partnership (OMOP) using the Technology Acceptance Model. According to the Davis model [17-20], these factors include external variables (eg, user interface design, data quality, and health information within the EMR system), perceived usefulness, perceived ease of use, attitudes toward the system, and the behavioral intention to use the EMR system. In China, several EMR systems are widely used, driven by significant government investments and initiatives to enhance health care information technology infrastructure. Notable systems include WeDoctor (WeDoctor Holdings Co Ltd), AliHealth (Alibaba Health/Alibaba Group), Ping An Good Doctor (Ping An Healthcare and Technology), Winning Health Technology Group, and Neusoft Corporation. WeDoctor and AliHealth dominate the market by offering comprehensive services with a strong focus on interoperability and data integration. Ping An Good Doctor integrates its EMR system with its online health platform, facilitating remote consultations. Winning Health Technology Group and Neusoft Corporation focus on hospital management and clinical information systems, offering specialized solutions for various health care settings. We herein examined the AliHealth EMR system and surveyed 98 users and practitioners to evaluate its effectiveness and user satisfaction.

In this study, 3 distinct methodologies are used to explore various aspects of health care information systems. First, factor analysis (FA) is used to investigate the correlation between EMR systems and attitudes toward the OMOP. This approach aims to uncover underlying relationships and dependencies between these 2 key elements. Second, the best-worst method (BWM) is applied to determine the weight of criteria and subcriteria, offering a structured and quantitative assessment of their significance. Finally, the decision-making trial and evaluation laboratory (DEMATEL) method is applied to illustrate the interactions and interdependencies among the identified criteria, shedding light on the complex relationships within the health care information system landscape. Together, these methodologies provide a comprehensive understanding of the complex dynamics and factors influencing EMRs and health care information management. This approach differs from earlier studies in several ways. First, it significantly reduces the number of survey questions required to create a systematic causality diagram that decision makers can use. In the context of enhancing health care quality through the implementation of the OMOP shared data model in EMR systems, the integration of the BWM and DEMATEL is crucial for a thorough analysis. The BWM is first used to determine the relative importance of various criteria by comparing the best and worst criteria against all others, generating a weighted set of criteria. These weights are then used as inputs for the DEMATEL method, which maps out the cause-and-effect relationships among the criteria. Specifically, the numerical results from BWM prioritize the importance of factors, and DEMATEL quantifies the degree of influence each factor has over others, thereby creating a structured network of interdependencies. This integration allows for a nuanced understanding by using BWM-derived weights to adjust the influence degrees determined by DEMATEL, resulting in a refined model that better predicts and enhances health care outcomes through the OMOP model. The research contribution is summarized in 3 main stages:

- Determining the correlation between EMR systems and attitudes toward OMOP using FA.
- Establishing the weight of criteria and subcriteria using the BWM.
- Illustrating the interactions and dependencies among the criteria using the DEMATEL method.

The subsequent sections of this document are organized as follows: The "Literature Review" provides a concise summary of the current literature on EMRs, with a specific focus on the OMOP. The "Methods" section details the research methodology used in this study, including a comprehensive description of the chosen strategies and methodologies. The "Results" section presents demographic information and summarizes the conclusions of the study. The "Discussion" section examines the findings of the research. The "Conclusions" section presents the final outcomes of the article.

Literature Review

One of the most discussed aspects of eHealth today is EMRs. EMRs form the foundation of eHealth applications by storing patients' medical histories. They also include legal documents created in both in-house and outpatient settings [21]. The

electronic health record (EHR) system relies on these files for its data. Despite the widespread use of EMR systems in hospitals, many in the medical field still lack confidence in them [22]. Research on EMR systems in hospital settings remains limited. To address the personal, privacy, and security aspects affecting EMR adoption and utilization, Enaizan et al [23] introduced a decision support review framework. This framework is based on a multicriteria approach and K-means clustering, derived from insights gathered from Malaysian health care professionals. Although EHRs represent a significant technological advancement for health care, their adoption has been slow. Liou et al [24] highlighted this issue and proposed a theoretical framework to investigate and improve EHR utilization. Their framework uses the Technology-Organization-Environment model and the DEMATEL approach to create an Influence Network Relationship Map. This map integrates the core concepts of the Analytic Network Process with a modified Vlsekriterijumska Optimizacija i Kompromisno Resenje (VIKOR) method. This approach helps us better understand and utilize EHR technology. Oja et al [10] presented the process of converting EHR, claims, and prescription data into the OMOP format. They detailed the challenges and solutions associated with this conversion process.

EMRs are computerized medical information systems that capture, store, and display patient information. They assist doctors in conducting better, safer, and more efficient work, ultimately improving patient well-being. However, their adoption is still limited globally. Therefore, management information system scholars should investigate the increasing use of EMRs in the health care sector. Despite the potential of EMR systems to reduce administrative costs and medical errors, their adoption rates in physician practices have been slow. To address this, Zaidan et al [25] conducted a comparative study using multicriteria decision-making (MCDM) methods to assess and select open-source EMR software. The medical field is undergoing unprecedented transformation in many countries. Health care organizations are leveraging EMRs to enhance technology utilization, decision-making, and the search for medical solutions. There are employment opportunities for health care workers involved in the transition from paper to electronic records. EMR systems and other health information technologies rely heavily on critical users, particularly doctors. Therefore, the benefits of EMRs cannot be fully realized without user acceptance and approval. According to the literature review, effective criteria and subcriteria for evaluating attitudes toward the use of the OMOP system are listed in Table 1. The questionnaire used to assess attitudes toward the OMOP system included these components. Table 1 presents the criteria and subcriteria relevant to users' attitudes toward the EMR system. The questions were designed using a 5-point Likert scale, ranging from 1=low to 5=high importance, allowing respondents to rate each subcriterion based on their perspective. The criteria include the ease of use, the usefulness of OMOP, the EMR system itself, and the quality of care, each with associated subcriteria reflecting the system's usability, user-friendliness, and overall effectiveness.

Table 1. Effective criteria and subcriteria on the attitude of using OMOP[a].

Effective criteria and subcriteria on the attitude of using OMOP	Code	Mean
Ease of use [12,21,22,26]		
EMR[b] resolution	E1	4.61
EMR ease of use	E2	8.65
Easy to remember EMR	E3	9.66
User-friendliness of EMR	E4	6.64
Getting started with EMR is easy	E5	5.61
The usefulness of OMOP [12,21,22,26]		
EMR screen character resolution	U1	7.61
Appropriateness and consistency of terms used in EMR	U2	5.57
Appropriateness and consistency of the information used	U3	3.53
Ease of learning the operation of the OMOP system	U4	9.57
Features of the OMOP system	U5	5.52
EMR systems [12,21,22,26]		
Using an EMR is a good idea	EM1	3.64
Satisfaction with the use of EMR	EM2	1.65
Does EMR save money?	EM3	6.34
Does EMR save time?	EM4	1.68
The use of EMR is useful for users	EM5	6.35
Quality of care [12,21,22,26]		
OMOP quality	A1	4.62
Usability of OMOP	A2	2.62
Level of satisfaction with OMOP	A3	5.6
The flexibility of the OMOP system	A4	2.59
The power of the OMOP system	A5	7.59

[a]OMOP: Observational Medical Outcomes Partnership.

[b]EMR: electronic medical record.

Methods

Study Design and Overview

This methodology differs from previous studies in several key aspects. First, it significantly reduces the number of survey questions required to create a systematic causality diagram that decision makers can use. To thoroughly investigate the effectiveness of EMR systems and their relationship with attitudes toward the OMOP, a structured research methodology is essential. The research process is outlined in 3 key stages: The initial step involves establishing the relationship between EMR systems and users' attitudes toward OMOP. FA will be applied to identify the underlying variables that influence users' perceptions and satisfaction with OMOP within EMR systems. This method is crucial as it allows for the reduction of data complexity by identifying latent constructs that represent correlated variables. By understanding these correlations, we can better comprehend how different aspects of the EMR system influence users' attitudes toward OMOP, thereby providing a clearer picture of user satisfaction and areas needing improvement.

The second stage involves determining the importance of various criteria and subcriteria related to EMR systems using the BWM. BWM is an MCDM approach that derives criteria weights efficiently by comparing the best and worst criteria against all others. This step is essential for prioritizing the factors identified in the first stage and understanding their relative importance in shaping user attitudes. By assigning precise weights, BWM quantifies the impact of each criterion, enabling targeted improvements and informed strategic decision-making. The final stage involves analyzing the interactions and dependencies among the criteria using the DEMATEL method. DEMATEL is a powerful tool for visualizing and understanding causal relationships among complex criteria. This method is crucial as it reveals how different criteria influence one another, providing insights into the systemic structure of the EMR system. Understanding these interdependencies is essential for identifying key leverage points and developing strategies to enhance overall system performance and user satisfaction. To

summarize, the assessment framework is divided into 3 main stages:

- Determine the correlation between EMR systems and attitudes toward OMOP using FA.
- Determine the weight and importance of criteria and subcriteria using the BWM.
- Show how the criteria interact or depend on each other using the DEMATEL method.

The integration of BWM and DEMATEL results involves a systematic approach where the numerical outputs from each method inform and enhance the other. First, BWM is used to determine the relative importance (weights) of criteria and subcriteria by identifying the best and worst criteria through pairwise comparisons. These weights are then used as inputs for the DEMATEL method, which analyzes the interdependencies and causal relationships among the criteria. This combined approach allows for a comprehensive understanding of how different criteria influence each other and their overall impact on the system. By using the weighted criteria from BWM as a basis, DEMATEL provides a clearer understanding of how the most and least important criteria influence each other, refining the causal map of the system. The combined results offer a comprehensive view, where the weighted importance of criteria (from BWM) is contextualized within the network of influences and interactions (from DEMATEL). This integrated approach ensures that strategic decisions are both data driven and holistically informed. It allows for prioritizing factors not only based on their individual significance but also considering their dynamic interactions within the system.

DEMATEL is used to identify interrelationships and influences among criteria, revealing cause-effect chains, while BWM focuses on ranking and assigning precise weights to the criteria based on decision makers' preferences for the best and worst options. For example, DEMATEL helps determine which criteria have the most significant influence on others, while BWM quantifies these influences into specific weights. The weights derived from BWM and the influence matrix from DEMATEL are combined using a weight aggregation method. This process involves normalizing the weights from both methods and integrating them to form a consolidated weight for each criterion. This approach ensures that both the hierarchical influence identified by DEMATEL and the precision of BWM are captured. Once the combined weights are obtained, they can be applied to MCDM models. For example, in a selection problem, the integrated model uses the combined weights to evaluate and rank alternatives, ensuring a balanced consideration of interrelationships and priority weights. The combined results must be validated against other decision-making methods and through performance analysis tests to ensure their reliability and applicability in real-world scenarios. This step involves comparing the outcomes of the integrated model with those from using DEMATEL or BWM alone, highlighting improvements in decision accuracy and robustness.

Factor Analysis

FA is a statistical technique used to explore the latent structure of a data set. The primary goal of this study is to identify and examine the underlying factors that influence the observed variables. FA is commonly used in fields such as psychology, sociology, economics, and other social sciences [27].

FA is based on the principle that observed variables are influenced by a smaller set of latent factors. These latent factors are not directly observable but are inferred from patterns of correlations among variables. FA is a statistical technique that helps researchers understand the relationships between observed variables and the underlying factors that drive these relationships [25,28].

The BWM is an MCDM technique used to assess the relative weights or significance of selection criteria. This method involves selecting the best and worst criteria, with the best criteria representing those of greatest significance and the worst criteria representing those of least significance. In this study, BWM is used to determine the local weights for each criterion, as opposed to the analytic hierarchy process. Below, we provide a detailed explanation of the extended BWM calculation process [29-32].

BWM Method

Introducing the Best-Worst Method: A More Efficient Approach to Pairwise Comparisons

This new method requires fewer pairwise comparisons compared with the analytic hierarchy process. In hierarchical analysis, the number of pairwise comparisons is given by the formula $[m \times (m-1)]/2$, where m is the number of criteria or indicators being compared. By contrast, the BWM reduces the number of pairwise comparisons to $(2 \times m) - 3$, significantly decreasing the number of comparisons needed. The steps of this method are described in the following sections.

Step 1: Determining Research Criteria

In the first step, the decision matrix for the research problem is established, followed by the identification of the factors influencing the problem's objective.

Step 2.1: Choosing the Set of Decision Criteria

During this step, it is essential to identify the most significant and least significant criteria from among all the indicators, termed as the best and worst criteria, respectively. Next, comparisons should be made between the best criteria and the other criteria, as well as between the other criteria and the worst criteria, using 2 separate matrices. Responses to these comparisons should be provided on a numerical scale ranging from 1 to 9. Overall, experts or decision makers establish evaluation criteria that align with the decision-making problem $\{c_1, c_2, \ldots, c_n\}$.

Step 2.2: Eliminating the Best and Worst Candidates

Once experts or decision makers have established the q standards in step 1, the best and worst criteria are selected. This step is crucial as it significantly impacts the analysis and outcomes.

Step 2.3: Creating the Best-to-Others Vector

Pairwise comparisons with other criteria should be performed using the best criterion [33]. The best-to-others vector is expressed as follows [34]:

$$A_b=(a_{b1}, a_{b2}, ..., a_{bn})$$ **(1)**

where the value of the best criterion b, which is superior to criterion j, is represented by a_{bj}. By itself, the best criterion pairwise comparison must have a value of 1, or a_{bb}=1.

Step 2.4: Creating the Other-to-Worst Vector

This step involves generating the other-to-worst vector using the worst criterion as a reference for comparisons with the worst [35]. The decision maker evaluates the remaining criteria on a scale of 1-9 relative to the worst criterion. The other-to-worst vector is then formulated, as shown in equation 2, based on these comparisons between the worst criterion and the other criteria.

$$A_w=(a_{1w}, a_{2w}, \ldots, a_{nw})^T$$ **(2)**

where a_{jw} denotes how much the remaining criterion j is more important than the least important criterion w. By itself, the worse criterion pairwise comparison must have a value of 1, a_{ww}=1.

*Step 3: Determining the Ideal Weight for Each Criterion (w*1, w*2, …, w*n)*

During this stage, we construct the nonlinear optimization model of the BWM approach using the following equation [24,29]:

In this model,  .

When there are n criteria in total, the model can only compare pairs within a set of 4n–5 constraints in the solution of equation 6. Ultimately, the weights of all criteria combined are constrained to sum up to 1. In a nonempty collection, by assigning an appropriate value to ξ, a feasible solution space is created [29].

DEMATEL

Understanding and Applying the DEMATEL Technique for Complex Problem Solving

The DEMATEL technique, developed between 1971 and 1976, is designed to address intricate and complex problems. Its purpose is to enhance the understanding of complex issues and interrelated problems, ultimately offering a clear solution through a hierarchical structure [17-20,24,29-44]. The DEMATEL technique involves the following key procedures. Its primary objective is to identify causal relationship patterns among a set of criteria. This method assesses the strength of communication through scoring, explores feedback and its significance, and recognizes relationships that are not easily transferable [24,36].

Step 1: Creating a Direct Relation Matrix

The first step is to create a direct relationship matrix. In this step, the effectiveness of each criterion is evaluated individually. When opinions from multiple people are used, their arithmetic mean is calculated. The dimensions of the evaluation scale are then established to accurately reflect the magnitude of the impact. Semantic operational definitions and values are categorized into a scale from 0 to 4, representing varying degrees of influence. The potential values are low impact (1), medium impact (2), high impact (3), and very high impact (4) [24].

Experts (evaluators) complete this questionnaire (matrix), rating the relative importance of pairs of criteria. The values specified in step 1 are used to generate the direct relationship matrix. A direct relationship matrix $A = [a_{ij}]_{n*n}$ is then created by integrating the responses from the different experts. The initial average matrix, denoted as J, is constructed by calculating the mean scores provided by the respondents. This process determines each element in the matrix a_{ij}. The diagonal elements of the matrix are assigned a value of 0 [24].

Step 2: The Direct Influence Matrix Normalization

To normalize the direct correlation matrix, the formula $N=A/Z$ (ie, equation 4) is used. To calculate Z, the sum of all rows and columns is first found. The largest resulting number is chosen as Z. Then, all the entries of the direct correlation matrix are divided by Z.

Step 3: Computing the T (Matrix of Total Influence)

To calculate the total correlation matrix $T = [t_{ij}]_{n*n}$, an identity matrix I is formed We then subtract the identity matrix from the normalized matrix that  . Finally, we multiply the inverted matrix by the normal matrix. The total communication matrix is calculated from the relationship $T=N\times(I - N) - 1$. In other words

$$T = D + D^2 + \cdots + D^h = D(I - D)^{-1}$$ **(5)**

Step 4: Drawing the Network Relation Map Relationships

Column vectors R and S represent the T's column and row sums as equations 6 and 7 [24]:

where $[S_j]^T_{1*n} = [s_i]n*1$. If r_i represents the sum of matrix T's ith row, then r_i represents the total of factor i's that affects every other factor. The total of the direct and indirect effects that factor i has gotten from all of the other factors is represented by s_i, if s_i represents the column sum from matrix T. Additionally, $(r_i + s_i)$ provides an index of the strength of the effects that are imparted and received, meaning that $(r_i + s_i)$ represents the extent of factor i's overall effect in this system. The other factors are therefore influenced by factor i if $(r_i - s_i)$ is positive; conversely, if $(r_i - s_i)$ is negative, then factor i is generally influenced by the other factors [24,45,46].

Step 5: Creating a Causal Diagram

The summation of elements in each row (D) for every factor represents the impact of that factor on other elements within the system, reflecting the extent of influence exerted by the factor. Similarly, the summation of elements in each column

(R) for each factor indicates the degree to which that factor is influenced by other elements in the system, showing the extent of influence received by the factor.

Consequently, the horizontal vector ($D + R$) represents the overall influence of the specific factor within the system. In simpler terms, a higher $D + R$ value for a factor implies a greater level of interaction that the factor has with other factors in the system.

By contrast, the vertical vector ($D – R$) illustrates the influence of each factor. Generally, if $D – R$ is positive, the factor is considered a causal variable; if it is negative, it is regarded as an effect.

Step 6: Drawing the Cartesian Coordinate Diagram

In conclusion, a Cartesian coordinate system is constructed where the longitudinal axis represents $D + R$ values, and the transverse axis represents $D – R$ values. Each factor's position is defined by a point with coordinates ($D + R$, $D – R$) within this system. This approach results in a graphical diagram that visually represents the relationships and interactions among the factors in the system.

Ethics Approval

This study did not involve human participants or animals, and therefore, no ethics approval was required.

Implementation (Results)

Demographic Information

In the first phase of our study, we performed an FA on data collected from 98 hospital visitors, all of whom are users of the AliHealth EMR system. The demographic details of these participants are presented in Table 2. In the second phase, we conducted an MCDM survey with 10 experienced physicians and specialists, each with over 10 years of experience and advanced degrees. They were knowledgeable about electronic health topics, EMR systems, and OMOP. Our research aimed to investigate the impact of the OMOP on AliHealth users. OMOP is a public-private initiative designed to enhance methods and tools for analyzing health care data, particularly for postmarket surveillance of medical products. Its primary goal is to develop a common data model that standardizes the structure and content of observational health data, facilitating more efficient and reliable analysis across different data sources [47]. This model facilitates the integration of various data sets, enabling researchers to perform large-scale analyses and generate evidence on the safety and effectiveness of medical products. By examining the role of OMOP within the EMR system, we aim to understand how this standardized model enhances data integration and analysis capabilities. Our findings will contribute to ongoing efforts to improve health care outcomes through better data management and analytical tools, ultimately benefiting both health care providers and patients by ensuring safer and more effective medical treatments. Descriptive statistical measures have been used to assess the demographic characteristics of the participants, as illustrated in Table 2.

Table 2. Demographic characteristics of the respondents (N=98).

Characteristics	Values
Gender, n (%)	
Male	69 (70)
Female	29 (30)
Age (years), n (%)	
≤25	25 (26)
26-35	46 (47)
36-45	15 (15)
46-55	9 (9)
≥56	3 (3)
Marital status, n (%)	
Single	35 (36)
Married	63 (64)
Education, n (%)	
High school	17 (17)
Diploma	24 (24)
Bachelor	37 (38)
Masters	17 (17)
PhD	3 (3)
Use the online app, n/N (%)	
Everyday	18/55 (33)
2 or 3 times a week	29/89 (33)
Once a week	21/64 (33)
Once every 2 weeks	9/27 (33)
Once a month or less	21/64 (33)
See a doctor and update the medical record , n/N (%)	
Very little	10/30 (33)
Low	11/34 (32)
Medium	52/57 (91)
Much	14/43 (33)
Very much	12/37 (32)

FA Finding

Data are considered unsuitable for FA if the Kaiser-Meyer-Olkin (KMO) value is below 0.5. When the KMO value ranges between 0.5 and 0.69, FA should be approached with caution. However, if the KMO score is above 0.7, the correlations in the data are deemed suitable for FA. In this research, a KMO score of 0.766 indicates that the data are appropriate for analysis. Figure 1 illustrates the results of the partial least squares analysis, while Figure 2 displays the *t* value.

The path coefficients and their significance are given in Table 3.

Figure 1. Constructing a comprehensive research model using the partial least squares.

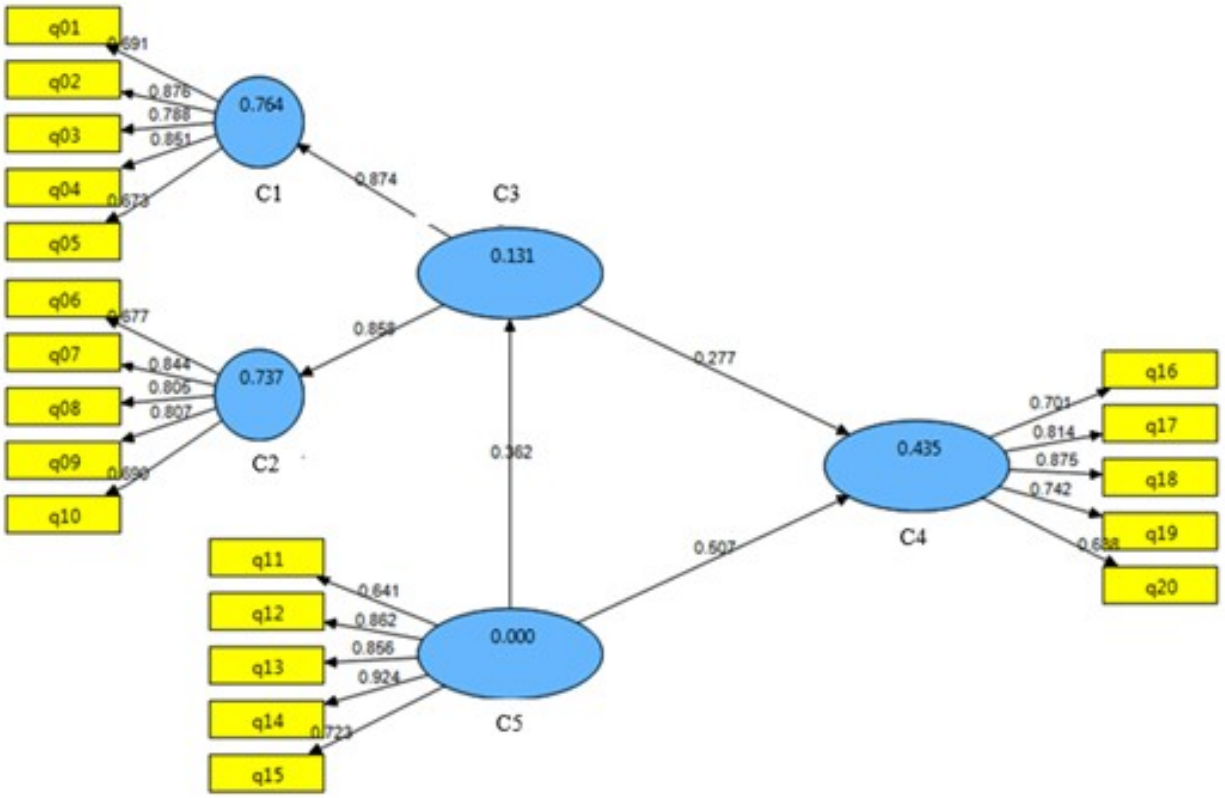


Figure 2. Utilizing the bootstrapping technique to derive T-statistics for the research model.

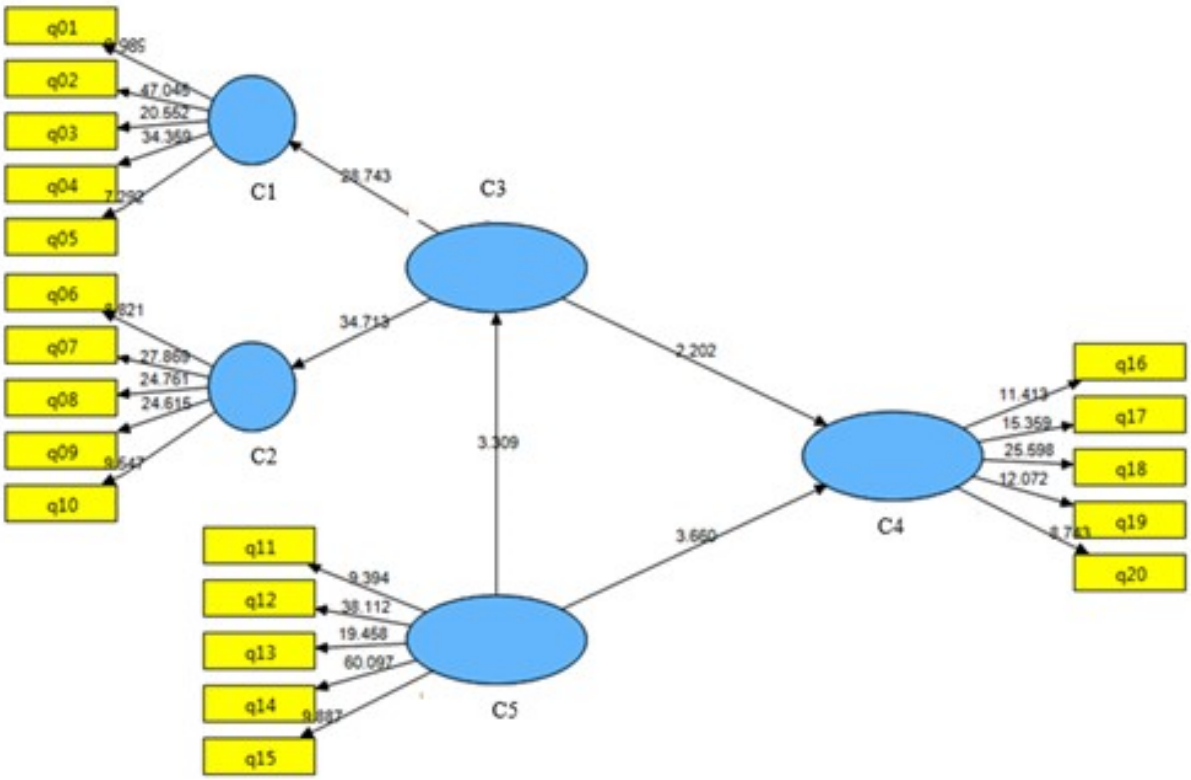


Table 3. Path coefficients.

t Value	Impact rate	Direction of the path
2.202	0.277	The effect of attitude toward the use of OMOP[a] on quality of care
34.713	0.858	The effect of attitude toward the use of OMOP on the usefulness of OMOP
28.743	0.874	The effect of attitude toward the use of OMOP on ease of use of OMOP
3.66	0.507	The effect of EMR[b] systems on quality of care
3.309	0.362	The effect of EMR systems on the attitude toward the use of OMOP

[a]OMOP: Observational Medical Outcomes Partnership.

[b]EMR: electronic medical record.

Table 3 reveals the strength of the relationships identified in the analysis. The correlation between EMR systems and attitudes toward OMOP usage is quantified at 0.362, with a test statistic of 3.309. This statistic exceeds the critical *t* value of 1.96 at the 5% significance level, confirming the statistical significance of the observed relationship. Therefore, with a 95% confidence level, a significant relationship between EMR systems and attitudes toward OMOP usage is established.

Similarly, the analysis reveals a relationship strength of 0.507 between EMR systems and quality of care, supported by a test statistic of 3.660. This result confirms a significant relationship between EMR systems and quality of care with 95% confidence.

Likewise, the correlation between the attitude toward OMOP usage and quality of care is 0.277, with a test statistic of 2.202. This statistic indicates a significant relationship between the attitude toward OMOP usage and quality of care with 95% confidence. Furthermore, the relationship strength between the attitude toward OMOP usage and the usefulness of OMOP is 0.858, with a test statistic of 34.713, which exceeds the critical *t* value at the 5% error level (1.96), confirming a significant relationship with 95% confidence. Therefore, a significant relationship is observed between the attitude toward OMOP usage and the usefulness of OMOP with 95% confidence. Additionally, the correlation between the attitude toward OMOP usage and the ease of use of OMOP is quantified at 0.874, with a test statistic of 28.743. This confirms a significant relationship between the attitude toward OMOP usage and the ease of use of OMOP with 95% confidence.

DEMATEL Findings

As outlined in the paper, the research questionnaire was developed using the DEMATEL technique and then administered to the participants. Table 4 presents the average opinions of the experts regarding the impact of each criterion (rows) on the other criteria (columns).

We use the formula  to normalize the components of Table 5.

Table 4. Average opinion of all experts.

Opinion	C1	C2	C3	C4	C5
C1	1.492	0.72	0.725	0.738	0.648
C2	0.867	1.707	0.96	1	0.779
C3	0.475	0.571	1.47	0.711	0.454
C4	0.693	0.825	0.872	1.723	0.852
C5	0.639	0.712	0.672	0.746	1.493

Table 5 shows the matrix after normalization.

Table 5. Normalized matrix.

Matrix	C1	C2	C3	C4	C5
C1	0.182	0.152	0.182	0.213	0
C2	0.152	0.273	0.273	0	0.303
C3	0.03	0.273	0	0.152	0.091
C4	0.333	0	0.242	0.214	0.121
C5	0	0.182	0.121	0.215	0.182

After computing the matrices mentioned above, the total relations matrix is derived using the following formula: 

In this formula, I is the unity matrix. The calculation results of the T matrix are shown in Table 6.

Table 6. The total relationship matrix (T).

Matrix	C1	C2	C3	C4	C5
C1	0.648	0.738	0.725	0.72	0.492
C2	0.779	1	0.96	0.707	0.867
C3	0.454	0.711	0.47	0.571	0.475
C4	0.852	0.723	0.872	0.825	0.693
C5	0.493	0.746	0.672	0.712	0.639

The next step is to obtain the sum of the rows and columns of the matrix *T*. We obtain the sum of rows and columns according to the following formula: and .

In the provided equation, *D* and *R* represent matrices of dimensions *n*×1 and 1×*n*, respectively. The next phase is evaluating the importance of indicators ($D_i + R_i$) and the correlation between criteria ($D_i - R_i$). If the difference between D_i and R_i is more than 0, the corresponding criterion is regarded as effective; conversely, if the difference between D_i and R_i is less than zero, the corresponding criterion is judged effective.

Table 7 and Figure 3 show the $D_i + R_i$ and $D_i - R_i$ and *XY* plots for importance, with their influence shown in Figure 4.

Table 7. Obtaining the importance and influence of criteria.

Matrix	Criteria	*D*	$D_i + R_i$	$D_i - R_i$
C1	The usefulness of OMOP[a]	3.324	6.49	0.157
C2	Ease of use of OMOP	4.314	7.849	0.778
C3	Attitude toward the use of OMOP	2.681	6.38	-1.018
C4	Quality of care	3.964	7.883	0.046
C5	Electronic medical record systems	3.262	6.489	0.036

[a]OMOP: Observational Medical Outcomes Partnership.

Figure 3. XY plot for D+R and D-R.

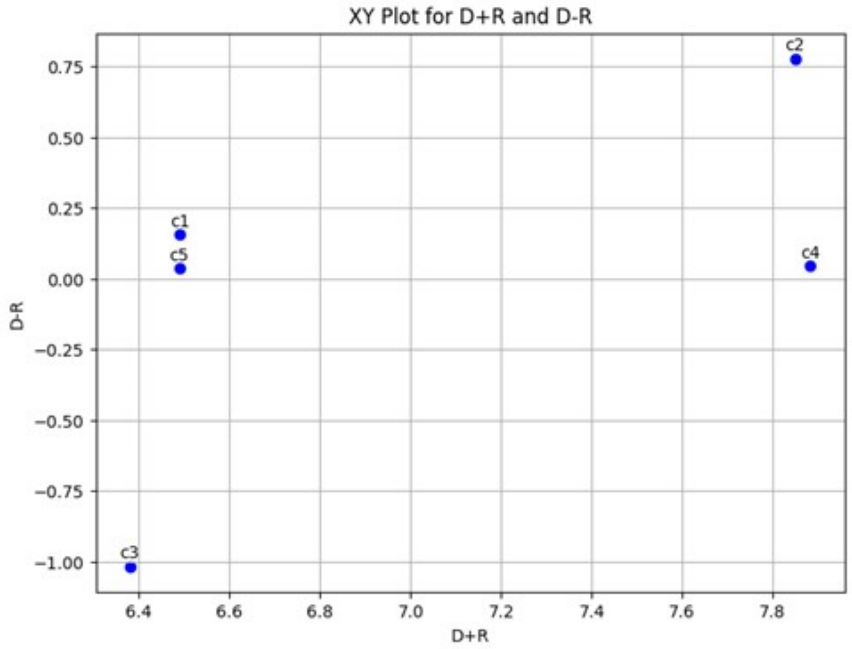
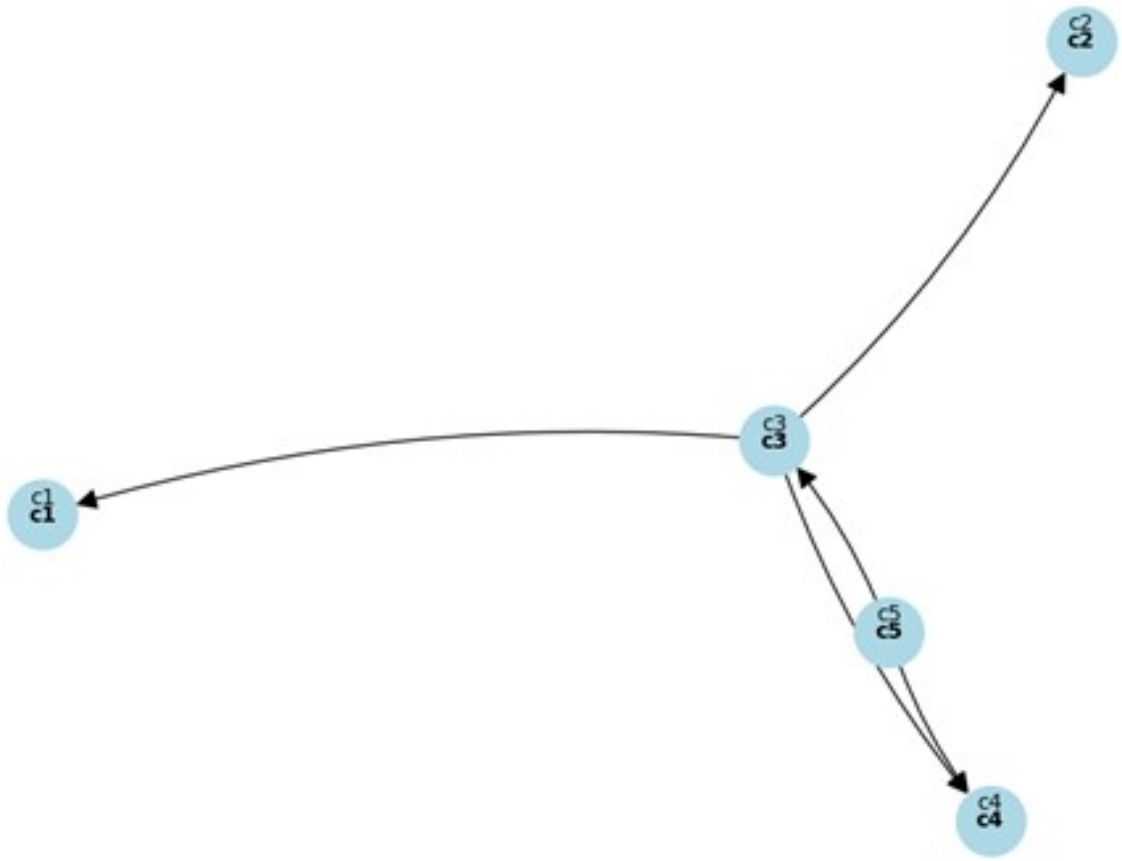


Figure 4. XY plot for importance and influence.



BWM Finding

The BWM is an effective MCDM technique used to determine the weights of criteria. The process begins by selecting the best and worst criteria based on expert judgment or specific research objectives. In this study, "The usefulness of OMOP (C1)" is identified as the best criterion, and "Electronic Medical Record Systems (C5)" is identified as the worst criterion. Following this, pairwise comparisons are performed between these criteria and all other criteria. The next step involves constructing a pairwise comparison matrix where comparisons are made between the best criterion (C1) and the other criteria, as well as between the worst criterion (C5) and the remaining criteria. Experts assign values based on how much better or worse one criterion is compared with another. An optimization model is then developed to minimize the maximum absolute differences between the derived weights and the optimal consistency ratio, ultimately determining the weights of the criteria. The weights assigned to the criteria in this analysis are as follows: "Usefulness of OMOP (C1)" has a weight of 0.3971, "Ease of use of OMOP (C2)" has a weight of 0.1985, "Attitude toward the use of OMOP (C3)" has a weight of 0.1489, "Quality of care (C4)" has a weight of 0.1538, and "Electronic medical record systems (C5)" has a weight of 0.1017. These weights reflect the relative importance of each criterion in the context of the research. The detailed results of the BWM analysis are presented in Table 8.

Table 8 shows the weights assigned to various factors related to the OMOP framework. The weights are as follows: W1 for "Usefulness of OMOP," W2 for "Ease of use of OMOP," W3 for "Attitude toward the use of OMOP," W4 for "Quality of care," and W5 for "Electronic medical record systems." Analysis of these weights reveals that W1 (Usefulness of OMOP) has the highest weight of 0.3971, indicating that this factor is considered the most important in the evaluation. By contrast, W5 (Electronic medical record systems) has the lowest weight of 0.1017, suggesting that it is considered less critical in this context. The weights provide a quantitative measure of the relative importance of each factor, with a higher weight indicating greater significance in the overall assessment of the OMOP framework. The process begins by selecting the best and worst criteria based on their relative importance, as reflected in their weights. After identifying these criteria, pairwise comparisons are conducted between the best and worst criteria and all other criteria. Following the pairwise comparisons, a linear programming model is formulated and solved to minimize the maximum deviation between the derived and optimal consistency ratios. Expert evaluations were aggregated using the simple weighted average method. The weights determined by this process are presented in Table 8.

Table 8. BWM_a result for subcriteria weight.

Criteria and subcriteria	Values
The usefulness of OMOP[b]	0.39708431
U1	0.31007783
U2	0.1860467
U3	0.23255837
U4	0.15503891
U5	0.11627919
Attitude toward the use of OMOP	0.14890674
Ease of use of OMOP	0.19854216
E1	0.19460294
E2	0.19041054
E3	0.12694036
E4	0.10722507
E5	0.38082109
Quality of care	0.15375285
A1	0.38793627
A2	0.19707721
A3	0.12694036
A4	0.10722507
A5	0.18082109
Electronic medical record systems	0.10171394
EM1	0.29528318
EM2	0.07827498
EM3	0.17854757
EM4	0.24996659
EM5	0.19792768

[a]BWM: best-worst method.

[b]OMOP: Observational Medical Outcomes Partnership.

These weights are shown in Multimedia Appendices 1 and 2 (the subcriteria weight is per Multimedia Appendix 2).

Table 8, along with Multimedia Appendices 1 and 2, assigns weights to various components related to the usefulness of OMOP. Among these components, “EMR resolution” has the highest weight, indicating its significant role in the evaluation. Conversely, “EMR ease of use” has the lowest weight, suggesting it is considered less critical in this context. The specific weight values are as follows: “EMR resolution” is assigned a weight of 0.31007783, while “EMR ease of use” has a weight of 0.1860467. This suggests that, in the assessment of EMR systems, stakeholders place the highest importance on the resolution aspect, while comparatively less weight is given to the ease of use of the EMR system. For the ease of use of OMOP, “EMR screen character resolution” holds the highest weight, highlighting its substantial importance in the assessment. Conversely, “Features of the OMOP system,” “Appropriateness and consistency of the information used,” and “Ease of learning the operation of the OMOP system” share the lowest weight, indicating a relatively lower significance for these aspects. The specific weight values are as follows: “EMR screen character resolution” has a weight of 0.19460294, while “Features of the OMOP system,” “Appropriateness and consistency of terms used in EMR,” and “Ease of learning the operation of the OMOP system” each have a weight of 0.12694036. This indicates that, in evaluating the EMR system within the OMOP framework, stakeholders consider the clarity and quality of screen character resolution to be the most important factor. By contrast, the features, terminology, and ease of learning are viewed as less influential.

For “Quality of care,” the component “Using an EMR is a good idea” has the highest weight, indicating that stakeholders place significant importance on the overall concept of using an EMR system. “Satisfaction with the use of EMR” follows closely, suggesting that user satisfaction is also a key consideration in the evaluation. By contrast, “Does EMR save time?” has the lowest weight, implying that the time-saving aspect is considered less critical in this context. The specific weight values are as

follows: "Using an EMR is a good idea" is assigned a weight of 0.38793627, indicating the highest priority. "Satisfaction with the use of EMR" has a weight of 0.19707721, reflecting its importance in stakeholder evaluations. "The use of EMR is useful for users" is assigned a weight of 0.18082109. By contrast, "Does EMR save money?" has a weight of 0.12694036, and "Does EMR save time?" has a weight of 0.10722507, suggesting that considerations related to time and cost savings are less emphasized compared with the perceived value and user satisfaction associated with EMR.

For "Electronic medical record systems," the component "OMOP quality" has the highest weight, reflecting the significant importance stakeholders place on the overall quality of the OMOP system. "The power of the OMOP system" is the next highest, indicating that the system's capability and strength are also crucial factors. Conversely, "Usability of OMOP," "Level of satisfaction with OMOP," and "The flexibility of the OMOP system" all share lower weights. This suggests that while usability, user satisfaction, and system flexibility are relevant, they are considered less critical compared with overall quality and system power in this context. The specific weight values are as follows: "OMOP quality" has a weight of 0.29528318, indicating a high priority for the overall quality of the OMOP system. "The power of the OMOP system" is assigned a weight of 0.19792768, reflecting its significant importance in the evaluation. By contrast, "Usability of OMOP," "Level of satisfaction with OMOP," and "The flexibility of the OMOP system" each have a weight of 0.07827498. This distribution suggests that stakeholders place greater emphasis on the overall quality and power of the OMOP system, while usability, satisfaction, and flexibility are considered relatively less critical in the evaluation process.

Discussion

Principal Findings

The primary objective of this research was to evaluate key components related to the usefulness, ease of use, quality of care, and EMR systems within the OMOP framework. The study aimed to identify the most critical factors influencing stakeholders' perceptions and priorities in these areas. The results, presented in Table 8 and Multimedia Appendices 1 and 2, highlight that "EMR resolution" and "OMOP quality" received the highest weights, indicating their significant importance in the assessment. Conversely, aspects such as "EMR ease of use" and "flexibility of the OMOP system" were considered less critical. Information technology plays a crucial role in enhancing data quality, which in turn improves patient care and safety. It is important for experts to define and explain the characteristics of health data quality so that system designers can identify appropriate methods to evaluate and improve these aspects during system revisions. This will contribute to better information quality in predictive and embedded systems. The findings of this study, particularly the emphasis on "EMR resolution" and "OMOP quality" as critical factors in evaluating the usefulness, ease of use, and quality of care within the OMOP framework, align well with existing literature. The focus on technical quality and the significance of detailed, high-resolution data for effective EMR use are consistently supported by multiple studies. For instance, the study by Kohane et al [48] on the impact of EMR systems on primary care practices underscores the importance of structural and process-related benefits that depend on high-resolution and quality EMR data to improve clinical outcomes. Additionally, the study by Lánczky and Győrffy [49] highlighted the usability of EMRs among health care professionals across various sectors, emphasizing the critical role of technical quality in facilitating efficient clinical workflows and decision-making processes. Effective EMR implementation in mental health settings depends on usability, acceptance, and alignment with clinical needs and workflows, although long-term outcomes remain unclear. Moreover, our study's conclusion that "EMR ease of use" and "flexibility of the OMOP system" were deemed less critical aligns with broader literature, which often underscores that user satisfaction is more strongly influenced by the quality and resolution of EMR data than by its ease of use. For example, a study [50] assessing the implementation of EMRs in mental health settings found that while ease of use is important, the perceived quality and accuracy of the data captured by the system were more significant in influencing user satisfaction and overall system acceptance.

The results related to the user interface reveal that users rated the clarity and quality of the EMR system screen, the appropriateness and consistency of the terms and information used, and the ease of learning its functions and capabilities as crucial factors. This indicates that, during the design and implementation of the EMR system, considerable attention has been given to external factors influencing system acceptance. When designing a new system, it is important to consider not only its functionality but also its appearance and all aspects that impact user interaction and satisfaction. Previous studies have found that more than 96.6% of doctors and midwives believe the EMR system is easy to operate, learn, and use [38-42]. In our study, the average user score for the ease of use of the system was 64.25. Regarding user attitudes toward EMR usage, polyclinic users exhibited a positive attitude with an average score of 66.84. This score is notably higher than the score (55.75) reported previously, indicating a favorable disposition toward the EMR system among polyclinic users. This positive attitude suggests that, with attention to other factors, there is potential to further enhance user satisfaction with the EMR system. In our study, the average behavioral preference score for the EMR system is 61.39. This score reflects users' preferences for the system, including its attractiveness, excitement, flexibility in adapting to changes, and overall power. These factors are related to users' perceptions of the system and are influenced by external factors associated with its implementation and use.

Conclusions

Based on the average scores for external factors such as data quality, clarity, and the user interface, it is evident that significant attention was given to these aspects during the design of the EMR system to ensure user satisfaction. Most users view the use of OMOP within the EMR system as beneficial, noting that this technology enhances productivity and fosters a positive attitude by improving task performance. This improvement can

be attributed to reducing task completion times and providing timely information. Over half of the EMR system users find it easy to use, believing that learning to operate the system requires minimal mental effort. Most users hold a positive attitude toward the EMR system. Consistent with these positive attitudes, the behavioral tendency to use the EMR system is also favorable. Future research should consider using the proposed integrated model to assess potential obstacles in the EMR system.

Data Availability

Data for this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Best-worst method result for criteria weight.
[PNG File , 50 KB - medinform_v12i1e58498_app1.png]

Multimedia Appendix 2
Best-worst method result for subcriteria weight.
[PNG File , 113 KB - medinform_v12i1e58498_app2.png]

References

1. Miller RH, Sim I. Physicians' use of electronic medical records: barriers and solutions. Health Aff (Millwood) 2004 Mar;23(2):116-126. [doi: 10.1377/hlthaff.23.2.116] [Medline: 15046136]
2. Jung H, Yoo S, Kim S, Heo E, Kim B, Lee H, et al. Patient-level fall risk prediction using the observational medical outcomes partnership's common data model: pilot feasibility study. JMIR Med Inform 2022 Mar 11;10(3):e35104 [FREE Full text] [doi: 10.2196/35104] [Medline: 35275076]
3. Spotnitz M, Ostropolets A, Castano VG, Natarajan K, Waldman GJ, Argenziano M, et al. Patient characteristics and antiseizure medication pathways in newly diagnosed epilepsy: feasibility and pilot results using the common data model in a single-center electronic medical record database. Epilepsy Behav 2022 Apr;129:108630. [doi: 10.1016/j.yebeh.2022.108630] [Medline: 35276502]
4. Kwong M, Gardner HL, Dieterle N, Rentko V. Optimization of electronic medical records for data mining using a common data model. Top Companion Anim Med 2019 Dec;37:100364 [FREE Full text] [doi: 10.1016/j.tcam.2019.100364] [Medline: 31837755]
5. Henke E, Zoch M, Kallfelz M, Ruhnke T, Leutner LA, Spoden M, et al. Assessing the use of German claims data vocabularies for research in the observational medical outcomes partnership common data model: development and evaluation study. JMIR Med Inform 2023 Nov 07;11:e47959 [FREE Full text] [doi: 10.2196/47959] [Medline: 37942786]
6. Williams N. Building the observational medical outcomes partnership's T-MSIS analytic file common data model. Inform Med Unlocked 2023;39:101259 [FREE Full text] [doi: 10.1016/j.imu.2023.101259] [Medline: 37305615]
7. Maier C, Kapsner LA, Mate S, Prokosch H, Kraus S. Patient cohort identification on time series data using the OMOP common data model. Appl Clin Inform 2021 Jan 27;12(1):57-64 [FREE Full text] [doi: 10.1055/s-0040-1721481] [Medline: 33506478]
8. Ji H, Kim S, Yi S, Hwang H, Kim J, Yoo S. Converting clinical document architecture documents to the common data model for incorporating health information exchange data in observational health studies: CDA to CDM. J Biomed Inform 2020 Jul;107:103459. [doi: 10.1016/j.jbi.2020.103459] [Medline: 32470694]
9. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
10. Oja M, Tamm S, Mooses K, Pajusalu M, Talvik HA, Ott A, et al. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) common data model: lessons learned. JAMIA Open 2023 Dec;6(4):ooad100-ooad102 [FREE Full text] [doi: 10.1093/jamiaopen/ooad100] [Medline: 38058679]
11. Raventós B, Fernández-Bertolín S, Aragón M, Voss EA, Blacketer C, Méndez-Boo L, et al. Transforming the information system for research in primary care (SIDIAP) in Catalonia to the OMOP common data model and its use for COVID-19 research. CLEP 2023 Sep;Volume 15:969-986. [doi: 10.2147/clep.s419481]
12. Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart J, et al. Transforming anesthesia data into the observational medical outcomes partnership common data model: development and usability study. J Med Internet Res 2021 Oct 29;23(10):e29259 [FREE Full text] [doi: 10.2196/29259] [Medline: 34714250]

13. Yoon D, Ahn EK, Park MY, Cho SY, Ryan P, Schuemie MJ, et al. Conversion and data quality assessment of electronic health record data at a Korean tertiary teaching hospital to a common data model for distributed network research. Healthc Inform Res 2016 Jan;22(1):54-58 [FREE Full text] [doi: 10.4258/hir.2016.22.1.54] [Medline: 26893951]
14. Papez V, Moinat M, Payralbe S, Asselbergs FW, Lumbers RT, Hemingway H, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure. JAMIA Open 2021 Jul;4(3):ooab001 [FREE Full text] [doi: 10.1093/jamiaopen/ooab001] [Medline: 34514354]
15. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre M, et al. Transforming French electronic health records into the Observational Medical Outcome Partnership's common data model: a feasibility study. Appl Clin Inform 2020 Jan;11(1):13-22 [FREE Full text] [doi: 10.1055/s-0039-3402754] [Medline: 31914471]
16. Lynch KE, Deppen SA, DuVall SL, Viernes B, Cao A, Park D, et al. Incrementally transforming electronic medical records into the observational medical outcomes partnership common data model: a multidimensional quality assurance approach. Appl Clin Inform 2019 Oct 23;10(5):794-803 [FREE Full text] [doi: 10.1055/s-0039-1697598] [Medline: 31645076]
17. Park K, Cho M, Song M, Yoo S, Baek H, Kim S, et al. Exploring the potential of OMOP common data model for process mining in healthcare. PLoS One 2023 Jan 3;18(1):e0279641 [FREE Full text] [doi: 10.1371/journal.pone.0279641] [Medline: 36595527]
18. Haberson A, Rinner C, Schöberl A, Gall W. Feasibility of mapping Austrian health claims data to the OMOP common data model. J Med Syst 2019 Sep 07;43(10):314 [FREE Full text] [doi: 10.1007/s10916-019-1436-9] [Medline: 31494719]
19. Voss E, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564. [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
20. Paris N, Lamer A, Parrot A. Transformation and evaluation of the MIMIC database in the OMOP common data model: development and usability study. JMIR Med Inform 2021 Dec 14;9(12):e30970 [FREE Full text] [doi: 10.2196/30970] [Medline: 34904958]
21. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021 Mar 18;39(3):275-285 [FREE Full text] [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
22. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. Create: cohort retrieval enhanced by analysis of text from electronic health records using OMOP common data model. arXiv. Preprint posted online on January 22, 2019 2019.
23. Enaizan O, Zaidan AA, Alwi NHM, Zaidan BB, Alsalem MA, Albahri OS, et al. Electronic medical record systems: decision support examination framework for individual, security and privacy concerns using multi-perspective analysis. Health Technol 2018 Nov 28;10(3):795-822. [doi: 10.1007/s12553-018-0278-7]
24. Liou J, Lu M, Hu S, Cheng C, Chuang Y. A hybrid MCDM model for improving the electronic health record to better serve client needs. Sustainability 2017 Oct 10;9(10):1819. [doi: 10.3390/su9101819]
25. Zaidan A, Zaidan B, Hussain M, Haiqi A, Mat Kiah M, Abdulnabi M. Multi-criteria analysis for OS-EMR software selection problem: a comparative study. Decision Support Systems 2015 Oct;78:15-27. [doi: 10.1016/j.dss.2015.07.002]
26. Esfahani AA, Ahmadi H, Nilashi M, Alizadeh M, Bashiri A, Abbasi Farajzadeh M, et al. An evaluation model for the implementation of hospital information system in public hospitals using multi-criteria-decision-making (MCDM) approaches. IJET 2017 Dec 04;7(1):1 [FREE Full text] [doi: 10.14419/ijet.v7i1.8404]
27. Ahmadi H, Salahshour Rad M, Nilashi M, Ibrahim O, Almaee A. Ranking the macro-level critical success factors of electronic medical record adoption using fuzzy AHP method. International Journal of Innovation and Scientific Research 2014 Sep;8(1):35-42 [FREE Full text]
28. Ahmadi H, Salahshour Rad M, Nilashi M, Ibrahim O, Almaee A. Ranking the micro level critical factors of electronic medical records adoption using TOPSIS method. Health Informatics 2013 Nov;2(4):19-32 [FREE Full text] [doi: 10.5121/hiij.2013.2402]
29. Alimohammadlou M, Alinejad S. Challenges of blockchain implementation in SMEs' supply chains: an integrated IT2F-BWM and IT2F-DEMATEL method. Electron Commer Res 2023 Apr 18:1-43 [FREE Full text] [doi: 10.1007/s10660-023-09696-3]
30. Kumar G, Bhujel RC, Aggarwal A, Gupta D, Yadav A, Asjad M. Analyzing the barriers for aquaponics adoption using integrated BWM and fuzzy DEMATEL approach in Indian context. Environ Sci Pollut Res Int 2023 Apr 07;30(16):47800-47821. [doi: 10.1007/s11356-023-25561-0] [Medline: 36749509]
31. Yazdi M, Khan F, Abbassi R, Rusli R. Improved DEMATEL methodology for effective safety management decision-making. Safety Science 2020 Jul;127:104705. [doi: 10.1016/j.ssci.2020.104705]
32. Dişkaya F, Şenol EMIR. Evaluation of factors affecting logistics performance in a global crisis environment with DEMATEL and BWM. Akıllı Ulaşım Sistemleri ve Uygulamaları Dergis 2003;6(2):300-325. [doi: 10.51513/jitsa.1261018]
33. Alimohammadlou M, Sharifian S. Industry 4.0 implementation challenges in small- and medium-sized enterprises: an approach integrating interval type-2 fuzzy BWM and DEMATEL. Soft Comput 2022 Nov 07;27(1):169-186. [doi: 10.1007/s00500-022-07569-9]
34. Kumar A, Mangla SK, Luthra S, Ishizaka A. Evaluating the human resource related soft dimensions in green supply chain management implementation. Production Planning & Control 2019 Apr 25;30(9):699-715. [doi: 10.1080/09537287.2018.1555342]

35. Feller D. An evaluation of computational methods to support the clinical management of chronic disease populations. Doctor of Philosophy Thesis. Columbia University. ProQuest. 2020. URL: https://www.proquest.com/openview/baa72b2b1d16bf89ad6ec1161e708bf5/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2020-06-22]
36. Ahmadi H, Darvishi M, Nilashi M, Almaee A, Ibrahim O, Zolghadri AH, et al. Evaluating the critical factors for electronic medical record adoption using fuzzy approaches. International Journal of Innovation and Scientific Research 2014 Sep;9(2):268-284 [FREE Full text]
37. Kwong M, Gardner HL, Dieterle N, Rentko V. Optimization of electronic medical records for data mining using a common data model. Top Companion Anim Med 2019 Dec;37:100364 [FREE Full text] [doi: 10.1016/j.tcam.2019.100364] [Medline: 31837755]
38. Kim J, Kim S, Ryu B, Song W, Lee H, Yoo S. Transforming electronic health record polysomnographic data into the Observational Medical Outcome Partnership's common data model: a pilot feasibility study. Sci Rep 2021 Mar 29;11(1):7013 [FREE Full text] [doi: 10.1038/s41598-021-86564-w] [Medline: 33782494]
39. Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. Implementation of a cohort retrieval system for clinical data repositories using the Observational Medical Outcomes Partnership common data model: proof-of-concept system validation. JMIR Med Inform 2020 Oct 06;8(10):e17376 [FREE Full text] [doi: 10.2196/17376] [Medline: 33021486]
40. Sathappan SMK, Jeon YS, Dang TK, Lim SC, Shao Y, Tai ES, et al. Transformation of electronic health records and questionnaire data to OMOP CDM: a feasibility study using SG_T2DM dataset. Appl Clin Inform 2021 Aug 11;12(4):757-767 [FREE Full text] [doi: 10.1055/s-0041-1732301] [Medline: 34380168]
41. Yu Y, Jiang G, Brandt E, Forsyth T, Dhruva SS, Zhang S, et al. Integrating real-world data to assess cardiac ablation device outcomes in a multicenter study using the OMOP common data model for regulatory decisions: implementation and evaluation. JAMIA Open 2023 Apr;6(1):ooac108 [FREE Full text] [doi: 10.1093/jamiaopen/ooac108] [Medline: 36632328]
42. Glicksberg B, Oskotsky B, Thangaraj PM, Giangreco N, Badgeley MA, Johnson KW, et al. PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. Bioinformatics 2019 Nov 01;35(21):4515-4518 [FREE Full text] [doi: 10.1093/bioinformatics/btz409] [Medline: 31214700]
43. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP common data model: experience from three pulmonary hypertension databases. BMC Med Res Methodol 2021 Nov 02;21(1):238 [FREE Full text] [doi: 10.1186/s12874-021-01434-3] [Medline: 34727871]
44. Makadia R, Ryan PB. Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) common data model. EGEMS (Wash DC) 2014 Nov 11;2(1):1110 [FREE Full text] [doi: 10.13063/2327-9214.1110] [Medline: 25848597]
45. Liu PCY, Lo H, Liou JJH. A combination of DEMATEL and BWM-based ANP methods for exploring the green building rating system in Taiwan. Sustainability 2020 Apr 16;12(8):3216. [doi: 10.3390/su12083216]
46. Bakhshi M, Karkehabadi A, Razavian S. Revolutionizing medical diagnosis with novel teaching-learning-based optimization. 2024 Presented at: 2024 International Conference on Emerging Smart Computing and Informatics (ESCI); March 5-7, 2024; Pune, India p. 1-6 URL: https://www.researchgate.net/publication/379903283_Revolutionizing_Medical_Diagnosis_with_Novel_Teaching-Learning-Based_Optimization [doi: 10.1109/ESCI59607.2024.10497216]
47. Mohammad Hassan H, Mina A, Ghazal R, Seyed Behnam R. Investigation of the dynamic system of providing medical services in the hospital for Covid-19 disease patients. Acad J Health Sci 2022;37(3):29-34 [FREE Full text]
48. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, Consortium For Clinical Characterization Of COVID-19 By EHR (4CE), et al. What every reader should know about studies using electronic health record data but may be afraid to ask. J Med Internet Res 2021 Mar 02;23(3):e22219 [FREE Full text] [doi: 10.2196/22219] [Medline: 33600347]
49. Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. J Med Internet Res 2021 Jul 26;23(7):e27633 [FREE Full text] [doi: 10.2196/27633] [Medline: 34309564]
50. Zurynski Y, Ellis LA, Tong HL, Laranjo L, Clay-Williams R, Testa L, et al. Implementation of electronic medical records in mental health settings: scoping review. JMIR Ment Health 2021 Sep 07;8(9):e30564 [FREE Full text] [doi: 10.2196/30564] [Medline: 34491208]

Abbreviations

BWM: best-worst method
DEMATEL: Decision-Making Trial and Evaluation Laboratory
EHR: electronic health record
EMR: electronic medical record
FA: factor analysis
KMO: Kaiser-Meyer-Olkin
MCDM: multicriteria decision-making

OMOP: Observational Medical Outcomes Partnership
VIKOR: Vlsekriterijumska Optimizacija i Kompromisno Resenje

Edited by J Klann; submitted 17.03.24; peer-reviewed by C Cai, M Gul; comments to author 12.07.24; revised version received 12.08.24; accepted 20.08.24; published 27.09.24.

Please cite as:
Luo M, Gu Y, Zhou F, Chen S
Implementation of the Observational Medical Outcomes Partnership Model in Electronic Medical Record Systems: Evaluation Study Using Factor Analysis and Decision-Making Trial and Evaluation Laboratory-Best-Worst Methods
JMIR Med Inform 2024;12:e58498
URL: https://medinform.jmir.org/2024/1/e58498
doi:10.2196/58498
PMID:

Original Paper

The Effects of Electronic Health Records on Medical Error Reduction: Extension of the DeLone and McLean Information System Success Model

Bester Chimbo[1*], MSc, PhD; Lovemore Motsi[1*], DEd, PhD

Department of Information Systems, University of South Africa, Johannesburg, South Africa

[*]all authors contributed equally

Corresponding Author:
Bester Chimbo, MSc, PhD
Department of Information Systems
University of South Africa
Cnr of Christiaan de Wet Road & Pioneer Avenue Florida
Johannesburg, 1709
South Africa
Phone: 27 82 333 8815
Email: chimbb@unisa.ac.za

Abstract

Background: Medical errors are becoming a major problem for health care providers and those who design health policies. These errors cause patients' illnesses to worsen over time and can make recovery impossible. For the benefit of patients and the welfare of health care providers, a decrease in these errors is required to maintain safe, high-quality patient care.

Objective: This study aimed to improve the ability of health care professionals to diagnose diseases and reduce medical errors.

Methods: Data collection was performed at Dr George Mukhari Academic Hospital using convenience sampling. In total, 300 health care professionals were given a self-administered questionnaire, including doctors, dentists, pharmacists, physiologists, and nurses. To test the study hypotheses, multiple linear regression was used to evaluate empirical data.

Results: In the sample of 300 health care professionals, no significant correlation was found between medical error reduction (MER) and knowledge quality (KQ) (β=.043, *P*=.48). A nonsignificant negative relationship existed between MER and information quality (IQ) (β=–.080, *P*=.19). However, a significant positive relationship was observed between MER and electronic health records (EHR; β=.125, 95% CI 0.005-0.245, *P*=.042).

Conclusions: Increasing patient access to medical records for health care professionals may significantly improve patient health and well-being. The effectiveness of health care organizations' operations can also be increased through better health information systems. To lower medical errors and enhance patient outcomes, policy makers should provide financing and support for EHR adoption as a top priority. Health care administrators should also concentrate on providing staff with the training they need to operate these systems efficiently. Empirical surveys in other public and private hospitals can be used to further test the validated survey instrument.

(JMIR Med Inform 2024;12:e54572) doi:10.2196/54572

KEYWORDS

medication error; patient safety; information system; information systems; electronic health record; service quality

Introduction

Background

Worldwide, the delivery of health care has been altered and improved through health information technology. In health care systems, patient administration and management have been facilitated by health information technology. The electronic health record (EHR) system is frequently cited as a vital piece of health information technology to raise the standard of patient care [1]. In the early 1970s, computerized patient EHR were first used to collect, save, and display patient data [2,3]. The ordering of tests, consultations, electronic prescriptions, decision support systems, digital imaging, telemedicine, and other clinical service units can all be included in EHRs, while preserving patient privacy and confidentiality [2].

According to Tegegne et al [4], both high-income and resource-constrained nations have the implementation of the EHR system on their priority agenda. Implementing an EHR is necessary to improve clinical judgment, patient information security, and privacy [1]. The EHR is thought to have the following potential advantages for the health care system: safety, patient information organization, care coordination, communication, patient history, quick access to medical information, and effectiveness of care [5,6]. Evidence also shows that EHR can improve data quality by storing patient data and performing medical tasks [7]. Furthermore, current studies on health information system (HIS) success are generally limited to exploring the driving factors at the HIS adoption level. However, the adoption of an HIS is not indicative of implementation success, as the value and potential of an HIS can only be realized when it is fully absorbed into the workflow by the organization and its users [8]. By comparing EHR adoption and assimilation, Upadhyay and Hu [9] provided empirical evidence that organizational assimilation over adoption can significantly improve patient treatment efficiency. Thus, this study aimed to explore how EHRs contribute to a decrease in medical errors using the expanded DeLone and McLean (D&M) Information System (IS) Success Model as the theoretical framework.

Thousands of people seek medical attention every day from a small group of doctors working in public health to improve the health of communities. Patient health care services must be provided more quickly to cater for the health needs for these communities. There are many problems with the public health care system, including long patient wait times, poor health care service, and inadequate infrastructure. Communities report that services provided by facilities fall short of fundamental standards of care and patient expectations despite government efforts to improve the quality of health care services [10]. In South Africa, more than half of the public health care facilities keep their records on paper, even though the country has adopted several EHR systems [11]. According to Rumball-Smith et al [5], using EHR to manage patient documentation could improve health care services. Despite their importance, medical records are common to be mishandled, resulting in patient files having medical records missing and the inability to get the correct treatment. Misplaced or absent records may have a negative effect on patients' quality of life [12]. It is currently acknowledged that modern productivity, efficiency, and effectiveness are necessary for the facilitation of medical care [13].

Losing a patient's medical records would be the worst-case situation because it could result in more issues, a wrong diagnosis, or, in severe circumstances, the patient's death [14]. In 1 case, where the woman gave birth to twins in the hospital, the Pietermaritzburg High Court ordered a KwaZulu-Natal district hospital to turn over medical information to the plaintiff's attorney in July 2006. In addition, hospital incompetence is said to have resulted in the patient losing 1 of the twins, and the surviving twin developing cerebral palsy. Medical records may occasionally be difficult to locate in the filing department due to a variety of issues, including documents being misfiled or misplaced [15]. These records, which are subsequently discovered in their offices, are occasionally misplaced by medical personnel. Furthermore, although the patient's status would have changed by the time the replacement record is found, the duplicate record would continue to exist.

This research developed a theoretical framework and a survey instrument consisting of questions to evaluate the efficacy of organizational EHR in day-to-day operations from the viewpoint of the nursing staff in residential adult care facilities. System quality, information quality (IQ), service quality (SQ), usability, user happiness, and net benefits were the 6 variables that made up the updated D&M IS Success Model that was part of the recommended study model.

The DeLone & McLean IS Success Model

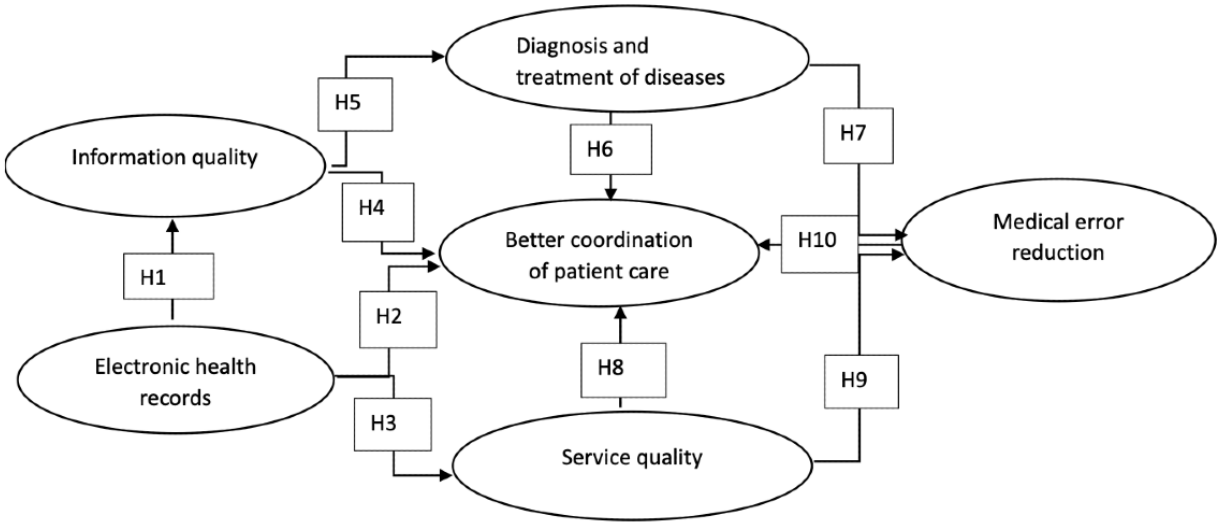
The D&M IS Success Model [16] has been 1 of the most popular measurement models in the IS industry when it comes to evaluating IS success. Numerous topics pertaining to the ongoing use of ISs have been investigated, with a focus on the model of IS success, which is theoretically based on DeLone and McLean's [16] work. Studies [17,18] have indicated that the D&M model is a reasonably developed theoretical model that is frequently used to predict people's behavior in a range of situations. Six interrelated IS success dimensions were identified using the original model. DeLone and McLean [16] proposed several factors that could be used to define IS success: system quality, output IQ, consumption (use) of the output, user response (user satisfaction), impact of the IS on user behavior (individual impact), and impact of the IS on organizational performance (organizational impact). This model indicated the causal and temporal relationships between the 6 characteristics and offered a system for categorizing the wide range of IS success measures. The D&M IS Success Model was updated 10 years later by DeLone and McLean, who combined organizational and individual impacts into a single impact variable known as "net benefit" and included SQ as a new dimension of measuring IS success [19]. The updated model [20] places strong emphasis on the value of gauging the effectiveness of IS variables.

The D&M IS Success Model [21] has been used to assess EHR implementations in numerous studies. It comprises elements of system quality, IQ, intention to use, satisfaction, actual use, and individual and organizational impacts. In the domain of health care, Jeyaraj [22] combined the Technology Acceptance Model (TAM) with the D&M IS Success Model. Consequently, he learned that having enough information available, having a well-designed interface, and having up-to-date information about the system are all crucial factors for an IS to succeed. The system's design is the most crucial element since success is determined by these factors. Noushi and Bedos [23] discovered certain guidelines that need to be followed while creating a new IS that may be used for patient work and diagnostics in a dental clinic setting. Based on the reviewed literature, this study sought to offer a research model to determine the effects of EHR on better coordination of patient care (BCP) in public hospitals. In the context of an HIS in a low- to middle-income country, the study validated the updated D&M IS Success Model. Following is a description of the research model and the proposed research model's hypotheses.

Research Model

The proposed research model is applied only to determine the effects of EHR on BCP in public hospitals. According to the objective of the study, the review of the literature, and the framework or TAM, the research model, as presented in Figure 1, adopted and incorporated the constructs of the D&M IS Success Model. A variable is defined as anything that has a quantity or quality that varies [24]. BCP was the dependent variable in this study's research model, which consisted of 3 independent variables or predictor factors. The independent variables in the study were EHR, IQ, medical error reduction (MER), diagnosis and treatment of diseases (DTD), and SQ. The model constructs are described in Figure 1.

Figure 1. Research model. H: hypothesis.



Hypothesis Development

Based on Figure 1, the 3 independent or predictor factors in the research model were:

- EHR: Digital copies of comprehensive patient records kept by health care professionals, containing information such as patients' medical history, diagnoses, and prescriptions
- IQ: The quality of information contained in the EHR system
- SQ: The quality of service provided by the health care facility, which can be influenced by factors such as the EHR system and the quality of information it contains

Every health care system aims to provide high-quality care [25]. Digital copies of thorough patient records that a health care professional keeps are called EHR or electronic medical records. According to Mitchell [26], EHR provide a comprehensive picture of a patient's medical history, improving care coordination. This history includes information about past medical conditions, prescription medications, allergies, and test results. The goal of an EHR is to guarantee that the patient receives care that is both effective and efficient. EHRs improve efficiency and accuracy, while making patient medical records easier to access and share. Furthermore, it improves treatment quality by allowing clinicians to quickly communicate, analyze data, and identify trends [9]. The following hypotheses were presented to investigate these relationships:

- Hypothesis (H)1: EHR have a significant positive influence on IQ.
- H2: EHR have a positive significant influence on BCP.
- H3: EHR have a positive significant influence on SQ.
- H4: IQ has a significant positive influence on BCP.
- H5: IQ has a significant positive influence on DTD.

The accurate diagnosis provided by this crucial information helps save time and money [12]. Patient data from the EHR system can be coordinated across numerous organizations because they are universally available. It is easier to read data between different EHR systems, since standardization of data according to a common set of standards is encouraged [27]. This major advantage of electronic records probably boosts the effectiveness of all health systems worldwide. Another crucial issue that health care facilities must address is preventive care [9]. More preventive treatment is expected to significantly improve patient outcomes. EHR systems are an excellent tool for helping preventive care initiatives. The following hypotheses were presented to investigate these relationships:

- H6: DTD has a significant positive influence on BCP.
- H7: DTD has a significant positive influence on MER.

Furthermore, EHR systems provide medical practitioners with more data so they can create campaigns for preventive health care [9,28]. EHR systems improve operational effectiveness and reduce error rates, immediately enhancing the standard and safety of patient care. EHR frequently encourage collaboration between organizations and the development of more robust institutions. EHR systems encourage improved hospital-wide communication and collaboration. For health care professionals, the issue of medical errors as an element of SQ has become crucial. According to Fraser et al [29] the SQ technique helps understand patient expectations and facilitates changes in medical practices, increasing patient happiness and compliance, while also improving the quality of the medical services

provided. The following hypotheses were presented to investigate these relationships:

- H8: SQ has a significant positive influence on BCP.
- H9: SQ has a significant positive influence on MER.
- H10: BCP has a significant positive influence on MER.

Methods

Research Design

The Dr George Mukhari Academic Hospital (DGMAH), a tertiary hospital located approximately 30 km north of Tshwane (Pretoria, South Africa), was the site of the study. Students in health sciences from the Sefako Makgatho Health Sciences University (SMU) use the DGMAH as a training hospital. It consists of 39 wards grouped together according to therapeutic specialties. A cross-sectional analytical research design was used for this investigation. Connelly [30] stated that all the data for a cross-sectional study should be gathered at once. Such studies are useful for recording the state of phenomena or the relationships between phenomena at a particular moment in time, since the phenomenon being studied is captured during a single data collection session. Data for the study were gathered at the DGMAH between August 2020 and July 2021. The sample size for this research was 300 medical health care professionals.

In this study, convenience sampling was the approach used in the participant selection process. Convenience samples are used in surveys where respondents are offered the choice to participate or not participate. This sampling is not probabilistic in any way. Probabilistic sampling involves selecting a sample using a probabilistic method without consulting the individuals selected [31]. Convenience sampling, according to Etikan et al [32], is a type of nonrandom sampling in which participants are selected from the target population only if they meet a specific set of pertinent practical requirements. Most often, convenience sampling is used to obtain information from subjects who are easy for the researcher to enroll into the study [32]. There are a number of inherent disadvantages to the convenience sampling technique. With this kind of sampling, biases in the sampling process and systematic errors could arise. In this sense, bias resulting from self-selection and noncoverage taints the convenience samples. Even though noncoverage is avoided and a sampling frame with a random pool of subjects is obtained, if empirical studies use nonprobability convenience samples, the researchers typically are unable to discharge self-selection, because individuals choose whether to complete the survey or participate in the interview at their own discretion. Furthermore, it is not possible to interpret the *P* value in a meaningful way. Alvi [33] further contended that the target population groups should be sufficiently inclusive to be further subdivided into an infinite number of categories that are relatively distinct from one another and therefore not representative of one another.

Teclaw et al [34] found that survey participants occasionally give up before completing the questionnaire. Therefore, it is imperative that the demographic information part of the survey be the first to be completed. Demographics are essential for comparison and descriptive reasons in any study. According to Teclaw et al [34], starting the survey with demographics enhances the item response rate. The goal of this research was to use EHR as a system to determine how IQ, MER, DTD, and SQ, which are independent variables, influence the dependent variable, BCP. Data were collected from the DGMAH, and the constructs were determined using the reviewed literature. Data were collected using a 5-point Likert scale (strongly disagree, disagree, neutral, agree, and strongly Agree). with items made up of demographic and background information, as well as D&M model variables. For the purposes of this study, the questionnaire items were adapted from conventional forms of the TAM, drawing on the relevant literature.

Ethical Considerations

The study was approved by the University of South Africa–College of Agriculture and Environmental Sciences (UNISA-CAES) Health Research Committee (reference number 2019/CAES/075). The DGMAH, Office of the Director of Clinical Services, gave its approval for the study to be carried out at the hospital. Each participant provided signed informed consent, which included details about the researcher, the purpose of the survey, its length, privacy protection protocols, and other information. The study's data were anonymized and deidentified.

Multiple Linear Regression

The rationale behind the selection of multiple linear regression in this study was its ability to evaluate the relationships between a single continuous dependent variable and multiple independent variables. This fit well with the study's goal of determining how different factors affect the reduction in medical errors. For that reason, multiple linear regression was used to investigate the relationship between the dependent variable (MER) and multiple independent variables (EHR, IQ, DTD, BCP, SQ). This method is appropriate when there is 1 continuous dependent variable hypothesized to be associated with 2 or more independent variables.

The specific variables included in the regression model were selected based on the research model and hypotheses developed from the literature review. The model aimed to test how EHR implementation influences MER, both directly and through potential mediating variables, such as IQ, DTD, SQ, and BCP.

Before conducting the regression, necessary assumption checks were performed:

- The Durbin-Watson statistic (2.119) indicated no significant autocorrelation in the residuals.
- The variance inflation factor (VIF) values were all below 5, suggesting no problematic multicollinearity.
- The histogram of standardized residuals followed a reasonably normal distribution.
- The Bartlett sphericity test ($P<.05$) and Kaiser-Meyer-Olkin (KMO) value ($0.727>0.5$) indicated sampling adequacy for factor analysis.

Data Analysis

SPSS Statistics (IBM Corp) was used to perform data analysis. A reliability test on the entire data set resulted in a Cronbach α value of .94, which confirmed the reliability of the data for further analysis. Data analysis and results were divided into 3

sections to cover descriptive analysis, factor analysis, and multiple regression analysis. In addition, quantitative data were analyzed to determine the causal relationship between the independent variables (EHR, IQ, MER, SQ, DTD) and the dependent variable (BCP). Inferential analysis was conducted that involved examining the nature of relationships between the variables under study using the Pearson correlation coefficient. Correlation and regression analyses were used in the inferential analysis. The data analyzed were presented using tables, correlation, regression, and ANOVA.

Results

Demographic Information

Table 1 displays the demographic profile of the survey participants. The sample was made up of 89 (29.7%) males and 211 (70.3%) females. Approximately 34% (n=102) of the respondents were in the 31-40–year age range, while 8.7% (n=26) were younger than 25 years. In addition, 243 (81.4%) were nurses, and 57 (18.6%) were medical professionals. Furthermore, 23 (5.0%) had less than a year's experience, 39 (13.0%) had 2-5 years' experience, 162 (5.4%) had 6-10 years' experience, and 76 (28.0%) had more than 10 years' experience. The structural model was put to the test on 35 items using the Bartlett sphericity test and KMO sample adequacy. Kaiser [35] argued that a KMO value below 0.5 is insufficient. The KMO value for this study was 0.727, indicating that the sample was sufficient and that factor analysis could be carried out. Furthermore, the Bartlett sphericity test was performed, and the result of $P<.05$ indicated that there was a statistically significant association between the variables.

Table 1. Demographic information.

Demographics	Participants (N=300), n (%)
Gender	
Male	89 (29.7)
Female	211 (70.3)
Age (years)	
<25	26 (8.7)
25-30	98 (32.6)
31-40	102 (34.0)
41-50	57 (19.0)
>50	17 (5.7)
Occupation	
Medical doctor	16 (5.0)
Pharmacist	12 (4.0)
Radiologist	10 (3.3)
Physiotherapist	9 (3.0)
Nurse	243 (81.4)
Dentist	10 (3.3)
Work experience (years)	
<1	23 (5.0)
2-5	39 (13.0)
6-10	162 (54.0)
>10	76 (28.0)

Reliability and Validity

To determine the internal consistency and relationship of the items on the scale, a reliability analysis was performed. Cronbach α was used to evaluate the dependability of 47 items. Cronbach α values exceeding 0.5 were considered as being within the acceptable range. The Cronbach α values of 7 variables were over 0.5 based on the indicated value, indicating strong consistency for those items. Items with values less than 0.5 were eliminated [36]. The reliability analysis of the 6 constructs is shown in Table 2.

Table 2. Reliability analysis.

Variable	Items, n	Cronbach α
DTD[a]	6	.783
BCP[b]	5	.852
MER[c]	5	.741
SQ[d]	5	.752
EHR[e]	5	.789
IQ[f]	5	.858

[a]DTD: diagnosis and treatment of diseases.

[b]BCP: better coordination of patient care.

[c]MER: medical error reduction.

[d]SQ: service quality.

[e]EHR: electronic health record.

[f]IQ: information quality.

Multiple Linear Regression

Ten variables were subjected to multiple linear regression to measure the success of the structural model in determining the effects of EHR on the reduction in medical errors in DTD in public hospitals. This resulted in an R^2 change, which showed an increase in variance accounted for by the new interaction term. The R^2 change increased by 0.159, indicating a 15.9% increase in the amount of variation that the extra interaction term could explain. It is important to note that the increase in variation was statistically significant (P<.05), indicating that EHR, SQ, and IQ all significantly have a significant positive influence on MER. Table 3 outlines the results of ANOVA for IQ, DTD, BCP, SQ, and EHR as mediating variables of MER.

The relative importance of each construct was represented by the model's standardized coefficients. The findings showed that there is no statistically significant relationship between MER and knowledge quality (KQ; β=.043, t=0.705, P<.05). MER and IQ had a negative and statistically insignificant relationship according to the predictor variables (β=–.080, t=–1.320, P<.05). However, there was a statistically significant relationship between MER and EHR (β=.125, t=2.043, P<.05). In general, the results showed a strong statistically significant correlation between the dependent variable MER and the predictors IQ, DTD, BCP, SQ, and EHR.

In this study, IQ, SQ, DTD, BCP, and EHR were all analyzed using hierarchical multiple regression. Table 4 represents the outcomes of the moderated regression analysis that are displayed in Figure 2, which shows the histogram of residuals in the regression model for the dependent variable, MER. The residual histogram was found to be reasonably normal and to be close to the normal curve.

Table 3. Summary of the regression model for analysis of success factors for model 1.[a]

Success factor	Value
R	0.399
R^2	0.159
Adjusted R^2 (SE)	0.154 (0.47074)
Change statistics	
R^2 change	0.159
F change	28.055
df1	2
df2	295
Significant F change	0
Durbin-Watson autocorrelation	2.119

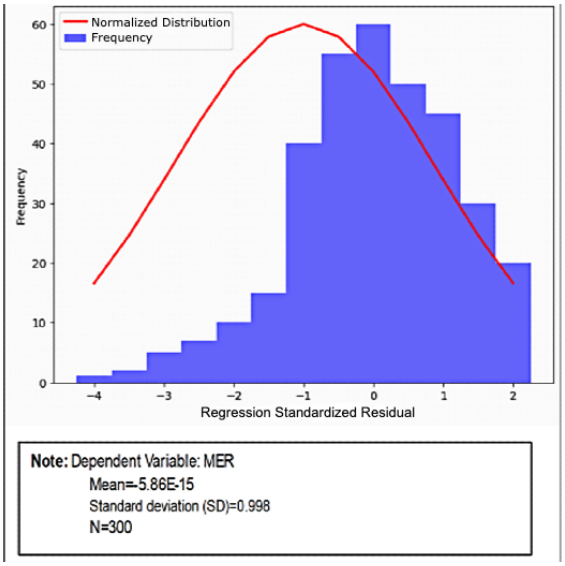
[a]Predictors: (constant), information quality (IQ), electronic health record (EHR), diagnosis and treatment of (DTD), service quality (SQ), and better coordination of patient care and (BCP); dependent variable: medical error reduction (MER).

Table 4. Regression coefficients of variables included in the optimized model.

Variable[a]	Unstandardized coefficient, B (SE)	Standardized coefficient, β	*t* test (*df*)	Significance	95% CI	Collinearity statistics	
						Tolerance	VIF[b]
(Constant)	3.903 (0.482)		8.100	0.000	2.954 to 4.851		
SQ[c]	0.035 (0.050)	.043	0.705 (0.600)	0.481	–0.064 to 0.135	0.976	1.024
IQ[d]	–0.061 (0.046)	–.080	–1.320 (–0.988)	0.188	–0.154 to 0.030	0.986	1.014
EHR[e]	0.136 (0.066)	.125	2.043 (2.890)	0.042	–0.005 to 0.067	0.975	1.026
BCP[f]	0.030 (0.207)	.016	0.144 (2.548)	0.085	–0.077 to 0.133	0.964	1.011
DTD[g]	0.181 (0.149)	.137	2.216 (3.118)	0.025	–0.112 to 0.171	0.971	1.028

[a]Dependent variable: medical error reduction (MER).

Figure 2. Histogram of standardized residuals for MER. MER: medical error reduction.



Hypotheses

Multiple regression testing was performed to examine the effects of EHR on BCP in public hospitals, as well as to evaluate the statistical significance of each hypothesis. Of the 10 hypotheses, 6 (H1, H2, H4, H5, H7, and H10) were statistically significant in determining the effects of EHR on BCP in public hospitals according to the study's findings, with $P<.05$. The results of the hypothesis testing are summarized in Table 5.

Table 5. Summary of the results of hypothesis testing.

Hypothesis (H)		Results	Outcome
H1	EHR[a]→IQ[b]	*P*=.028<.05, β=.354	Accepted
H2	EHR→BCP[c]	*P*=.010<.05, β=–.391	Accepted
H3	EHR→SQ[d]	*P*=.226>.05, β=.109	Rejected
H4	IQ→BCP	*P*=.010<.05, β=–.391	Accepted
H5	IQ→DTD[e]	*P*=.021<.05, β=.329	Accepted
H6	DTD→BCP	*P*=.229>.05, β=–.129	Rejected
H7	DTD→MER[f]	*P*=.002<.05, β=.415	Accepted
H8	SQ→BCP	*P*=.224>.05, β=.100	Rejected
H9	SQ→MER	*P*=.990>.05, β=.001	Rejected
H10	BCP→MER	*P*=.021<.05, β=.329	Accepted

[a]EHR: electronic health record.

[b]IQ: information quality.

[c]BCP: better coordination of patient care.

[d]SQ: service quality.

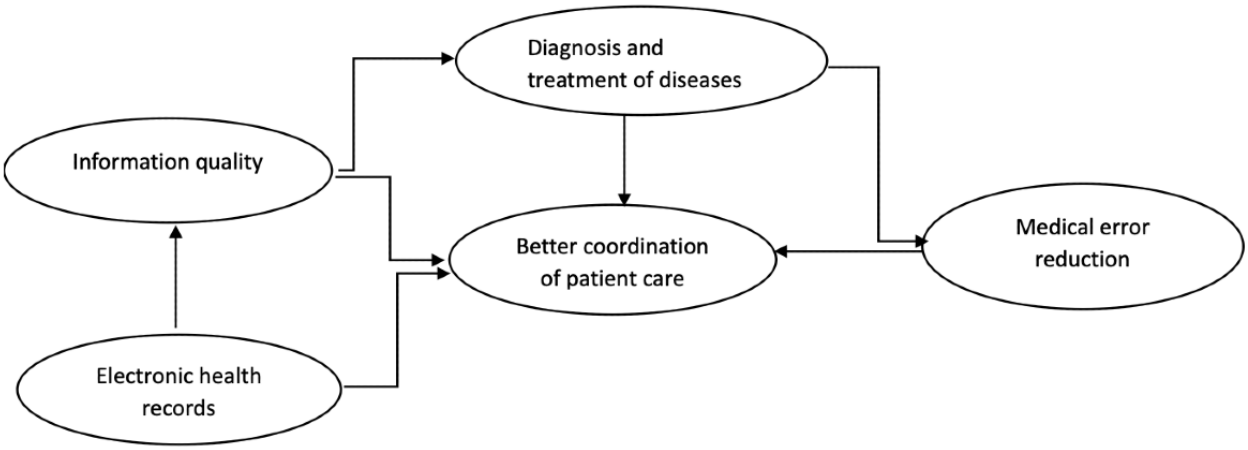
[e]DTD: diagnosis and treatment of diseases.

[f]MER: medical error reduction.

Final Research Model

The final proposed research model was reviewed by removing the rejected hypotheses that were not significant and keeping the variables supported by the accepted hypotheses based on the findings of the hypothesis testing and the significance level. These variables included MER rates, EHR, and IQ. Figure 3 shows the revised recommended model to examine the variables to determine the effects of EHR on BCP in public hospitals.

Figure 3. Revised research model.



Discussion

Principal Findings

Developing a proposed model for reducing medical errors based on the updated D&M IS Success Model as the underpinning theory was the goal of this study. Empirical research was conducted on the suggested model, with medical professionals chosen from the DGMAH as participants. The 5 components of the suggested conceptual model were as follows: MER as the dependent variable, improved BCP, DTD, SQ, and EHR. Based on the validation procedure, the model was updated and changed. Of 10 hypotheses, 6 (H1, H2, H4, H5, H7, and H10) were statistically significant in predicting the effects of EHR on BCP in public hospitals according to the study's results (*P*<.05). The use of EHRs was found to improve patient care coordination and IQ in a statistically significant way. EHR systems are intended to facilitate effective coordination of patient care by practitioners and support evidence-based decision-making [19]. The benefits of implementing EHRs, such as better patient outcomes, more patient safety measures, and lower costs, have also been noted in the literature [19].

The results of this study underline the notion that EHR-based clinical data offer several benefits over traditional medical records. As a result, they greatly improve overall health quality.

They also become readily accessible through a range of communication channels [37]. This ensures that medical personnel treat patients correctly and greatly improves patient outcomes (H1, H2, H4). Additionally, the findings are consistent with those of previous studies [9], which indicated that EHR systems improve patient care quality and increase patient safety, specifically through gains in operational efficiency and a reduction in errors. Fraser et al [29] indicated that the EHR system also tends to promote stronger institutions and cooperation between organizations.

However, other studies have supported the findings of H5 and H7 by demonstrating that EHR systems encourage greater information exchange and interorganizational cooperation [38]. This outcome was in line with most of the earlier research [25]. Preventive treatment is considered crucial to improving patient outcomes and reducing medical expenditures for both individuals and the health system, particularly in regions that are susceptible to certain disease epidemics. The findings aligned with those of Tsai et al [39], who suggested that EHR systems can facilitate the creation of novel and evidence-based treatment objectives, in addition to enhancing data analytics and identifying strategies to improve patient outcomes. The study's conclusions imply that by enhancing the capacity of health care organizations to communicate with patients, particularly those who require preventive care measures, the use of EHR systems can help improve preventive care. EHR in clinics can help medical professionals diagnose patients, treat them, and perform other tasks more accurately [29]. According to research by O'Donnell et al [40], physicians who are older and less tech savvy would likely be against the adoption of new technology. In this study, 5.6% of participants were older than 50 years and 19% were 41-50 years old. Therefore, it may be said that most of the participants were in the age range of 21-40 years. Based on the results, it appears that most participants were eager to use EHRs.

In supporting the findings of H10, Motsi and Chimbo [15] found that their survey results were in line with the conclusions, stating that the primary duty of all health care providers is to provide medical care. A well-known indicator of the effectiveness of hospital health services is patient satisfaction. Patient satisfaction is a critical indicator used to evaluate the quality of health care services rendered [25]. These results also support those of a study by Mohd and Chakravarty [41], who showed that patient-perspective evaluation of health service delivery has gained popularity and is now a fundamental component of all health systems because it serves as a useful gauge of service delivery effectiveness, especially in public health facilities. In this investigation, H3, H6, H8, and H9 were rejected. To further validate the results of the rejected hypotheses, more studies should be conducted in other public hospitals.

The insignificant relationship between SQ and BCP (H8 rejected) could potentially be explained by the fact that SQ encompasses many factors beyond just care coordination. Although high SQ is desirable, it may not directly lead to better coordination if other systemic issues exist. This highlights the need for a multifaceted approach targeting different aspects of health care delivery. The lack of a significant relationship between SQ and MER (H9 rejected) is somewhat surprising, as one would expect higher SQ to correspond with fewer errors. However, it underscores the fact that medical errors can stem from a complex interplay of factors, such as staffing levels, training, and communication protocols, rather than just SQ perceptions. Dedicated interventions focused on error prevention may be needed. The rejected H6 suggests that although accurate DTD is crucial, it alone may not automatically translate into BCP if there are disconnects in the overall continuum of care processes. Improving DTD must go hand in hand with enhancements in teamwork, information sharing, and smooth care transitions.

To effectively leverage the benefits of EHR for improving patient care coordination, a multipronged systems-based approach is recommended. This entails concurrently targeting electronic records, IQ, care coordination processes, accurate diagnosis/treatment, and error reduction through integrated efforts rather than siloed initiatives. Providing comprehensive training and clinical decision support tools can enable health care professionals to optimize EHR use for precise diagnosis, treatment planning, and error prevention. Clear protocols and accountability measures must be implemented to ensure seamless flow of information from EHR across the entire continuum of care, enabling truly coordinated services. Regular assessment of SQ from the patient's perspective through surveys is crucial, and the feedback obtained should drive quality improvement initiatives. Interdisciplinary quality assurance teams should be established to conduct root cause analyses of medical errors and devise preventive strategies that go beyond just SQ aspects. Furthermore, updating health care policies and funding models to incentivize the adoption of integrated EHR systems and prioritize care coordination activities is vital for sustainable progress in this domain.

Limitations

Although the results offer useful information for assessing the impact of EHR on BCP in public hospitals, the generalizability of the findings was hampered by the study's use of a single academic hospital as its study unit. In this study, the effects of organizational factors were not considered. Even though the results are significant, it is imperative that the proposed framework be assessed in light of many theories in further studies, even if it only includes constructs from 1 model. Further research is required to determine their impact on EHRs and on BCP in public hospitals. Organizational culture, managerial support, and implementation readiness are additional variables that should be considered.

The researchers were unable to compare the findings of the study in private hospitals, as the sample only consisted of health care professionals working at a public hospital. A comparative analysis between public and private hospitals may shed light on how and the extent to which organizational and environmental factors influence the implementation of EHR adoption. The fact that a self-report questionnaire was used to gather data may also have limited the accuracy of the responses. Credibility concerns could have been raised by the health care professionals' answers if they had an unclear understanding of EHR. Interviews ought to be used in future studies. The capacity to collect rich, thorough data; elicit and explain participant

responses; customize the interview to the requirements of the research project; build rapport and trust with participants; and be careful when researching sensitive subjects are just a few of the benefits that come with conducting interviews.

Conclusion

The purpose of this study was to determine how EHRs' impact improves patient care coordination in public hospitals. The study proposed a model to determine the factors associated with improved patient care coordination. The study examined data collected from 300 health care professionals at the DGMAH using a cross-sectional analytical research design, and 6 of the 10 hypotheses were found to be supported by the data. The study's findings indicate that EHR are statistically significant in 2 areas: better IQ and BCP. It was found that better DTD, as well as BCP, are significantly impacted by the quality of information. However, it was observed that improved patient care coordination has a positive and considerable influence on reducing medical errors but has no discernible effect on disease diagnosis and treatment. In terms of SQ, it was found that there is no correlation between decreased medical errors and BCP. The government should move more quickly to put policies into effect to increase the eagerness of medical personnel to practice.

In this study the D&M IS Success Model served as the underpinning theory for the development of a framework for MER influenced by the integration of EHR. To address the current issues of health care costs for treatment patients, this framework offers a solution that enables quick access to patient records for more coordinated, efficient care in low- to -middle-income countries, especially in Africa. In addition, this study's findings will assist stakeholders in better understanding the importance behind the integration of the eHealth system with the full implementation of electronic records in South African public hospitals. This understanding will help the department of health and stakeholders to make informed decisions regarding the integration of eHealth with electronic records, which has been implemented at a snail's pace. In addition, by expanding the body of knowledge, the study advances the field of academia in eHealth and health informatics. Furthermore, the study also contributes to the body of knowledge in both the fields of e-HIS governance.

Acknowledgments

We would like to express our deepest gratitude to every single health care professional who filled out the survey and contributed to this research at Dr George Mukhari Academic Hospital.

Conflicts of Interest

None declared.

References

1. Dutta B, Hwang H. The adoption of electronic medical record by physicians: a PRISMA-compliant systematic review. Medicine (Baltimore) 2020 Feb;99(8):e19290 [FREE Full text] [doi: 10.1097/MD.0000000000019290] [Medline: 32080145]
2. Oumer A, Muhye A, Dagne I, Ishak N, Ale A, Bekele A. Utilization, determinants, and prospects of electronic medical records in Ethiopia. Biomed Res Int 2021;2021:2230618 [FREE Full text] [doi: 10.1155/2021/2230618] [Medline: 34790816]
3. Yehualashet DE, Seboka BT, Tesfa GA, Demeke AD, Amede ES. Barriers to the adoption of electronic medical record system in Ethiopia: a systematic review. JMDH 2021 Sep;14:2597-2603. [doi: 10.2147/jmdh.s327539]
4. Tegegne MD, Wubante SM, Kalayou MH, Melaku MS, Tilahun B, Yilma TM, et al. Electronic medical record system use and determinants in Ethiopia: systematic review and meta-analysis. Interact J Med Res 2023 Jan 11;12:e40721 [FREE Full text] [doi: 10.2196/40721] [Medline: 36630161]
5. Rumball-Smith J, Ross K, Bates DW. Late adopters of the electronic health record should move now. BMJ Qual Saf 2020 Mar;29(3):238-240 [FREE Full text] [doi: 10.1136/bmjqs-2019-010002] [Medline: 31732701]
6. Tilahun B, Fritz F. Comprehensive evaluation of electronic medical record system use and user satisfaction at five low-resource setting hospitals in ethiopia. JMIR Med Inform 2015 May 25;3(2):e22 [FREE Full text] [doi: 10.2196/medinform.4106] [Medline: 26007237]
7. Biruk S, Yilma T, Andualem M, Tilahun B. Health professionals' readiness to implement electronic medical record system at three hospitals in Ethiopia: a cross sectional study. BMC Med Inform Decis Mak 2014 Dec 12;14:115 [FREE Full text] [doi: 10.1186/s12911-014-0115-5] [Medline: 25495757]
8. Trout KE, Chen LW, Wilson FA, Tak HJ, Palm D. The impact of meaningful use and electronic health records on hospital patient safety. Int J Environ Res Public Health 2022 Sep 30;19(19):e68053 [FREE Full text] [doi: 10.3390/ijerph191912525] [Medline: 36231824]
9. Upadhyay S, Hu H. A qualitative analysis of the impact of electronic health records (EHR) on healthcare quality and safety: clinicians' lived experiences. Health Serv Insights 2022;15:1-7 [FREE Full text] [doi: 10.1177/11786329211070722] [Medline: 35273449]
10. Maphumulo WT, Bhengu BR. Challenges of quality improvement in the healthcare of South Africa post-apartheid: a critical review. Curations 2019;42(1):a1901. [doi: 10.4102/curationis. v42i1.1901] [Medline: 20044162]

11. Katurura MC, Cilliers L. Electronic health record system in the public health care sector of South Africa: a systematic literature review. Afr J Prim Health Care Fam Med 2018 Nov 20;10(1):e1-e8 [FREE Full text] [doi: 10.4102/phcfm.v10i1.1746] [Medline: 30456963]
12. Marutha NS, Ngoepe M. The role of medical records in the provision of public healthcare services in the Limpopo province of South Africa. S Afr J Inf Manag 2017 Sep 27;19(1):a873. [doi: 10.4102/sajim.v19i1.873]
13. Taiwo Adeleke I, Hakeem Lawal A, Adetona Adio R, Adisa Adebisi A. Information technology skills and training needs of health information management professionals in Nigeria: a nationwide study. Health Inf Manag 2015;44(1):30-38. [doi: 10.1177/183335831504400104] [Medline: 27092467]
14. Kama Z. An evaluation of access to health care: Gugulethu Community Health Clinic. Cape Peninsula University of Technology. 2017. URL: https://etd.cput.ac.za/handle/20.500.11838/2456 [accessed 2024-09-11]
15. Motsi L, Chimbo B. Success factors for evidence-based healthcare practice adoption. S Afr J Inf Manag 2023 May 30;25(1):a1622. [doi: 10.4102/sajim.v25i1.1622]
16. DeLone WH, McLean ER. Information systems success: the quest for the dependent variable. Inf Syst Res 1992 Mar;3(1):60-95. [doi: 10.1287/isre.3.1.60]
17. Lee M, Lee SA, Jeong M, Oh H. Quality of virtual reality and its impacts on behavioral intention. Int J Hosp Manag 2020 Sep;90:102595. [doi: 10.1016/j.ijhm.2020.102595]
18. Mustafa SZ, Kar AK, Janssen M. Understanding the impact of digital service failure on users: integrating Tan's failure and DeLone and McLean's success model. Int J Inf Manag 2020 Aug;53:102119. [doi: 10.1016/j.ijinfomgt.2020.102119]
19. Stefanovic D, Marjanovic U, Delić M, Culibrk D, Lalic B. Assessing the success of e-government systems: an employee perspective. Inf Manag 2016 Sep;53(6):717-726. [doi: 10.1016/j.im.2016.02.007]
20. DeLone WH, McLean ER. The DeLone and McLean model of information systems success: a ten year update. J Manag Inf Syst 2014 Dec 23;19(4):9-30. [doi: 10.1080/07421222.2003.11045748]
21. DeLone WH, McLean ER. Information systems success measurement. FNT in Information Systems 2016;2(1):1-116. [doi: 10.1561/2900000005]
22. Jeyaraj A. Int J Inf Manag 2020 Oct;54:102139. [doi: 10.1016/j.ijinfomgt.2020.102139]
23. Noushi N, Bedos C. Developing person-centred dental care: the perspectives of people living in poverty. Dent J (Basel) 2020 Aug 03;8(3):82 [FREE Full text] [doi: 10.3390/dj8030082] [Medline: 32756307]
24. Freire-González J, Decker C, Hall J. The economic impacts of droughts: a framework for analysis. Ecol Econom 2017 Feb;132:196-204 [FREE Full text] [doi: 10.1016/j.ecolecon.2016.11.005]
25. Kruk ME, Gage AD, Arsenault C, Jordan K, Leslie HH, Roder-DeWan S, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. Lancet Global Health 2018 Nov;6(11):e1196-e1252. [doi: 10.1016/s2214-109x(18)30386-3]
26. Mitchell J. History of electronic health records. Study.com. URL: https://tinyurl.com/2s35nw2k [accessed 2024-09-11]
27. Martin P, Sbaffi L. Electronic health record and problem lists in leeds, united kingdom: variability of general practitioners' views. Health Inform J 2020 Sep;26(3):1898-1911 [FREE Full text] [doi: 10.1177/1460458219895184] [Medline: 31875417]
28. Stellefson M, Paige SR, Chaney BH, Chaney JD. Evolving role of social media in health promotion: updated responsibilities for health education specialists. Int J Environ Res Public Health 2020 Feb 12;17(4):1153 [FREE Full text] [doi: 10.3390/ijerph17041153] [Medline: 32059561]
29. Fraser H, Adedeji TA, Amendola P. ICT in Health: Survey on the Use of Information and Communication Technologies in Brazil Healthcare Facilities. COVID-19 Edition. São Paulo: Brazilian Network Information Center; 2021:273-296.
30. Connelly LM. Cross-sectional survey research. MedSurg Nursing 2016 Sep;25(5):369-370.
31. Fricker RJ. In: Hughes J, editor. Sampling Methods for Web and E-mail Surveys, SAGE Internet Research Methods. London, UK: SAGE Publications; 2012:195-216.
32. Etikan I. Comparison of convenience sampling and purposive sampling. AJTAS 2016;5(1):1-4. [doi: 10.11648/j.ajtas.20160501.11]
33. Alvi M. A manual for selecting sampling techniques in research. Munich Personal RePEc Archive. URL: https://tinyurl.com/5caf54jn [accessed 2024-09-11]
34. Teclaw R, Price MC, Osatuke K. Demographic question placement: effect on item response rates and means of a Veterans Health Administration Survey. J Bus Psychol 2011 Dec 1;27(3):281-290. [doi: 10.1007/s10869-011-9249-y]
35. Kaiser HF. An index of factorial simplicity. Psychometrika 1974 Mar;39(1):31-36. [doi: 10.1007/bf02291575]
36. Saidi S, Siew N. Investigating the validity and reliability of survey attitude towards statistics instrument among rural secondary school students. Int J Educ Methodol 2019 Nov 15;5(4):651-661 [FREE Full text] [doi: 10.12973/ijem.5.4.651]
37. Tertulino R, Antunes N, Morais H. Privacy in electronic health records: a systematic mapping study. J Public Health (Berl.) 2023 Jan 23;32(3):435-454. [doi: 10.1007/s10389-022-01795-z]
38. Kruse CS, Mileski M, Vijaykumar AG, Viswanathan SV, Suskandla U, Chidambaram Y. Impact of electronic health records on long-term care facilities: systematic review. JMIR Med Inform 2017 Sep 29;5(3):e35 [FREE Full text] [doi: 10.2196/medinform.7958] [Medline: 28963091]

39. Tsai CH, Eghdam A, Davoody N, Wright G, Flowerday S, Koch S. Effects of electronic health record implementation and barriers to adoption and use: a scoping review and qualitative analysis of the content. Life (Basel) 2020 Dec 04;10(12):327 [FREE Full text] [doi: 10.3390/life10120327] [Medline: 33291615]
40. O'Donnell A, Kaner E, Shaw C, Haighton C. Primary care physicians' attitudes to the adoption of electronic medical records: a systematic review and evidence synthesis using the clinical adoption framework. BMC Med Inform Decis Mak 2018 Nov 13;18(1):101 [FREE Full text] [doi: 10.1186/s12911-018-0703-x] [Medline: 30424758]
41. Mohd A, Chakravarty A. Patient satisfaction with services of the outpatient department. Med J Armed Forces India 2014 Jul;70(3):237-242 [FREE Full text] [doi: 10.1016/j.mjafi.2013.06.010] [Medline: 25378776]

Abbreviations

BCP: better coordination of patient care
D&M: DeLone and McLean
DGMAH: Dr George Mukhari Academic Hospital
DTD: diagnosis and treatment of diseases
EHR: electronic health record
HIS: health information system
IQ: information quality
IS: information system
KMO: Kaiser-Meyer-Olkin
KQ: knowledge quality
MER: medical error reduction
SQ: service quality
VIF: variance inflation factor

Edited by C Perrin; submitted 15.11.23; peer-reviewed by J Walsh, Y Zhang, GK Gupta; comments to author 17.04.24; revised version received 06.06.24; accepted 23.07.24; published 16.10.24.

Please cite as:
Chimbo B, Motsi L
The Effects of Electronic Health Records on Medical Error Reduction: Extension of the DeLone and McLean Information System Success Model
JMIR Med Inform 2024;12:e54572
URL: https://medinform.jmir.org/2024/1/e54572
doi:10.2196/54572
PMID:39412857

Original Paper

Evaluating the Bias in Hospital Data: Automatic Preprocessing of Patient Pathways Algorithm Development and Validation Study

Laura Uhl[1], MSc, ING; Vincent Augusto[1], Prof Dr; Benjamin Dalmas[1], PhD; Youenn Alexandre[2], PhD; Paolo Bercelli[2], MD; Fanny Jardinaud[3], PhD; Saber Aloui[4], PhD

[1]Mines Saint-Etienne Centre Ingénierie Santé, Unité Mixte de Recherche (UMR) 6158 Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS), Centre national de la recherche scientifique (CNRS), Saint-Etienne, France

[2]Groupe Hospitalier Bretagne Sud, Lorient, France

[3]Direction Anticipation & Usages, Enovacom, Marseille, France

[4]Inserm, UMR 1085 Ester, Centre Hospitalier Universitaire Angers, Angers, France

Corresponding Author:
Laura Uhl, MSc, ING
Mines Saint-Etienne Centre Ingénierie Santé
Unité Mixte de Recherche (UMR) 6158 Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS)
Centre national de la recherche scientifique (CNRS)
158 cours Fauriel
Saint-Etienne, 42000
France
Phone: 33 477420123
Email: l.uhl@emse.fr

Abstract

Background: The optimization of patient care pathways is crucial for hospital managers in the context of a scarcity of medical resources. Assuming unlimited capacities, the pathway of a patient would only be governed by pure medical logic to meet at best the patient's needs. However, logistical limitations (eg, resources such as inpatient beds) are often associated with delayed treatments and may ultimately affect patient pathways. This is especially true for unscheduled patients—when a patient in the emergency department needs to be admitted to another medical unit without disturbing the flow of planned hospitalizations.

Objective: In this study, we proposed a new framework to automatically detect activities in patient pathways that may be unrelated to patients' needs but rather induced by logistical limitations.

Methods: The scientific contribution lies in a method that transforms a database of historical pathways with bias into 2 databases: a labeled pathway database where each activity is labeled as relevant (related to a patient's needs) or irrelevant (induced by logistical limitations) and a corrected pathway database where each activity corresponds to the activity that would occur assuming unlimited resources. The labeling algorithm was assessed through medical expertise. In total, 2 case studies quantified the impact of our method of preprocessing health care data using process mining and discrete event simulation.

Results: Focusing on unscheduled patient pathways, we collected data covering 12 months of activity at the Groupe Hospitalier Bretagne Sud in France. Our algorithm had 87% accuracy and demonstrated its usefulness for preprocessing traces and obtaining a clean database. The 2 case studies showed the importance of our preprocessing step before any analysis. The process graphs of the processed data had, on average, 40% (SD 10%) fewer variants than the raw data. The simulation revealed that 30% of the medical units had >1 bed difference in capacity between the processed and raw data.

Conclusions: Patient pathway data reflect the actual activity of hospitals that is governed by medical requirements and logistical limitations. Before using these data, these limitations should be identified and corrected. We anticipate that our approach can be generalized to obtain unbiased analyses of patient pathways for other hospitals.

(JMIR Med Inform 2024;12:e58978) doi:10.2196/58978

KEYWORDS

preprocessing; framework; health care data; patient pathway; bed management

Introduction

Context

Bed management is a critical task for hospitals to provide coherent care pathways. Daily bed management consists of finding beds for patients coming from the emergency department (ED) in appropriate medical units without canceling planned hospitalizations. Therefore, bed management involves 2 distinct flows: unscheduled flow (life-threatening emergencies and patients coming to the ED) and scheduled flow (planned hospitalizations). Despite the complexity of the task, bed management is most often organized without the help of any decision support tools and involves multiple phone calls to find a bed in a medical unit matching the patient's needs [1]. When medical units are facing high occupation rates, it is not always possible to find a bed to match patient needs.

In these situations, patients are either kept in the short-stay hospitalization unit of the ED or transferred to an overflow medical unit to wait for a bed. Consequently, the medical units visited by a patient do not always correspond to their medical needs. For example, a patient from the ED can be transferred to a surgery unit and then to a cardiology unit. This is the pathway observed in the data. The patient did not receive any surgical treatment. He was admitted to the surgery unit waiting for a cardiology unit bed. Therefore, the location of the patient does not always match the cause of hospitalization. The succession of medical units is called a patient pathway. Unscheduled pathways describe the pathways of patients coming from the ED. In this work, we only considered patients who visited the ED and were subsequently hospitalized.

The study of patient pathways reveals several challenges due to the variety of pathways, the lack of complete guidelines and references, and the heterogeneity of patient management between hospitals (due to equipment and organizational differences). Unscheduled pathways are difficult to explain because management rules or clear indicators are not available to identify them. In addition, the high number of pathway variants makes individual studies of each pathway impossible (eg, >1000 variants for French hospitals of average size) [2]. Process mining is an interesting tool for studying a set of pathways with several variants because a pathway can be seen as a patient care process [2,3]. Nevertheless, a large variance in pathways leads to uninterpretable process graphs. Strategies exist to make a process graph easier to read, such as trace clustering or graph size reduction using filters or aggregation [2], but these methods cannot identify which activities are relevant and which activities are induced by logistical limitations.

In this study, we sought to develop a method to assess observed pathways extracted from a hospital information system. We wanted to identify which medical units matched the cause of hospitalization (relevant) or not (irrelevant) in a patient pathway. The medical relevance or the relevance of treatments was not evaluated, nor was the choice of the bed manager. Only the relevance of the patient's location was evaluated. An *irrelevant* medical unit means that the patient would have been hospitalized in another unit if there were an infinite number of beds. The identification of such irrelevance is important to avoid any misinterpretation of further analysis results. In this paper, we often use the word *bias* to denote a wrong, inaccurate, or incomplete interpretation of a real situation because the data do not represent reality. We use the expression *bias in pathways* or *data bias* to refer to data that represent pathways that do not always correspond to patients' medical needs.

Related Work

We did not find proper literature on the task of assessing pathways but, rather, heterogeneous papers dealing with bias or phases of a pathway. In 1989, Selker et al [4] designed the "Delay Tool," which detects medically unnecessary hospital days. It is based on a taxonomy of delays. Each stay was manually evaluated using patient records with the Delay Tool method. In an article on the prediction of the disposition of ED patients, El-Bouri et al [5] considered the fact that ED patients can be admitted to an inpatient unit inappropriate for their diagnosis. Patients were filtered according to whether their primary diagnosis code for the ED visit clearly corresponded to the admission inpatient unit. Their aim was to avoid learning from biased data. These methods require a thesaurus of all possible diagnoses linked to appropriate wards. To study patient pathways, Franck et al [6] designed a generic framework to model pathways and distinguished 3 different phases: (1) a waiting phase—the patient waits in the ED (unscheduled) or at home (scheduled) to be admitted to the relevant medical unit, (2) an acute phase—the patient receives care in the medical unit, and (3) a rehabilitative phase—rehabilitative care of the patient. They also differentiated scheduled patients from unscheduled patients. To analyze the clinical pathways, they defined relevant pathways for each type of patient by considering only the acute phase and substitution options. To identify the relevant pathways and substitutions, they used process mining on administrative data. This method is very accurate but time-consuming given that a relevant pathway and substitutions must be defined for each pathology. They applied this method exclusively to patients with stroke.

Data quality in health research is a shared problem, and solutions have been proposed to improve several dimensions of quality [7]. However, methods are often not suggested to correct specific bias in health care data due to missing details about a piece of information.

Patient pathways can be seen as processes, with the succession of medical units being the succession of events. Therefore, pathways can be studied using process mining techniques. "The goal of process mining is to use event data to extract process-related information" [8]. The first rough representation of patient pathways using process discovery algorithms provides a spaghettilike process model. Indeed, process discovery algorithms are not successful with event logs that involve numerous variants and many events [2]. A typical solution to untangle a spaghettilike model is to cluster the whole set of traces (trace clustering) and represent each cluster using a process model that should be smaller and more comprehensive. The main challenges of the clustering of patient pathways are the integration of medical knowledge (medical logic) and the evaluation of the resulting clusters. In the literature, several

methods for trace clustering have been proposed. Some of these methods are distance-based clustering algorithms. The core of these methods is to compute distances between traces to apply classic clustering algorithms (trace clustering [9], trace clustering based on conserved patterns [10], context-aware clustering [11], and the method by Delias et al [12]). Others are model based; these methods gradually build a process model that represents a cluster, and a trace is assigned to the cluster with the nearest process model (sequence clustering [13], active trace clustering [14], disjunctive workflow schema [15], graph-based approach and Markov models [16], and behavioral topic analysis [17]). In terms of cluster evaluation, different metrics are used. Some metrics analyze cluster intrahomogeneity, and others analyze the complexity of the process model of each cluster. There is no consensus on these metrics, and they do not guarantee that the clusters computed using the algorithm have an expert logic. Hence, trace clustering does not give us complete satisfaction in characterizing pathways. Another approach for simplifying process models was proposed by Fahland and Van der Aalst [18] based on unfolding.

Some data preparation techniques can also reduce the "spaghettiness" of process models. Data preparation is an unavoidable step in a process mining project and impacts on the resulting process graph, as highlighted by De Roock and Martin [19] in their most recent state-of-the-art study. Several methods have been suggested in the literature to simplify process models. Semantic log purging was proposed by Ly et al [20] in 2012 to clean log data. This method is based on the identification of "fundamental constraints that a process has to obey" thanks to experts. Only a qualitative evaluation and 1 experiment using 1 dataset were performed. Van Zelst et al [21] reviewed the literature on event abstraction in process mining. However, this technique is not related to the problem addressed in this study because our dataset did not provide information on the granularity of events. Several papers address the issue of time stamp inaccuracy.

Martin et al [22] proposed interactive data cleaning. Dixit et al [23] created a method to detect and repair event ordering mistakes. Rogge-Solti et al [24] presented a similar approach to repairing missing events based on alignment. In addition, these researchers created a method for time repairing.

To rigorously prepare the data and event log, different frameworks have been developed. Andrews et al [25] applied the Cross-Industry Standard Process for Data Mining method to identify data quality issues. The data quality dimensions used in data mining are also useful for assessing data quality in process mining. Nevertheless, the researchers do not consider dimensions specific to processes, such as trace coherence. Therefore, the fourth step, namely, prestudy process mining analysis, is important to assess this dimension. Bose et al [26] noted 27 event log quality issues based on 4 categories (missing data, incorrect data, imprecise data, and irrelevant data) and 9 components of an event log (case, event, belongs to, case attributes, event attributes, position, activity name, time stamp, and resource). The researchers also distinguished 4 process characteristics: (1) voluminous data, which refers to a large number of cases or events; (2) case heterogeneity, which refers to a large number of distinct traces; (3) event granularity, which refers to a large number of distinct activities; and (4) process flexibility and concept drifts. The issues caused by case heterogeneity are a part of the problem we attempted to address in this study. Van Eck et al [27] suggested PM^2, a process mining project methodology. Data processing is the third step and consists of creating views (creating the event log), aggregating events, enriching logs (addition of attributes), and filtering logs. Vanbrabant et al [28] presented a data quality framework based on 3 previous frameworks and applied it to a case study—pretreatment of ED data before simulation. These researchers divided quality problems into hierarchical classes. Verhulst [29] defined very precisely the different data quality dimensions for process mining and their scoring methods. All these papers on data preparation are general to process mining datasets and do not answer the question of pathway bias.

Process mining is not the only method used to analyze patient pathways, and this method can be combined with other methods such as discrete event simulation (DES). Prodel et al [30,31] developed a framework to automatically convert a process model discovered using process mining into a simulation model of clinical pathways. Abohamad et al [32] used process mining to discover ED processes and then used DES to study bottlenecks. Wood and Murch [33] modeled patient pathways using Markov chains to study transfer delays between medical units and discharge delays. Karakra et al [34] also used a DES to model an ED and added a real-time connection to real-time patient data to create a digital twin. The digital twin of the patient enables the monitoring of their pathway and activities as well as near-future predictions. Some models reproduce an entire hospital. Holm et al [35] used a DES to model an entire hospital and patient flows through the wards and determine bed use. Demir et al [36] used a similar model to anticipate an increase in the number of patients and adapt resources. Ordu et al [37] achieved an even more complete model of a hospital and patient flows.

To conclude, this review of the literature reveals that process mining and simulation are the principal methods used to study patient pathways. Process mining is a standard tool for discovering patient pathways (or other health care processes), but important limitations are noted in the literature, especially the complexity of the model graphs. Data preparation techniques and clustering methods are suggested to compensate for this issue. Some methods are based on expert interviews or expert knowledge integration. This is similar to our construction of rules. Several papers focus on time stamp correction or missing events or labels, but none of the studies focus on our problem of biased events. Simulations do not consider input data quality. In this study, we focused on enriching the log by adding an attribute that can define an event as relevant or irrelevant.

Objectives

The objective of this paper was twofold: (1) building a framework to model and analyze patient pathways and (2) proposing a method to automatically identify bias in patient pathway data. In other words, this method turns a database of observed pathways with bias into 2 databases: one database includes labeled pathways (identified bias), and one database

includes corrected pathways (without bias). Such a method is intended to ease the preprocessing of real data for data analysts or hospital managers who seek a clean database with unbiased medical pathways.

The scientific contributions of this paper are as follows:

1. This study provides a new framework to model patient pathways considering hospital management constraints (eg, bed occupancy and resource availability). This framework is used to assess patient pathways.
2. This study develops a new method to automatically label and correct pathways based on hospital data. Pathway labeling aims to identify the steps in a patient pathway due to a difficult bed management, and path correction aims to correct irrelevant activities. The labeling algorithm was assessed by comparing its outputs with experts' answers.
3. Two case studies are reported: (1) a quantitative comparison of the observed pathway database and the corrected pathway database was performed based on process mining using process model comparison and classical process mining indicators, and (2) a DES model based on the observed pathway database and the corrected pathway database was used to evaluate the impacts of data correction on the occupation of medical units.

Methods

Unscheduled Hospital Pathway Modeling Framework

Formal Definition of the Framework for the Study Patient Pathway

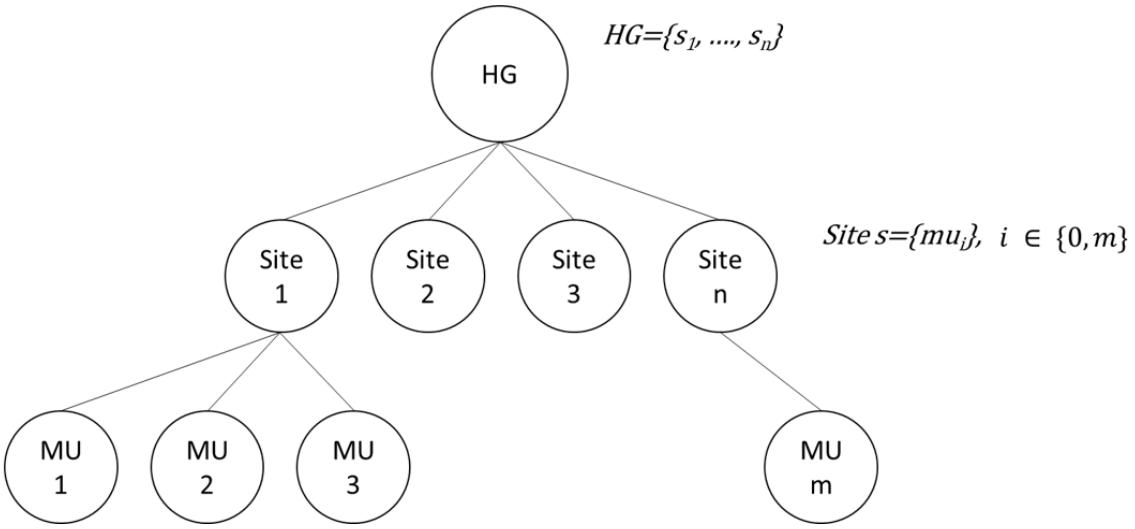
Overview

In this section, we propose a set of definitions that will be used to formalize the unscheduled hospital pathway modeling framework.

In this study, we were interested in the medicine, surgery, obstetrics, and odontology (MCO). In French, the initials MCO stand for *Médecine, Chirurgie, Obstétrique et Odontologie* (medicine, surgery, obstetrics, and odontology). The medical units belong to a *hospital* that itself can belong to a *hospital group*. Figure 1 represents the dependencies among the hospital group, site, and medical unit. Here, we are interested in the pathways inside the same hospital group, which we call the *MCO-stay*. Multimedia Appendix 1 provides detailed definitions of the aforementioned concepts.

The hospital pathways are defined using a process mining formalism [8].

Figure 1. Hospital group (HG; {s1,..., sn}) structure. MU: medical unit.



Definition 1 (Event)

Let E be the event universe (ie, the set of all possible event identifiers), E^* be the set of all sequences over E, and T be the time domain. We assume that events are defined by several attributes; however, the case ID, time stamp, and activity name are mandatory for case identification, trace ordering, and event labeling, respectively.

Let AN be a set of attribute names. For any event $e \in E$ and name $z \in AN$, $\#_n(e)$ is the value of the attribute z for event e. We consider $\#_{activity} \in E \rightarrow A$ and $\#_{time} \in E \rightarrow T$ functions that assign an activity name from a finite set of process activities A and a time stamp, respectively, to each event. For convenience, we assume the following standard attributes: (1) $\#_{activity}(e)$ is the activity associated with event e, (2) $\#_{time}(e)$ is the time stamp of event e, and (3) $\#_{trans}(e)$ is the transaction type associated with event e (eg, schedule, start, complete, and suspend).

The transaction type attribute $\#_{trans}(e)$ refers to the life cycle of activities. In most situations, activities take time. Therefore, events may point out, for example, the start or completion of activities.

Definition 2 (Trace)

A trace is a finite sequence of events denoted as $\sigma = < e_1, e_2, ..., e_n > \forall e_i \in E^*$ such that each event appears only once: $e_i \neq e_j$ for $1 \leq i < j \leq |\sigma|$.

Specifically, $|\sigma|$ denotes the length of the trace. In this case, $|\sigma| = n$.

Definition 3 (Stage)

In a trace, several events can have the same activity name. In the following, the subset of events with the same activity name that contains a single start event and a single completion event subsequent to the start event is referred to as a stage.

Let e_i and e_m such that (1) $\#_{activity}(e_i) = \#_{activity}(e_m)$=a; (2) $\#_{trans}(e_i)$=start and $\#_{trans}(e_m)$=complete; (3) $\#_{time}(e_i)=t_i \leq \#_{time}(e_m)=t_m$; and (4) e_j such that $\#_{time}(e_j)=t_j \geq t_i$, $t_j \leq t_m$, $\#_{activity}(e_j)$=a, and $\#_{trans}(e_j)$=complete or $\#_{trans}(e_j)$=start.

$$\text{Stage} = \{e_j | \#_{activity}(e_j) = a \text{ and } t_i \leq t_j \leq t_m\}$$

The duration of a stage is defined as the time between the start of the stage and its completion:

$$\#_{duration}(s) = \#_{time}(e_m) - \#_{time}(e_i)$$

In the following, we will note e_{x1}, e_{x2},..., e_{xi} as all the events that compose stage *x*.

Furthermore, the duration of the trace σ of length *n* is determined as follows:

$$duration(\sigma) = \#_{time}(e_n) - \#_{time}(e_1)$$

In our study framework, a trace always begins at the ED stage and ends at the last unit of the MCO stay (the unit before the MCO discharge).

Definition 4 (Event Log)

An event log is a set of traces representing the execution of the underlying process. An event can only occur in 1 trace; however, events from different traces can share the same activity.

Definition 5 (Patient Pathway)

A patient pathway describes the succession of medical events inside a health care facility. In this work, each pathway is linked to an MCO stay.

A patient pathway is a set (p, s, σ, d) where $p \in$ N is the identifier of the patient, $s \in$ N is the identifier of the MCO stay, σ is the trace of the MCO stay, and d is the MCO discharge disposition.

We consider two types of pathways:

1. Scheduled pathways
   : these pathways are planned before patient admission.
2. Unscheduled pathways
   : neither the admissions nor the pathways are planned. The patients are hospitalized from the ED or admitted to a specific unit for life-threatening emergencies.

Definition 6 (Relevance of Stage)

A stage of a patient pathway is relevant if it is adequate that, at this moment, the patient is still hospitalized (first condition) and if the patient is in the medical unit intended to care for their pathology (second condition).

We consider 3 levels of relevance: level 2 (both conditions are met), level 1 (the second condition is not met; ie, the patient is not hospitalized in the ideal medical unit for their pathology), and level 0 (no condition is met, and there is no medical reason that justifies the patient being still hospitalized in this discipline).

Definition 7 (Bias)

In our context, a bias in the data is noted when some details about a piece of information are missing, which leads to a misinterpretation of a situation.

For example, a pathway {ED, Surgery, Geriatrics} without additional information suggests that the patient needs surgery after the ED followed by geriatric care. The bias is that the patient just stays in surgery while waiting for a bed in geriatrics.

Definition 8 (Activity Labels)

In this study, we consider 2 levels of activity names.

In level 1, *U* is the set of labels corresponding to all the names or IDs of the medical units constituting the hospital group. Consequently, an event activity is a medical unit that a patient has visited.

$$GH = \{mu_1, ..., mu_n\}$$

$$U = \{id(mu_i)\} \text{ with } i \in [1, n]$$

$$\#_{activity}(e_x) = id(mu_m)$$

In level 2, let *L* be the set of labels corresponding to the relevance levels. *A* is the set of labels corresponding to the product of *U* and *L*:

$$L = \{\text{level 0, level 1, level 2}\}$$

$$A = U \times L$$

$$\#_{activity}(e_x) = (id(mu_m), \text{level l})$$

Hence, level 1 characterizes the activity of an event based on the ID of the medical unit, and level 2 adds a level of relevance. For more convenient reading, in the following sections, the activity of an event will be noted using the name of the medical unit.

Motivation

The pathway of a patient is governed not only by pure medical logic (health care needs) but also by logistical limitations. In other words, the pathway of a patient depends not only on their medical needs but also on the availability of inpatient beds and the possibility of discharge. Therefore, a patient can go to an unsuitable medical unit (unit *b*) because of a lack of beds in the suitable unit (unit *a*). The patient can later be transferred to unit *a*. Discharge also has an impact on patient pathways. Indeed, patients do not always immediately leave the hospital when they are medically fit for discharge because they are waiting for a discharge disposition. The challenge is to automatically identify these irrelevant steps in any MCO pathway. Indeed, these pathways are not clearly identified in the electronic health records (EHRs), and there is no generally applicable thesaurus of ideal pathways and no clear indicator of the adequacy of a unit in the EHR. The same medical unit can have different functions (see Textbox 1 for an example), and identical patients in terms of pathology can have different pathways according to hospital occupancy [6].

Our objective was to find a function (an algorithm) that evaluates the relevance of each stage. It is important to understand the word *relevance* as defined in the previous sections (*Definition 6: Relevance of Stage* section). Medical practices or medications were not judged here. Only the relevance of the patient's location was evaluated. In our framework, an input trace with event labels comprising only the activity name is converted into an output trace with event labels comprising the activity name and the level of relevance.

Textbox 1. Example of the different roles that a medical unit can play in a patient pathway.

Examples

- Emergency department (ED) to neurology: acute care in neurology
- ED to neurology to neurovascular intensive care: waiting in neurology for a bed in neurovascular intensive care
- ED to neurovascular intensive care to neurology: waiting for a discharge solution in neurology

Definition of the Function Evaluating the Relevance of Stages

Let $\sigma =< e_{11}, e_{12}, e_{nx} > \forall e_{xi}, \#_{activity}(e_{xi}) \in U$, and $\sigma' =< e'_{11}, e'_{12},..., e'_{nx} > \forall e'_{xi}, \#_{activity}(e'_{xi}) \in A$.

σ and σ' are 2 traces of the same case: σ is the *historical trace* and σ' is the *labeled trace*.

$f: \sigma \rightarrow \sigma'$ with $|\{e'_i|\#_{act}e'_i = a\}|_{\forall e'_j \in \sigma'} \geq |\{e_i|\#_{act}e_i = a\}|_{\forall ei \in \sigma}$

The function *f* identifies the relevance levels of each stage in a trace. A stage can be divided into several phases with distinct levels of relevance. Therefore, the number of events that correspond to activity *a* in the trace σ is smaller or equal to the number of events that correspond to the activity *a* in trace σ'.

Example

This paragraph illustrates the definitions and the transformation of a pathway using the function *f*. In the following fictive example, the hospital group is named Groupe Hospitalier Bretagne Sud (GHBS) and is composed of 2 sites, named Scorff and Villeneuve. One patient arrived on January 4 at 5:36 AM at the ED of the Scorff Hospital. At 10:13 AM, he was admitted to the observation unit (OU), but the patient was actually waiting for a bed in the geriatric unit. On January 5 at 9:45 AM, the patient was transferred to the geriatric unit of the Villeneuve Hospital, another site of the hospital group. He arrived at 10:15 AM. On January 10 at 2 PM, the patient was medically fit for discharge. On January 12 at 1:30 PM, the patient was discharged, and he returned home with additional community nursing services. Figure 2 illustrates the pathway of the patient according to the framework defined previously.

The pathway of patient *00000056098* can be formalized as follows:

$\sigma =< e_{11}, e_{12}, e_{21}, e_{22}, e_{31}, e_{32} >$ with

$\#_{activity}(e_{11}) = ED, \#_{trans}(e_{11}) = start$

$\#_{activity}(e_{12}) = ED, \#_{trans}(e_{12}) = end$

$\#_{activity}(e_{21}) = OU, \#_{trans}(e_{21}) = start$

$\#_{activity}(e_{22}) = OU, \#_{trans}(e_{22}) = end$

$\#_{activity}(e_{31}) = GERIATRICS, \#_{trans}(e_{31}) = start$

$\#_{activity}(e_{32}) = GERIATRICS, \#_{trans}(e_{32}) = end$

The function *f* takes the trace σ as input and returns the trace $\sigma' =< e'_{11}, e'_{12}, e'_{21}, e'_{22}, e'_{31}, e'_{32} >$ with the following features:

$\#_{activity}(e'_{11}) = (ED, level2), \#_{trans}(e'_{11}) = start$

$\#_{activity}(e'_{12}) = (ED, level3), \#_{trans}(e'_{12}) = end$

$\#_{activity}(e'_{21}) = (OU, level1), \#_{trans}(e'_{21}) = start$

$\#_{activity}(e'_{22}) = (OU, level1), \#_{trans}(e'_{22}) = end$

$\#_{activity}(e'_{31}) = (GERIATRICS, level2), \#_{trans}(e'_{31}) = start$

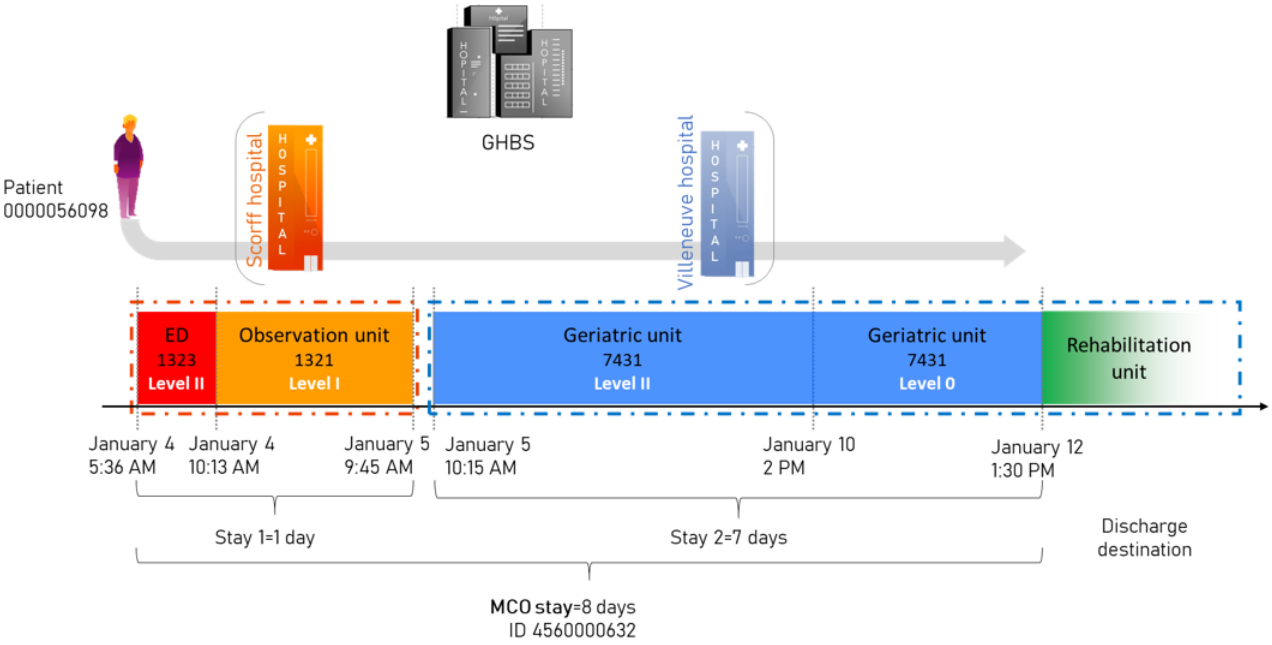
$\#_{activity}(e'_{32}) = (GERIATRICS, level2), \#_{trans}(e'_{32}) = end$

$\#_{activity}(e'_{41}) = (GERIATRICS, level0), \#_{trans}(e'_{41}) = start$

$\#_{activity}(e'_{42}) = (GERIATRICS, level0), \#_{trans}(e'_{42}) = end$

The succession of stages is described by level-2 activity labels. The initial stage *Geriatrics* has been divided into a relevant phase (level 2) and an irrelevant phase (level 0).

Figure 2. Illustration of the framework using a fictive pathway and patient. ED: emergency department; GHBS: Groupe Hospitalier Bretagne Sud; MCO: medicine, surgery, obstetrics, and odontology.

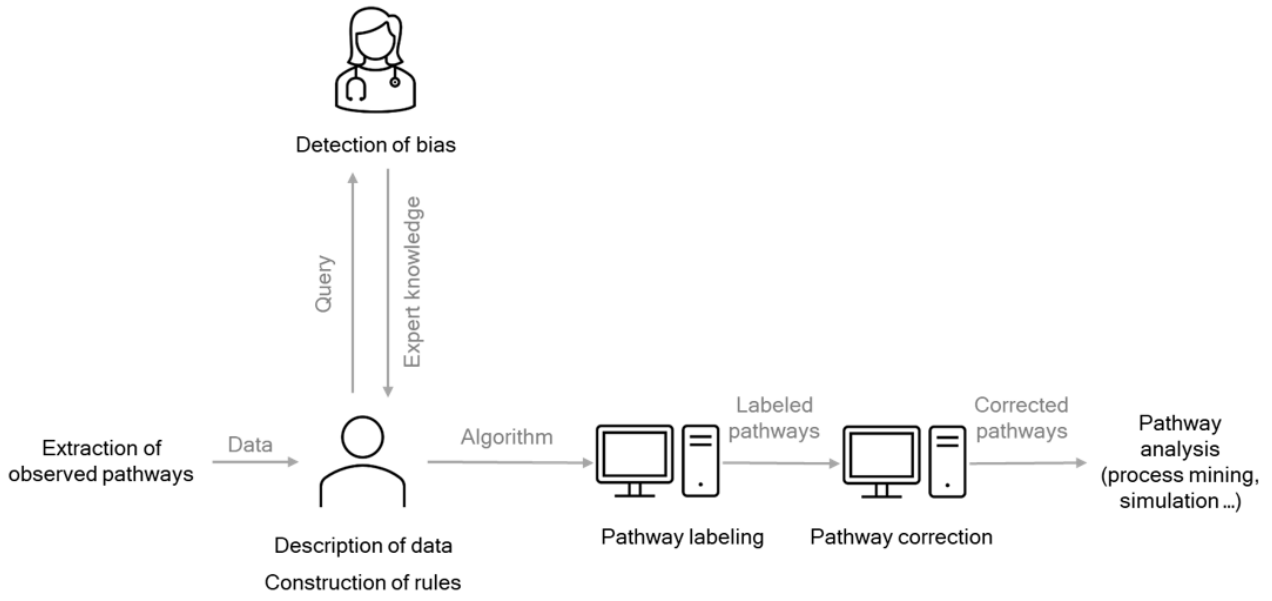


Automatic Pathway Labeling and Correction

Overview

In this section, the method for identifying bias in patient pathway data and the method for building an algorithm to label the stages of the pathways as relevant or irrelevant are described. We then present an algorithm for automatically transforming a historical trace into a labeled trace and an algorithm for automatically transforming a labeled trace into a corrected trace. These algorithms are based on a symbolic approach. In other words, the rules are *if*, *then* propositions. Our approach can be visualized in Figure 3.

Figure 3. Pathway-labeling method.



Identification of Bias and Rule Definition

Overview

The methodology for identifying the bias and then defining the rules consists of three steps, which are described as follows:

1. Description of the pathway dataset: the objective was to distinguish the different pathways and identify frequent and rare patterns.
2. Identification of bias with expert interpretation: the objective was to analyze the patterns with experts to determine the bias.
3. Definition of rules: the rules are defined based on the experts' analysis.

The first step can be achieved by computing the frequency and representativeness of each pathway variant and the number of events per variant, grouping some events, and computing the

new variants to identify repetitive patterns. The second step enables us to identify bias through discussions with experts. Preferably, the discussions are conducted with several physicians to obtain different opinions. The third step requires the translation of the analyses of physicians into rules. A rule deduces whether an event is relevant from the EHR data. According to our definition of relevance (cf Definition 6), we consider three levels: (1) level 0, in which the stage is completely irrelevant (none of the conditions are met); (2) level 1, in which the stage is not totally relevant (only the first condition is met); and (3) level 2, in which the stage is relevant (both conditions are met). The aim of the rules is to identify in a pathway the phases with various levels of relevance. We remind the reader that only the hospitalization ward (patient location) is evaluated.

Example

A rule states that, if the first stage of the trace lasts <10 units, then the stage is completely irrelevant (level 0).

Let $\sigma =< e_{11}, e_{12}, e_{21}, e_{22} >$ be a trace with 2 activities a_1 and a_2 such that the following conditions are met: (1) *#activity*(e_{11}) = *#activity*(e_{12})=a1; (2) *#activity*(e_{21}) = *#activity*(e_{22})=a2; and (3) *#time*(e_{11})=0, (4) *#time*(e_{12})=7, and (5) *#time*(e_{21})=7, *#time*(e_{22})=20.

The first stage is a_1 and lasts 7 units; therefore, e_{11} and e_{12} are labeled as level 0.

Algorithm 1: Pathway Labeling

Once the rules have been defined, algorithm 1 (Textbox 2) labels a pathway according to these rules (lines 6-9). A level of 1 or 0 is assigned if a stage is identified as irrelevant, and a level of 2 is assigned if the stage is relevant (lines 7-9). A stage can be labeled through several rules. In this case, the worst label is applied (ie, level 0 has a priority over level 1, and level 1 has a priority over level 2 [lines 10-12]).

We want to emphasize that this algorithm is purely based on logic and administrative rules. It does not include medical reasoning and is, therefore, inaccurate. However, the aim of the next section is to evaluate this inaccuracy (ie, the number of errors between an algorithm with simple rules and the complex reasoning of an expert [expert knowledge]).

Textbox 2. Algorithm 1—pathway labeling.

1: Let e_x be an event of the historical pathway

2: Let e'_x be an event of the labeled pathway

3: Let t_1 be the start date of the stay.

4: Let t_n be the end date of the stay.

5: Let $\sigma =< e_{11}, e_{12},..., e_{n1}, e_{n2} >$ be the trace representing the historical pathway.

6: Let L be a list that stores the result of each rule.

7: for each rule r_k do

8: Add $r_k(\sigma)$ to L

9: end for

10: for each $e_x \in \sigma$ do

11: Apply the modification of each rule that has changed e_x. The lowest relevance level has priority.

12: end for

13: Return $\sigma' =< e'_{11}, e'_{12},..., e'_{m1}, e'_{m2} >$, the labeled trace.

Algorithm 2: Pathway Correction

We also implemented an algorithm to correct the pathways labeled using algorithm 1 (Textbox 3). The idea is to transform the observed pathway into a theoretical pathway by correcting irrelevant stages. The different corrections applied to a labeled pathway are deduced from the rules. The irrelevant activities are replaced with the relevant activities (lines 4-6). Only the label of an event is changed, and the time stamp remains the same. At the end of the correction, subsequent identical activities are merged (lines 7-13). For example, let us note a stage (activity name, relevance level, start, or end). The pathway *<(ED, level2, t1, t2), (OU, level1, t2, t3), (Geriatrics, level2, t3, t4)>* is corrected and becomes *<(ED, t1, t2), (Geriatrics, t2, t3), (Geriatrics, t3, t4)>*. In addition, the pathway can be merged to become *<(ED, t1, t2), (Geriatrics, t2, t4)>*.

Textbox 3. Algorithm 2—pathway correction.

1: Let $\sigma' = < e'_{11}, e'_{12}, ..., e'_{m1}, e'_{m2} >$ be the labeled trace.

2: Let r be the rule applied at e'_x.

3: Let $\sigma'' = < e''_{11}, e''_{12}, ..., e''_{m1}, e''_{m2} >$ be a copy of σ'.

4: for each $e''_x \in \sigma''$ do

5: $\#_{activity}(e''_x)$ = the corrected activity according to the rules of correction

6: end for

7: for each $e''_x \in \sigma''$ with $t_x < t_m$ do

8: if $\#_{activity}(e''_{x1}) = \#_{activity}(e''_{x+1,1})$ then

9: $\#_{time}(e''_{x2}) = \#_{time}(e''_{x+1,2})$ with $\#_{trans}(e''_{x+1,2})$ = complete and all the events of stage $x+1$ are deleted from σ''

10: end if

11: end for

12: Return $\sigma'' = < e''_{11}, e''_{12}, ..., e''_{p1}, e''_{p2} >$, the corrected trace.

Evaluation of the Performance of the Labeling Algorithm

This subsection describes the method used to assess the labeling algorithm. Because there is no reference to compare the results of the algorithm with a ground truth, the evaluation of the algorithm has to be made by comparing its results with the analyses of experts. The methodology used for this study is inspired by the framework developed by the French think tank Ethik IA for its humane oversight board (Ethik-IA, unpublished data, April 2021). A representative sample of patient pathways was analyzed by 2 experts. They had access to information from the electronic patient records. Each expert performed the analysis separately. The results of the first expert were compared with those of the second expert. When the results did not match, the medical experts discussed them to find a common answer. Their answers were then compared with the algorithms' answers. For each difference, a discussion with the experts allowed us to determine whether the algorithm was wrong and, if so, qualify the errors.

The method presented in this paper is general and can be applied at any hospital. In the next section, we apply these methods to a real case study to create rules to label and correct a real dataset extracted from a hospital database, and we evaluate the accuracy of these rules.

Ethical Considerations

This study was approved by the French Data Protection Authority (*Commission Nationale de l'Informatique et des Libertés*) under the number 922243. French and European rules about access to health care data for research were respected and ethical standards also.

Results

Data

This work was performed with the GHBS, a French hospital group located in the Lorient area. It has 2 general hospital sites with an ED and 6 other sites. In total, there were 89,791 ED visits and 108,875 hospitalizations and sessions (values for 2021). This study was based on data collected at the GHBS. Data were retrospectively collected for the period from July 2020 to July 2021. The data cover 12 months of activity and only adults, 54,850 different ED visits and 41,161 unique patients, including 19,905 MCO stays. Multiple MCO stays of the same patient were treated as separate instances. Three sources of data were used: (1) electronic patient records, (2) administrative health care databases, and (3) data from the software used for rehabilitation and home hospitalization. Only structured data were used to save time in the data analyses; no plug-and-play natural language processing tool was available for our data. The pediatric and obstetric pathways were excluded from the study dataset, as were the pathways with only a visit to the ED.

Identification of Bias and Rule Definition

In this section, we detail the results obtained using our method to identify bias from our data.

Results of Step 1: Description of the Pathway Dataset

In our dataset, there were 19,905 pathways and 1013 trace variants. Some variants were very frequent, such as (ED, OU) representing 22.75% (4528/19,905) of the pathways, and others were very rare, such as (ED, neurology intensive care, cardiology) occurring just once.

We observed that most pathways had only 1 stage after the ED visit (60/1013, 5.92% of the variants and 15,699/19,905, 78.87% of the pathways). Pathways with >3 stages after the ED were rare. They appeared between 1 and 3 times in the dataset and represented 0.98% (195/19,905) of the pathways but 18.85% (191/1013) of the variants. Therefore, the diversity of pathways was mainly due to pathways with many activities. We identified five types of pathways at the GHBS: (1) mono-disease pathways, which include 1 necessary medical unit; (2) pathways for patients who were seriously ill, which include transfer to an intensive care unit (ICU); (3) older person pathways, which include geriatric units; (4) frequent and processed pathways,

which include strokes; and (5) multi-disease and complex pathways, which include several medical units.

The most frequent medical units were OUs, polyvalent medicine units, geriatric medicine units, postemergency units, surgery units, and specialized medical units. By categorizing the medical units into 4 groups (ED, medicine, surgery, and ICU), we obtained 165 patterns, and the 10 most frequent structures of the pathways are listed in Table 1.

Within the pathways (MEDICINE, MEDICINE), several patterns were frequently observed. Pathways such as heart failure or stroke pathways were normally composed of 2 steps after the ED: admission to a cardiology (or neurology, respectively) ICU followed by cardiology (or neurology, respectively). The second type of pattern is the pathway with an admission to a polyvalent unit before an admission to a specialized unit or another polyvalent unit. In this case, a polyvalent unit is a medical unit, such as an OU, polyvalent medicine unit, or postemergency unit, where patients with multiple diseases or who do not require specialized treatment are treated. We also observed a few transfers between specialized units. Finally, patients could be transferred between the weekly hospitalization unit and the full hospitalization unit.

For most of the pathways that followed the pattern (SURGERY, MEDICINE), the activity of surgery was noted as an overflow bed. For the few others, a surgical act was performed before a transfer for medical reasons to a medicine unit.

The pattern (MEDICINE, SURGERY) mainly concerned an admission to an OU (while waiting for surgery) followed by a surgery unit. In the other pathways, patients were first admitted to a specialized unit (eg, hepatogastroenterology) before surgery.

Hence, we obtained four patterns for the pathways in 2 steps: (1) a polyvalent unit followed by a specialized unit (pattern 1), (2) a surgery unit followed by a medical unit (pattern 2), (3) a daily or weekly hospitalization unit followed by a full hospitalization unit of the same specialty (pattern 3), and (4) a specialized unit followed by another specialized unit (pattern 4).

Table 1. Main structures of historical pathways and occurrence differences for corrected pathways.

Variant	Occurrence (%)	
	Historical	Corrected
ED[a], MEDICINE	68.68	68.57
ED, MEDICINE, MEDICINE	10.67	7.24
ED, SURGERY	9.66	10.90
ED, SURGERY, MEDICINE	2.88	0.33
ED, MEDICINE, SURGERY	1.7	0.63
ED, MEDICINE, MEDICINE, MEDICINE	1.2	0.80
ED, ICU[b], MEDICINE	0.88	1.05
ED, ICU	0.53	0.56
ED, SURGERY, SURGERY	0.47	0.16
ED, ED, MED	0.39	0.22
ED	—[c]	7.18

[a]ED: emergency department.

[b]ICU: intensive care unit.

[c]Not present in historical pathways.

Results of Step 2: Interpretation of Patterns by Experts

We discussed the patterns identified in the first step with the experts. According to the experts, pattern 1 (a polyvalent unit followed by another unit) usually indicates that the polyvalent unit is used as a buffer. In fact, when beds are lacking, patients can be admitted to a polyvalent unit to begin their treatment while waiting for a bed in the ideal unit. The second pattern has the same explanation—a patient requiring medical treatment is placed in a surgery unit while waiting for a bed in the ideal unit. Pattern 3 (transfers between daily or weekly hospitalization units and full hospitalization units) is explained by a lack of beds in the full hospitalization unit. The last pattern (transfers between specialized units) has no general explanation. Occasionally, a specialized unit can also be used as a buffer, but the transfer can also be medically explained.

The length of stay (LoS) was also an aspect of the pathway discussed with the experts. Some pathways are too long because patients cannot be discharged as soon as they are medically fit for MCO discharge. The delay is mainly due to a back home impossible without community care or a lack of beds in rehabilitation centers. LoS can be compared with a national reference. In France, the national reference is the average LoS used for hospital stay invoicing (*durée moyenne de séjour* in French). It is defined for each diagnosis-related group. However, each pathway is unique, and an LoS above the national reference does not guarantee that the stay was too long because of discharge difficulties. For example, longer stays for patients

receiving palliative care are frequent and normal. Occasionally, the diagnosis-related group does not correctly report the seriousness of the patient because the patient was transferred to another hospital.

The LoS in the ED was also discussed. According to the experts, the LoS in the ED is occasionally too long because patients are waiting to be hospitalized. The ED LoS should not exceed 5 to 10 hours.

Results of Step 3: Rule Construction

From these observations and discussions, we deduced several dimensions to investigate in a pathway:

1. Time spent in the ED: ED LoS can be prolonged because of a lack of inpatient beds in acute care units. Rule 1 evaluates whether the time spent in the ED is too long.
2. LoS: the LoS can be prolonged because of a delayed discharge. Rules 2 and 3 evaluate this condition.
3. Overflow bed: occasionally, patients are admitted to a medical unit but are treated by the physicians of another unit. The location of the patient is entered (ie, the activity name) in the hospital data, and the unit medically responsible for the patient is also indicated. For example, when a patient is in surgery and awaits a bed in cardiology, the activity is surgery and the medically responsible unit is cardiology.
4. Sequence of activities: a typical pattern is the transfer between a polyvalent unit and a specialized unit. According to physician experience, when the transfer occurs within 1 week, in general, the polyvalent unit stage is irrelevant. In rule 5, the threshold is 7 and a half days to consider the transfer time. Rule 6 is dedicated to the OU because, in our dataset, a transfer in the OU is labeled as "back home within 24 h," "observation," or "awaiting bed." The label "awaiting bed" indicates an irrelevant stage because the patient should have been transferred immediately to the appropriate unit.

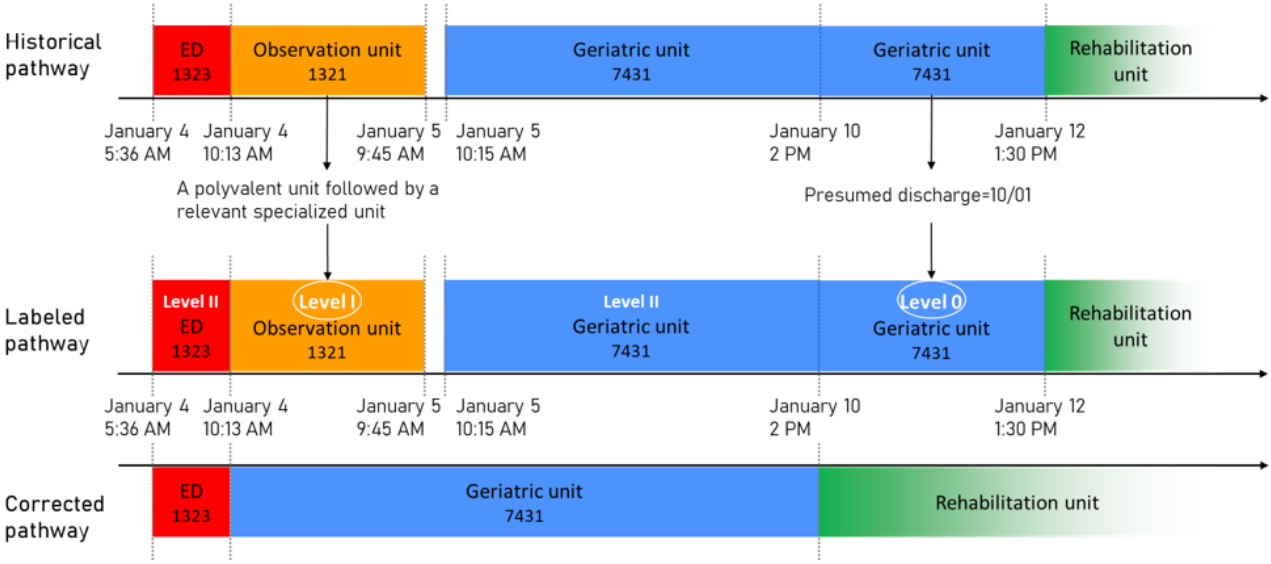
The challenge was to identify which structured data could be used to investigate these different dimensions and, therefore, to create the rules. We detail the rules obtained from our dataset in Multimedia Appendix 2, as well as the algorithms of the 7 rules.

From these rules, we deduced how to correct the labeled pathways. The different corrections are listed in Table S1 in Multimedia Appendix 2.

Example

Figure 4 illustrates the labeling and correction of a pathway. In total, 2 phases are considered irrelevant by the algorithm: the OU (rule 6) and the end of the stay in the geriatric unit (rule 3). To correct the pathway, the time spent in the OU is replaced by the time spent in the geriatric unit, which is the relevant stage following the OU, and the patient is discharged earlier at the presumed discharge date.

Figure 4. Example of the correction of a pathway. ED: emergency department.



Accuracy of the Labeling Algorithm

A total of 118 different pathways were analyzed by 6 different duos of physicians from the GHBS (only 1/12, 8% had previously participated in the discussion to define the rules) and compared with the algorithm output. Only rules 3, 4, 5, 6, and 7 could be evaluated because the physicians could not reach a consensus on either the ideal LoS or the ideal length of ED visit within the allotted time. The physicians had access to the patient's pathway, their age, the chief complaint, the diagnoses, the national LoS reference for the diagnosis-related groups, the discharge destination, and requests for rehabilitation and home hospitalization. They also had access to patient records if more information was needed.

We counted the number of errors per stage (except for the first ED visit) and per pathway. For example, the pathway (ED, OU, CARDIOLOGY) has 2 stages (we do not count the ED), and the experts assessed that OU was not relevant and that CARDIOLOGY was relevant, but the algorithm assessed that the 2 stages were relevant. Therefore, we identified 1 pathway error per 1 pathway and 1 stage error per 2 stages. Table 2 presents the results of the evaluation of the 118 pathways. The algorithm failed to evaluate 19.5% (23/118) of the pathways

and 12.9% (30/232) of the stages. Among these errors, false-positive errors were 5 and 3 times more important, respectively, than false-negative errors. This means that the main error of the algorithm was considering a pathway or a stage as relevant when it was actually irrelevant.

The main errors were related to medical knowledge. When patients were hospitalized in a specialized medical unit (eg, oncology) but pertained to another specialized medical unit (eg, cardiology), the algorithm did not detect this irrelevance. Similarly, hospitalization in the general ICU before transfer to the cardiology ICU was occasionally irrelevant but not detected by the algorithm. In addition, the algorithm considered that a polyvalent unit following a specialized unit was irrelevant; however, in some cases, it was wrong.

Table 2. Performance of the algorithm based on rules 3, 4, 5, 6, and 7.

	Actual values		Precision	Recall
	Positive [a]	Negative [b]		
Pathways			0.75	0.93
Predicted values				
Positive	56	19		
Negative	4	39		
Stages			0.88	0.95
Predicted values				
Positive	162	22		
Negative	8	40		

[a]Relevant pathway or stage (level 2).

[b]Irrelevant pathway or stage (level 0 and level 1).

Preprocessing of Pathway Data

In this section, we use only rules 3 to 7 to preprocess our data given that rules 1 and 2 could not be assessed. We evaluated 19,832 pathways, including 24,989 stages (excluding the first ED stage). Of these 19,832 pathways, 2669 (13.46%) were evaluated as irrelevant, and 11.21% (2802/24,989) of stages were also evaluated as irrelevant. Considering the error margin, between 2162 and 3176 pathways were irrelevant, and between 2438 and 3166 stages were irrelevant. The main irrelevant movements detected by the algorithm were overflow bed in surgery, overflow bed in polyvalent units, and waiting time in the OU. The 19,832 historical pathways included 986 (4.97%) variants. Once corrected, 792 variants were noted. Table 1 details the distribution of the structures of the variants. We observed that the *ED, MEDICINE, MEDICINE* and *ED, SURGERY, MEDICINE* traces were less frequent than before correction, which is due to the correction of the overflow beds. The trace ED appeared because the OU stage was assessed as irrelevant for several *ED, OU* (included in *ED, MEDICINE*) traces.

Statistical Analysis of Relevant and Irrelevant Pathways

Once the pathways were labeled, we compared the relevant pathways with the irrelevant pathways to understand the causes of irrelevance in the pathways. Several causes were already known among medical and administrative staff: bed occupation rates, ED crowding, age, and discharge destination. We tested these hypotheses using 4 bivariate analyses. The 4 variables to explain were an ED LoS of >5 hours, an ED LoS of >10 hours, the presence of overflow beds, and delayed discharge. For categorical variables, the proportions were compared using a chi-square test. For quantitative variables, the distributions were compared using a 2-tailed Student *t* test. We studied different explanatory variables: weekday corresponds to the start of the ED visit or admission to the inpatient unit; the arrival period is divided into 4 periods (morning from 7 AM to noon, afternoon from noon to 5 PM, night from 5 PM to 11 PM, and deep night from 11 PM to 7 AM); the next historical stage is the inpatient unit where the patient was admitted, and the next corrected stage is where the patient should have been admitted (based on the evaluation of the pathways); the last stage is the medical unit from which the patient was discharged; and the ED crowds are the number of patients present in the ED when the patient arrives.

Table 3 reports the bivariate analyses. Tables S2 and S3 in Multimedia Appendix 3 provide the detailed results. Several observations can be made from the statistical analysis. First, the seasons and the irrelevance of the next stage are not significant for a delay in the ED of >5 hours, but they are significant for a delay of >10 hours. Second, counterintuitively, fewer irrelevant post-ED admissions occur for long ED delays. This finding is probably because patients who stay in the ED for a long time are ultimately admitted to a relevant unit. Third, the weekday of patient arrival influences the ED delay. On Mondays, more patients wait >5 hours, and the proportion decreases throughout the week and increases again on Sundays. This phenomenon is caused not only by the greater number of patients admitted on Monday but also by the difficulty in hospitalizing patients during the weekend. Therefore, on Sunday, many patients are waiting for hospitalization. Fourth, this is the same observation and explanation for the overflow beds; more patients are admitted to an irrelevant unit on Sunday. Fifth, crowding in the ED was less important for the longest ED delays. Indeed, patients arriving at night or late at night are less

likely to be transferred to an inpatient unit, and this is also the period during which fewer patients arrive at the ED. Sixth, a greater number of patients are admitted to irrelevant units when the ED is more crowded. Seventh, increased age is a factor of long delays in accessing the ED. Eighth, similar features are noted for the occupation rate of the next stage. Ninth, the next corrected stages had an occupation rate (95%) higher than that of the next historical stages (92%). Tenth, age does not impact the risk of being in an overflow bed. Eleventh, the season has an impact on discharge delays. Specifically, in summer, more discharges are delayed, perhaps because the health care supply is lower during the summer holidays. Twelfth, the proportion of delayed discharges varies according to the destination of the discharge. Discharges at a psychiatric center have the highest rate of delay (97/203, 47.8% of delayed stays), followed by discharges at rehabilitation centers (1285/3294, 39.01% of delayed stays). Delayed discharges for death correspond to requests for palliative care at home or at another center that were not accepted in time. Thirteenth, age does not affect the risk of delayed discharge.

The period of study was impacted by the COVID-19 pandemic. These results could be more robust with access to a longer period of study (3 years instead of 1), and the seasons variably could be assessed several times during a longer time frame. Furthermore, the quality of the data was imperfect, especially for the computation of the occupation rate. Therefore, the results should not be extrapolated to other periods or hospitals. However, the analysis allowed us to compare the relevant and irrelevant pathways because they were derived from the same dataset. Hence, we can conclude that significant differences (P<.001 for most of the features) were observed between relevant and irrelevant pathways. Logistic factors such as the day of the week, the hour of arrival, medical unit occupation, and the discharge destination influence the risk of overflow.

Table 3. Bivariate analysis.

Variable to explain and features	*P* value
ED[a,b] visit of >5 h	
Age	<.001
ED crowds	<.001
Occupation rate historical next stage[c]	<.001
Occupation rate corrected next stage[c]	<.001
Weekday	<.001
Season	.51
Arrival period	<.001
Next stage irrelevant	.05
Next historical stage	<.001
Next corrected stage	<.001
ED[a] visit of >10 h	
Age	<.001
ED crowds	<.001
Occupation rate historical next stage[c]	<.001
Occupation rate corrected next stage[c]	<.001
Weekday	<.001
Season	<.001
Arrival period	<.001
Next stage irrelevant	<.001
Next historical stage	<.001
Next corrected stage	<.001
Overflow beds	
Age[d]	.03
ED crowds[e]	<.001
Occupation rate corrected unit[f]	<.001
Arrival hour[f]	.37
Weekday	<.001
Season	.78
Arrival period	<.001
Delayed discharge	
Age	.72
Discharge destination	<.001
Last stage	<.001
Season	<.001

[a]The analysis was performed exclusively using the data from the principal site (Scorff) because emergency department crowds and age differ between the principal site and the smaller site (Villeneuve).

[b]ED: emergency department.

[c]The next historical stage is the inpatient unit where the patient was admitted, and the next corrected stage is where the patient should have been admitted (based on the evaluation of the pathways). Occupation rate is the number of patients already present in the unit over its capacity.

[d]We compared the set of pathways without an overflow stage and the set of pathways with at least one overflow stage.

[e]In each pathway, the ED crowds were only computed for the first medical unit subsequent to the ED stage.

[f]We compared the sets of relevant and irrelevant stages.

Synthesis of Patient Pathway Labels

To summarize this section, from the analysis of the structure of the patient pathways and expert knowledge, we built 7 rules that detect irrelevant stages in a patient pathway (description of the dataset and construction of rules). On the basis of these rules, we used a pathway-labeling algorithm that labels the stages of a pathway according to 3 levels of relevance (pathway labeling). We then used a pathway correction algorithm that transforms a labeled pathway into an ideal pathway (pathway correction). The evaluation of our algorithm showed that it exhibits 87% accuracy. In our dataset, 13.46% (2669/19,832) of the pathways were labeled as irrelevant. Finally, a statistical comparison between relevant and irrelevant pathways demonstrated that logistic constraints influence the quality of patient pathways.

The next 2 sections show the importance of this preprocessing step before analyzing patient pathways using process mining and of using these data for hospital management.

Case Study 1: Analysis of Patient Pathways Using Process Mining

Motivation

The first case study investigates the impact of our preprocessing technique on process discovery. We evaluated the ability of our preprocessing method to simplify process graphs. We compared the process graph of an event log comprising historical traces with the process graph of an event log comprising corrected traces. We used the ProM framework (version 6.12; ProM Tools) [38] to discover the graphs and estimate different metrics. The process graphs were computed using the Fodina algorithm [39], which outputs a causal graph that was transformed into a Petri net with the plug-in "Convert Causal net (C-Net) to Petri net" (F. Mannhardt).

Metrics

The metrics were computed using the plug-in "Show Petri-net Metrics" (HMW Verbek). This plug-in computes (1) the extended Cardoso metric (ECaM), (2) the extended Cyclomatic metric (ECyM), (3) the structuredness [40], and (4) the density [41].

The ECaM counts the splits (XOR, OR, or AND) in the net and penalizes each of them. The ECyM is the difference between the number of edges and vertices plus the number of strongly connected components. According to Lassen and Van der Aalst [40], a high ECaM score can be caused by a "high degree of fan-out from places," and numerous parallelisms can increase the ECyM. Structuredness recognizes different types of structures and scores each structure by giving it a penalty value. Finally, the density relates the number of arcs to the number of all possible arcs for a given number of nodes. Therefore, these 4 metrics quantify the different structural characteristics of a graph.

Quantitative Analysis

The process graph discovered from the whole dataset of pathways is spaghettilike because the number of variants is large with or without correction (986 vs 792). To avoid spaghettilike effects, we reduced the analysis to 1 medical unit (ie, the event log included only the traces that contained this activity). Table 4 shows the metrics calculated for 5 medical units. As expected, the corrected event log contains fewer variants and activities than the historical event log. Consequently, the number of arcs, places, and transitions on the graph also decreases. The historical graph density is greater than the corrected graph density. This finding indicates that the corrected graph is more compact than the historical graph. The ECaM and ECyM of the corrected graph are lower than those of the historical graph. Indeed, we can observe more places with many output transitions and more parallelisms in the historical Petri net. Finally, the structuredness of the historical graph is also greater than the structuredness of the corrected graph except for the neurology unit. This means that more complex structures or more unstructured components are observed in the historical graphs than in the corrected graph. The neurology exception can be explained by the fact that a state machine is identified in the historical graph but only an unstructured component is identified in the corrected graph. In conclusion, the corrected Petri nets can be considered simpler than the historical Petri nets.

Table 4. Comparison of the historical and corrected process graph.

Process graph	Variants, N	Activities, N	Arcs, N	Places, N	Transitions, N	Density	ECaM[a]	ECyM[b]	Structuredness
Cardiology									
Historical	13	10	32	7	16	0.14	11	16	64
Corrected	9	7	22	6	11	0.17	8	11	22
Visceral surgery									
Historical	19	15	42	8	21	0.13	12	20	42
Corrected	8	6	20	6	10	0.17	8	10	20
Polyvalent medicine									
Historical	12	13	76	16	38	0.06	33	27	76
Corrected	7	6	28	9	14	0.11	13	11	56
Geriatric medicine									
Historical	5	6	14	4	7	0.25	3	7	9.5
Corrected	3	4	14	4	5	0.25	3	5	6.5
Neurology									
Historical	24	13	52	11	26	0.09	21	26	208
Corrected	16	8	37	10	18	0.10	17	24	1602

[a]ECaM: extended Cardoso metric.

[b]ECyM: extended Cyclomatic metric.

Qualitative Analysis

A qualitative analysis can also be performed (see Multimedia Appendix 4 for the pictures of the graphs). The cardiology historical graph (Figure S1 in Multimedia Appendix 4) shows that several activities can occur before admission to cardiology, but it is not easy to distinguish between groups of patients. The corrected graph is easier to read, and four groups of patients can be identified: (1) serious patients who need intensive or continuous care before being admitted to cardiology, (2) patients who need to be permanently monitored for cardiac examination, (3) patients who need pulmonology care before cardiology care, and (4) patients who do not need other care before transfer and are directly admitted to cardiology (with eventually a step in the OU before).

The historical graph of the polyvalent medicine unit (Figure S2 in Multimedia Appendix 4) is complex to read, and several specialized units are present but not related to the polyvalent unit in the graph. The corrected graph is much simpler to read. Three groups of stays are identified: (1) stays with intensive or continuous care before admission to polyvalent medicine, (2) stays with direct admission, and (3) stays with a step in the OU or seasonal unit before admission to polyvalent medicine.

Equivalent analyses can be performed on other medical units (Figures S3, S4 and S5 in Multimedia Appendix 4). The neurology graph (Figure S3 in Multimedia Appendix 4) is less simple than the other graphs, possibly because patients going to neurology have complex pathways or because the correction of neurology pathways requires particular rules. However, the corrected graph is again more interpretable than the historical one.

Case Study 2: Estimation of Ward Capacity Through Simulation

Motivation

Computer simulations can be used to estimate the number of beds necessary in each medical unit to admit unscheduled patients and help solve capacity planning problems. Indeed, the actual number of patients admitted to each medical unit does not include all the patients not admitted because of a lack of beds, and the number of the patients who should not be admitted to each unit is considered. Hence, it does not reflect the real need for beds. To solve this problem, pathway correction can be used. This method can be useful for estimating capacities when building or renovating a hospital or for organizing medical teams.

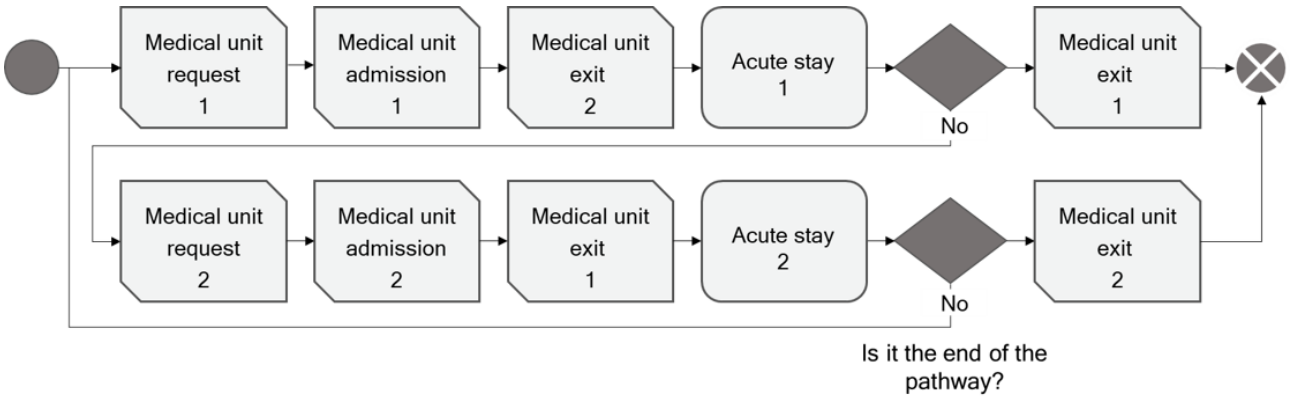
DES Model

In this case study, we simulate patient flow through the medical units of one hospital to compute the level of occupation of the medical units. We compared a simulation with the historical pathways (scenario 1) and a simulation with the corrected pathways (scenario 2). To do so, we modeled the medical units and patient flow using the AnyLogic software (The AnyLogic Company). Figure 5 shows the DES model.

The model represents 1 general hospital and 20 full hospitalization units. Each patient has a succession of units to follow. The model simulates the admissions of patients to medical units, their stays, and their discharges. A patient exits the simulation when they have completed their pathway. To simulate the source of patients, we used the dataset described in section Data (only patients from the main general hospital site [named Scorff] were included). We filtered the dataset based on the following criterion: only the pathways with at most 3

activities (ED visits plus 1 or 2 units) were included. We excluded pathways with weekly or daily units, the rare variants (coverage percentage of <0.001), and patients from sites other than Scorff. The corrected pathway dataset was filtered to keep only the stays included.

Figure 5. Modeling of patient flows through medical units.



Experimental Settings

The capacities of the medical units were set to infinity to calculate how many unscheduled patients needed to be admitted to each unit each day. The LoS in each unit was randomly generated according to a probability law. To choose this probability law, we compared several distributions (normal, β, γ, log-normal, Weibull, and exponentiated Weibull) and fitted them on the stays of our dataset (between 150 and 5000 stays per medical unit). The log-normal distribution best fit the LoSs in each unit. The parameters of the log-normal distributions were adjusted for each unit by fitting the distribution to the real values. The log-normal distribution tends to generate more extreme values than those observed in reality; thus, the LoS was limited to 28 days in a unit and 24 hours in the ED. The simulated patients' arrival dates were fixed and equal to the real patients' arrival dates. The simulated patients either followed the historical pathways (scenario 1) or the corrected pathways (scenario 2). The simulation duration was 1 year, and the simulation run included 13,366 pathways.

To choose the warm-up time, we monitored the mean number of present patients in each medical unit over the simulation time. After 50 days, a stable situation was reached (Figure S6 in Multimedia Appendix 5). The warm-up time was set to 2 months. The number of replications was chosen to have an error of <10%. In total, 15 replications allowed this target to be reached and a reasonable simulation time to be reached—approximately 1 minute is required for 1 run, and the 15 replications take 10 minutes. The results are the mean values of 15 replications.

The simulation model was validated by comparing the mean LoS of each medical unit from the simulation results to those from the real dataset. This was the only source of randomness in the model as the simulated patients arrive according to the real dataset and follow a deterministic pathway. The mean absolute error was 0.6 days, which is <10% of the mean length of a hospital stay (Table S4 in Multimedia Appendix 5).

Results of DES

Table 5 shows the mean number of acute patients present in each unit in both scenarios. For several units, the number of patients differed between scenario 1 and scenario 2. For example, the polyvalent medicine unit had, on average, 2 patients less with corrected pathways, and the neurovascular unit had 1 more patient. We also observed that, globally, there were fewer patients present at the same time with the corrected pathways compared with the historic pathways. Indeed, there were fewer stages in the corrected pathways because some were judged as irrelevant; therefore, in the simulation, the patients stayed less time in the hospital.

In conclusion, preprocessing pathway data is important for addressing capacity planning problems in hospitals. In this example, we observed that using historic pathways can lead to biased numeric interpretations for the capacity planning of medical units.

Table 5. Mean number of acute patients in medical units over 1 year.

Medical unit	Scenario 1	Scenario 2	Difference
3O surgery[a], mean (95% CI)	7.6 (7.3-7.9)	1.1 (0.9-1.2)	–6.5
Orthopedic surgery, mean (95% CI)	10.1 (9.9-10.3)	13.3 (13.0-13.5)	3.1
Visceral surgery, mean (95% CI)	6.5 (6.0-7.1)	5.9 (5.4-6.3)	–0.7
Pulmonology, mean (95% CI)	14.2 (14.0-14.4)	14.3 (14.1-14.6)	0.1
Cardiology ICU[b], mean (95% CI)	2.0 (1.8-2.2)	2.0 (1.8-2.2)	0.0
Cardiology, mean (95% CI)	10.1 (9.8-10.3)	10.1 (9.9-10.3)	0.1
Postemergency unit, mean (95% CI)	17.0 (16.7-17.3)	14.5 (14.3-14.7)	–2.5
Polyvalent medicine, mean (95% CI)	35.5 (35.5-35.5)	33.9 (33.9-33.9)	–1.7
Neurology ICU, mean (95% CI)	3.0 (2.7-3.3)	3.0 (2.7-3.3)	0.0
Neurovascular, mean (95% CI)	3.9 (3.8-4.1)	5.1 (4.8-5.5)	1.2
Neurology, mean (95% CI)	3.7 (3.7-3.7)	3.1 (3.1-3.1)	–0.7
Hepatogastroenterology, mean (95% CI)	13.3 (12.9-13.6)	13.7 (13.5-13.8)	0.4
Rheumatology, mean (95% CI)	11.4 (11.2-11.6)	11.3 (11.3-11.3)	–0.1
Observation unit, mean (95% CI)	8.7 (8.4-9.0)	6.5 (6.1-6.9)	–2.2
Geriatric medicine, mean (95% CI)	39.3 (39.0-39.5)	38.7 (38.5-38.9)	–0.5
ICU, mean (95% CI)	1.8 (1.5-2.1)	2.0 (1.7-2.3)	0.2
Seasonal unit[c], mean (95% CI)	2.9 (2.5-3.2)	2.0 (1.7-2.3)	–0.9
Oncology hematology, mean (95% CI)	5.0 (4.7-5.3)	5.2 (4.8-5.6)	0.2
Nephrology endocrinology, mean (95% CI)	3.5 (2.9-4.0)	3.4 (3.0-3.8)	–0.1
CCU[d], mean (95% CI)	0.6 (0.3-0.9)	0.6 (0.3-0.9)	0.0
Total patients, N	200	190	–10

[a]Ear, nose, and throat; ophthalmologic; and orthopedic surgery.

[b]ICU: intensive care unit.

[c]The seasonal unit is only open during the winter months. Therefore, the occupation figures computed over a year do not reflect reality.

[d]CCU: continuing care unit.

Discussion

Principal Findings

A framework and a methodology to study patient pathways were presented in this paper. They were used to develop a pathway-labeling algorithm that automatically detects whether a patient pathway is irrelevant (ie, contains stages due to resource limitations [as defined in the *Definition 6: Relevance of Stage* section]). Two main methods are available to achieve such a task: (1) building a thesaurus with medical experts (or using supervised learning) that links the main diagnosis (or the chief complaint) with an ideal pathway and (2) building a symbolic algorithm. The first method is the most accurate but is very time-consuming. This method would require hours of work with experts to build a thesaurus or annotate data for training a machine. None of these methods provide general results because the thesaurus and rules need to be adapted to each hospital. We chose the second option, an algorithm based on logic and administrative data, because it can be built quickly and is easily adaptable to organizational changes. We provided a general method to build this algorithm. We applied our algorithm to our dataset, and we were able to estimate the gap between our algorithm and an expert assessment. Our results demonstrate that a nonnegligible gap exists (13% to 19% of errors); however, we believe that the error rate was small enough for globally evaluated pathways. The estimation of this error also enabled us to identify the source of errors of the algorithm. On the basis of this labeling, a correction of the pathways was then performed to represent pathways that would be considered ideal.

We also demonstrated that resource limitations impact the choice of pathway using a statistical analysis that compared relevant and irrelevant pathways. The factors identified as increasing difficulties in managing patient flows could be included in hospitals' strategies to improve patient pathways.

The 2 case studies illustrate the importance of preprocessing patient pathway data before any analysis. Studying and representing patient pathways using process mining is complicated (*Related Work* section). By focusing on pathways with a common medical unit, we demonstrated that a corrected graph is more interpretable than a historical graph. Hence, our algorithm is an efficient preprocessing tool for the analysis of

patient pathways using process mining. The simulation of patient pathways is useful for testing bed management changes, but numeric results can be false if the input data include bias. In our example, the determination of the mean number of beds required for acute patients differed for the historic and corrected pathways. Some medical units need fewer beds, and others need more beds.

Pathway labeling should be applied before any analysis, such as process mining (case study 1), simulation (case study 2), or training of machine learning models to predict hospital pathways. In another work, we studied the prediction of the medical unit where a patient will be admitted after an ED visit. If raw data are used to train a machine learning model, the training will be biased. Indeed, the model will learn, for example, that some patients who do not need surgical treatment should be transferred to surgery. In contrast, if the model is trained using relevant pathways, it will learn the ideal medical unit for the patients [42].

Limitations

We proposed a general method to study patient pathways and identify bias in the data. However, our approach could only be tested in 1 dataset because of legal constraints. Therefore, additional studies with other hospitals should be performed to validate the generalizability of our approach. Our labeling algorithm is not 100% accurate. To avoid errors, more rules could be created by exploiting textual data using natural language processing. Indeed, to build our algorithm, we only used structured data because of the unavailability of an adequate tool to treat textual data in our hospital.

Conclusions

This work suggests a new approach to preprocess data on pathways of unscheduled patients. To our knowledge, there are no other studies that have evaluated nonspecific disease pathways. Our approach has the advantages of being explicable, simple to implement, and adaptable to each hospital.

Future research could develop process discovery techniques that consider the relevance labels of the activities.

Acknowledgments

The authors thank the Groupe Hospitalier Bretagne Sud, especially the Groupe Hospitalier Bretagne Sud Information System Department teams and the SIB Company Data Department teams, for providing access to electronic health records for this study. The authors would like to thank the physicians who helped them construct and validate their algorithms, including Dr Bry, Dr Chevallier, Dr Dollon, Dr La Combe, Dr Le Corf, Dr Girard, Dr Henry, Dr Laurichesse, Dr Le Merlay, Dr Lenoir, Dr Luquet, Dr Maigre, and Dr Vaillant. The authors also thank Dr Quiguer for her advice and Professor Saulnier for his help with the statistical analyses. This work was funded by the company Enovacom.

Conflicts of Interest

This work was funded by the company Enovacom where two authors are employed, Fanny Jardinaud and Laura Uhl. The authors have no other competing interests to declare that are relevant to the content of this paper.

Multimedia Appendix 1
Additional definitions.
[PDF File (Adobe PDF File), 163 KB - medinform_v12i1e58978_app1.pdf]

Multimedia Appendix 2
Algorithms of the rules.
[PDF File (Adobe PDF File), 574 KB - medinform_v12i1e58978_app2.pdf]

Multimedia Appendix 3
Satistical analysis results.
[PDF File (Adobe PDF File), 645 KB - medinform_v12i1e58978_app3.pdf]

Multimedia Appendix 4
Process models.
[PDF File (Adobe PDF File), 1064 KB - medinform_v12i1e58978_app4.pdf]

Multimedia Appendix 5
Simulation parameterization.
[PDF File (Adobe PDF File), 681 KB - medinform_v12i1e58978_app5.pdf]

References

1. Bernard A, Boichut M, Descouts A. Bed manager: mission, profil et activité? Ecole Des Hautes Études en Santé Publique (EHESP). 2019. URL: https://documentation.ehesp.fr/memoires/2019/mip/groupe%206.pdf [accessed 2024-09-05]
2. Munoz-Gama J, Martin N, Fernandez-Llatas C, Johnson OA, Sepúlveda M, Helm E, et al. Process mining for healthcare: characteristics and challenges. J Biomed Inform 2022 Mar;127:103994 [FREE Full text] [doi: 10.1016/j.jbi.2022.103994] [Medline: 35104641]
3. Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: a literature review. J Biomed Inform 2016 Jun;61:224-236 [FREE Full text] [doi: 10.1016/j.jbi.2016.04.007] [Medline: 27109932]
4. Selker HP, Beshansky JR, Pauker SG, Kassirer JP. The epidemiology of delays in a teaching hospital. The development and use of a tool that detects unnecessary hospital days. Med Care 1989 Feb;27(2):112-129 [FREE Full text] [doi: 10.1097/00005650-198902000-00003] [Medline: 2918764]
5. El-Bouri R, Eyre DW, Watkinson P, Zhu T, Clifton DA. Hospital admission location prediction via deep interpretable networks for the year-round improvement of emergency patient care. IEEE J Biomed Health Inform 2021 Jan;25(1):289-300 [FREE Full text] [doi: 10.1109/JBHI.2020.2990309] [Medline: 32750898]
6. Franck T, Bercelli P, Aloui S, Augusto V. A generic framework to analyze and improve patient pathways within a healthcare network using process mining and discrete-event simulation. In: Proceedings of the 2020 Winter Simulation Conference. 2020 Presented at: WSC 2020; December 14-18, 2020; Orlando, FL. [doi: 10.1109/wsc48552.2020.9384021]
7. Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data quality in health research: integrative literature review. J Med Internet Res 2023 Oct 31;25:e41446 [FREE Full text] [doi: 10.2196/41446] [Medline: 37906223]
8. van der Aalst · W. Process Mining: Data Science in Action. Berlin, Germany: Springer; 2016.
9. Song M, Günther CW, van der Aalst WM. Trace clustering in process mining. In: Proceedings of the Business Process Management Workshops. 2008 Presented at: BPM 2008; September 1-4, 2008; Milano, Italy. [doi: 10.1007/978-3-642-00328-8_11]
10. Bose RP, van der Aalst WM. Trace clustering based on conserved patterns: towards achieving better process models. In: Proceedings of the Business Process Management Workshops. 2009 Presented at: BPM 2009; September 7, 2009; Ulm, Germany. [doi: 10.1007/978-3-642-12186-9_16]
11. Bose RP, van der Aalst WM. Context aware trace clustering: towards improving process mining results. In: Proceedings of the 2009 SIAM International Conference on Data Mining. 2009 Presented at: SDM 2009; April 30-May 2, 2009; Sparks, NV. [doi: 10.1137/1.9781611972795.35]
12. Delias P, Doumpos M, Grigoroudis E, Manolitzas P, Matsatsinis N. Supporting healthcare management decisions via robust clustering of event logs. Knowl Based Syst 2015 Aug;84:203-213. [doi: 10.1016/j.knosys.2015.04.012]
13. Veiga GM, Ferreira DR. Understanding spaghetti models with sequence clustering for ProM. In: Proceedings of the Business Process Management Workshops. 2009 Presented at: BPM 2009; September 7, 2009; Ulm, Germany. [doi: 10.1007/978-3-642-12186-9_10]
14. De Weerdt J, vanden Broucke S, Vanthienen J, Baesens B. Active trace clustering for improved process discovery. IEEE Trans Knowl Data Eng 2013 Dec;25(12):2708-2720. [doi: 10.1109/TKDE.2013.64]
15. Greco G, Guzzo A, Pontieri L, Sacca D. Discovering expressive process models by clustering log traces. IEEE Trans Knowl Data Eng 2006 Aug;18(8):1010-1027. [doi: 10.1109/TKDE.2006.123]
16. Elghazel H, Deslandres V, Kallel K, Dussauchoy A. Clinical pathway analysis using graph-based approach and Markov models. In: Proceedings of the 2nd International Conference on Digital Information Management. 2007 Presented at: ICDIM 2007; October 28-31, 2007; Lyon, France. [doi: 10.1109/icdim.2007.4444236]
17. Huang Z, Dong W, Duan H, Li H. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. IEEE J Biomed Health Inform 2014 Jan;18(1):4-14. [doi: 10.1109/JBHI.2013.2274281] [Medline: 24403398]
18. Fahland D, van der Aalst WM. Simplifying mined process models: an approach based on unfoldings. In: Proceedings of the 9th International Conference on Business Process Management. 2011 Presented at: BPM 2011; August 30-September 2, 2011; Clermont-Ferrand, France. [doi: 10.1007/978-3-642-23059-2_27]
19. De Roock E, Martin N. Process mining in healthcare - an updated perspective on the state of the art. J Biomed Inform 2022 Mar;127:103995 [FREE Full text] [doi: 10.1016/j.jbi.2022.103995] [Medline: 35077900]
20. Ly LT, Indiono C, Mangler J, Rinderle-Ma S. Data transformation and semantic log purging for process mining. In: Proceedings of the 24th International Conference on Advanced Information Systems Engineering. 2012 Presented at: CAiSE 2012; June 25-29, 2012; Gdansk, Poland. [doi: 10.1007/978-3-642-31095-9_16]
21. van Zelst SJ, Mannhardt F, de Leoni M, Koschmider A. Event abstraction in process mining: literature review and taxonomy. Granul Comput 2020 May 27;6(3):719-736. [doi: 10.1007/S41066-020-00226-2]
22. Martin N, Martinez-Millana A, Valdivieso B, Fernández-Llatas C. Interactive data cleaning for process mining: a case study of an outpatient clinic's appointment system. In: Proceedings of the International Workshops on Business Process Management Workshops. 2019 Presented at: BPM 2019; September 1-6, 2019; Vienna, Austria. [doi: 10.1007/978-3-030-37453-2_43]
23. Dixit PM, Suriadi S, Andrews R, Wynn MT, ter Hofstede AH, Buijs JC, et al. Detection and interactive repair of event ordering imperfection in process logs. In: Proceedings of the 30th International Conference on Advanced Information

Systems Engineering. 2018 Presented at: CAiSE 2018; June 11-15, 2018; Tallinn, Estonia. [doi: 10.1007/978-3-319-91563-0_17]
24. Rogge-Solti A, Mans RS, van der Aalst WM, Weske M. Repairing event logs using timed process models. In: Proceedings of the On the Move to Meaningful Internet Systems. 2013 Presented at: OTM 2013; September 9-13, 2013; Graz, Austria. [doi: 10.1007/978-3-642-41033-8_89]
25. Andrews R, Wynn MT, Vallmuur K, Ter Hofstede AH, Bosley E, Elcock M, et al. Leveraging data quality to better prepare for process mining: an approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in Queensland. Int J Environ Res Public Health 2019 Mar 29;16(7):1138 [FREE Full text] [doi: 10.3390/ijerph16071138] [Medline: 30934913]
26. Bose RP, Mans RS, van der Aalst WM. Wanna improve process mining results? In: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining. 2013 Presented at: CIDM 2013; April 16-19, 2013; Singapore, Singapore. [doi: 10.1109/cidm.2013.6597227]
27. van Eck ML, Lu X, Leemans SJ, van der Aalst WM. PM²: a process mining project methodology. In: Proceedings of the Advanced Information Systems Engineering. 2015 Presented at: CAiSE 2015; June 8-12, 2015; Stockholm, Sweden. [doi: 10.1007/978-3-319-19069-3_19]
28. Vanbrabant L, Martin N, Ramaekers K, Braekers K. Quality of input data in emergency department simulations: framework and assessment techniques. Simul Model Pract Theory 2019 Feb;91:83-101. [doi: 10.1016/j.simpat.2018.12.002]
29. Verhulst R. Evaluating quality of event data within event logs: an extensible framework. Eindhoven University of Technology. 2016 Aug 31. URL: https://research.tue.nl/en/studentTheses/evaluating-quality-of-event-data-within-event-logs [accessed 2024-09-05]
30. Prodel M, Augusto V, Jouaneton B, Lamarsalle L, Xie X. Optimal process mining for large and complex event logs. IEEE Trans Automat Sci Eng 2018 Jul;15(3):1309-1325. [doi: 10.1109/tase.2017.2784436]
31. Prodel M, Augusto V, Xie X, Jouaneton B, Lamarsalle L. Stochastic simulation of clinical pathways from raw health databases. In: Proceedings of the 13th IEEE Conference on Automation Science and Engineering (CASE). 2017 Presented at: CASE 2017; August 20-23, 2017; Xi'an, China. [doi: 10.1109/coase.2017.8256167]
32. Abohamad W, Ramy A, Arisha A. A hybrid process-mining approach for simulation modeling. In: Proceedings of the 2017 Winter Simulation Conference. 2017 Presented at: WSC 2017; December 3-6, 2017; Las Vegas, NV. [doi: 10.1109/wsc.2017.8247894]
33. Wood RM, Murch BJ. Modelling capacity along a patient pathway with delays to transfer and discharge. J Oper Res Soc 2019 May 28;71(10):1530-1544. [doi: 10.1080/01605682.2019.1609885]
34. Karakra A, Lamine E, Fontanili F, Lamothe J. HospiT'Win: a digital twin framework for patients' pathways real-time monitoring and hospital organizational resilience capacity enhancement. In: Proceedings of the 9th International Workshop on Innovative Simulation for Healthcare. 2020 Presented at: IWISH 2020; September 16-18, 2020; Online. [doi: 10.46354/i3m.2020.iwish.012]
35. Holm LB, Lurås H, Dahl FA. Improving hospital bed utilisation through simulation and optimisation: with application to a 40% increase in patient volume in a Norwegian General Hospital. Int J Med Inform 2013 Feb;82(2):80-89. [doi: 10.1016/j.ijmedinf.2012.05.006] [Medline: 22698645]
36. Demir E, Gunal MM, Southern D. Demand and capacity modelling for acute services using discrete event simulation. Health Syst 2017 Dec 19;6(1):33-40. [doi: 10.1057/hs.2016.1]
37. Ordu M, Demir E, Tofallis C, Gunal MM. A comprehensive and integrated hospital decision support system for efficient and effective healthcare services delivery using discrete event simulation. Healthc Anal 2023 Dec;4:100248. [doi: 10.1016/j.health.2023.100248]
38. ProM Tools homepage. ProM Tools. URL: https://promtools.org/ [accessed 2022-11-22]
39. vanden Broucke SK, De Weerdt J. Fodina: a robust and flexible heuristic process discovery technique. Decis Support Syst 2017 Aug;100:109-118. [doi: 10.1016/j.dss.2017.04.005]
40. Lassen KB, van der Aalst WM. Complexity metrics for workflow nets. Inf Softw Technol 2009 Mar;51(3):610-626. [doi: 10.1016/j.infsof.2008.08.005]
41. Mendling J. Testing density as a complexity metric for EPCs. Vienna University of Economics and Business Administration. 2006. URL: https://www.researchgate.net/publication/228347008_Testing_density_as_a_complexity_metric_for_EPCs [accessed 2024-09-05]
42. Uhl L, Augusto V, Lemaire V, Alexandre Y, Jardinaud F, Bercelli P, et al. Progressive prediction of hospitalisation and patient disposition in the emergency department. In: Proceedings of the IEEE International Conference on Big Data. 2022 Presented at: IEEE BigData 2022; December 17-20, 2022; Osaka, Japan.

Abbreviations

DES: discrete event simulation
ECaM: extended Cardoso metric
ECyM: extended Cyclomatic metric

ED: emergency department
EHR: electronic health record
GHBS: Groupe Hospitalier Bretagne Sud
ICU: intensive care unit
LoS: length of stay
MCO: medicine, surgery, obstetrics, and odontology
OU: observation unit

Edited by J Hefner; submitted 29.03.24; peer-reviewed by M Ordu, JL Raisaro; comments to author 14.05.24; revised version received 08.07.24; accepted 10.07.24; published 23.09.24.

Please cite as:
Uhl L, Augusto V, Dalmas B, Alexandre Y, Bercelli P, Jardinaud F, Aloui S
Evaluating the Bias in Hospital Data: Automatic Preprocessing of Patient Pathways Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e58978
URL: https://medinform.jmir.org/2024/1/e58978
doi:10.2196/58978
PMID:39312289

Pediatric Sedation Assessment and Management System (PSAMS) for Pediatric Sedation in China: Development and Implementation Report

Ziyu Zhu[1,*], MBBS; Lan Liu[2,*], MD; Min Du[2,*], MD; Mao Ye[2], MD; Ximing Xu[1], PhD; Ying Xu[2,*], PhD

[1]

[2]

[*]these authors contributed equally

Corresponding Author:
Ying Xu, PhD

Abstract

Background: Recently, the growing demand for pediatric sedation services outside the operating room has imposed a heavy burden on pediatric centers in China. There is an urgent need to develop a novel system for improved sedation services.

Objective: This study aimed to develop and implement a computerized system, the Pediatric Sedation Assessment and Management System (PSAMS), to streamline pediatric sedation services at a major children's hospital in Southwest China.

Methods: PSAMS was designed to reflect the actual workflow of pediatric sedation. It consists of 3 main components: server-hosted software; client applications on tablets and computers; and specialized devices like gun-type scanners, desktop label printers, and pulse oximeters. With the participation of a multidisciplinary team, PSAMS was developed and refined during its application in the sedation process. This study analyzed data from the first 2 years after the system's deployment.

Implementation (Results): From January 2020 to December 2021, a total of 127,325 sedations were performed on 85,281 patients using the PSAMS database. Besides basic variables imported from Hospital Information Systems (HIS), the PSAMS database currently contains 33 additional variables that capture comprehensive information from presedation assessment to postprocedural recovery. The recorded data from PSAMS indicates a one-time sedation success rate of 97.1% (50,752/52,282) in 2020 and 97.5% (73,184/75,043) in 2021. The observed adverse events rate was 3.5% (95% CI 3.4% - 3.7%) in 2020 and 2.8% (95% CI 2.7%-2.9%) in 2021.

Conclusions: PSAMS streamlined the entire sedation workflow, reduced the burden of data collection, and laid a foundation for future cooperation of multiple pediatric health care centers.

(JMIR Med Inform 2024;12:e53427) doi:10.2196/53427

KEYWORDS

electronic data capture; information systems; pediatric sedation; sedation management; workflow optimization

Introduction

Context

Procedural sedation in infants and children is in great demand for anxiety, pain, and motor control. Over the past decades, pediatric sedation has evolved into a multispecialty practice, with guidelines established by the American Society of Anesthesiologists (ASA), the American Academy of Pediatrics, and the International Committee for the Advancement of Procedural Sedation . However, there is currently no universally accepted optimal strategy for pediatric procedural sedation. Factors such as medical resources, patient volume, and health insurance systems may contribute to variations in criteria for administering sedation, methods for monitoring depth of sedation, qualifications of professionals performing sedation, and the choice of sedative agents [1-3]. The Children's Hospital of Chongqing Medical University (CHCMU) is the largest pediatric center in Southwest China. The CHCMU had 1400 inpatient beds in 2019, which increased to 2480 in 2021, with more than 3 million outpatient visits annually [4]. The huge patient volume at CHCMU highlights the need for an electronic data capture and management system (EDCMS) that seamlessly integrates multiple sedation steps. Such a system ensures reliability and consistency, reduces documentation time, and allows more time and resources to be devoted to sedation services [4,5]. In addition, the system can flag high-risk patients (eg, ASA III patients) and monitor their sedation process in real time, facilitating quicker detection of potential risks. Smaller hospitals may also find this system beneficial, as it can enhance work efficiency and enable proper data management for comprehensive assessment and tracking of all records.

After a 2-month pilot deployment, the feasibility of the current system was proven, and it has been fully implemented since January 2020.

Problem Statement

The demand for procedural sedation has significantly risen due to increased awareness of its importance [6-8]. Nevertheless, the current anesthesia system employed inside the operating room is incompatible with pediatric sedation outside the operating room. Other challenges include insufficient data collection and nonuniform management, which were initially deemed unavoidable owing to the large number of patients and limited pediatric personnel and resources for pediatric care. Moreover, a low ratio of medical personnel to patient is a long-standing issue for procedural sedation in many limited-income countries in Africa and Asia [9] as well as in many European countries [10].

Similar Interventions

Traditional tools, such as Research Electronic Data Capture (REDCap) [11], were developed for data collection and lacked the features necessary to manage sedation services. Similarly, anesthesia systems designed for the surgical settings do not align with the requirements of outpatient sedation services. The newly implemented Pediatric Sedation Assessment and Management System (PSAMS) was designed to meet the unique needs of pediatric sedation services, optimizing both data management and clinical workflow.

Methods

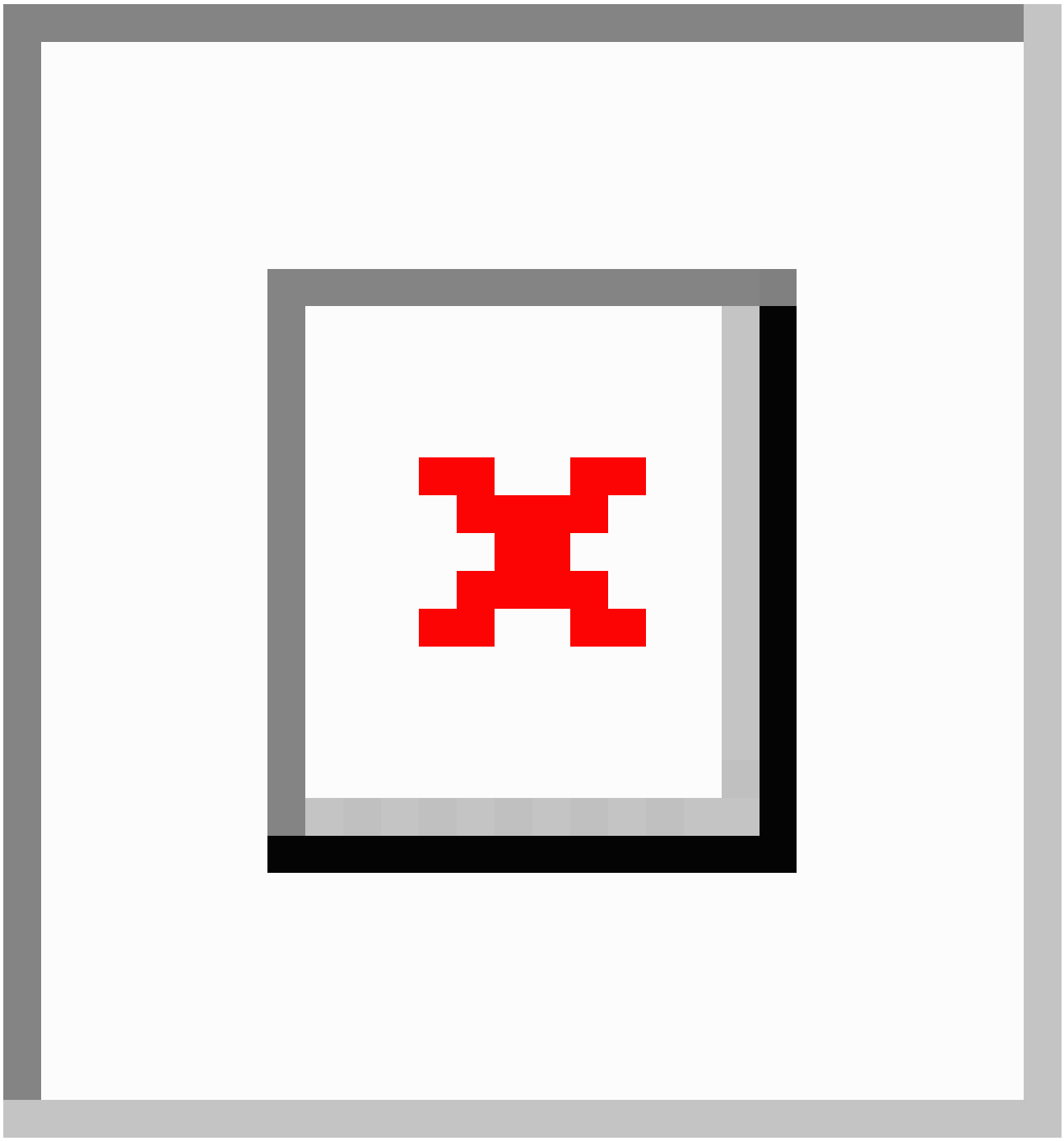
Aims and Objectives

In this study, we aimed to develop a system that could meet the needs of pediatric sedation services and optimize both data management and clinical workflow. First, we introduced the development and implementation of the PSAMS. Second, we analyzed the one-time sedation success rate and the incidence of adverse events based on the data collected by the PSAMS.

Blueprint Summary

PSAMS was developed at the CHCMU based on the following principles: (1) user-friendly interfaces and features for anesthesiologists and nurses; (2) high accuracy with carefully designed error-proofing techniques; (3) interoperability and integration with the Hospital Information System (HIS); (4) compatible scalability reserved for future updates and to accommodate different deployments; and (5) solid security assurance on patients' privacy. This system comprises 3 main parts: the software hosted on a server; clients distributed on portable tablets and computers; and devices, including gun-type scanners, desktop label printers, and pulse oximeters (Figure 1). This study adhered to the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) guidelines (Checklist 1) [12].

Figure 1. Devices and user interfaces of the Pediatric Sedation Assessment and Management System (PSAMS). All tasks are assigned to 4 roles by PSAMS: nurse A, anesthesiologist, nurse B, and nurse C. PSAMS is distributed on computers and portable tablets, and every user interface is specifically designed based on the different tasks of each user role. A large display in the sedation center provides a real-time update summary and statistics generated by PSAMS.



Technical Design

The in-depth technical details of the construction of the PSAMS are available from the correspondence author upon reasonable request. Briefly, most of the system was written in C++; the Qt framework was used for application and graphical user interface development, and Java was used for back-end data processing. MySQL was applied as the database management system.

Target

PSAMS has demonstrated great feasibility since its full implementation in 4 sedation centers of CHCMU in 2 districts of Chongqing from January 2020. PSAMS can be easily customized and adapted to other health care systems to provide sedation services for all pediatric patients (≤18 years of age). As of December 2022, it has been adopted by other health care centers in China, including Shenzhen Children's Hospital in Guangdong Province and Zhengzhou Children's Hospital in Henan Province, following team training and pilot implementation.

Ethical Considerations

This study was approved by the Institutional Review Board of the Children's Hospital of Chongqing Medical University (file

number 2022, 220), and all data analyzed in this study were collected in accordance with its regulations. Informed consent was waived due to the retrospective observational nature of the study. Data will be stored on the hospital's internal servers and managed by the hospital.

Data

A comprehensive approach was implemented to ensure data quality. For data validation and verification, we designed and used prestructured data entry as drop-down menus or checkboxes. Other mandatory basic information fields were automatically synchronized with the HIS. Each data entry was immediately validated against predefined ranges or criteria. If an anomalous entry was detected (eg, 32 °C for body temperature or 400 kg for body weight), PSAMS would send notifications and remind the user to confirm the entry while recording the incident in the system's running log, which also served as a dataset for analyzing error patterns and system performance issues over time. To further improve data integrity and support clinical decision making, PSAMS included advanced tools for automated dose calculation and sedation regimen recommendations, while aligning with best clinical practices. For example, the regimen recommendation tool was developed based on sedation practice and expert guidelines, allowing the anesthesiologist to accept or modify recommendations based on real-world scenarios. The regimen recommendation tool incorporated expert knowledge through a set of predefined rules considering factors such as age, weight, diagnosis, and procedure type. All tools were regularly updated and maintained based on user feedback and updated guidelines (Figure S1 in Multimedia Appendix 1). For adverse events reporting, the adverse event reporting tool from the World SIVA International Sedation Task Force [13] was integrated into PSAMS to facilitate early detection and provide standardized records.

Interoperability

The PSAMS is integrated with the HIS, which applies the *International Classification of Diseases, Tenth Revision*. This interoperability ensures seamless data exchange and standardized coding for diagnoses, enhancing the accuracy and efficiency of medical records management within the system.

Participating Entities

The multidisciplinary team included anesthesiologists and nurses from the Department of Anesthesiology, staff from the Department of Health Information Management, and software engineers. The Department of Health Information Management provided server, network, and HIS integration support. The software engineers created a prototype and refined the system based on feedback from the anesthesiologists and nurses during clinical practice. The funding body did not participate in the study design, implementation, or data governance.

Budget Planning

Approximately 54% of the budget was spent on software development, 10% on change management, 25% on project management, 5% on user training, and 6% on product deployment. The development phase took 1 year in addition to a 2-month pilot deployment.

Sustainability

PSAMS has been integrated into the HIS and used in several pediatric centers for procedural sedation. Therefore, PSAMS was maintained through the hospital budget and the remaining funds were dedicated to the active development and improvement of PSAMS.

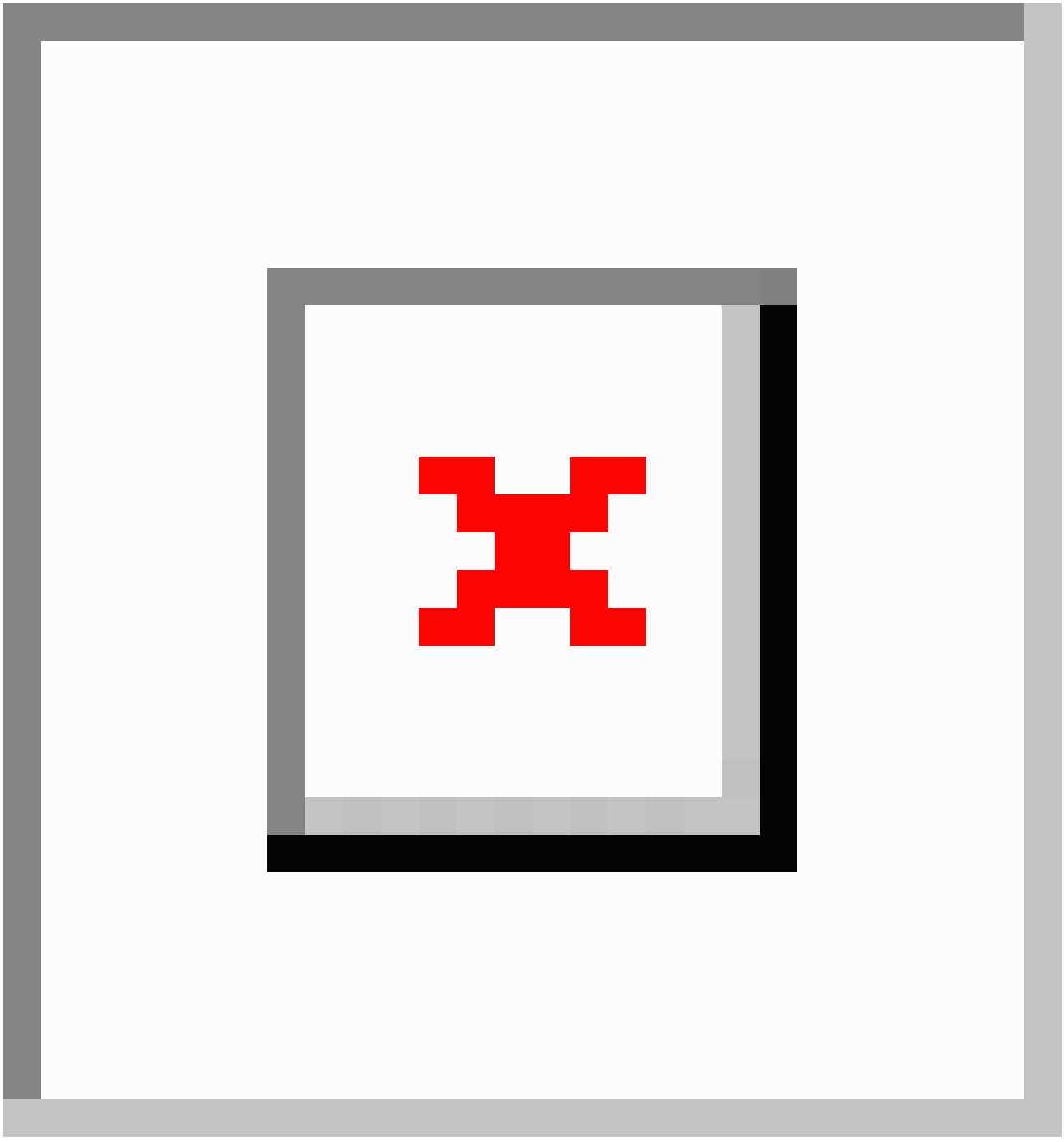
Implementation (Results)

Implementation of PSAMS

Figure 2 shows the PSAMS-mediated sedation workflow. Patients arriving at the nurses' station were registered by nurse A using a gun-type barcode scanner. This allowed automatic registration and synchronization of basic patient information (eg, name, age, gender, and ID) from the HIS. Previously, this process required manual entry of multiple forms, which typically took at least several minutes per patient. Following registration, nurse A quickly measured the patients' body weight, temperature, pulse, and saturation of peripheral oxygen (SpO_2), and completed the guidance sheet in PSAMS. The anesthesiologist then performed a presedation assessment based on physical examination and directed history (Table S1 in Multimedia Appendix 1). All relevant medical records were readily accessible via synchronization with the HIS, which was a significant improvement over the previous workflow, where data were isolated and difficult to retrieve. Patients deemed suitable for sedation who had signed informed consent were then directed to designated waiting or sedation areas.

From sedation to postprocedure recovery, pulse rate, and oxygen saturation were routinely recorded for ASA I and II patients at 4 time points: drug administration, sedation depth evaluation, procedure completion, and recovery from the procedure. Patients with ASA III and above received continuous monitoring of pulse rate and oxygen saturation by an experienced nurse with advanced life-support skills. PSAMS also tracked each step of the sedation process, displaying patient status on nurse dashboard panels, enabling nurse B to efficiently manage the patient queue and address any adverse events or complications promptly. In the case of an adverse event occurrence, PSAMS instantly alerted all screens within the sedation center, allowing the emergency team to quickly identify the potential reason and take appropriate measures to manage the adverse events. Meanwhile, the sedation records of the patients were immediately archived, synchronized, and reported to the department. Nurse C administered the drugs and logged the details into PSAMS. After the procedure, nurse C repeated the physical examinations and evaluated the recovery status. Overall, PSAMS had integrated previously independent sedation steps into a unified workflow, while maintaining and prioritizing important information, including vital parameters and sedation status at each stage.

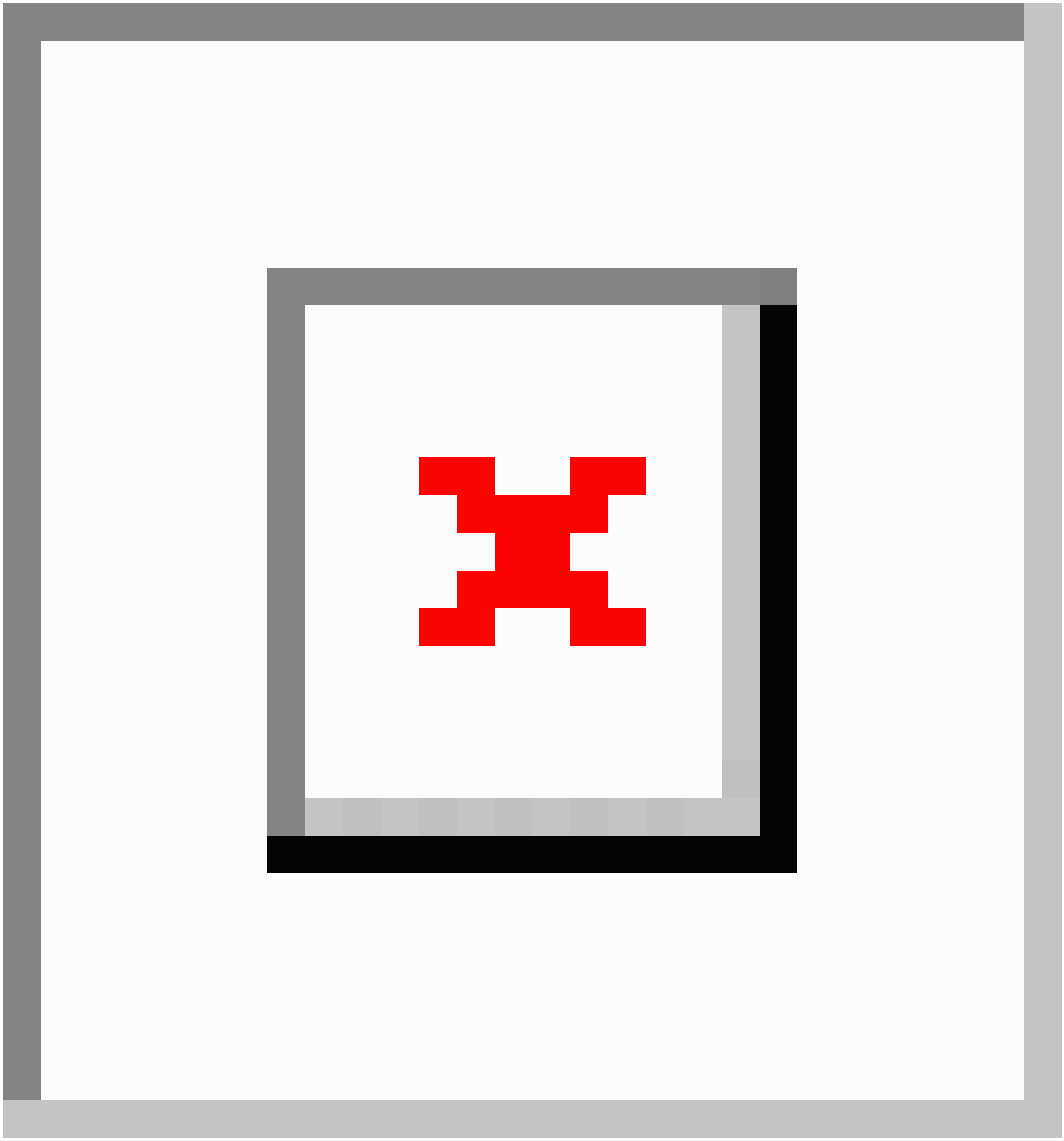
Figure 2. Complete sedation workflow mediated by the Pediatric Sedation Assessment and Management System in Children's Hospital of Chongqing Medical University. HIS: Hospital Information System; PR:pulse rate; SpO_2: saturation of peripheral oxygen.



Overview of the Current Database of PSAMS

For a 2-year deployment from 2020 to 2021, a total of 127,325 sedations were performed in 85,281 patients (Figure 3A and B). In all, patients came from 31 provinces and municipalities in China (Table S2 in Multimedia Appendix 1). The year 2020 saw fewer sedations due to the COVID-19 pandemic, and February was notably affected by the Chinese lunar new year. We included 43 variables from presedation assessment (eg, body weight, temperature, and pulse) to postprocedural recovery (eg, postprocedural pulse, SpO_2, and recovery time) for every single record (Table S3 in Multimedia Appendix 1). Of the 43 variables, 19 were complete in all records (Figure 3). High missing rates in some variables were due to the fact that most sedations did not require further intervention.

Figure 3. Overview of the sedation database from 2020 to 2021. (A) Calendar heatmap showing daily performed sedations from 2020 to 2021; (B) monthly summary of sedations from 2020 to 2021; (C) missing rate of each included variable in the database of the Pediatric Sedation Assessment and Management System. ASA: American Society of Anesthesiologists; SpO_2: saturation of peripheral oxygen.



The demographic characteristics and ASA levels of the patients are summarized in Table 1. The mean age was 2.2 (SD 1.7) years, and the mean weight was 11.7 (SD 4.8) kg. The proportion of male patients (n=76,001, 59.7%) was higher than that of female patients (n=51,324, 40.3%). .

Table . Demographic characteristics and the American Society of Anesthesiologists (ASA) levels.

Demographic characteristics		Total (N=127,325)	Year 2020 (n=52,282, 41.1%)	Year 2021 (n=75,043, 58.9%)
Age (year)				
	Mean (SD)	2.2 (1.7)	2.1 (1.8)	2.3 (1.7)
	Range	0.0-18.0	0.0-17.9	0.0-18.0
Gender, n (%)				
	Female	51,324 (40.3)	21,325 (40.8)	29,999 (40.0)
	Male	76,001 (59.7)	30,957 (59.2)	45,044 (60.0)
Weight (kg)				
	Mean (SD)	11.7 (4.8)	11.5 (4.9)	11.9 (4.8)
	Range	1.0-61.0	1.5-61.0	1.0-60.0
ASA, n (%)				
	I	10,495 (8.2)	9134 (17.5)	1361 (1.8)
	II	108,136 (84.9)	39,196 (75.0)	68,940 (91.9)
	III	8568 (6.7)	3845 (7.4)	4723 (6.3)
	IV	125 (0.1)	106 (0.2)	19 (0.0)
	V	1 (0.0)	1 (0.0)	—[a]

[a]Not applicable.

Sedative Choice, Procedures, Success Rate, and Adverse Events

PSAMS comprehensively collected and maintained information regarding the one-time sedation success rate and adverse events. The one-time sedation success was defined as achieving appropriate sedation depth with a single drug administration without adverse events. The most commonly used sedation regimen was dexmedetomidine combined with chloral hydrate (n=85,337, 67%) with a one-time sedation success rate of 96.6% (82,427/85,337; Table S4 in Multimedia Appendix 1). The most common procedure performed after sedation was magnetic resonance imaging (MRI; n=27,416, 21.5%), with the highest one-time sedation success rate of 99.7% (27,327/27,416; Figure S2 in Multimedia Appendix 1). The overall one-time sedation success rate was 97.1% (50,752/52,282) in 2020, which increased to 97.5% (73,184/75,043) in 2021. However, we were unable to summarize the one-time sedation success rate in 2019 and earlier because of the large volume of disorganized documents generated from the paper-based workflow and limited availability of personnel. Adverse events were recorded and reported using the adverse event reporting tool from the World SIVA International Sedation Task Force [13]. Most minimal risk adverse events, such as vomiting and hypersalivation, were not tracked and recorded by the PSAMS due to incompatibility with the monitoring devices. Minor risk adverse events were reported in 3397 sedations, where only 2 major adverse events were reported (Table 2).

Table . Adverse events reported in the Pediatric Sedation Assessment and Management System from 2020 to 2021.

Adverse events		Total (N=3962), n (%; 95% CI)	2020 (n=1843), n (%; 95% CI)	2021 (n=2119), n (%; 95% CI)
Minimal adverse events				
	Vomiting/retching	2 (0.1; 0-0.1)	1 (0; 0-0.2)	1 (0.1; 0-0.1)
	Subclinical respiratory depression	—[a]	—	—
	Hypersalivation	2 (0.1; 0-0.1)	2 (0.1; 0-0.3)	—
	Paradoxical response	—	—	—
	Recovery agitation	8 (0.2; 0.1-0.3)	8 (0.4; 0.1-0.7)	—
	Prolonged recovery	551 (13.9; 12.8-15.0)	295 (16; 14.3-17.7)	256 (12.1; 10.7-13.5)
Minor adverse events				
	Oxygen desaturation (75% - 90%) for <60 s	4 (0.1; 0-0.2)	2 (0.1; 0-0.3)	2 (0.1; 0-0.2)
	Apnea—not prolonged	—	—	—
	Airway obstruction	2 (0.1; 0-0.1)	2 (0.1; 0-0.3)	—
	Failed sedation	3389 (85.5; 84.4-86.6)	1530 (83; 81.3-84.7)	1859 (87.7; 86.3-89.1)
	Allergic reaction without asphyxia	—	—	—
	Bradycardia	1 (0; 0-0.1)	1 (0.1; 0-0.2)	—
	Tachycardia	1 (0; 0-0.1)	1 (0.1; 0-0.2)	—
	Hypotension	—	—	—
	Hypertension	—	—	—
	Seizure	—	—	—
Major adverse events				
	Oxygen desaturation—severe (<75% at any time) or prolonged (<90% for >60 s)	2 (0.1; 0-0.1)	1 (0.1; 0-0.2)	1 (0.1; 0-0.1)
	Apnea—prolonged (>60 s)	—	—	—
	Cardiovascular collapse/shock	—	—	—
	Cardiac arrest/absent pulse	—	—	—

[a]Not applicable.

Discussion

Principal Results

We developed the PSAMS through a multidisciplinary collaboration. To the best of our knowledge, This is the first EDCMS, specifically designed and used for pediatric sedation in China. PSAMS is deeply integrated with HIS. Thus, anesthesiologists could systematically determine a patient's medical status and optimize the sedation techniques. In case of an adverse outcome or complication, PSAMS immediately called for backup assistance and reported the event. All the generated data were recorded and tracked by PSAMS. When transitioning to PSAMS, the biggest challenge for users was adapting to the new system, which was resolved after a 2-week training program. During this time, users encountered specific problems, such as unfamiliarity with the interface navigation, difficulty entering and saving data, or not knowing how to use the built-in tools. To address these challenges, we provided targeted support in the form of quick reference guides and on-demand lectures to facilitate a smoother transition. Meanwhile, PSAMS continues to be updated to improve usability and better meet the needs of users based on their feedback. Due to the previous poorly documented paper-based system, quantitatively comparison was unavailable, and a preliminary survey was conducted to evaluate the improvement after PSAMS implementation. Most users (19 out of 22 nurses and all 5 anesthesiologists) reported a reduction in administrative burdens, convenient access to patient data, and a smoother workflow, which allowed more focus on patient care per se. Overall, PSAMS has streamlined the entire workflow and aligned each member of the sedation team within a cohesive and structured process to ensure high-quality and efficient pediatric sedation.

Limitations and Lessons Learned

PSAMS has several limitations. First, we did not perform real-time monitoring of every patient's vital parameters, such as body temperature, respiration rate, pulse, and blood pressure, during the entire sedation process. In our clinical practice, current physiological monitoring systems failed to function while transferring patients and during examinations. To overcome this hurdle, new wearable monitoring devices, a highly networked information system, and a robust computing resource are necessary to achieve real-time monitoring. We are actively engaged in developing innovative wearable devices that integrate seamlessly with PSAMS and improve server performance with advanced algorithms and models. Second, many minimal adverse events were not recorded, and we believe that PSAMS can fill in this missing information after the implementation of new devices. Third, this study was performed at a single center; however, efforts are being made to extend PSAMS to other health care centers in China to prove its generalizability. Nevertheless, necessary adjustments and training are still required to ensure successful deployment of PSAMS in various settings. Additionally, PSAMS is vulnerable to mishandled data from the HIS, which means that the current system cannot yet verify the data imported from other databases.

In summary, valuable lessons have been learned from the PSAMS program. The key success factor is ensuring smooth communication within the multidisciplinary team, which keeps development on track and enables the realization of practical and user-friendly features. Additionally, allocating sufficient time for user training in advance is essential, as it accelerates the transition from the previous paper-based workflow. Finally, maintaining device redundancy is crucial, as devices such as pulse oximeters and tablets can be accidentally damaged during long-term deployment.

Comparison With Prior Work

The advent of EDCMS and its integration with HIS has transitioned paper-based clinical workflow to an electronic one, significantly improved efficiency and data quality [14]. However, commonly used EDCMSs, such as REDCap, which provides hundreds of preconfigured forms to facilitate data capture, unfortunately, are not suitable for clinical sedation practice [15]. In our previous paper-based workflow, each step was separate, with no system for integrating and processing the information in one place. Nurses manually signed piles of medical paperwork, recorded patient status on notepads, and entered the information into a computer. Anesthesiologists performed sedation and spent more time preparing presedation assessment because of the cumbersome information exchange. The previous paper-based workflow had several limitations, such as the lack of a computerized system to integrate and process information in one place, cumbersome information exchange, and the inability to have a complete view of the entire sedation process. Consequently, nurses and anesthesiologists were compelled to allocate a significant proportion of their time on administrative tasks rather than patient care, and adverse events could be difficult to identify and were even ignored. Therefore, the application of well-computerized systems has been proposed as an effective way to reduce errors and improve the health care services [16,17].

Moreover, previous sedation databases only recorded data regarding sedation implementation, and the data were updated after a complete sedation workflow [18]. PSAMS included data from presedation assessment to postprocedural recovery and discharge and updated the database in a real-time manner. Consequently, the quickly accumulating data collected by PSAMS provided new insights into pediatric sedation.

At our institution, the most common procedure performed after sedation was MRI (n=27,416, 21.5%). Similarly, a retrospective analysis of 109,947 entries for MRI from the Pediatric Sedation Research Consortium, a large multicenter corroboration database containing more than 600,000 sedations, found that MRI was one of the most common imaging procedures requiring sedation [19]. Procedures like MRI and computed tomography (CT) scans require patients to remain still during imaging to ensure scanning quality, which can be difficult for children [20]. In fact, dexmedetomidine is used as a sole drug for noninvasive procedures, such as MRI and CT scans [21]. However, at our institution, we found that the dexmedetomidine alone cannot induce ideal sedation depth and often causes the children to awaken during or even before the MRI procedure. Therefore, a combination of rapidly titratable drugs and rapid onset drugs has been proposed as a standard regimen to improve sedation efficacy in China [4].

Conclusions

In summary, we developed and applied PSAMS for pediatric sedation to meet the increasing demands. This is the first assessment and management system tailored for pediatric sedation. We are actively maintaining and improving PSAMS to optimize individualized sedation protocols, despite the heavy workload.

Acknowledgments

This study was supported by the Chongqing Medical Scientific Research Project (Joint Project of Chongqing Health Commission and Science and Technology Bureau; 2023DBXM003), the General Project of Technology Innovation and Application Development of Chongqing Science and Technology Bureau (CSTB2022TIAD-GPX0007), and the Program for Youth Innovation in Future Medicine, Chongqing Medical University.

Data Availability

The in-depth technical details of the construction of the PSAMS are available from the correspondence author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional Tables and Figures.
[DOCX File, 2021 KB - medinform_v12i1e53427_app1.docx]

Checklist 1
iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations).
[DOCX File, 24 KB - medinform_v12i1e53427_app2.docx]

References

1. Cravero JP, Havidich JE. Pediatric sedation—evolution and revolution. Pediatr Anesth 2011 Jul;21(7):800-809. [doi: 10.1111/j.1460-9592.2011.03617.x] [Medline: 21585616]
2. Fagin A, Palmieri TL. Considerations for pediatric burn sedation and analgesia. Burns Trauma 2017;5:28. [doi: 10.1186/s41038-017-0094-8] [Medline: 29051890]
3. Green SM, Leroy PL, Roback MG, et al. An international multidisciplinary consensus statement on fasting before procedural sedation in adults and children. Anaesthesia 2020 Mar;75(3):374-385. [doi: 10.1111/anae.14892] [Medline: 31792941]
4. Yuen VMY, Li BL, Xue B, Xu Y, Tse JCK, Lee RSM. Paediatric sedation: the Asian approach—current state of sedation in China. In: Pediatric Sedation Outside of the Operating Room: Springer; 2021:601-613. [doi: 10.1007/978-3-030-58406-1_29]
5. Mason KP. The pediatric sedation service: who is appropriate to sedate, which medications should I use, who should prescribe the drugs, how do I bill? Pediatr Radiol 2008 May;38 (Suppl 2):218-224. [doi: 10.1007/s00247-008-0769-1] [Medline: 18401615]
6. Krauss B, Green SM. Procedural sedation and analgesia in children. Lancet 2006 Mar 4;367(9512):766-780. [doi: 10.1016/S0140-6736(06)68230-5] [Medline: 16517277]
7. American Academy of Pediatrics, American Academy of Pediatric Dentistry, Coté CJ, Wilson S, Work Group on Sedation. Guidelines for monitoring and management of pediatric patients during and after sedation for diagnostic and therapeutic procedures: an update. Pediatrics 2006 Dec;118(6):2587-2602. [doi: 10.1542/peds.2006-2780] [Medline: 17142550]
8. Coté CJ, Wilson S, American Academy of Pediatrics, American Academy of Pediatric Dentistry. Guidelines for monitoring and management of pediatric patients before, during, and after sedation for diagnostic and therapeutic procedures. Pediatrics 2019 Jun;143(6):e20191000. [doi: 10.1542/peds.2019-1000] [Medline: 31138666]
9. Bösenberg AT. Pediatric anesthesia in developing countries. Curr Opin Anaesthesiol 2007 Jun;20(3):204-210. [doi: 10.1097/ACO.0b013e3280c60c78] [Medline: 17479022]
10. Sahyoun C, Cantais A, Gervaix A, et al. Pediatric procedural sedation and analgesia in the emergency department: surveying the current Eeuropean practice. Eur J Pediatr 2021 Jun;180(6):1799-1813. [doi: 10.1007/s00431-021-03930-6] [Medline: 33511466]
11. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
12. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694. [doi: 10.2196/46694] [Medline: 37163336]
13. Mason KP, Green SM, Piacevoli Q, International Sedation Task Force. Adverse event reporting tool to standardize the reporting and tracking of adverse events during procedural sedation: a consensus document from the World SIVA International Sedation Task Force. Br J Anaesth 2012 Jan;108(1):13-20. [doi: 10.1093/bja/aer407] [Medline: 22157446]
14. Hawley S, Yu J, Bogetic N, et al. Digitization of measurement-based care pathways in mental health through REDCap and electronic health record integration: development and usability study. J Med Internet Res 2021 May 20;23(5):e25656. [doi: 10.2196/25656] [Medline: 34014169]
15. Nichols BN, Pohl KM. Neuroinformatics software applications supporting electronic data capture, management, and sharing for the neuroimaging community. Neuropsychol Rev 2015 Sep;25(3):356-368. [doi: 10.1007/s11065-015-9293-x] [Medline: 26267019]
16. Devine EB, Hansen RN, Wilson-Norton JL, et al. The impact of computerized provider order entry on medication errors in a multispecialty group practice. J Am Med Inform Assoc 2010;17(1):78-84. [doi: 10.1197/jamia.M3285] [Medline: 20064806]
17. Radley DC, Wasserman MR, Olsho LE, Shoemaker SJ, Spranca MD, Bradshaw B. Reduction in medication errors in hospitals due to adoption of computerized provider order entry systems. J Am Med Inform Assoc 2013 May 1;20(3):470-476. [doi: 10.1136/amiajnl-2012-001241] [Medline: 23425440]

18. Cravero JP, Blike GT, Beach M, et al. Incidence and nature of adverse events during pediatric sedation/anesthesia for procedures outside the operating room: report from the pediatric sedation research consortium. Pediatrics 2006 Sep;118(3):1087-1096. [doi: 10.1542/peds.2006-0313] [Medline: 16951002]
19. Mallory MD, Travers C, Cravero JP, Kamat PP, Tsze D, Hertzog JH. Pediatric sedation/anesthesia for MRI: results from the pediatric sedation research consortium. J Magn Reson Imaging 2023 Apr;57(4):1106-1113. [doi: 10.1002/jmri.28392] [Medline: 36173243]
20. Bailey MA, Saraswatula A, Dale G, Softley L. Paediatric sedation for imaging is safe and effective in a district general hospital. Br J Radiol 2016;89(1061):20150483. [doi: 10.1259/bjr.20150483] [Medline: 26959609]
21. Mahmoud M, Mason KP. Dexmedetomidine: review, update, and future considerations of paediatric perioperative and periprocedural applications and limitations. Br J Anaesth 2015 Aug;115(2):171-182. [doi: 10.1093/bja/aev226] [Medline: 26170346]

Abbreviations

ASA: American Society of Anesthesiologists
CHCMU: Children's Hospital of Chongqing Medical University
CT: computed tomography
EDCMS: electronic data capture and management system
HIS: Hospital Information System
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
MRI: magnetic resonance imaging
PSAMS: Pediatric Sedation Assessment and Management System
REDCap: Research Electronic Data Capture
SpO2: saturation of peripheral oxygen

Edited by C Perrin; submitted 06.10.23; peer-reviewed by C Chen, M Randriambelonoro, T Hao; revised version received 20.06.24; accepted 26.06.24; published 07.08.24.

Please cite as:
Zhu Z, Liu L, Du M, Ye M, Xu X, Xu Y
Pediatric Sedation Assessment and Management System (PSAMS) for Pediatric Sedation in China: Development and Implementation Report
JMIR Med Inform 2024;12:e53427
URL: https://medinform.jmir.org/2024/1/e53427
doi:10.2196/53427

Impact of an Electronic Health Record–Based Interruptive Alert Among Patients With Headaches Seen in Primary Care: Cluster Randomized Controlled Trial

Apoorva Pradhan[1], BAMS, MPH; Eric A Wright[1,2], MPH, PharmD; Vanessa A Hayduk[1], BS; Juliana Berhane[3], MS; Mallory Sponenberg[4], MSN, RN; Leeann Webster[5], MBA, CDCES, RPh; Hannah Anderson[1], MS; Siyeon Park[6], PharmD; Jove Graham[1], PhD; Scott Friedenberg[7], MD

1

2

3

4

5

6

7

Corresponding Author:
Apoorva Pradhan, BAMS, MPH

Abstract

Background: Headaches, including migraines, are one of the most common causes of disability and account for nearly 20% - 30% of referrals from primary care to neurology. In primary care, electronic health record–based alerts offer a mechanism to influence health care provider behaviors, manage neurology referrals, and optimize headache care.

Objective: This project aimed to evaluate the impact of an electronic alert implemented in primary care on patients' overall headache management.

Methods: We conducted a stratified cluster-randomized study across 38 primary care clinic sites between December 2021 to December 2022 at a large integrated health care delivery system in the United States. Clinics were stratified into 6 blocks based on region and patient-to–health care provider ratios and then 1:1 randomized within each block into either the control or intervention. Health care providers practicing at intervention clinics received an interruptive alert in the electronic health record. The primary end point was a change in headache burden, measured using the Headache Impact Test 6 scale, from baseline to 6 months. Secondary outcomes included changes in headache frequency and intensity, access to care, and resource use. We analyzed the difference-in-differences between the arms at follow-up at the individual patient level.

Results: We enrolled 203 adult patients with a confirmed headache diagnosis. At baseline, the average Headache Impact Test 6 scores in each arm were not significantly different (intervention: mean 63, SD 6.9; control: mean 61.8, SD 6.6; *P*=.21). We observed a significant reduction in the headache burden only in the intervention arm at follow-up (3.5 points; *P*=.009). The reduction in the headache burden was not statistically different between groups (difference-in-differences estimate –1.89, 95% CI –5 to 1.31; *P*=.25). Similarly, secondary outcomes were not significantly different between groups. Only 11.32% (303/2677) of alerts were acted upon.

Conclusions: The use of an interruptive electronic alert did not significantly improve headache outcomes. Low use of alerts by health care providers prompts future alterations of the alert and exploration of alternative approaches.

Trial Registration: ClinicalTrials.gov NCT05067725; https://clinicaltrials.gov/study/NCT05067725

(JMIR Med Inform 2024;12:e58456) doi:10.2196/58456

KEYWORDS

headache management; migraine management; electronic health record–based alerts; primary care; clinician decision support tools; electronic health record; EHR

Introduction

Headache disorders are a major public health concern, with migraine being the second most common cause of disability [1]. The frequency of headaches is nearly twice as common in women, adults aged 18 - 44 years, people with low family income (<US $35,000), and Native Americans [2]. Patients with migraines and chronic headaches experience greater than normal use of emergency department (ED) services, with headaches accounting for approximately 3% of all ED visits annually [2]. Headaches also directly impact workplace productivity, with an economic impact estimated to be $19.3 billion in 2019 [3].

Headaches account for up to 30% of referrals from primary care to neurology. Of these, nearly 50% of patients are ultimately diagnosed with migraines, and the majority have never adequately trialed first-line therapy prior to referral [4-7]. Barriers to primary care management of headaches include time constraints, access challenges, and lack of expertise to accurately diagnose and give appropriate treatment [8,9]. These barriers result in high referrals to specialized care, an easier option than drug initiation titration, monitoring, and adjustments. This leads to prolonged wait times, resulting in patients often remaining unmanaged and seeking care in urgent or emergent settings for episodic management [10,11].

The Geisinger system was an early adopter of the electronic health record (EHR) system (beginning in 1996), with its EHR implemented across all sites of care. The use of clinician decision support tools in the form of alerts within the EHR has increased exponentially in recent years [12]. These alerts have been used across different disease states within primary care to provide useful information to clinicians, shape their behaviors, and positively impact patient safety and outcomes [12-14]. Considering the utility of alerts and the challenges described above, we instituted an EHR-based, clinician-facing, interruptive alert that provides real-time guidance on managing patients with headaches to improve headache care before neurology referral. This includes assistance in collecting information about headache characteristics, guidance on medication management, and the opportunity to e-consult with neurology providers. The purpose of this pragmatic randomized controlled trial was to evaluate the impact of the alert on the management of primary headache disorders in primary care settings. We hypothesized that an electronic alert for the management of headaches improves patient-reported headache burden at 6 months among patients with headache managed in primary care.

Methods

Setting

Geisinger is an integrated health care delivery system located in central and northeastern Pennsylvania, serving more than 1 million patients yearly. Geisinger maintains 44 primary care and over 100 specialty clinics, which include 8 neurology practices with 43 neurology physicians all using the same EHR system.

Study Design

We conducted a prospective stratified cluster-randomized controlled trial across 38 primary care sites to assess the impact of an interruptive electronic alert on the management of primary headache disorders. Eligible patients were enrolled and followed for 6 months post index encounter to evaluate headache management-related outcomes. It was uploaded to ClinicalTrials.gov and can be found using the registration number NCT05067725. The trial followed CONSORT (Consolidated Standards of Reporting Trials) reporting guidelines for randomized trials. We have also reported the e-CONSORT per the journal requirements [15].

Site Randomization and Patient Enrollment

In total, 38 of 44 primary care clinic sites were selected for randomization based on their ability to participate. The sites were first stratified into 9 possible blocks based on 2 criteria: geographical location (west, central, or northeast) and patient-to–health care provider ratio (low, moderate, or high). Note that 3 of the 9 possible blocks did not contain any sites, leaving only 6 blocks for this study. Within each block, simple randomization was then used to assign each site to either the intervention or control arm, using a 1:1 ratio. A Microsoft Excel 365 (Microsoft Corporation) random number generator was used to generate the simple randomization allocation sequence. Patients received the treatment to which their site had been assigned, making this a cluster-randomized study with sites as clusters (Multimedia Appendix 1).

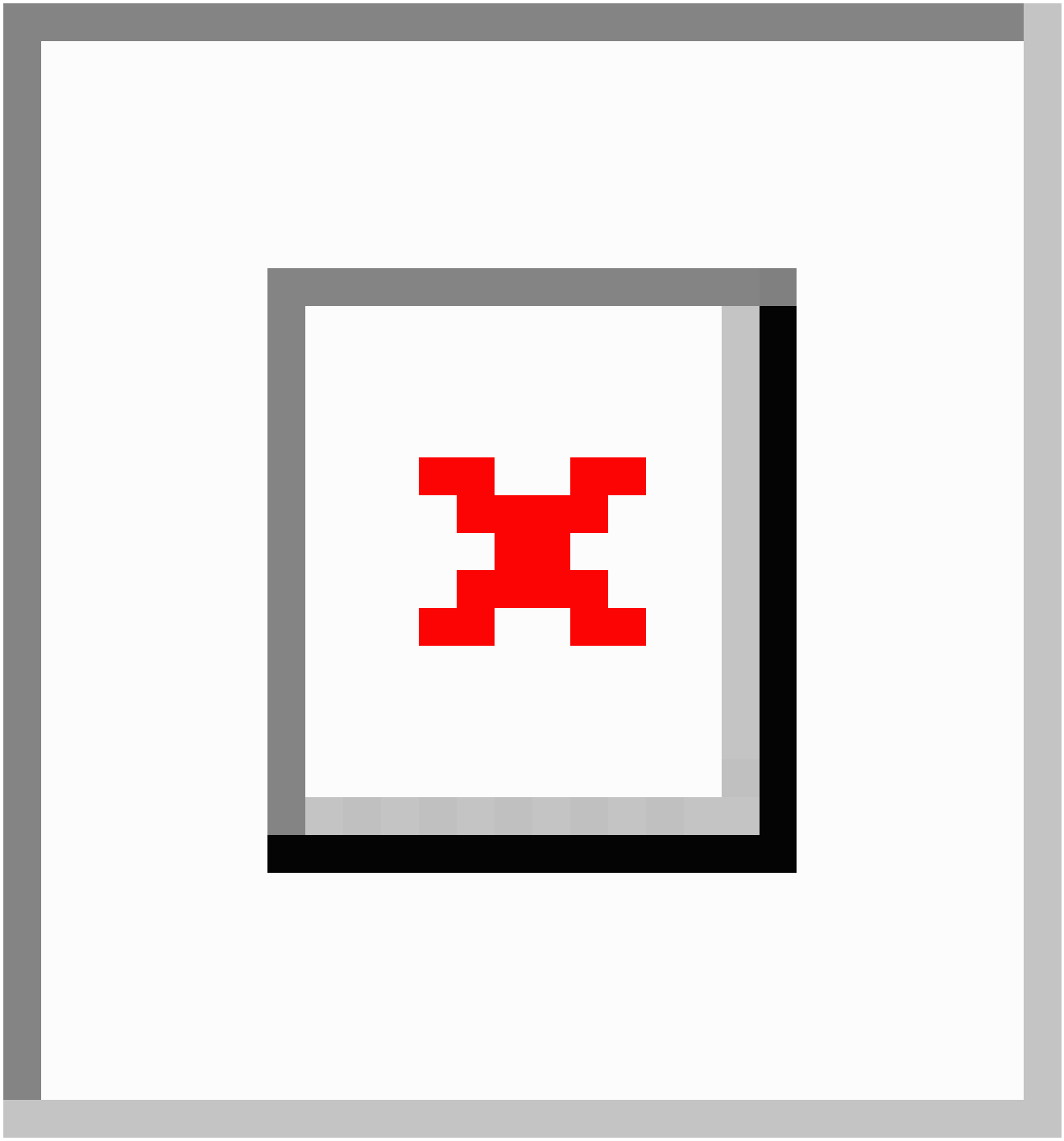
Patient Selection and Enrollment

Patients were eligible if (1) the criteria for the alert were met, which included an in-person or telemedicine encounter and a headache diagnosis or chief complaint of headache (Multimedia Appendix 2); (2) they were aged 18 years or older; and (3) postvisit criteria were met, which included a confirmed diagnosis of primary headache disorder by manual medical record review and a minimum baseline Headache Impact Test 6 (HIT-6) score of 50 or higher or headache frequency of 12 days or greater in the last 3 months. Clinical input and the scale guidance document were used to determine thresholds for HIT-6 scores and headache frequency [16]. Manual medical record reviews were conducted by study team members (HA, AP, and SP) under the guidance of the neurologist (SF) to confirm headache diagnosis. Types of headaches in the analytic sample included migraine, chronic daily, tension-type, cluster, or medication overuse headaches, collectively grouped as “headaches.” Patients were excluded if they had a serious systemic illness (eg, headaches due to hypertensive emergencies, hepatic or renal failure, or cardiac failure), secondary headaches (eg, concussion), were pregnant, or were actively followed with neurology [17]. Patients with serious systemic illness were excluded from our study as it would have been difficult to distinguish whether headaches observed among them were due to primary or secondary factors.

Within 2 weeks after their scheduled visit, eligible patients were contacted by the Survey Research and Recruitment Core to collect baseline measures (which determined minimum headache eligibility) and to gain verbal consent. Patient contact was

attempted 7 times. Figure 1 represents this study's process of patient selection, enrollment, and follow-up.

Figure 1. Patient selection, enrollment, and follow-up.



Intervention

A clinician-facing interruptive EHR-based alert in the form of a Best Practice Advisory was built based on input from an expert panel of neurologists, primary care providers (PCPs), pharmacists, and informaticians. PCPs include primary care physicians and advanced practitioners such as nurse practitioners and physician assistants. The first iteration of the alert was developed based on input received from the needs assessment conducted with the PCPs. This version was piloted at 2 primary care sites over 3 months [18]. Face-to-face education of PCPs at the pilot sites was undertaken to familiarize health care providers with the utility of the alert [18]. A simple pre-post assessment of the 900 eligible patient encounters during this pilot phase demonstrated a significant reduction in neurology referrals post implementation of the electronic alert. Following the pilot implementation, feedback was collected from 10 PCPs at these sites about the structure, verbiage, and utility of the electronic alert. Subsequent modifications were made based on this feedback before full implementation.

The Best Practice Advisory consisted of four parts: (1) a questionnaire to assist the PCP in characterizing the patient's headache (Multimedia Appendix 3); (2) a "SmartSet" guide for ordering medications, imaging, laboratories, and referrals using evidence-based guidelines (Multimedia Appendix 3); (3) a

hyperlink to allow for electronic consultation with a neurologist; and (4) a link to the "synopsis" which provided a summary of prior patient headache care. The "SmartSet" guide could be accessed directly within the alert by clicking on the "Open Express Lane" button to facilitate quick ordering for health care providers. The alert appeared within the EHR when a PCP entered a headache diagnosis for an in-person or telemedicine encounter. All actions taken in the alert were recorded, although health care providers could dismiss them without any action. Suppression criteria were built to lock out and avoid repeated firing of the alert. The alert was suppressed for 24 hours if the dismissal reason was headache-controlled and for 365 days for PCP long-term management. The alert could also be dismissed by acknowledging that the health care provider was previewing the patients' medical record, in this case, the alert would refire during the same encounter.

An invisible or silent alert was activated at control sites that would be triggered by identical criteria to the intervention arm, however, this alert was not visible to the clinicians, so no suppression criteria were developed for control sites. The silent alert was developed to help generate a list of eligible patients for the control group that would be comparable to those identified in the intervention arm.

Before full implementation, the primary care leadership team was encouraged to disseminate information about the tool to PCPs in the intervention cohort through local departmental meetings and direct emails, including text and short video instructions.

Data Collection

Data sources included the EHR and survey results from patients at baseline, 3 months, and 6 months. The surveys conducted by the Survey Research and Recruitment Core included 6 questions from the HIT-6 questionnaire, which measures headache impact on a participant's ability to function in day-to-day life [19]; 1 question on pain intensity; 1 on the frequency of headaches; 3 from the Migraine Treatment Optimization Questionnaire (M-TOQ), which determines if a medication treatment for a participant is optimal; and 2 asking for sociodemographic information that could not be pulled from the EHR [20]. The M-TOQ questions were modified for our study by replacing the word "migraine" with "headache." Prior to study enrollment, the survey was piloted among a sample (n=10) of patients from pilot sites to ensure appropriateness and check for face validity. Patients had the option to opt-out at any time during this study. See Multimedia Appendix 4 for the full survey guide.

Study Outcomes

Primary Outcome—Headache Burden

Headache burden was assessed using the HIT-6 scale [7]. The scale ranged from 36 to 78, and a score of 50 or greater represented a greater impact and disruption of life caused by headache [21,22]. The primary end point was the comparison of changes in HIT-6 scores between arms from baseline to 6 months.

Secondary Outcomes

Secondary outcomes included headache frequency, defined as the self-reported frequency of headache days over the previous 3 months; average headache pain intensity defined by pain on a 10-point analog scale over the same period (0=no pain at all to 10=the worst pain ever); access to care, defined as the proportion of patients referred to neurology; resource utilization, defined as the proportion of patients with an ED-visit (all-cause and headache specific); and medication management, defined as the proportion of patients that initiated preventive or abortive medications for headache [7,23]. Referral to neurology was used as a surrogate measure to represent access to neurological care considering that the availability of a timely neurology appointment impacts the volume of referrals.

Implementation Outcomes

We assessed the adoption of the alert among health care providers through actions within the alert. Positive adoption was any activity where the health care provider interacted with the components of the alert such as completing the headache questionnaire, using the SmartSet, or using the synopsis. We also reviewed feedback provided by PCPs from reports generated by a committee that oversees all EHR-based alerts.

Several modifications to the initial analysis plan were implemented following go-live but before analysis. We omitted some secondary outcomes, including time to initiation of treatment in treatment naïve patients and M-TOQ outcomes. The data on the M-TOQ were only available at baseline eliminating any pre-post analysis, and the time to initiation of medication in naïve patients was dropped in favor of initiation of treatment in all patients as a binary outcome. We added a referral to neurology due to its relevance to clinical practice. All these changes were updated on the ClinicalTrials.gov website to ensure reporting transparency. We also updated our analytical approach from generalized linear models with robust SEs in the analytical plan to generalized estimating equations (GEE), a specialized generalized linear model approach. After further internal discussion, we felt given that the outcomes data were longitudinal and correlated due to clustering, GEE were better suited.

Sample Size Calculation

A minimum difference of equal to or greater than a 2.3-point (SD 4.3) reduction on the HIT-6 scale represents a clinically important and significant change [22]. We needed to enroll at least a total of 66 patients per arm to detect this change with 80% power at a significance level of .05 and an intraclass correlation of 0.02. The sample size was calculated using an online calculator developed by the University of California San Francisco [24].

Statistical Analysis

The total number of patients in each arm was counted to ensure a balanced distribution of patient numbers. Descriptive statistics were computed for continuous variables as means or medians with SD or IQRs, and as frequency and percentages for categorical variables when appropriate for baseline patient-reported outcomes and socioeconomic and demographic characteristics (ie, sex, age, race, ethnicity, insurance, education,

and income). Appropriate statistical tests were used to compare differences in these characteristics between the 2 groups and within groups: 2-tailed independent *t* tests or paired *t* tests were used for continuous data, χ^2 for categorical data, and Mann-Whitney for nonnormal categorical or continuous data.

We constructed GEE for all outcome comparisons between groups, adjusting for effects of treatment, phase (baseline, 6 months), treatment-by-phase interaction, and demographics while accounting for clustering of patients per primary care site. The GEE model with log link was used, considering the skewness of data observed for headache days and pain scores (Multimedia Appendix 5). For all the binary variables, a log link with binomial distribution was used, while an identity link with normal distribution was used for HIT-6 scores. As a first step, we performed an unadjusted GEE model with this study's groups as only a univariate analysis. Multivariable (adjusted) models only included demographic variables that were significantly different between groups (P<.05). For the GEE analysis, we reported odds ratios or estimates and associated CIs. All analyses were conducted using SAS Enterprise Guide software (version 8.3 for Windows; SAS Institute Inc).

The analysis was conducted based on the principle of "intention-to-treat (ITT)" analysis. All patients who consented and met the inclusion criteria were included in the analysis per the group they were recruited into. The analysis was performed to assess the difference in outcomes at the individual patient level. For missing data via survey collection, we applied a last observation carried forward (LOCF) technique to maximize the number of observations available for analysis [25]. The results from the LOCF technique were further compared to those from the measured data to assess the impact of loss to follow-up.

Ethical Considerations

This study was an evaluation of a quality improvement initiative that was undertaken at Geisinger with the intent to affect clinical practice. As a result, it was determined as "not human subjects research" by Geisinger's institutional review board (2021 - 0729).

Results

Overview

The alert was fired for a total of 9239 times in 3183 patients (intervention arm: 2677 times in 1507 patients; control arm: 6562 times in 1676 patients) between December 2021 and February 2022. We excluded 1828 patients for not meeting eligibility criteria. We manually verified the diagnosis of primary headache among 729 of the remaining 1355 patients who were eligible for consent and baseline survey assessment. We obtained verbal consent from 221 patients at baseline but only enrolled the 203 patients that either had an HIT-6 score of ≥50 or a headache frequency of >12 days, as per our inclusion criteria (Figure 1). At the end of 6 months, follow-up patient-reported data were available for 88/203 (43.3%) patients.

All baseline characteristics were similar between groups for patients enrolled in this study except for mean age, which was higher in the intervention group (mean 43, SD 14.4 y vs mean 39, SD 14.4 y, P=.04; Table 1). There was no difference in the HIT-6 scores, headache frequency and intensity, or patient-reported use of medications at baseline between groups (HIT-6 scores: mean 63, SD 6.9 vs mean 61.8, SD 6.6; headache days in the past 3 months: median 20.5, IQR 10.0 - 45.0 vs median 20, IQR 10.0 - 45.0; pain scores: median 7, IQR 6.0 - 8.0 vs median 7, IQR 5.0 - 8.0; reported use of preventive or abortive medications: 63/98, 64% vs 57/105, 54.2%). The majority of the patients were female, were White, were non-Hispanic, had a commercial insurance plan, had an educational qualification of greater than high school, and had an annual household income between US $50,000 to US $100,000. Similarly, there was no statistically significant difference in the baseline characteristics observed between groups for patients that completed the 6-month follow-up.

Table . Patient characteristics at baseline.

Demographics		Intervention arm (n=98)	Control arm (n=105)	*P* value
Age (years), mean (SD)		43.2 (14.4)	39.1 (14.4)	.04
Sex, n (%)				.19
	Males	28 (28.6)	21 (20.0)	
	Females	70 (71.4)	84 (80.0)	
Race, n (%)				.56
	African American	2 (2.0)	2 (1.9)	
	White	95 (97)	98 (94.2)	
	Other	1 (1.0)	4 (3.8)	
Ethnicity, n (%)				.60
	Non-Hispanic	92 (93.9)	95 (91.4)	
	Hispanic	6 (6.1)	9 (8.6)	
Type of insurance at the most recent encounter (members), n (%)				.32
	Commercial	59 (60.2)	56 (53.3)	
	Medicare	13 (13.3)	12 (11.4)	
	Medicaid	26 (26.7)	31 (29.5)	
	Other	0 (0)	6 (5.8)	
Highest level of education, n (%)				.98
	<High school or GED[a]	9 (9.2)	9 (8.6)	
	High school or GED	32 (32.7)	38 (36.2)	
	Some college or technical program	29 (29.6)	26 (24.8)	
	4 y college (BS or BA)	16 (16.3)	20 (19.0)	
	Master's degree (MS, MA, or MPH)	8 (8.2)	9 (8.6)	
	Doctorate (PhD or ScD or professional—MD, DO, or JD)	3 (3.0)	2 (1.9)	
	Refuse	1 (1.0)	1 (0.9)	
Annual household income (US $), n (%)				.12
	25,000 or less	20 (20.4)	26 (24.8)	
	Over 25,000-50,000	21 (21.4)	27 (25.7)	
	Over 50,000-100,000	34 (34.7)	21 (20.0)	
	Over 100,000	14 (14.3)	13 (12.4)	
	Refuse	9 (9.2)	18 (17.1)	
HIT-6[b] scores, mean (SD)		63 (6.9)	61.8 (6.6)	.21
Number of headache days, median (IQR)		20.5 (10.0 - 45.0)	20 (10.0 - 45.0)	.87
Pain scores, median (IQR)		7 (6.0 - 8.0)	7 (5.0 - 8.0)	.91
Patient-reported use of preventive or abortive medications, n (%)		63 (64.3)	57 (54.2)	.14

[a]GED: General Educational Development.

[b]HIT-6: Headache Impact Test 6.

Primary Outcome—Headache Burden

HIT-6 scores improved significantly from baseline in the intervention arm at 6 months (3.5 points; *P*=.009) but not in the control group (1.40 points; *P*=.23); additionally, the improvements from baseline did not differ significantly between groups (Table 2). The intracluster correlation coefficient observed was 0.411 for HIT-6 scores.

Table . Adjusted primary and secondary outcomes of headache burden, frequency, and intensity.

Outcome and arm	Phase	n	Mean (SD)	Change, mean (SD)	D-I-D[a] estimate, regression coefficients (95% CI)	*P* value
HIT-6[b] score					–1.84 (–5 to 1.31)	.25
Intervention				–3.5 (7.9)[c]		
	Baseline	98	63.0 (6.9)			
	Month 6	38	58.7 (8.6)			
Control				–1.40 (8.1)		
	Baseline	105	61.8 (6.5)			
	Month 6	50	60.1 (8.7)			
Headache days					1.22 (0.82 to 1.81)	.24
Intervention				–9.2 (30.5)		
	Baseline	98	30.9 (25.8)			
	Month 6	38	22.8 (26.2)			
Control				–9.3 (16.3)[c]		
	Baseline	105	28.9 (23.5)			
	Month 6	50	18.4 (18.2)			
Pain					1 (0.9 to 1.11)	.99
Intervention				–0.395 (1.59)		
	Baseline	98	6.9 (1.8)			
	Month 6	38	6.5 (1.8)			
Control				–0.340 (1.92)		
	Baseline	105	6.9 (1.9)			
	Month 6	50	6.4 (2.1)			

[a]D-I-D: difference-in-differences.

[b]HIT-6: Headache Impact Test 6.

[c]*P*=.01.

Secondary Outcomes

Change in the number of headache days and pain score did not differ between groups (Table 2). Similarly, compared to baseline, there was no difference in the proportion of patients being referred to neurology, using ED for all-cause or headache-specific reasons, or initiating new abortive or preventative treatment in the 6 months post intervention (Table 3).

Table . Percent change in the proportion of patients from baseline to 6 months and adjusted odds ratio estimates for resource use and access to care (note: a negative or positive percentage denotes a decrease or increase post recruitment).

Outcome and arm	Phase	Use, n (%)	Percent change	D-I-D[a] estimate, odds ratio (95% CI)	*P* value
ED [b] use (all-cause)				1.4 (0.58 to 3.36)	.46
Intervention (n=98)			–2.0		
	Baseline	19 (19.4)			
	Month 6	17 (17.4)			
Control (n=105)			–4.8		
	Baseline	15 (14.3)			
	Month 6	10 (9.5)			
Neurology referral				0.52 (0.064 to 4.29)	.55
Intervention (n=98)			12.3[c]		
	Baseline	2 (2.04)			
	Month 6	14 (14.3)			
Control (n=105)			21.0[d]		
	Baseline	2 (1.9)			
	Month 6	24 (22.9)			
Medication initiation				0.91 (0.51 to 1.60)	.73
Intervention (n=98)			25.5[d]		
	Baseline	50 (51.0)			
	Month 6	75 (76.5)			
Control (n=105)			28.5[d]		
	Baseline	49 (46.7)			
	Month 6	79 (75.2)			
ED use (headache-specific)				0.49 (0.051 to 4.72)	.54
Intervention (n=98)			–2.10		
	Baseline	4 (4.1)			
	Month 6	2 (2.0)			
Control (n=105)			0.0		
	Baseline	4 (3.81)			
	Month 6	4 (3.81)			

[a]D-I-D: difference-in-differences.

[b]ED: emergency department

[c]*P*=.002.

[d]*P*<.001.

Results from the LOCF models for HIT-6 scores, headache days, and pain intensity are listed in Table 4. Findings from the LOCF models were similar to those from the base models presented in Table 2. It was observed that there was a similar significant improvement in HIT-6 scores in the intervention arm, and a reduction in the headache days in the control arm from baseline to 6 months, however, the difference between groups was not statistically significant.

Table . Adjusted primary and secondary outcomes of headache burden, frequency, and intensity using the LOCF[a] method.

Outcome and arm	Phase	N	Mean (SD)	Change, mean (SD)	D-I-D[b] estimate, regression coefficients (95% CI)	*P* value
HIT-6 score					–1.01 (–2.87 to 0.85)	.29
Intervention				–1.36 (5.2)[c]		
	Baseline	98	63 (6.9)			
	Month 6	98	61 (8.2)			
Control				–0.67 (5.6)		
	Baseline	105	61.8 (6.5)			
	Month 6	105	60.8 (8.4)			
Headache days					1.07 (0.87 to 1.32)	.49
Intervention				–3.57 (19.3)		
	Baseline	98	30.9 (25.8)			
	Month 6	98	27.5 (22.9)			
Control				–4.41 (17.6)[c]		
	Baseline	105	28.9 (23.5)			
	Month 6	105	24 (22.6)			
Pain					1.02 (0.95 to 1.10)	.60
Intervention				–0.153 (1.00)		
	Baseline	98	6.9 (1.82)			
	Month 6	98	6.6 (2.01)			
Control				–0.162 (1.30)		
	Baseline	105	6.9 (1.9)			
	Month 6	105	6.4 (2.1)			

[a]LOCF: last observation carried forward.

[b]D-I-D: difference-in-differences.

[c]*P*=.01.

Implementation Outcomes

Of the 2677 alert firings in the intervention arm, 2228 (83.23%) were overridden or dismissed, 146 (5.45%) were ignored, and 303 (11.32%) action was taken. Of the 205 PCPs exposed to the intervention alert, 15 provided feedback, 3 found the alert helpful, and 12 expressed dissatisfaction due to alert firing on inappropriate patients or noting the alert was disruptive to workflow. A manual medical record review of 1355 patients revealed that 729 (53.80%) had a confirmed diagnosis of primary headache disorder. The remaining 626 (46.20%) either did not have a confirmed diagnosis of primary headache disorder or were already followed by neurology for headaches.

Discussion

In this pragmatic clinical trial across 38 clinics, we found no benefit of the electronic alert as designed on headache outcomes. As the alert was based on evidence-based guidelines for diagnosis and treatment, it seems unlikely that the content of the alert was poor and more likely that the lack of impact was due to gaps in implementation.

This is the first study to offer a clinician decision support tool that guided PCPs in diagnosing and appropriately treating patients with different types of primary headache disorders using an episodic electronic alert. Most of the research studies to date focus on a single type of headache disorder, chronic headaches, or migraines [23,26]. As initial preventative treatment for the most common headache types is similar, our alert was designed to be used across them as a one-stop access to help diagnose and treat headaches and e-consult with neurologists if required.

Even though we did notice improvements in several outcome measures in both the intervention and control arms from baseline to 6-month follow-up, on a between-groups comparison, the effects did not hold. A similar trend for improvement from baseline in outcomes such as HIT-6 scores and headache days was observed when an LOCF approach was used, but the effect

did not hold for between-group comparisons. Using our pragmatic design, we were able to distinguish between expected improvements in headaches over time and those attributed to the intervention using a comparator group, with largely similar characteristics except by age at baseline [27]. Past research studies that have evaluated the impact of headache management programs and found improvement in patients' headache burden and resource use have often used a pre-post study design which makes attribution of success to the intervention difficult [17,23]. Our results have substantial implications, as identification of the ineffectiveness of these types of alerts is necessary to determine which should remain active and which should be either deactivated or revised. Only a well-designed test of comparison with a suitable counter-factual arm (ie, control) can distinguish improvements due to bias (eg, regression to the mean) versus the intervention itself.

The total number of unique patients for whom the alert fired was relatively similar between groups, but the alert fired 2.2 times more frequently in the control arm. This is to be expected, as in the interventional arm, the interruptive alert that was visible to the health care providers could be suppressed for the duration of the encounter by acknowledging a reason for dismissal, as opposed to the silent or invisible alert in the control arm. Health care providers also noted several challenges with alert firing such as confusion with multiple components of the alert, concerns over alert firing for inappropriate patients and appearing at times other than during direct patient contact, and trouble dismissing the alert, highlighting flaws with the existing alert design. Medical record reviews illustrated that health care providers often list headaches as not just chief complaints but also visit diagnoses, due to headaches often being a symptom of other underlying conditions. The inaccurate use of headache as a visit diagnosis led to the alert's misfiring in nearly 40% of our patients. It also complicates the firing logic for decision support tools such as ours. Research shows that the nonspecificity of alerts may desensitize clinicians and lead to habituation, thereby lessening their likelihood to follow alert recommendations or guidance [28,29]. Furthermore, even when firing appropriately, health care providers reported the location of the alert to be disruptive to their normal workflow. Some of the challenges mentioned could be attributed to the health care providers not having received sufficient education about the alert. The majority of the health care providers did not recall having seen or received any education about the alert despite efforts to disseminate this information through emails, fact sheets, and videos. This highlights an implementation gap and the need for a more engaged educational outreach in primary care settings.

Much of the feedback and low adoption found in our study affirms prior reports of alert fatigue experienced by health care providers in primary care settings and is likely to be the main contributor to the lack of significant impact [30]. With only 11.32% (303/2677) of alerts being acted upon, we can expect only this fraction to benefit from the intervention. Yet, research does not support the notion that prior alert exposure has a carry-forward impact on the care of future patients; suggesting that future recognition of patients and bypassing of alerts does not have a strong persistence of effect on health care provider action behaviors [28,29,31].

The lack of positive outcomes within the intervention arm was unexpected, considering the detailed human-centered design-build, the engagement of pilot PCPs, and the early feedback received, which was affirmative for the intervention. Slight modifications to the verbiage within the alert were undertaken based on feedback from the pilot health care providers before its full-scale implementation. Despite this, we noted low use of the features present within the alert and additional adoption challenges that were not fully uncovered in the pilot period. These results are similar to other studies published using human-centered design, which failed to translate into adoption and outcome differences [32-35]. These studies have found that better engagement by pilot health care providers does not always translate to all health care providers at the time of scaling interventions. Additionally, differences in the mechanism used for delivering training between the pilot and full implementation phase also impact the uptake of the intervention [32-35]. Furthermore, the positive findings and feedback from the pilot phase could be influenced by commitment and congruence bias in addition to observation and measurement bias resulting in a Hawthorne effect of the intervention [36].

Considering the challenges identified in this study, post completion we made significant modifications to the tool. First, we redesigned the identification of the patients to make it more specific (but less sensitive) to the targeted patient population. We repositioned the alert to fire at the time of neurology referral as opposed to when assigning a visit diagnosis or chief complaint. We also recognize that an EHR-only intervention is not the only mechanism to address incomplete headache care, and a pilot initiative involving pharmacist-based medication management has been launched. Research is currently underway to assess the impact of this pilot.

We recognize several limitations to our study. First, given the single-site design of this study, generalizability to other sites is limited and likely influenced by variations in practice environments and standards of practice. Our use of an alert was largely driven by our integrative health care environment, the ubiquitous use of EHR across primary care and specialty health care providers, and the low cost and light maintenance of the alert. Second, while an LOCF method was used to reduce the effect of a type II error, the high attrition rate observed might have affected the magnitude of the impact. We recognize that other interventions to address the optimal treatment of headaches do exist and are worthy of consideration. Programs that involved greater manpower and dedicated professionals for follow-up and management observed significant improvement in their patients' headache burden and headache-related disability or reduced resource use, and attributed their success to the dedicated personnel [23,26]. Alternatively, programs that solely relied on physician education or leveraging existing clinical pathways, while at times cost-saving, did not significantly improve patients' headache burden or quality of life [7,21]. Most of these existing studies have been standalone initiatives that focused on headache-specific work groups with limited generalizability of results, some of which are no longer in effect

due to resource and infrastructure constraints. Our objective was to sustainably change the culture of practice surrounding headache care without creating new standalone avenues. In conclusion, our study found that the implementation of an interruptive electronic alert in primary care to aid diagnosis and management of primary headache disorders did not improve headache care for patients. Low adoption of the tool by the health care providers has prompted the development of alternative population health–based approaches to improve headache management.

Acknowledgments

We would like to thank Neil R Holland, MBBS, MBA, MSMEd; Aubrielle C Smith-Masri, PharmD; Alex Skitolsky, BA, MFA; Maria Kobylinski, MD; the Survey Research and Recruitment Core (SRRC) at Geisinger; and the primary care providers and patients at Geisinger for their contributions to this study. We would also like to thank Duncan Dobbins, BSPF, PharmD for his creation of our paper's table of contents image.

Authors' Contributions

Conception and design were done by AP, EAW, LW, and SF. Acquisition of data was handled by AP, VAH, HA, and SP. Analysis and interpretation of data were performed by AP, JB, MS, and JG. All authors (AP, EAW, VAH, JB, MS, LW, HA, SP, JG, and SF) drafted this paper, revised it for intellectual content, and gave final approval of this completed paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Site-level distribution based on stratification.
[DOCX File, 24 KB - medinform_v12i1e58456_app1.docx]

Multimedia Appendix 2
Project inclusion and exclusion *ICD* codes. *ICD*: *International Classification of Diseases*.
[DOCX File, 177 KB - medinform_v12i1e58456_app2.docx]

Multimedia Appendix 3
Snapshots of the electronic alert questionnaire and Express Lane or SmartSet tool.
[DOCX File, 624 KB - medinform_v12i1e58456_app3.docx]

Multimedia Appendix 4
Survey administered by Geisinger’s SRRC. SRRC: Survey Research and Recruitment Core.
[DOCX File, 31 KB - medinform_v12i1e58456_app4.docx]

Multimedia Appendix 5
Distribution of pain scores and the number of headache days observed in the intervention and control arm of this study.
[DOCX File, 104 KB - medinform_v12i1e58456_app5.docx]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 392 KB - medinform_v12i1e58456_app6.pdf]

References

1. Stovner LJ, Hagen K, Linde M, Steiner TJ. The global prevalence of headache: an update, with analysis of the influences of methodological factors on prevalence estimates. J Headache Pain 2022 Apr 12;23(1):34. [doi: 10.1186/s10194-022-01402-2] [Medline: 35410119]
2. Burch R, Rizzoli P, Loder E. The prevalence and impact of migraine and severe headache in the United States: figures and trends from government health studies. Headache 2018 Apr;58(4):496-505. [doi: 10.1111/head.13281] [Medline: 29527677]
3. Yucel A, Thach A, Kumar S, Loden C, Bensink M, Goldfarb N. Estimating the economic burden of migraine on US employers. Am J Manag Care 2020 Dec 1;26(12):e403-e408. [doi: 10.37765/ajmc.2020.88547] [Medline: 33315334]
4. Migraine and other headache disorders. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/headache-disorders [accessed 2024-08-02]

5. Tepper SJ, Dahlöf CGH, Dowson A, et al. Prevalence and diagnosis of migraine in patients consulting their physician with a complaint of headache: data from the landmark study. Headache 2004 Oct;44(9):856-864. [doi: 10.1111/j.1526-4610.2004.04167.x] [Medline: 15447694]
6. Fernandes L, Tanti M, Randall M, Cosgrove J. 111 an audit of headache referrals from primary care, and subsequent management in the outpatient clinic. J Neurol Neurosurg Psychiatry 2019 Dec;90(12):e33. [doi: 10.1136/jnnp-2019-ABN-2.110]
7. Rua T, Mazumder A, Akande Y, et al. Management of chronic headache with referral from primary care to direct access to MRI compared with neurology services: an observational prospective study in London. BMJ Open 2020 Oct 16;10(10):e036097. [doi: 10.1136/bmjopen-2019-036097] [Medline: 33067273]
8. Matchar DB, Harpole L, Samsa GP, et al. The headache management trial: a randomized study of coordinated care. Headache 2008 Oct;48(9):1294-1310. [doi: 10.1111/j.1526-4610.2007.01148.x] [Medline: 18547268]
9. Minen M, Shome A, Halpern A, et al. A migraine management training program for primary care providers: an overview of a survey and pilot study findings, lessons learned, and considerations for further research. Headache 2016 Apr;56(4):725-740. [doi: 10.1111/head.12803] [Medline: 27037903]
10. Roy S, Keselman I, Nuwer M, Reider-Demer M. Fast Neuro: a care model to expedite access to neurology clinic. Neurol Clin Pract 2022 Apr;12(2):125-130. [doi: 10.1212/CPJ.0000000000001152] [Medline: 35747888]
11. Dall TM, Storm MV, Chakrabarti R, et al. Supply and demand analysis of the current and future US neurology workforce. Neurology 2013 Jul 30;81(5):470-478. [doi: 10.1212/WNL.0b013e318294b1cf] [Medline: 23596071]
12. Powers EM, Shiffman RN, Melnick ER, Hickner A, Sharifi M. Efficacy and unintended consequences of hard-stop alerts in electronic health record systems: a systematic review. J Am Med Inform Assoc 2018 Nov 1;25(11):1556. [doi: 10.1093/jamia/ocy112] [Medline: 30239810]
13. Konerman MA, Thomson M, Gray K, et al. Impact of an electronic health record alert in primary care on increasing hepatitis C screening and curative treatment for baby boomers. Hepatology 2017 Dec;66(6):1805-1813. [doi: 10.1002/hep.29362] [Medline: 28714196]
14. Downing NL, Rolnick J, Poole SF, et al. Electronic health record-based clinical decision support alert for severe sepsis: a randomised evaluation. BMJ Qual Saf 2019 Sep;28(9):762-768. [doi: 10.1136/bmjqs-2018-008765] [Medline: 30872387]
15. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126. [doi: 10.2196/jmir.1923] [Medline: 22209829]
16. QualityMetric Inc, GlaxoSmithKline Group of Companies. HIT-6. National Headache Foundation. URL: https://headaches.org/wp-content/uploads/2018/02/HIT-6.pdf [accessed 2023-05-04]
17. Blumenfeld A, Tischio M. Center of excellence for headache care: group model at Kaiser Permanente. Headache 2003 May;43(5):431-440. [doi: 10.1046/j.1526-4610.2003.03087.x] [Medline: 12752747]
18. Patel AD, Sponenberg M, Webster L, et al. Using design thinking to understand the reason for headache referrals and reduce referral rates. Neurol Clin Pract 2024 Aug 16;14(6):e200336. [doi: 10.1212/CPJ.0000000000200336]
19. Yang M, Rendas-Baum R, Varon SF, Kosinski M. Validation of the Headache Impact Test (HIT-6TM) across episodic and chronic migraine. Cephalalgia 2011 Feb;31(3):357-367. [doi: 10.1177/0333102410379890] [Medline: 20819842]
20. Lipton RB, Kolodner K, Bigal ME, et al. Validity and reliability of the Migraine-Treatment Optimization Questionnaire. Cephalalgia 2009 Jul;29(7):751-759. [doi: 10.1111/j.1468-2982.2008.01786.x] [Medline: 19239676]
21. Smelt AFH, Blom JW, Dekker F, et al. A proactive approach to migraine in primary care: a pragmatic randomized controlled trial. CMAJ 2012 Mar 6;184(4):E224-E231. [doi: 10.1503/cmaj.110908] [Medline: 22231680]
22. Coeytaux RR, Kaufman JS, Chao R, Mann JD, Devellis RF. Four methods of estimating the minimal important difference score were compared to establish a clinically significant change in Headache Impact Test. J Clin Epidemiol 2006 Apr;59(4):374-380. [doi: 10.1016/j.jclinepi.2005.05.010] [Medline: 16549259]
23. Harpole LH, Samsa GP, Jurgelski AE, Shipley JL, Bernstein A, Matchar DB. Headache management program improves outcome for chronic headache. Headache 2003;43(7):715-724. [doi: 10.1046/j.1526-4610.2003.03128.x] [Medline: 12890125]
24. Kohn MA, Senyak J. Sample Size Calculators. UCSF CTSI. 2024. URL: https://www.sample-size.net/ [accessed 2024-04-17]
25. Liu X. Methods and Applications of Longitudinal Data Analysis: Academic Press; 2016.
26. Veenstra P, Kollen BJ, de Jong G, Baarveld F, van den Berg JP. Nurses improve migraine management in primary care. Cephalalgia 2016 Jul;36(8):772-778. [doi: 10.1177/0333102415612767] [Medline: 26487468]
27. Austrian J, Mendoza F, Szerencsy A, et al. Applying A/B testing to clinical decision support: rapid randomized controlled trials. J Med Internet Res 2021 Apr 9;23(4):e16651. [doi: 10.2196/16651] [Medline: 33835035]
28. Kawamanto K, Flynn MC, Kukhareva P, et al. A pragmatic guide to establishing clinical decision support governance and addressing decision support fatigue: a case study. AMIA Annu Symp Proc 2018 Dec;2018:624-633. [Medline: 30815104]
29. Kawamoto K, McDonald CJ. Designing, conducting, and reporting clinical decision support studies: recommendations and call to action. Ann Intern Med 2020 Jun 2;172(11 Suppl):S101-S109. [doi: 10.7326/M19-0875] [Medline: 32479177]
30. Backman R, Bayliss S, Moore D, Litchfield I. Clinical reminder alert fatigue in healthcare: a systematic literature review protocol using qualitative evidence. Syst Rev 2017 Dec 13;6(1):255. [doi: 10.1186/s13643-017-0627-z] [Medline: 29237488]

31. Flottorp S, Håvelsrud K, Oxman AD. Process evaluation of a cluster randomized trial of tailored interventions to implement guidelines in primary care--why is it so hard to change practice? Fam Pract 2003 Jun;20(3):333-339. [doi: 10.1093/fampra/cmg316] [Medline: 12738704]
32. Beets MW, Weaver RG, Ioannidis JPA, et al. Identification and evaluation of risk of generalizability biases in pilot versus efficacy/effectiveness trials: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2020 Feb 11;17(1):19. [doi: 10.1186/s12966-020-0918-y] [Medline: 32046735]
33. Holeman I, Kane D. Human-centered design for global health equity. Inf Technol Dev 2019 Sep 29;26(3):477-505. [doi: 10.1080/02681102.2019.1667289] [Medline: 32982007]
34. Das M, Eichner J. Challenges and barriers to clinical decision support (CDS) design and implementation experienced in the Agency for Healthcare Research and Quality CDS demonstrations (prepared for the AHRQ National Resource Center for Health Information Technology under Contract No. 290-04-0016). AHRQ Publication No. 10-0064-EF. Agency for Healthcare Research and Quality. 2010 Mar. URL: https://digital.ahrq.gov/sites/default/files/docs/page/CDS_challenges_and_barriers.pdf [accessed 2024-08-03]
35. Byrne C, Sherry D, Mercincavage L, Johnston D, Pan E, Schiff G. Advancing clinical decision support—key lessons in clinical decision support implementation. Office of the National Coordinator for Health IT, Office of Policy and Planning. 2012. URL: https://www.healthit.gov/sites/default/files/acds-lessons-in-cds-implementation-deliverablev2.pdf [accessed 2024-08-03]
36. Berkhout C, Berbra O, Favre J, et al. Defining and evaluating the Hawthorne effect in primary care, a systematic review and meta-analysis. Front Med (Lausanne) 2022;9:1033486. [doi: 10.3389/fmed.2022.1033486] [Medline: 36425097]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
ED: emergency department
EHR: electronic health record
GEE: generalized estimating equations
HIT-6: Headache Impact Test 6
ITT: intention-to-treat
LOCF: last observation carried forward
M-TOQ: Migraine Treatment Optimization Questionnaire
PCP: primary care provider

Edited by C Lovis; submitted 25.03.24; peer-reviewed by B Littenberg, R Marshall; revised version received 21.06.24; accepted 23.06.24; published 29.08.24.

Please cite as:
Pradhan A, Wright EA, Hayduk VA, Berhane J, Sponenberg M, Webster L, Anderson H, Park S, Graham J, Friedenberg S
Impact of an Electronic Health Record–Based Interruptive Alert Among Patients With Headaches Seen in Primary Care: Cluster Randomized Controlled Trial
JMIR Med Inform 2024;12:e58456
URL: https://medinform.jmir.org/2024/1/e58456
doi:10.2196/58456

Scalable Approach to Consumer Wearable Postmarket Surveillance: Development and Validation Study

Richard M Yoo[1,*], MBI, PhD; Ben T Viggiano[1,*], BS; Krishna N Pundi[2,*], MD; Jason A Fries[1], PhD; Aydin Zahedivash[3], MBA, MD; Tanya Podchiyska[4], MS; Natasha Din[4], MBBS, MAS; Nigam H Shah[1,5,6], MBBS, PhD

1

2

3

4

5

6

[*]these authors contributed equally

Corresponding Author:
Richard M Yoo, MBI, PhD

Abstract

Background: With the capability to render prediagnoses, consumer wearables have the potential to affect subsequent diagnoses and the level of care in the health care delivery setting. Despite this, postmarket surveillance of consumer wearables has been hindered by the lack of codified terms in electronic health records (EHRs) to capture wearable use.

Objective: We sought to develop a weak supervision–based approach to demonstrate the feasibility and efficacy of EHR-based postmarket surveillance on consumer wearables that render atrial fibrillation (AF) prediagnoses.

Methods: We applied data programming, where labeling heuristics are expressed as code-based labeling functions, to detect incidents of AF prediagnoses. A labeler model was then derived from the predictions of the labeling functions using the Snorkel framework. The labeler model was applied to clinical notes to probabilistically label them, and the labeled notes were then used as a training set to fine-tune a classifier called Clinical-Longformer. The resulting classifier identified patients with an AF prediagnosis. A retrospective cohort study was conducted, where the baseline characteristics and subsequent care patterns of patients identified by the classifier were compared against those who did not receive a prediagnosis.

Results: The labeler model derived from the labeling functions showed high accuracy (0.92; F_1-score=0.77) on the training set. The classifier trained on the probabilistically labeled notes accurately identified patients with an AF prediagnosis (0.95; F_1-score=0.83). The cohort study conducted using the constructed system carried enough statistical power to verify the key findings of the Apple Heart Study, which enrolled a much larger number of participants, where patients who received a prediagnosis tended to be older, male, and White with higher CHA_2DS_2-VASc (congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category) scores (P<.001). We also made a novel discovery that patients with a prediagnosis were more likely to use anticoagulants (525/1037, 50.63% vs 5936/16,560, 35.85%) and have an eventual AF diagnosis (305/1037, 29.41% vs 262/16,560, 1.58%). At the index diagnosis, the existence of a prediagnosis did not distinguish patients based on clinical characteristics, but did correlate with anticoagulant prescription (P=.004 for apixaban and P=.01 for rivaroxaban).

Conclusions: Our work establishes the feasibility and efficacy of an EHR-based surveillance system for consumer wearables that render AF prediagnoses. Further work is necessary to generalize these findings for patient populations at other sites.

(JMIR Med Inform 2024;12:e51171) doi:10.2196/51171

KEYWORDS

consumer wearable devices; atrial fibrillation; postmarket surveillance; surveillance; monitoring; artificial intelligence; machine learning; natural language processing; NLP; wearable; wearables; labeler; heart; cardiology; arrhythmia; diagnose; diagnosis; labeling; classifier; EHR; electronic health record; electronic health records; consumer; consumers; device; devices; evaluation

Introduction

Background

Consumer-facing devices such as the Apple Watch [1] and Fitbit [2] now have the capability to notify users with a *prediagnosis* such as atrial fibrillation (AF). As these notifications may incentivize patients to seek follow-up medical care, wearables now have the potential to affect diagnosis rates and initiate cascades of medical care [3,4]. Although these devices undergo premarket validation to obtain Food and Drug Administration (FDA) clearance [5], limited information exists on their postmarket use and clinical utility.

To conduct *postmarket surveillance* on consumer wearables, electronic health records (EHRs) should capture wearable use, in particular those incidents where patients received prediagnosis notifications. However, EHRs are often built around medical diagnosis codes used for billing purposes [6,7], which do not contain terms for describing wearable use. Prescription wearables should have ordering information, but this does not capture how the wearables are used. Therefore, unstructured data such as clinical notes must be parsed to obtain the wearable use information.

Deep learning–based natural language processing (NLP) methods [8-10] have been shown to outperform traditional approaches on clinical note classification tasks [11,12]. However, these deep learning–based classifiers require large, hand-labeled training sets that are costly to generate. For EHR-based postmarket surveillance to be widely implemented, a scalable approach is necessary to reduce the labeling burden.

Objectives

We aimed to demonstrate the feasibility and efficacy of postmarket surveillance on consumer wearables that render AF prediagnoses. The first aim of this study was to evaluate the efficacy of a weakly supervised approach to heuristically generate labels for a training set. A *labeler model* derived from programmatically expressed heuristics probabilistically assigns labels to clinical notes regarding whether the note contains a mention of the patient receiving a prediagnosis from a wearable. The second aim was to evaluate the performance of a *classifier* fine-tuned on the training set labeled by the labeler model, which identifies mentions of an AF prediagnosis in a note. The third aim was to summarize the clinical characteristics of patients identified by the classifier and compare them to patients who were not alerted to a prediagnosis.

Methods

Cohort Identification

We used the Stanford Medicine Research Data Repository (STARR) data set [13], which contains EHR-derived records from the inpatient, outpatient, and emergency department visits at Stanford Health Care and the Lucile Packard Children's Hospital. We retrieved all clinical notes from the STARR data set that contain a mention of a wearable device (Textbox 1), resulting in 86,260 notes from 34,329 unique patients. Following the FDA guidance for pertinent cardiovascular algorithms [5], we excluded patients younger than 22 years of age when the note was written, leaving 78,323 notes from 30,133 unique patients. We further limited the data set to notes written on or after January 1, 2019, since the first consumer-facing AF detection feature became available in December 2018 [14]. The resulting cohort comprised 56,924 notes from 21,332 unique patients.

Textbox 1. Search terms for wearable devices.

Apple watch, iwatch, applewatch, fitbit, fit bit, fit-bit, galaxy watch, samsung watch, google watch, kardia, alivecor, alive cor, wearable, smart watch, and smartwatch

To evaluate the performance of the labeler model and the classifier, we constructed a test set by manually labeling 600 notes. Specifically, we randomly selected 600 unique patients and then selected 1 note for each patient that contained *action terms* (Textbox 2) in the vicinity (30 characters) of a wearable mention. This was to filter out nonrelevant wearable descriptions (eg, boilerplate texts recommending the use of wearables during meditation), so that resulting notes are enriched with relevant use cases.

Textbox 2. Action terms used to enrich sample relevance.

Alert, notify, warn, observe, identify, detect, note, record, capture, show, report, give, alarm, register, read, tell, have, had, see, saw, receive, get, got, notice, check, and confirm

These notes were then labeled independently by 2 data scientists, and differences were adjudicated by 2 physicians. A clinical note was labeled as positive when the patient received an automated AF notification from the wearable, or when the patient initiated an on-demand measurement (eg, electrocardiogram strip) that resulted in an AF prediagnosis. There were no instances where the 2 physicians disagreed on the label. The resulting test set contained 105 positive notes (prevalence=0.18).

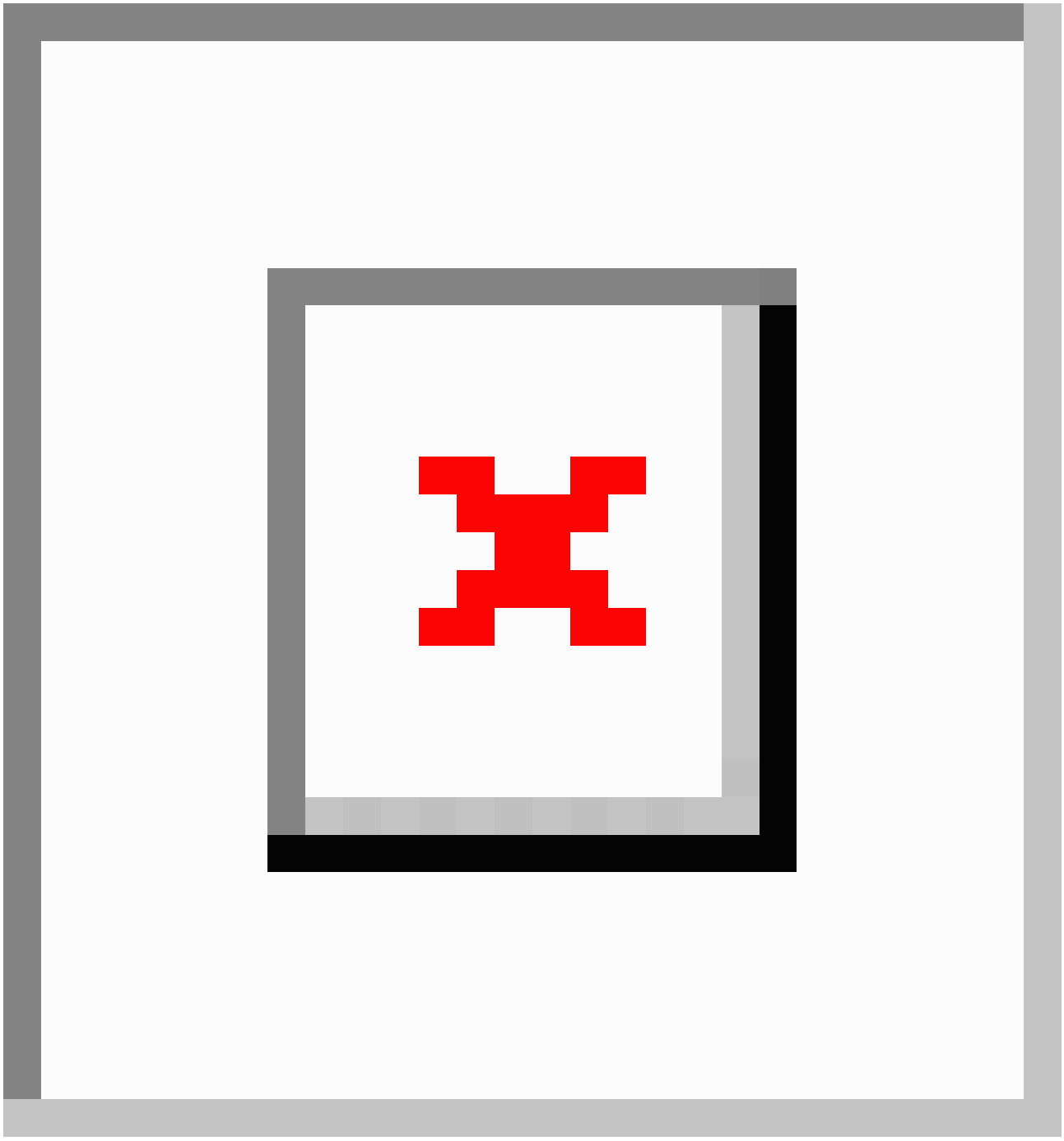
In addition to the test set, we prepared a development set of 600 notes that was used to aid the development of the labeler model. This set was manually labeled by a single data scientist, using a labeling guideline (Multimedia Appendix 1) that was developed as part of the test set generation. The development set contained 100 positive notes (prevalence=0.17).

Labeler Model Derivation

We then derived a labeler model that used weak supervision to probabilistically assign labels for the training set. Specifically,

as shown in Figure 1, we used data programming [15], where labeling heuristics are expressed as code-based *labeling functions*. Using the encoded heuristics, the labeling functions make predictions as to which label a clinical note should be assigned. Predictions from these labeling functions are then combined to develop a generative *labeler model*.

Figure 1. Labeler model generation process. Labeling heuristics were expressed as code-based labeling functions. Snorkel [16] then applied the labeling functions to the sample clinical notes and fit a generative model on the predictions of the labeling functions. The resulting labeler model probabilistically assigns a label to a clinical note based on whether the note mentions the patient receiving an AF prediagnosis from the wearable device. AF: atrial fibrillation.



We used the Snorkel framework [16] to implement data programming. A preprocessing framework [17] was applied to our notes to split them into sentences using the spaCy [18] framework, with a specialized tokenizer to recognize terms specific to medical literature. Thus parsed grammatical information was made available to the labeling functions as metadata.

We then used the development set to understand how the AF prediagnosis was described, and we expressed each pattern as a labeling function. The development process was iterative, where the Snorkel framework allowed us to observe the predictive values of the labeling functions on development set records. Each function could then be further optimized to reduce the differences between predictive values and actual labels, leading to overall performance improvement on the development

set. Textbox 3 shows all the terms that were identified as denoting AF. Negations were properly handled.

Once developed, we applied the labeling functions on the samples and then instructed Snorkel to fit a generative model on the output. Specifically, we used 10-fold cross-validation on the test set and chose the labeler model with the best F_1-score. This model was then applied to the entire corpus of 56,924 notes to probabilistically assign labels.

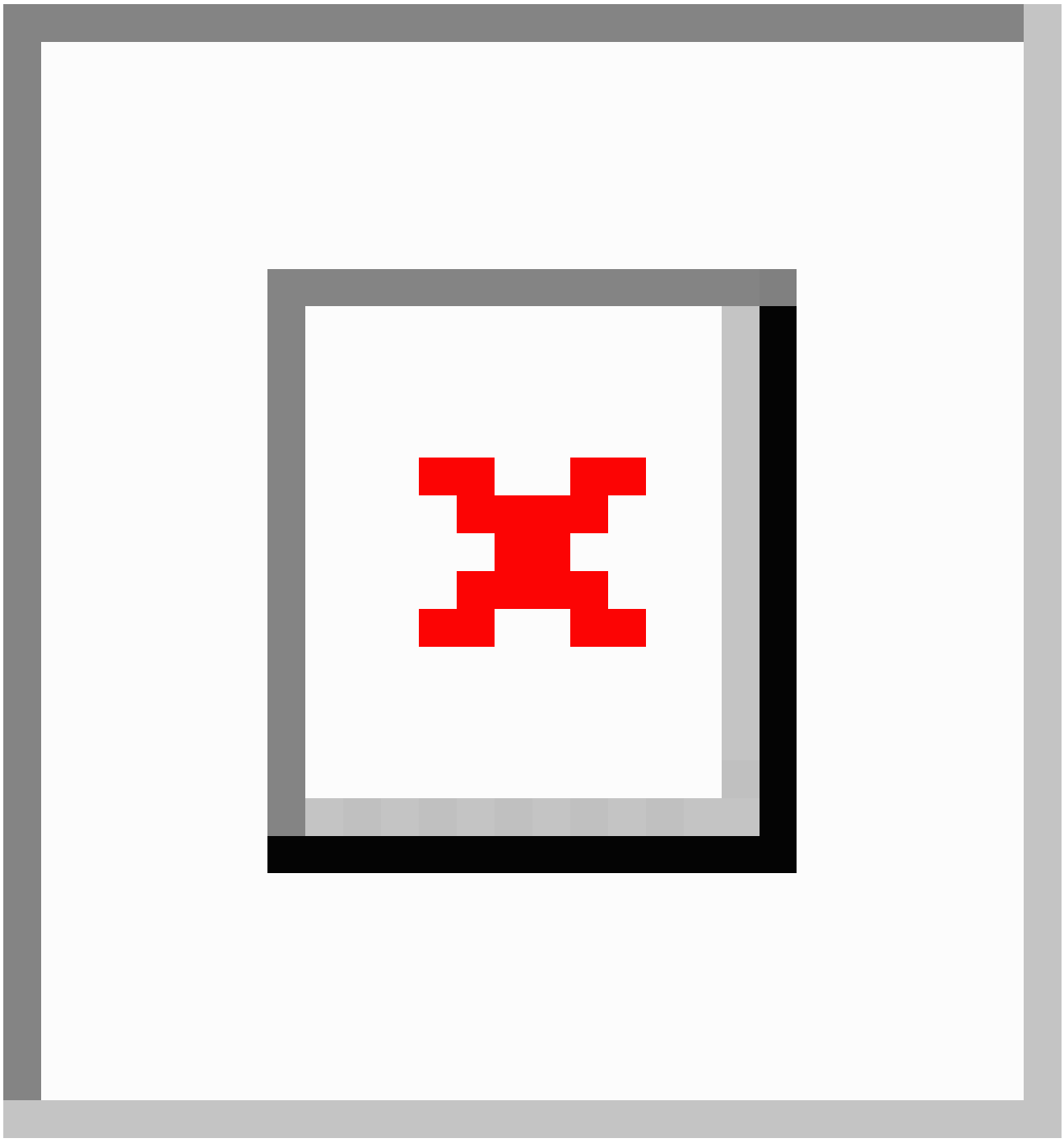
Textbox 3. Terms denoting atrial fibrillation.

Af, afib, a-fib, a.fib, arrhythmia, paf, atrial fibrillation, a. fib, a fib, atrial fib, atrial arrhythmia, irregular heartbeat, irregular hr, irregular rhythm, irregular pulse, irreg hr, irregular heart beat, irregular heart rhythm, irregular heart rate, irreg heart rhythm, irreg heart beat, irreg heart rate, abnormal ekg rhythm, paroxysmal atrial fibrillation, a. fib, a - fib, pafib, abnormal heart rhythm, abnormal rhythm, abnormal HR, and arrhythmia

Classifier Fine-Tuning

Notes that were probabilistically labeled by the labeler model were then used to fine-tune a large, NLP-based classifier: Clinical-Longformer [12] (Figure 2). The resulting classifier takes plain note text as the input and classifies the note as positive (ie, includes mention of a patient receiving an AF notification, or patient-initiated cardiac testing or electrocardiogram resulting in an AF prediagnosis) or negative. When a classifier is tuned on the labeler model output, it enables generalizing beyond the labeling heuristics encoded in the labeling functions, such that the classifier can recognize more patterns.

Figure 2. Classifier generation process. The labeler model was used to probabilistically assign labels for a large number of unlabeled clinical notes, which were then used to fine-tune a classifier to detect whether a patient received an AF prediagnosis from a wearable device. AF: atrial fibrillation; NLP: natural language processing.



Specifically, we fine-tuned the pretrained Clinical-Longformer for the sequence classification task, with varying training set sizes. For a single fine-tuning run, we chose the snapshot with the best F_1-score on the test set as the representative. The Adam optimizer was used, with the learning rate ramping up to 1×10^{-5} followed by linear decay over 3 epochs. Clinical-Longformer has a maximum input length of 4096 subword tokens: 94% (53,509/56,924) of our notes fit this criterion, and notes with more tokens were trimmed. Fine-tuning other NLP-based classifiers (eg, ClinicalBERT [11], which takes a smaller number of input tokens [512 or fewer]) resulted in abysmal performance numbers (F_1-score=0.21), hinting that they could not be properly fine-tuned on our lengthy clinical notes.

The test set was never presented to the classifier during the fine-tuning process. Since our data set was highly skewed toward negative samples, we stratified the training set to maintain a 1:2 ratio between the positive and negative notes. All samples were chosen randomly.

The classifier with the best F_1-score was then run across the entire set of 56,924 clinical notes to identify all incidents of AF prediagnoses.

Retrospective Cohort Study

Using the classifier, we identified patients who received an AF prediagnosis and performed 3 retrospective cohort studies comparing the characteristics of patients who received a prediagnosis to those who did not, using the same STARR data set.

First, we considered all the patients in the cohort regardless of their prior AF diagnosis. We compared the demographics, CHA_2DS_2-VASc (congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category) [19] score, and its related comorbidities on the date the index note was created. We defined the oldest note with a prediagnosis as the index note since it was the most likely to drive downstream medical intervention. When a patient had not received any prediagnosis, the oldest note with mention of a wearable was chosen as the index.

Second, we focused on patients who did not have a prior AF diagnosis. A patient was filtered out if the patient had received an AF diagnosis, defined as an ambulatory or inpatient encounter with SNOMED code 313217 and its descendants, prior to the index note. We then compared the same demographics and comorbidities between those who received a prediagnosis and those who did not, on the date the index note was created.

Lastly, we further confined the analysis to patients who received a clinician-assigned AF diagnosis within 60 days from the index note. Same as before, we excluded patients who had a prior AF diagnosis before the index note. Patients were then grouped based on whether they had received an AF prediagnosis from a wearable and characterized on the date they received the index AF diagnosis. In addition to the demographics and comorbidities, we also compared anticoagulant medication (Textbox 4), rhythm management medication (Textbox 5), and cardioversion rates between the 2 groups. Only the index prescription and procedure that took place within 60 days from the index diagnosis were considered.

Textbox 4. Anticoagulant medications analyzed in this study.

Warfarin

Direct oral anticoagulants

- Apixaban, dabigatran, rivaroxaban, edoxaban, and betrixaban

Textbox 5. Rhythm management medications analyzed in this study.

Class I antiarrhythmics

- Propafenone, disopyramide, quinidine, mexiletine, and flecainide

Class II antiarrhythmics

- Metoprolol, carvedilol, labetalol, nadolol, propranolol, carteolol, penbutolol, pindolol, atenolol, betaxolol, bisoprolol, esmolol, nebivolol, and timolol

Class III antiarrhythmics

- Sotalol and dofetilide

Class IV antiarrhythmics

- Verapamil, diltiazem, nicardipine, amlodipine, felodipine, nifedipine, isradipine, and nisoldipine

Others

- Digoxin

Statistical Analysis

When compiling patient race and ethnicity information, we used the 5 categories of race defined by the US Census and denoted Hispanic as a dedicated ethnicity. A total of 11.12% (2371/21,327) of the patients were missing race and ethnicity information, so we categorized them as belonging to the *undisclosed* category.

For hypothesis testing, we used the 1-tailed Welch *t* test for continuous variables and χ^2 test for categorical variables. One-tailed tests were chosen over 2-tailed tests since clinical contexts helped establish the comparison direction, providing for a stricter analysis. Statistical analysis was performed using *Pandas* [20] 1.3.0 and *SciPy* [21] 1.7.0, running on Python 3.9.6 configured through Conda 4.5.11.

Ethical Considerations

The STARR data set is derived from consented patients only. Patients were not compensated for participation. Data analyzed in this study were not deidentified, but its analysis was conducted in a HIPAA (Health Insurance Portability and Accountability Act)–compliant, high-security environment. The Stanford University Institutional Review Board approved this study (62865).

Results

Labeler Model Performance

In total, 8 labeling functions were developed. Most (7/8, 88%) labeling functions used the grammatical information present in the metadata, whereas 1 (12%) used a simple dictionary-based lookup. Table 1 provides the performance of each labeling function, followed by the combined labeler model.

Since each labeling function was geared toward identifying positive samples that follow a specific pattern, each labeling function exhibited substantially higher precision than recall. By combining these labeling functions into 1 generative labeler model, we improved recall (0.72). The high labeler model accuracy (0.92) also showed that the model correctly classified negative samples. After running the labeler model on the set of 56,924 clinical notes, 5829 notes were flagged as positive, a substantial increase from the 105 positive notes identified through manual labeling.

Table . Labeling function (LF) and labeler model performance[a].

Function or model	Target pattern	Example	Precision[b]	Recall[c]	F_1-score[d]	Accuracy[e]
LF1	Simple dictionary lookup	"AF" and "wearable" and "notification"	0.90	0.33	0.51	0.87
LF2	AF[f]+verb+prep[g]+wearable	"AF noted on wearable"	0.78	0.12	0.24	0.84
LF3	Wearable+verb+AF	"Wearable notified AF"	0.91	0.42	0.55	0.89
LF4	Verb+wearable+verb+AF	"Observed wearable showing AF"	0.85	0.14	0.29	0.85
LF5	Verb+AF+prep+wearable	"Received AF from wearable"	0.81	0.15	0.31	0.85
LF6	Verb+event+prep+wearable+AF	"Got notification from wearable of AF"	0.67	0.02	0.20	0.83
LF7	Event+prep+wearable+AF	"Notified on wearable of AF"	0.74	0.10	0.27	0.84
LF8	Wearable+subject+verb+AF	"Per wearable, patient had AF"	*0.96*	0.22	0.38	0.86
Labeler model	N/A[h]	N/A	0.84	*0.72*	*0.77*	*0.92*

[a]Averages taken from 10-fold cross-validation on the test set of 600 manually labeled notes. Italic numbers indicate the best observed performance for each metric.

[b]Precision = true positive / (true positive + false positive).

[c]Recall = true positive / (true positive + false negative).

[d]F_1-score = 2 × precision × recall / (precision + recall).

[e]Accuracy = (true positive + true negative) / (positive + negative).

[f]AF: atrial fibrillation.

[g]Prep: preposition.

[h]N/A: not available.

Classifier Performance

Here, we report the performance of the classifier that was fine-tuned using the clinical notes labeled by the labeler model. Table 2 shows the average performance of the classifier on the test set, across varying training set sizes. The training set size was capped at 15,000 to maintain the 1:2 positive-to-negative ratio (the labeler model labeled 5829 notes as positive). Regardless of the training set size, the test set was excluded from the input to the fine-tuning process.

Table . Classifier performance across varying training set sizes[a].

Training set size	Precision[b]	Recall[c]	F_1-score[d]	Accuracy[e]
600	0.37	0.68	0.48	0.73
5000	0.79	*0.85*	0.81	0.93
10,000	0.84	0.81	*0.83*	*0.94*
15,000	*0.85*	0.81	*0.83*	*0.94*

[a]For each training set, average values are reported across 3 runs with different random seeds. For each run, the classifier snapshot with highest F_1-score was used. Italic numbers indicate the best performance observed for each metric.

[b]Precision = true positive / (true positive + false positive).

[c]Recall = true positive / (true positive + false negative).

[d]F_1-score = 2 × precision × recall / (precision + recall).

[e]Accuracy = (true positive + true negative) / (positive + negative).

Table 2 demonstrates how classifier performance benefits from the weakly supervised approach. In particular, a training set size of 600 emulated the hypothetical scenario where the size of the training set is limited due to manual labeling overhead. Such a small data set was not enough to adequately fine-tune Clinical-Longformer (F_1-score=0.48).

As the training set size increased, the classifier obtained better performance, reaching the best average F_1-score of 0.83. When compared to the labeler model in Table 1 (recall=0.72), the classifier significantly improved recall (0.81), demonstrating that the classifier managed to generalize beyond the rules specified by the labeling functions.

Figures 3 and 4 show the comparisons of the best-performing (by F_1-score) classifiers from each training set size.

Figure 3. Classifier receiver operating characteristic (ROC) curve across varying training set sizes. For each training set, the best-performing (by F_1-score) run was chosen among 3 runs with different random seeds. For each run, the best-performing classifier snapshot was chosen.

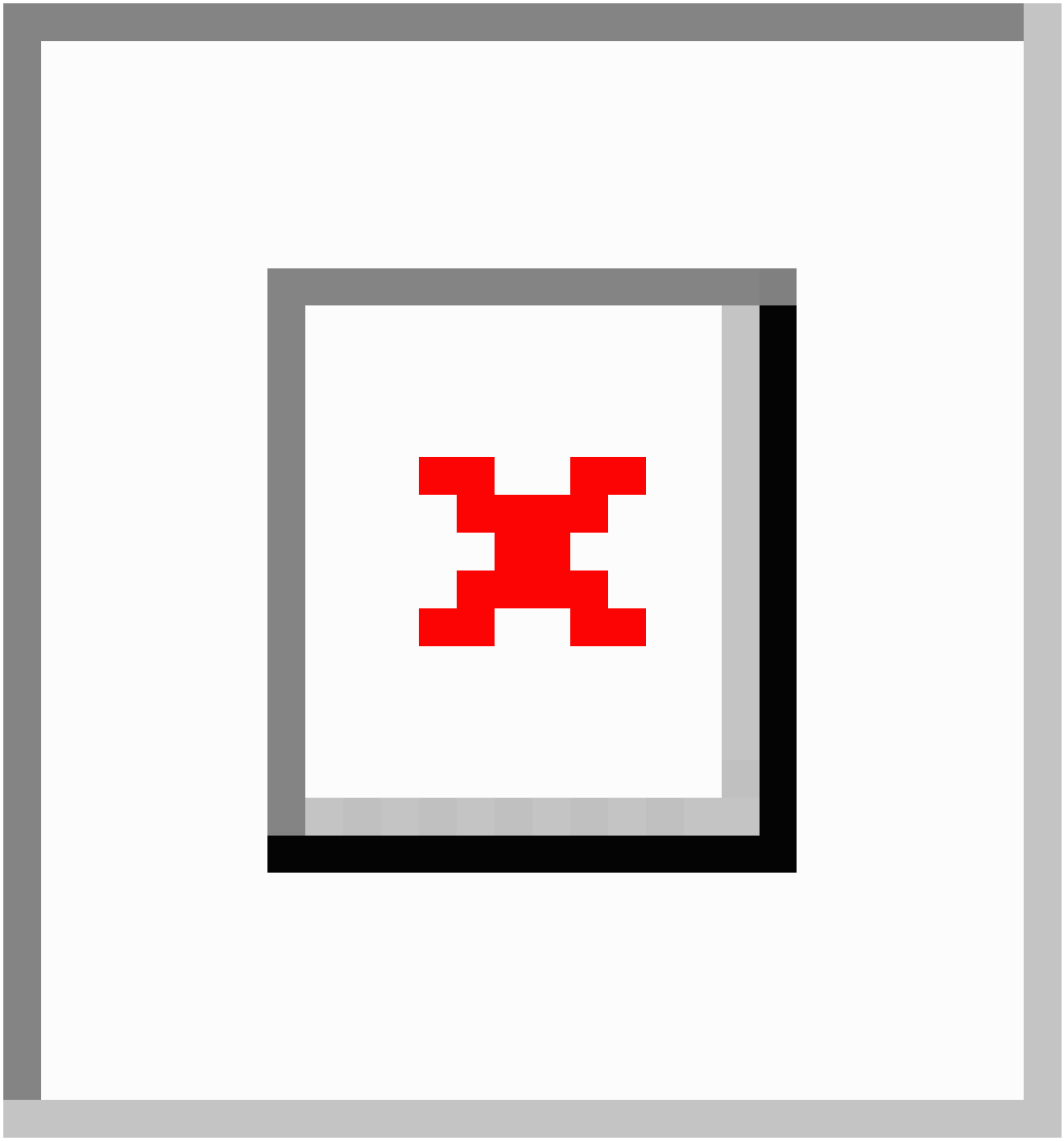
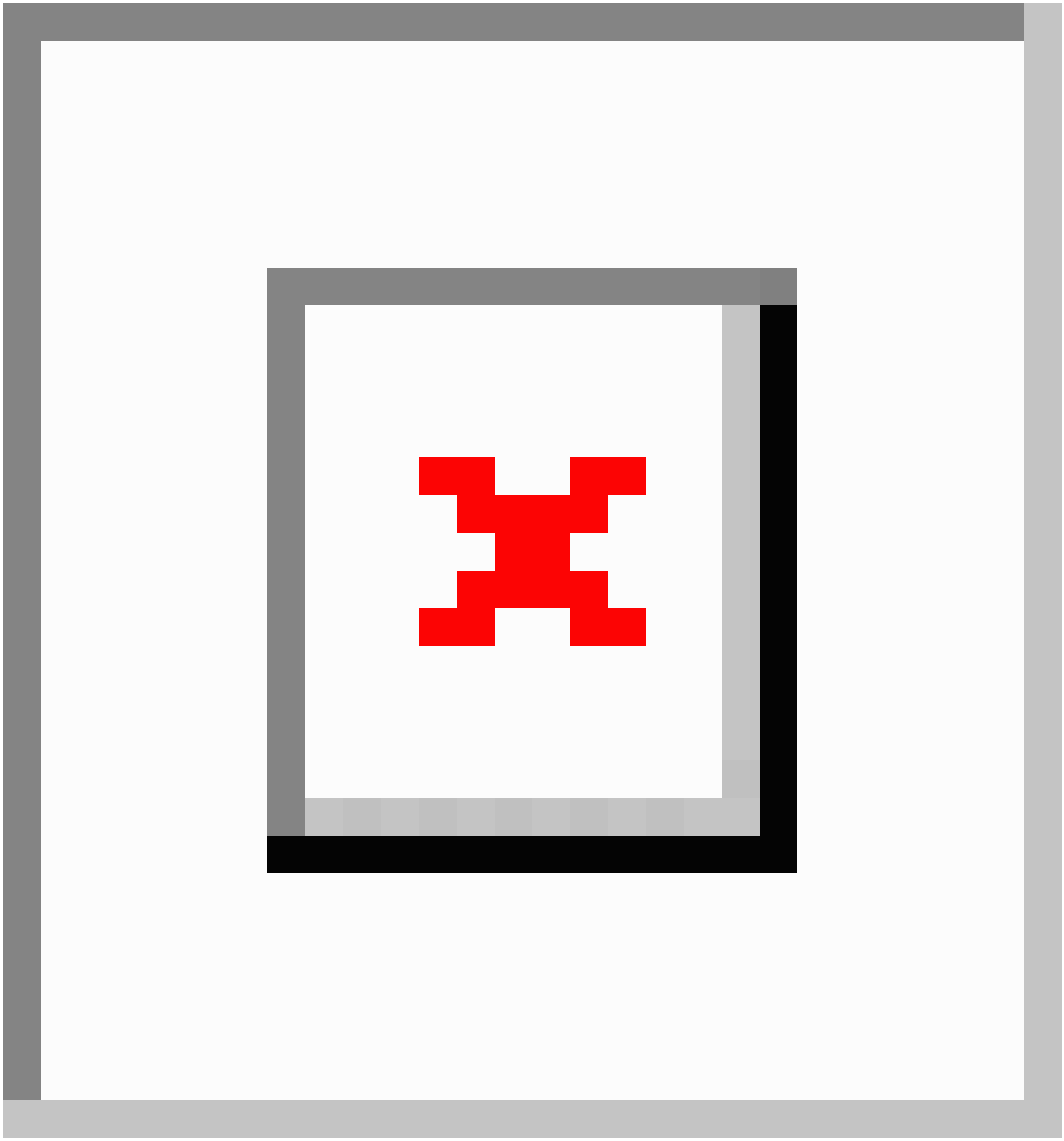


Figure 4. Classifier precision-recall curve across varying training set sizes. For each training set, the best-performing (by F_1-score) run was chosen among 3 runs with different random seeds. For each run, the best-performing classifier snapshot was chosen.



The receiver operating characteristic curve (Figure 3) shows that even the best classifier with a training set size of 600 performed worse than classifiers from larger data set sizes. In the precision-recall curve (Figure 4), the classifier lost substantial precision for small gains in recall, further hinting that the classifier was not properly trained.

Across all training set sizes and runs, the best-performing classifier achieved an F_1-score of 0.85 (accuracy=0.95). Running this classifier on 56,924 clinical notes identified 6515 notes as containing an AF prediagnosis across 2279 unique patients.

Cohort Study: All Patients

Table 3 summarizes the characteristics of the entire cohort regardless of their prior AF diagnosis, reflecting the characteristics of generic patients that used wearables. In all, 5 patients were missing sex information and were not included in the analysis.

Table . Characteristics of all patients[a].

Characteristics			With a prediagnosis (n=2279)	Without a prediagnosis (n=19,048)	*P* value
Demographics					
	Age (y), mean (SD)		63.85 (14.21)	53.53 (16.70)	<.001[b]
	Race and ethnicity, n (%)				<.001[b]
		Asian	295 (12.94)	3143 (16.5)	
		Black	53 (2.33)	619 (3.25)	
		Hispanic	96 (4.21)	1731 (9.09)	
		White	1613 (70.78)	11,240 (59.01)	
		Others	13 (0.57)	153 (0.8)	
		Undisclosed	209 (9.17)	2162 (11.35)	
	Sex, n (%)				<.001[b]
		Male	1384 (60.73)	7739 (40.63)	
		Female	895 (39.27)	11,309 (59.37)	
Comorbidities, n (%)					
	Congestive heart failure		341 (14.96)	1434 (7.53)	<.001[b]
	Hypertension		1267 (55.59)	6796 (35.68)	<.001[b]
	Diabetes mellitus		101 (4.43)	1018 (5.34)	.07
	Vascular disease		251 (11.01)	1582 (8.31)	<.001[b]
CHA_2DS_2-VASc[c] score, mean (SD)			2.12 (1.55)	1.61 (1.35)	<.001[b]

[a]Measured on the date of the index note.

[b]Statistically significant at α=.05.

[c]CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category.

Patients who received an AF prediagnosis from a wearable tended to be older, with more comorbidities except for diabetes mellitus. White and male individuals constituted a larger portion of patients with a prediagnosis, who also exhibited higher CHA_2DS_2-VASc scores.

Cohort Study: Patients Without a Prior AF Diagnosis

Table 4 then compares the characteristics of patients who had no AF diagnosis prior to the index note, highlighting the efficacy of wearables on the undiagnosed population.

Table . Characteristics of patients without a prior atrial fibrillation diagnosis[a].

Characteristics			With a prediagnosis (n=1037)	Without a prediagnosis (n=16,560)	*P* value
Demographics					
	Age (y), mean (SD)		60.16 (15.65)	51.54 (16.28)	<.001[b]
	Race and ethnicity, n (%)				<.001[b]
		Asian	127 (12.25)	2890 (17.45)	
		Black	28 (2.7)	553 (3.34)	
		Hispanic	55 (5.3)	1598 (9.65)	
		White	723 (69.72)	9414 (56.85)	
		Others	3 (0.29)	136 (0.82)	
		Undisclosed	101 (9.74)	1969 (11.89)	
	Sex, n (%)				<.001[b]
		Male	595 (57.38)	6241 (37.69)	
		Female	442 (42.62)	10,319 (62.31)	
Comorbidities, n (%)					
	Congestive heart failure		85 (8.2)	696 (4.2)	<.001[b]
	Hypertension		461 (44.46)	5082 (30.69)	<.001[b]
	Diabetes mellitus		42 (4.05)	805 (4.86)	.27
	Vascular disease		95 (9.16)	1090 (6.58)	.002[b]
CHA_2DS_2-VASc[c] score, mean (SD)			1.78 (1.44)	1.46 (1.23)	<.001[b]

[a]Measured on the date of the index note.

[b]Statistically significant at α=.05.

[c]CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category.

These patients exhibited similar characteristics to the overall cohort, where those who received an AF prediagnosis tended to be older, White, and male, with more comorbidities except for diabetes mellitus. In particular, 50.63% (525/1037) of the patients who received a prediagnosis had CHA_2DS_2-VASc scores of 2 or higher, warranting anticoagulation therapy [22]. In contrast, among the patients without a prediagnosis, only 35.85% (5936/16,560) had CHA_2DS_2-VASc scores of 2 or higher.

Cohort Study: Patients With a Clinician-Assigned AF Diagnosis

Among those patients who did not have a prior AF diagnosis, 29.41% (305/1037) of the patients with a wearable-assigned prediagnosis received a clinician-assigned AF diagnosis within 60 days from the index prediagnosis. The average duration from prediagnosis to diagnosis was 4.74 days. In contrast, only 1.58% (262/16,560) of those patients without a prediagnosis received a clinician-assigned AF diagnosis.

Table 5 compares the clinical characteristics of those patients who received an AF diagnosis, based on whether they had received a wearable-assigned prediagnosis prior to the diagnosis.

None of the patient characteristics reported in Table 5 differed significantly between those with an AF prediagnosis and those without (all *P*>.05). However, anticoagulant prescriptions differed based on AF prediagnoses, where more patients with a prediagnosis were prescribed apixaban and rivaroxaban.

Table . Characteristics of patients with a clinician-assigned atrial fibrillation diagnosis[a].

Characteristics		With a prediagnosis (n=305)	Without a prediagnosis (n=262)	*P* value
Demographics				
Age (y), mean (SD)		64.45 (14.16)	63.65 (14.29)	.75
Race and ethnicity, n (%)				.21
	Asian	35 (11.48)	27 (10.31)	
	Black	6 (1.97)	11 (4.20)	
	Hispanic	10 (3.28)	15 (5.73)	
	White	218 (71.48)	175 (66.79)	
	Others	1 (0.33)	5 (1.91)	
	Undisclosed	35 (11.48)	29 (11.07)	
Sex, n (%)				.86
	Male	193 (63.28)	163 (62.21)	
	Female	112 (36.72)	99 (37.79)	
Comorbidities, n (%)				
Congestive heart failure		14 (4.59)	21 (8.02)	.13
Hypertension		111 (36.39)	109 (41.6)	.24
Diabetes mellitus		10 (3.28)	4 (1.53)	.29
Vascular disease		24 (7.87)	30 (11.45)	.19
CHA_2DS_2-VASc[b] score, mean (SD)		1.76 (1.49)	1.81 (1.39)	.36
Diagnosis subtype, n (%)				.40
Generic		230 (75.41)	213 (81.3)	
Chronic		2 (0.66)	1 (0.38)	
Paroxysmal		68 (22.3)	45 (17.18)	
Persistent		5 (1.64)	3 (1.15)	
Anticoagulant, n (%)				
Warfarin		1 (0.33)	3 (1.15)	.51
Direct oral anticoagulants				
	Apixaban	76 (24.92)	39 (14.89)	.004[c]
	Rivaroxaban	29 (9.51)	10 (3.82)	.01[c]
Rhythm management, n (%)				
Class I antiarrhythmics				
	Propafenone	7 (2.3)	2 (0.76)	.26
	Flecainide	17 (5.57)	8 (3.05)	.21
Class II antiarrhythmics				
	Metoprolol	50 (16.39)	45 (17.18)	.89
	Carvedilol	1 (0.33)	3 (1.15)	.51
	Labetalol	6 (1.97)	4 (1.53)	.94
	Atenolol	3 (0.98)	5 (1.91)	.57
Class IV antiarrhythmics				
	Verapamil	3 (0.98)	2 (0.76)	>.99
	Diltiazem	15 (4.92)	9 (3.44)	.51

Characteristics		With a prediagnosis (n=305)	Without a prediagnosis (n=262)	*P* value
	Amlodipine	4 (1.31)	3 (1.15)	>.99
Others				
	Digoxin	3 (0.98)	1 (0.38)	.73
Procedures, n (%)				
Cardioversion		30 (9.84)	14 (5.34)	.07

[a]Measured on the date of the index atrial fibrillation diagnosis. Medications that were not prescribed are omitted.

[b]CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category.

[c]Statistically significant at α=.05.

Discussion

Principal Findings

In this study, we applied a weak supervision–based approach to demonstrate the feasibility and efficacy of an EHR-based postmarket surveillance system for consumer wearables that render AF prediagnoses.

We first derived a labeler model from labeling heuristics expressed as labeling functions, which showed high accuracy (0.92; F_1-score=0.77) on the test set. We then fine-tuned a classifier on labeler model output, to accurately identify AF prediagnoses (0.95; F_1-score=0.83).

Further, using the classifier output, we identified patients who received an AF prediagnosis from a wearable and conducted a retrospective analysis to compare the baseline characteristics and subsequent clinical treatment of these patients against those who did not receive a prediagnosis.

Across the entire cohort, patients with a prediagnosis were older with more comorbidities. The race and sex composition of these patients also differed from those who did not receive a prediagnosis (*P*<.001).

Focusing on the subgroup of patients without a prior AF diagnosis (Table 4), we observed that a higher percentage of patients (525/1037, 50.63% vs 5936/16,560, 35.85%) who received a wearable-assigned prediagnosis exhibited CHA_2DS_2-VASc scores that warranted a recommendation for anticoagulation therapy [22]. This increased likelihood for anticoagulation therapy could be attributed to an early prediagnosis from the wearable.

In the same subgroup, patients who received a prediagnosis were 18.61 times more likely to receive a clinician-assigned AF diagnosis than those who did not. The existence of a prediagnosis was not correlated with patient demographics, comorbidities, or AF subtype at the index diagnosis (Table 5) but did correlate with anticoagulant prescription, where patients with an AF prediagnosis were more frequently prescribed apixaban (*P*=.004) and rivaroxaban (*P*=.01).

Comparison With Prior Work

Given that more consumer wearables will be introduced with increasing prediagnostic capabilities, a surveillance framework for wearable devices is urgently needed to properly assess their impact on downstream health care [3,4]. However, publications sponsored by wearable vendors focused mostly on ascertaining the accuracy of the prediagnostic algorithm itself [1,2].

On the other hand, publications that sought to conduct postmarket surveillance relied solely on manual chart review [3,4], which is hard to scale. In a prior study on wearable notifications, clinician review of 534 clinical notes yielded only 41 patients with an AF prediagnosis [3]. With a weakly supervised approach, our clinician review of 600 notes (ie, the test set) allowed the subsequent identification of 2279 patients with a prediagnosis.

Such an improvement in recall enhanced the statistical power of our analysis. First, our cohort study findings that showed patients with an AF prediagnosis tended to be older, male, and White with higher CHA_2DS_2-VASc scores matches the key findings of the Apple Heart Study [1], which enrolled a much larger number of participants (n=419,297). Second, we were able to make a novel discovery in that a wearable-assigned prediagnosis increases the likelihood of patients receiving anticoagulation therapy and an eventual AF diagnosis, and we identified statistically meaningful anticoagulant prescription differences.

Prior work has applied various methods of weakly supervised learning to some form of medical surveillance [16,17,23-25]. Most relevantly, Callahan et al [23] implemented a surveillance framework for hip implants, and Sanyal et al [25] implemented one for insulin pumps. To the best of our knowledge, however, our work is the first to apply a weakly supervised approach to consumer wearable surveillance. Without prescription records, consumer wearable surveillance can be challenging to scale.

Limitations

We acknowledge that the STARR data set is confined to a small health care system in a single geographic region, which is known [13] to serve populations with higher percentages of male, White, and older individuals. We recommend other institutions to monitor their patient population by developing their own surveillance framework using our weakly supervised methodology. In fact, work is already underway to adapt this approach for use at Palo Alto Veterans Affairs.

We could not establish causality between prediagnoses and patient characteristics. The fact that patients who are older, with more comorbidities; White; and male had a higher likelihood

of receiving an AF prediagnosis may very well reflect that they are health conscious and use wearables more frequently.

Conclusions

By providing prediagnoses, consumer wearables have the potential to affect subsequent diagnoses and downstream health care. Postmarket surveillance of wearables is necessary to understand the impact but is hindered by the lack of codified terms in EHRs to capture wearable use. By applying a weakly supervised methodology to efficiently identify wearable-assigned AF prediagnoses from clinical notes, we demonstrate that such a surveillance system could be built.

The cohort study conducted using the constructed system carried enough statistical power to verify the key findings of the Apple Heart Study, which enrolled a much larger number of patients, where patients who received a prediagnosis tended to be older, male, and White with higher CHA_2DS_2-VASc scores. We also made a novel discovery in that a prediagnosis from a wearable increases the likelihood for anticoagulant prescription and an eventual AF diagnosis. At the index diagnosis, the existence of a prediagnosis from a wearable did not distinguish patients based on clinical characteristics but did correlate with anticoagulant prescription.

Our work establishes the feasibility and efficacy of an EHR-based surveillance system for consumer wearable devices. Further work is necessary to generalize these findings for patient populations at other sites.

Authors' Contributions

RMY, BTV, KNP, JAF, and NHS contributed to concept and design. JAF contributed to the acquisition of data. RMY, BTV, KNP, JAF, AZ, TP, and ND contributed to the analysis and interpretation of data. RMY and BTV contributed to the drafting of the manuscript. RMY, KNP, JAF, AZ, TP, ND, and NHS contributed to critical revision of the manuscript for important intellectual content. RMY contributed to statistical analysis. NHS contributed to the provision of patients or study materials, obtaining funding, and supervision. JAF and NHS contributed to administrative, technical, or logistic support.

Conflicts of Interest

KNP receives research grants from the American Heart Association and the American College of Cardiology and is a consultant for Evidently and 100Plus. JAF is a research consultant for Snorkel AI.

Multimedia Appendix 1
Labeling guideline developed as part of the test set generation.
[DOCX File, 46 KB - medinform_v12i1e51171_app1.docx]

References

1. Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019 Nov 14;381(20):1909-1917. [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
2. Lubitz SA, Faranesh AZ, Selvaggi C, et al. Detection of atrial fibrillation in a large population using wearable devices: the Fitbit Heart Study. Circulation 2022 Nov 8;146(19):1415-1424. [doi: 10.1161/CIRCULATIONAHA.122.060291] [Medline: 36148649]
3. Wyatt KD, Poole LR, Mullan AF, Kopecky SL, Heaton HA. Clinical evaluation and diagnostic yield following evaluation of abnormal pulse detected using Apple Watch. J Am Med Inform Assoc 2020 Jul 1;27(9):1359-1363. [doi: 10.1093/jamia/ocaa137] [Medline: 32979046]
4. Feldman K, Duncan RG, Nguyen A, et al. Will Apple devices' passive atrial fibrillation detection prevent strokes? estimating the proportion of high-risk actionable patients with real-world user data. J Am Med Inform Assoc 2022 May 11;29(6):1040-1049. [doi: 10.1093/jamia/ocac009] [Medline: 35190832]
5. Device classification under section 513(F)(2)(De Novo). US Food and Drug Administration. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN180044 [accessed 2023-03-05]
6. ICD-10. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/Medicare/Coding/ICD10 [accessed 2023-04-28]
7. List of CPT/HCPCS codes. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/medicare/fraud-and-abuse/physicianselfreferral/list_of_codes [accessed 2023-04-28]
8. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017): Curran Associates Inc; 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [accessed 2023-03-05]
9. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers): Association for Computational Linguistics; 2019:4171-4186. [doi: 10.18653/v1/N19-1423]
10. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv. Preprint posted online on Apr 10, 2020. [doi: 10.48550/arXiv.2004.05150]

11. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
12. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc 2023 Jan 18;30(2):340-347. [doi: 10.1093/jamia/ocac225] [Medline: 36451266]
13. Datta S, Posada J, Olson G, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv. Preprint posted online on Mar 17, 2020. [doi: 10.48550/arXiv.2003.10534]
14. ECG app and irregular heart rhythm notification available today on Apple Watch. Apple. 2018 Dec 6. URL: https://www.apple.com/newsroom/2018/12/ecg-app-and-irregular-heart-rhythm-notification-available-today-on-apple-watch/ [accessed 2023-3-5]
15. Ratner A, De Sa C, Wu S, Selsam D, Ré C. Data programming: creating large training sets, quickly. In: NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems: Curran Associates Inc; 2016:3574-3582. [doi: 10.5555/3157382.3157497]
16. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. VLDB J 2020;29(2):709-730. [doi: 10.1007/s00778-019-00552-1] [Medline: 32214778]
17. Fries JA, Steinberg E, Khattar S, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun 2021 Apr 1;12(1):2017. [doi: 10.1038/s41467-021-22328-4] [Medline: 33795682]
18. Industrial-strength natural language processing in Python. spaCy. URL: https://spacy.io/ [accessed 2023-03-05]
19. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro Heart Survey on Atrial Fibrillation. Chest 2010 Feb;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
20. McKinney W. Data structures for statistical computing in Python. Presented at: 9th Python in Science Conference (SciPy 2010); Jun 28 to Jul 3, 2010; Austin, Texas p. 56-61. [doi: 10.25080/Majora-92bf1922-00a]
21. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272. [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
22. The revised ACC/AHA/HRS guidelines for the management of patients with atrial fibrillation. American College of Cardiology. 2014 Oct 29. URL: https://www.acc.org/latest-in-cardiology/articles/2014/10/14/11/02/the-revised-acc-aha-hrs-guidelines-for-the-management-of-patients-with-atrial-fibrillation [accessed 2023-03-12]
23. Callahan A, Fries JA, Ré C, et al. Medical device surveillance with electronic health records. NPJ Digit Med 2019 Sep 25;2:94. [doi: 10.1038/s41746-019-0168-z] [Medline: 31583282]
24. Datta S, Roberts K. Weakly supervised spatial relation extraction from radiology reports. JAMIA Open 2023 Apr 22;6(2):ooad027. [doi: 10.1093/jamiaopen/ooad027] [Medline: 37096148]
25. Sanyal J, Rubin D, Banerjee I. A weakly supervised model for the automated detection of adverse events using clinical notes. J Biomed Inform 2022 Feb;126:103969. [doi: 10.1016/j.jbi.2021.103969] [Medline: 34864210]

Abbreviations

AF: atrial fibrillation
CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category
EHR: electronic health record
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
NLP: natural language processing
STARR: Stanford Medicine Research Data Repository

Edited by C Lovis; submitted 29.07.23; peer-reviewed by D Teo, L Wu; revised version received 15.01.24; accepted 04.02.24; published 04.04.24.

Please cite as:
Yoo RM, Viggiano BT, Pundi KN, Fries JA, Zahedivash A, Podchiyska T, Din N, Shah NH
Scalable Approach to Consumer Wearable Postmarket Surveillance: Development and Validation Study
JMIR Med Inform 2024;12:e51171
URL: https://medinform.jmir.org/2024/1/e51171
doi:10.2196/51171

Effect of Implementing an Informatization Case Management Model on the Management of Chronic Respiratory Diseases in a General Hospital: Retrospective Controlled Study

Yi-Zhen Xiao[1], MBBS; Xiao-Jia Chen[1], MBBS; Xiao-Ling Sun[1], MBBS; Huan Chen[1], MM; Yu-Xia Luo[1], MBBS; Yuan Chen[1], MBBS; Ye-Mei Liang[2], MM

1

2

Corresponding Author:
Ye-Mei Liang, MM

Abstract

Background: The use of chronic disease information systems in hospitals and communities plays a significant role in disease prevention, control, and monitoring. However, there are several limitations to these systems, including that the platforms are generally isolated, the patient health information and medical resources are not effectively integrated, and the "Internet Plus Healthcare" technology model is not implemented throughout the patient consultation process.

Objective: The aim of this study was to evaluate the efficiency of the application of a hospital case management information system in a general hospital in the context of chronic respiratory diseases as a model case.

Methods: A chronic disease management information system was developed for use in general hospitals based on internet technology, a chronic disease case management model, and an overall quality management model. Using this system, the case managers provided sophisticated inpatient, outpatient, and home medical services for patients with chronic respiratory diseases. Chronic respiratory disease case management quality indicators (number of managed cases, number of patients accepting routine follow-up services, follow-up visit rate, pulmonary function test rate, admission rate for acute exacerbations, chronic respiratory diseases knowledge awareness rate, and patient satisfaction) were evaluated before (2019 - 2020) and after (2021 - 2022) implementation of the chronic disease management information system.

Results: Before implementation of the chronic disease management information system, 1808 cases were managed in the general hospital, and an average of 603 (SD 137) people were provided with routine follow-up services. After use of the information system, 5868 cases were managed and 2056 (SD 211) patients were routinely followed-up, representing a significant increase of 3.2 and 3.4 times the respective values before use (U=342.779; P<.001). With respect to the quality of case management, compared to the indicators measured before use, the achievement rate of follow-up examination increased by 50.2%, the achievement rate of the pulmonary function test increased by 26.2%, the awareness rate of chronic respiratory disease knowledge increased by 20.1%, the retention rate increased by 16.3%, and the patient satisfaction rate increased by 9.6% (all P<.001), while the admission rate of acute exacerbation decreased by 42.4% (P<.001) after use of the chronic disease management information system.

Conclusions: Use of a chronic disease management information system improves the quality of chronic respiratory disease case management and reduces the admission rate of patients owing to acute exacerbations of their diseases.

(JMIR Med Inform 2024;12:e49978) doi:10.2196/49978

KEYWORDS

chronic disease management; chronic respiratory disease; hospital information system; informatization; information system; respiratory; pulmonary; breathing; implementation; care management; disease management; chronic obstructive pulmonary disease; case management

Introduction

Chronic obstructive pulmonary disease (COPD) and asthma are examples of common chronic respiratory diseases. The prevalence of COPD among people 40 years and older in China is estimated to be 13.7%, with the total number of patients reaching nearly 100 million. The lengthy disease cycle, recurrent acute exacerbations, and low control rate were found to have a significant impact on the prognosis and quality of life of middle-aged and older patients with COPD [1,2]. Therefore, to decrease the morbidity and disability rates and enhance the quality of life of all patients with chronic respiratory diseases, it is crucial to investigate effective prevention and treatment methods and establish a life cycle management model for chronic respiratory diseases.

Since the development of information technology, the internet and medical technology have been applied to the management of chronic diseases [3]. The chronic disease information systems adopted in hospitals and communities, along with mobile medical apps, can enhance the self-management capabilities of patients and play a significant role in disease prevention, control, and monitoring [4-9]. However, the existing platforms are generally isolated, the patient health information and medical resources are not effectively integrated, and the Internet Plus Healthcare technology model is not implemented throughout the patient consultation process [3,9].

Yulin First People's Hospital developed a chronic disease management information system based on the hospital information system (HIS) to fully and effectively utilize the medical resources in hospitals and to better support and adapt the system to the needs of patients with chronic diseases. In this study, we evaluated the impact of the use of this system on the efficacy of case management for patients with chronic respiratory diseases.

Methods

Chronic Respiratory Diseases Case Management Model Prior to Implementation of the Chronic Disease Management Information System

Yulin First People's Hospital is a public grade-3 general hospital with 2460 open beds, a specialty clinic in the Department of Pulmonary and Critical Care Medicine, and 180 beds in the Inpatient Department. Chronic respiratory diseases case management was initiated in 2019, which did not involve the use of an information system and was implemented by a chronic respiratory diseases case management team led by two nurses qualified as case managers, one chief physician, two supervisor nurses, and one technician. Under this system, patients with COPD, bronchial asthma, bronchiectasis, pulmonary thromboembolism, lung cancer, and lung nodules were managed using the traditional inpatient-outpatient-home chronic respiratory diseases case management model, including 1024 cases managed from 2019 to 2020. Except for medical prescriptions and electronic medical records, the patient case management information such as the basic information form, follow-up form, patient enrollment form, inpatient follow-up register, patient medication and inhalation device use records, smoking cessation and vaccination records, and pulmonary rehabilitation and health education records was managed using Microsoft Excel forms that were regularly printed for filing.

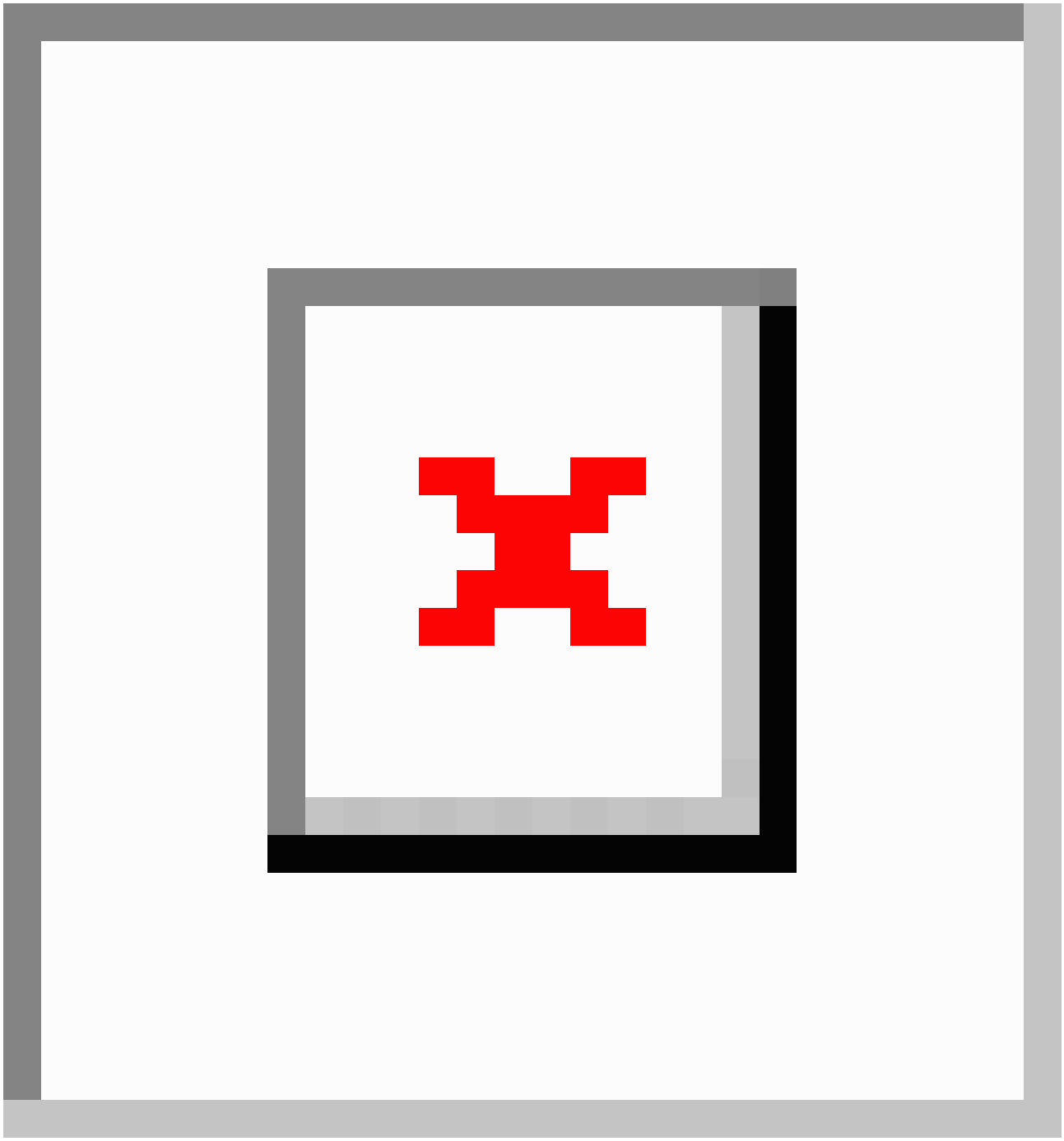
Establishment of a Management Information System for Chronic Diseases

The information carrier forming the basis of the management information system is constituted by the model of internet technology, chronic disease case management models, and overall quality management. The key technology is to establish a scientific, refined, and feasible follow-up pathway according to the methods and procedures of chronic disease case management based on the guidelines for the diagnosis and treatment of single chronic diseases. The closed-loop management of the clinical pathway was conducted in accordance with the Deming cycle (plan-do-check-act), and dynamic monitoring of single-disease health-sensitive and quality-sensitive indicators was carried out. The successfully developed system was installed on the hospital server to connect personal terminals (medical terminals and customer apps) to the existing HIS, which includes electronic medical records and medical advice.

Using the single-disease path assessment or plan scale as a framework, the system can automatically collect and integrate the majority of the medical information of patients with chronic respiratory diseases and provide these patients with inpatient, outpatient, and home intelligent medical services. Patients with chronic diseases who enroll in use of the system can use the app to schedule appointments for medical guidance, payment, and result queries; receive health guidance information; perform self-health assessments; write a treatment diary; and obtain medical communication materials.

The medical terminal consists of five functional modules: user entry, data statistics and query, quality control, knowledge base, and module management. As the core of the system, the user entry module can manage case information in seven steps: enrollment, assessment, planning, implementation, feedback, evaluation, and settlement [10-14]. Each step has a corresponding assessment record scale as well as the health-sensitive and quality-sensitive indicators. The structure of the HIS-based chronic disease management information system is shown in Figure 1.

Figure 1. Main structure of the hospital information system–based chronic disease management information system. *ICD*: *International Classification of Diseases*.



Implementation of the Chronic Disease Information Management System

Overview

Using the chronic disease management information system, two full-time case managers oversaw the case management of 2747 patients diagnosed with six diseases among chronic respiratory diseases between 2021 and 2022. The operation process was broken down into enrollment, assessment, planning, implementation, evaluation, feedback, and settlement stages.

Enrollment

Case managers entered the system through the medical app, selected a disease and an enrolled patient from the list of patients (the system automatically captures the patient's name and ID number according to the *International Classification of Diseases* [*ICD*] code) in accordance with the chronic respiratory diseases diagnostic criteria to sign the enrollment contract and determine the relationship between the personal information and data [15-19].

Assessment

The system can be seamlessly integrated with multiple workstations on the HIS to automatically capture the basic

information, electronic medical records, medical advice, and inspection materials, and can generate questionnaires or assessment scales for patients with chronic respiratory diseases such as the COPD Assessment Test, Asthma Control Test, modified Medical Research Council scale, form for lung function test results, inhalation device technique evaluation form, 6-minute walk test record, rehabilitation assessment form, health promotion form, and nutritional assessment form. The above materials can be added or removed based on the requirements for individual patients.

Planning

The case managers drafted the follow-up plan based on the patient assessment criteria and included the patients on the 1-, 3-, 6-, and 12-month follow-up lists. If the patient satisfied the self-management and indicator control requirements after follow-up, they could be settled and included in the annual follow-up cohort. Case managers can set up follow-up warning and treatment, involving the return visit plan, health education, follow-up content, pathway, and time, and notify the patients and nurses on day 7 and at months 1, 3, 6, and 12 after discharge. The nurses should promptly deal with patients who miss their scheduled follow-up visit.

Implementation

During the inpatient or outpatient care, supervising physicians, nurses, and patients collaborated with each other to implement the treatment. Case managers monitored the patients, evaluated them, documented the results, interpreted various test indicators, and provided health guidance. The chronic disease management information system acquired the corresponding data for chronic disease–sensitive indicators from outpatient and inpatient orders and medical records automatically. The chronic respiratory diseases management team reviewed the patients' conditions and the dynamics of chronic disease–sensitive indicators to make accurate decisions based on the current situation. The outpatient physicians obtained the single-disease package advice and personalized prescriptions to modify the diagnosis and treatment scheme.

Evaluation

Case managers highlighted evaluation and health education. First, they assessed and examined the content of the previous education and recorded and analyzed the patients' conditions, medication, diet, nutrition, rehabilitation exercises, and self-management. Second, they prepared the personalized health education plan, return visit plan, and rehabilitation plan, and used standardized courseware, educational videos, and health prescriptions to provide the patients with one-on-one health guidance. Finally, they sent the management tasks and educational contents to the phones of the patients for consolidating the learning in the hospital, as an outpatient, and at home.

Feedback

Patients can access their biochemical, physical, and chemical data as well as chronic disease–sensitive indicators in the hospital, as an outpatient, and at home for self-health management. Case managers can also perform online assessment, appraisal, and guidance via telephone, WeChat, and the chronic disease information system and record the data. Client mobile terminals can receive SMS text message alerts and the main interface of the chronic disease information system would display reminders of follow-up and return visits within ±7 days.

Settlement

If a patient was out of contact for 3 months, died, or refused to accept the treatment, case managers could settle the case.

Evaluation of the Effect of Implementing the Chronic Disease Management Information System

Evaluation Method

In accordance with case quality management indicators [20], two full-time case managers collected and evaluated data in the process of the follow-up procedure. To reduce the potential for evaluation bias, the case managers consistently communicated and learned to standardize the evaluation method. The cases were divided based on different chronic respiratory diseases case management models (ie, before and after use of the chronic disease information system). The following case management quality indicators were evaluated under the noninformation system management model (2019 - 2020) and under the chronic disease management information system model (2021 - 2022): number of managed cases, number of patients accepting routine follow-up services, follow-up visit rate, pulmonary function test rate, admission rate for acute exacerbations, chronic respiratory diseases knowledge awareness rate, and patient satisfaction. Excel sheets were used to acquire data prior to incorporation of the chronic disease management information system into the new information system.

Evaluation Indicators

The annual number of cases was calculated as the sum of the number of newly enrolled patients and the number of initially enrolled patients. The number of cases was calculated as the sum of the number of cases in different years. The number of routine follow-up visits represents the number of patients who completed the treatment plan. The follow-up visit rate was calculated as the number of completed follow-up visits in the year divided by the number of planned follow-up visits in the same year. The pulmonary function test rate was calculated as the number of pulmonary function tests completed for patients scheduled for follow-up during the year divided by the number of pulmonary function tests for patients scheduled for follow-up during the year. The admission rate for acute exacerbations was calculated as the number of recorded patients admitted to the hospital due to acute exacerbations divided by the total number of patients recorded. The chronic respiratory diseases knowledge awareness rate was determined by the number of people having sufficient knowledge divided by the total number of people tested. This knowledge indicator was based on the self-prepared chronic respiratory diseases knowledge test scale, which consists of 10 items determined using the Delphi method (following expert consultation) through review of the literature, including common symptoms, disease hazards, treatment medication, diet, living habits, exercise, negative habits affecting the disease, regular review items, effective methods for cough and sputum removal, appointments, and follow-ups. The content of the

questionnaire was refined by disease type, and the reviewers included 11 personnel with the title of Deputy Chief Nurse or above in the Internal Medicine Department of the hospital. The expert authority coefficients were 0.85 and 0.87 and the coordination coefficients were 0.50 and 0.67 for the two rounds of review, respectively; the χ^2 test showed a statistically significant value of P=.01. Patient satisfaction was assessed with a self-made questionnaire that showed good internal reliability (Cronbach α=0.78) and content validity (0.86). The questionnaire items included the reminder of return visits, practicability of health education content, and service attitude of medical staff; the full-time case managers surveyed the patients (or their caregivers) at the time of return visits after the third quarter of each year. Satisfaction items were rated using a 5-point Likert scale with a score of 1 - 5, and a mean ≥4 points for an individual indicated satisfaction. Patient satisfaction was then calculated as the number of satisfied patients divided by the total number of managed patients.

Statistical Analysis

SPSS 16.0 software was used for data analysis. The Mann-Whitney U test was performed to compare continuous variables between groups and the χ^2 test was performed to compare categorical variables between groups. P<.05 indicated that the difference was statistically significant.

Ethical Considerations

The study was conducted in accordance with the principles of the Declaration of Helsinki. This study received approval from the Ethics Committee of Yulin First People's Hospital (approval number: YLSY-IRB-RP-2024005). The study did not interfere with routine diagnosis and treatment, did not affect patients' medical rights, and did not pose any additional risks to patients. Therefore, after discussion with the Ethics Committee of Yulin First People's Hospital, it was decided to waive the requirement for informed consent from patients. Patients' personal privacy and data confidentiality have been upheld throughout the study.

Results

Characteristics of Patient Populations Before and After Implementation of the Information System

There was no significant difference in age and gender distributions in the patient populations that received care before and after implementation of the chronic disease management information system (Table 1).

Table . General characteristics of the patient populations under case management before and after use of the chronic disease management information system.

Characteristic		Before use (n=1024)	After use (n=2747)	χ^2 value	*df*	*P* value
Gender, n (%)				1.046	1	.31
	Men	677 (66.1)	1767 (64.3)			
	Women	347 (33.9)	980 (35.7)			
Age group (years), n (%)				0.997	3	.80
	<30	26 (2.6)	73 (2.7)			
	30-59	370 (36.1)	1013 (36.9)			
	60-79	510 (49.8)	118 (11.5)			
	>80	1322 (48.1)	339 (12.3)			

Comparison of Workload Before and After Implementation of the Information Management System

Before use of the system, 1808 cases were managed, with a mean of 603 (SD 137) cases having routine follow-up visits. After use of the system, 5868 cases were managed, with a mean of 2056 (SD 211) routine follow-up visits. Therefore, the number of managed cases and the number of follow-up visits significantly increased by 3.2 and 3.4 times, respectively, after use of the system (U=342.779; P<.001).

Comparison of Quality Indicators of Managed Cases Before and After Implementation of the Information System

The quality indicators in the two groups are summarized in Table 2. Compared with the corresponding indicators before use of the system, the follow-up visit rate increased by 50.2%, the pulmonary function test rate increased by 26.2%, the chronic respiratory diseases knowledge awareness rate increased by 20.1%, the retention rate increased by 16.3%, and the patient satisfaction increased by 9.6%; moreover, the admission rate for acute exacerbations decreased by 42.4%.

Table . Comparison of case management quality indicators before and after implementation of the chronic diseases information management system.

Quality indicators	Before use (n=1024), n (%)	After use (n=2747), n (%)	χ^2 value (*df*=1)	*P* value
Subsequent visit rate	209 (20.4)	1939 (70.6)	7.660	<.001
Lung function test achievement rate	190 (18.6)	1231 (44.8)	2.190	<.001
CRD[a] knowledge awareness rate	443 (43.3)	1742 (63.4)	1.243	<.001
Retention rate	787 (76.9)	2560 (93.2)	1.995	<.001
Acute exacerbation admission rate	663 (64.7)	613 (22.3)	5.999	<.001
Patient satisfaction	862 (84.2)	2577 (93.8)	86.190	.01

[a]CRD: chronic respiratory disease.

Discussion

Principal Findings

The main purpose of this study was to build a chronic disease management information system and apply it to the case management of chronic respiratory diseases. Our evaluation showed that the chronic disease management information system not only improves the efficiency and quality of case management but also has a benefit for maintaining the stability of the condition for patients with respiratory diseases, reduces the number of acute disease exacerbations, increases the rate of outpatient return, and improves patients' adherence with disease self-management. Thus, a chronic disease management information system is worth popularizing and applying widely.

Value of the HIS-Based Chronic Disease Management Information System

Chronic diseases constitute a significant public health issue in China. Public hospitals play important roles in the health service system, particularly large-scale public hospitals with the most advanced technologies, equipment, and enormous medical human resources, which can greatly aid in the diagnosis and treatment of diseases and also serve as important hubs for the graded treatment of chronic diseases. Moreover, a significant number of patients with chronic diseases visit large hospitals, making them important sources of big data on chronic diseases [21]. Adoption of an HIS-based chronic disease management information system can make full use of and exert the advantages of large-scale public hospitals in terms of labor, technology, and equipment in the diagnosis, treatment, and prevention of chronic diseases; enhance the cohesiveness of the case management team in chronic disease management; and achieve prehospital, in-hospital, and posthospital continuity of care for patients with chronic diseases. Overall, use of a chronic disease management information system can enhance the quality and efficiency of chronic disease management and lay a good foundation for teaching and research on chronic diseases.

Improved Efficiency of Case Management

China was relatively late in applying case management practices, and chronic disease management has traditionally been primarily conducted offline [14,20] or supplemented by management with apps and WeChat [7,8]. Traditional case management methods require case managers to manually search, record, store, query, count, and analyze information. This manual process necessitates substantial time and makes it challenging to realize a comprehensive, systematic, and dynamic understanding of patient information, resulting in a small number of managed cases and follow-up visits. With the application of information technology, use of an HIS-based chronic disease monitoring and case management system can automatically extract and integrate patient information, thereby increasing the efficiency of chronic disease management and reduce costs [4,22]. In this study, two case managers played leading roles both before and after implementation of the information system; however, compared with the situation before the use of the system, the numbers of both managed cases and of follow-up visits increased, reaching 3.2 and 3.4 times the preimplementation values, respectively. The information system can automatically obtain a patient's name and ID number based on the *ICD* code, which can expand the range of enrollment screening and appoint the register of patients as planned. In addition, the information system can automatically obtain outpatient, inpatient, and home medical information for the postillness life cycle management of patients. Moreover, the intuitive, clear, and dynamic indicator charts on the system can save a significant amount of time for diagnosis and treatment by medical staff, while the paperless office and online data-sharing functions can essentially solve the problem of managing files by case managers to ultimately enhance efficiency.

Improved Quality of Case Management

According to evidence-based medicine, the seven steps of case management represent the optimal clinical pathway [10-14,22]. In this study, the concept of an Internet-Plus medical service was introduced; that is, the chronic disease management information system was established based on the HIS data and case management model [22] and the function of a mobile medical terminal app was incorporated in the system [6,7]. Compared with the noninformation system case management model, this system has several advantages. First, owing to the swift management mode, it can overcome the limitations of time and space [4-8]. Second, multichannel health education and communication can enhance patients' knowledge and skills, as well as their compliance with self-management, based on diversified forms of image data such as graphics and audio [6,22]. Third, the use of intelligent management can remind

doctors and patients to complete management work and follow-up visits as planned, and to perform intelligent pushes of patient outcome indicators to improve confidence in the treatment [22]. Fourth, this system enables information sharing and big data analysis, as well as multidisciplinary diagnosis and treatment based on the matching of doctor-patient responsibility management, which can be more conducive to the precise health management of patients.

Compared with the traditional case management model, information-based case management significantly increased the follow-up visit rate, lung function test rate, chronic respiratory diseases knowledge awareness rate of patients, patient satisfaction rate, and retention rate. Among these indicators, the follow-up visit rate and lung function test rate represent aspects related to the patients' own management of their condition [1]. The results of this study are consistent with previous findings related to information-based management of chronic diseases in China, demonstrating that such a management system was more conducive to planned, systematic, and personalized education and follow-up by the case management team, thereby promoting the virtuous cycle of compliance with self-management and reducing the number of acute exacerbations among patients with chronic respiratory diseases, ultimately enhancing the precision of medical resource allocation and hospital management [22,23].

Helping Patients With COPD Maintain Stability of Their Condition

The admission rate for acute exacerbations serves as a common indicator of the quality of the treatment of chronic respiratory diseases [23]. The deployment of a clinical pathway–based hospital case management information system significantly reduced the admission rate for acute exacerbations and enhanced the quality of treatment for chronic respiratory diseases, indicating its high clinical significance. There are several reasons for these observed benefits. First, home care and self-management are essential in the management of chronic respiratory diseases. The information-based case management model improved the patients' knowledge and skills along with their compliance with self-management. Consequently, the standardized self-management process helped to reduce the number of acute exacerbations of chronic respiratory diseases and thus lowered the admission rate. Second, the information-based case management model increased the regular return rate, which allowed the medical staff to identify the potential risk factors for acute exacerbations in a timely manner, deal with them when they occur, and prepare personalized treatment plans and precise health management schemes. This consequently enabled adjustment of treatment schemes in real time, reduced the number of admissions due to acute exacerbations, and lowered the readmission rate. For hospitals interested in implementing a similar model, we suggest first conducting a detailed review of the current situation prior to making adequate changes based on the relevant disease and patient populations.

Consequently, the HIS-based case information management model could improve efficiency, enhance the quality of case management, and aid in stabilizing the conditions of patients with chronic respiratory diseases. In contrast to the hospital case management information system reported by Yuan et al [22], the system described in this study includes a personal terminal app. Previous studies confirmed that a stand-alone mobile health app could improve patient compliance and disease control [6-8]; thus, whether this system can be used to manage specialized disease cohorts for patients with chronic diseases remains to be determined. In this study, the effect on the retention rate of patients was confirmed; however, the overall operational indicators for the diagnosis and treatment of chronic diseases should be further determined.

Conclusion

With the advancement of information technology, the internet and medical technology have been applied to the management of chronic diseases. As an information-based platform for the case management of patients with chronic respiratory diseases, a newly developed chronic disease management information system was introduced in this study. This system is capable of designing the follow-up time registration, follow-up content, approaches, methods, quality control, and feedback process for a single chronic respiratory disease via the single-disease clinical pathway following the case management process (enrollment, assessment, planning, implementation, feedback, and evaluation). Use of this system can encourage patients with chronic respiratory diseases to adhere to regular follow-up and form an outpatient-inpatient-home chronic disease management strategy. This can help in reducing the admission rate for acute exacerbations, increase the return visit rate, and improve the correctness and compliance of home self-management of patients with chronic respiratory diseases. Owing to these benefits, wide adoption of such information systems for the management of chronic diseases can offer substantial economic and social value.

Acknowledgments

We are particularly grateful to all the people who provided help with our article. This study was supported by a grant from Yulin City Science and Technology Planning Project (20202002).

Data Availability

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

YML, YZX, XLS, and YXL designed this study. XLS and XJC wrote the draft of the paper. YML, YZX, and YC contributed final revisions to the article. XJC, HC, and YC collected the data. XJC, YML, YXL, and HC performed the statistical analysis. YML received funding support. All authors read and approved the final draft of the article.

Conflicts of Interest

None declared.

References

1. Anaev EK. Eosinophilic chronic obstructive pulmonary disease: a review. Ter Arkh 2023 Oct 11;95(8):696-700. [doi: 10.26442/00403660.2023.08.202316] [Medline: 38158908]
2. Shakeel I, Ashraf A, Afzal M, et al. The molecular blueprint for chronic obstructive pulmonary disease (COPD): a new paradigm for diagnosis and therapeutics. Oxid Med Cell Longev 2023 Dec;2023:2297559. [doi: 10.1155/2023/2297559] [Medline: 38155869]
3. Morimoto Y, Takahashi T, Sawa R, et al. Web portals for patients with chronic diseases: scoping review of the functional features and theoretical frameworks of telerehabilitation platforms. J Med Internet Res 2022 Jan 27;24(1):e27759. [doi: 10.2196/27759] [Medline: 35084355]
4. Donner CF, ZuWallack R, Nici L. The role of telemedicine in extending and enhancing medical management of the patient with chronic obstructive pulmonary disease. Medicina 2021 Jul 18;57(7):726. [doi: 10.3390/medicina57070726] [Medline: 34357007]
5. Wu F, Burt J, Chowdhury T, et al. Specialty COPD care during COVID-19: patient and clinician perspectives on remote delivery. BMJ Open Respir Res 2021 Jan;8(1):e000817. [doi: 10.1136/bmjresp-2020-000817] [Medline: 33414261]
6. Hallensleben C, van Luenen S, Rolink E, Ossebaard HC, Chavannes NH. eHealth for people with COPD in the Netherlands: a scoping review. Int J Chron Obstruct Pulmon Dis 2019 Jul;14:1681-1690. [doi: 10.2147/COPD.S207187] [Medline: 31440044]
7. Gokalp H, de Folter J, Verma V, Fursse J, Jones R, Clarke M. Integrated telehealth and telecare for monitoring frail elderly with chronic disease. Telemed J E Health 2018 Dec;24(12):940-957. [doi: 10.1089/tmj.2017.0322] [Medline: 30129884]
8. McCabe C, McCann M, Brady AM. Computer and mobile technology interventions for self-management in chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2017 May 23;5(5):CD011425. [doi: 10.1002/14651858.CD011425.pub2] [Medline: 28535331]
9. Briggs AM, Persaud JG, Deverell ML, et al. Integrated prevention and management of non-communicable diseases, including musculoskeletal health: a systematic policy analysis among OECD countries. BMJ Glob Health 2019;4(5):e001806. [doi: 10.1136/bmjgh-2019-001806] [Medline: 31565419]
10. Franek J. Home telehealth for patients with chronic obstructive pulmonary disease (COPD): an evidence-based analysis. Ont Health Technol Assess Ser 2012;12(11):1-58. [Medline: 23074421]
11. Shah A, Hussain-Shamsy N, Strudwick G, Sockalingam S, Nolan RP, Seto E. Digital health interventions for depression and anxiety among people with chronic conditions: scoping review. J Med Internet Res 2022 Sep 26;24(9):e38030. [doi: 10.2196/38030] [Medline: 36155409]
12. Sugiharto F, Haroen H, Alya FP, et al. Health educational methods for improving self-efficacy among patients with coronary heart disease: a scoping review. J Multidiscip Healthc 2024 Feb;17:779-792. [doi: 10.2147/JMDH.S455431] [Medline: 38410523]
13. Metzendorf MI, Wieland LS, Richter B. Mobile health (m-health) smartphone interventions for adolescents and adults with overweight or obesity. Cochrane Database Syst Rev 2024 Feb 20;2(2):CD013591. [doi: 10.1002/14651858.CD013591.pub2] [Medline: 38375882]
14. Reig-Garcia G, Suñer-Soler R, Mantas-Jiménez S, et al. Assessing nurses' satisfaction with continuity of care and the case management model as an indicator of quality of care in Spain. Int J Environ Res Public Health 2021 Jun 19;18(12):6609. [doi: 10.3390/ijerph18126609] [Medline: 34205373]
15. Aggelidis X, Kritikou M, Makris M, et al. Tele-monitoring applications in respiratory allergy. J Clin Med 2024 Feb 4;13(3):898. [doi: 10.3390/jcm13030898] [Medline: 38337592]
16. Seid A, Fufa DD, Bitew ZW. The use of internet-based smartphone apps consistently improved consumers' healthy eating behaviors: a systematic review of randomized controlled trials. Front Digit Health 2024;6:1282570. [doi: 10.3389/fdgth.2024.1282570] [Medline: 38283582]
17. Verma L, Turk T, Dennett L, Dytoc M. Teledermatology in atopic dermatitis: a systematic review. J Cutan Med Surg 2024;28(2):153-157. [doi: 10.1177/12034754231223694] [Medline: 38205736]
18. Tański W, Stapkiewicz A, Szalonka A, Głuszczyk-Ferenc B, Tomasiewicz B, Jankowska-Polańska B. The framework of the pilot project for testing a telemedicine model in the field of chronic diseases - health challenges and justification of the project implementation. Pol Merkur Lekarski 2023;51(6):674-681. [doi: 10.36740/Merkur202306115] [Medline: 38207071]

19. Popp Z, Low S, Igwe A, et al. Shifting from active to passive monitoring of Alzheimer disease: the state of the research. J Am Heart Assoc 2024 Jan 16;13(2):e031247. [doi: 10.1161/JAHA.123.031247] [Medline: 38226518]
20. Sagare N, Bankar NJ, Shahu S, Bandre GR. Transforming healthcare: the revolutionary benefits of cashless healthcare services. Cureus 2023 Dec;15(12):e50971. [doi: 10.7759/cureus.50971] [Medline: 38259368]
21. Noncommunicable Diseases, Rehabilitation and Disability (NCD), Surveillance, Monitoring and Reporting (SMR) WHO Team. Noncommunicable diseases progress monitor. : World Health Organization; 2017 URL: https://www.who.int/publications/i/item/9789241513029 [accessed 2024-05-09]
22. Yuan W, Zhu T, Wang Y, et al. Research on development and application of case management information system in general hospital. Nurs Res 2022;36(12):2251-2253.
23. 2020 GOLD report. Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease. : Global Initiative for Chronic Obstructive Lung Disease; 2020 URL: https://goldcopd.org/gold-reports/ [accessed 2024-05-09]

Abbreviations

COPD: chronic obstructive pulmonary disease
HIS: hospital information system
***ICD*:** *International Classification of Diseases*

Edited by C Lovis; submitted 15.06.23; peer-reviewed by KM Kuo; revised version received 14.04.24; accepted 17.04.24; published 19.06.24.

Please cite as:
Xiao YZ, Chen XJ, Sun XL, Chen H, Luo YX, Chen Y, Liang YM
Effect of Implementing an Informatization Case Management Model on the Management of Chronic Respiratory Diseases in a General Hospital: Retrospective Controlled Study
JMIR Med Inform 2024;12:e49978
URL: https://medinform.jmir.org/2024/1/e49978
doi:10.2196/49978

Addressing Hospital Overwhelm During the COVID-19 Pandemic by Using a Primary Health Care–Based Integrated Health System: Modeling Study

Jiaoling Huang[1], PhD; Ying Qian[2,3], PhD; Yuge Yan[1]; Hong Liang[4], PhD; Laijun Zhao[2,3], PhD

1

2

3

4

Corresponding Author:
Ying Qian, PhD

Abstract

Background: After strict COVID-19–related restrictions were lifted, health systems globally were overwhelmed. Much has been discussed about how health systems could better prepare for future pandemics; however, primary health care (PHC) has been largely ignored.

Objective: We aimed to investigate what combined policies PHC could apply to strengthen the health care system via a bottom-up approach, so as to better respond to a public health emergency.

Methods: We developed a system dynamics model to replicate Shanghai's response when COVID-19–related restrictions were lifted. We then simulated an alternative PHC-based integrated health system and tested the following three interventions: first contact in PHC with telemedicine services, recommendation to secondary care, and return to PHC for recovery.

Results: The simulation results showed that each selected intervention could alleviate hospital overwhelm. Increasing the rate of first contact in PHC with telemedicine increased hospital bed availability by 6% to 12% and reduced the cumulative number of deaths by 35%. More precise recommendations had a limited impact on hospital overwhelm (<1%), but the simulation results showed that underrecommendation (rate: 80%) would result in a 19% increase in cumulative deaths. Increasing the rate of return to PHC from 5% to 20% improved hospital bed availability by 6% to 16% and reduced the cumulative number of deaths by 46%. Moreover, combining all 3 interventions had a multiplier effect; bed availability increased by 683%, and the cumulative number of deaths dropped by 75%.

Conclusions: Rather than focusing on the allocation of medical resources in secondary care, we determined that an optimal PHC-based integrated strategy would be to have a 60% rate of first contact in PHC, a 110% recommendation rate, and a 20% rate of return to PHC. This could increase health system resilience during public health emergencies.

(JMIR Med Inform 2024;12:e54355) doi:10.2196/54355

KEYWORDS

hospital overwhelm; primary health care; modeling study; policy mix; pandemic; model; simulation; simulations; integrated; health system; hospital; hospitals; management; service; services; health systems; develop; development; bed; beds; overwhelm; death; deaths; mortality; primary care

Introduction

The World Health Organization (WHO) announced the end of the COVID-19 public health emergency of international concern on May 5, 2023. Over the past 3 years, the COVID-19 epidemic has resulted in more than 765 million infections and 6.92 million deaths globally and has involved ongoing outbreaks, infection control via restrictions, the lifting of restrictions, and large-scale infections [1]. The limitations of health care systems worldwide regarding the response to mass infections and admissions have been exposed, and these limitations exist in countries classified as high-performing and resilient countries, as well as in resource-limited countries [2-4]. Although most governments have prudently considered the appropriate time to relax restriction policies, health care systems have unavoidably been overwhelmed, and some even collapsed once restrictions were lifted [5].

Much has been discussed regarding how health care systems could have better prepared for COVID-19 and how to prepare for future pandemics. Topics of discussion include adequate health care workforces and facilities [6], better intensive care unit capacity [7], early intervention to avoid local transmission

[8], and the broader application of telemedicine [9]. Some scholars have advocated for the integration and coordination of the health system, including public health and clinical medicine. The role of primary health care (PHC) in COVID-19 management has received attention [10-13], but this attention is obviously insufficient when compared with the attention given to the professional treatment capabilities of large hospitals. In particular, there is a lack of empirical research on the role of PHC. After touring 5 cities in China, the WHO provided recommendations that were predominantly focused on secondary care and epidemiological tracking and control, and the role of PHC was missed again [14].

Distinguishing itself from secondary care for specialist treatment, PHC is regarded as the most inclusive, effective, and efficient approach to enhancing people's physical and mental health. PHC has great value in a strong, coordinated response to a public health crisis [10,15]. Recent studies show that a strong PHC foundation could effectively mitigate an epidemic. One such case is that of Singapore, which promptly instituted aggressive containment measures by establishing public health preparedness clinics that were supported in a sustained manner by the PHC network [16]. In contrast, PHC resources in the African Union are exceedingly scarce, which resulted in insufficient engagement when dealing with COVID-19 [17-20]. Even in countries with adequate PHC resources, such as the United Kingdom, the health system did not respond quickly and struggled to meet medical demands under a large-scale epidemic [21]. Legido-Quigley and colleagues [7] argued that well-developed integration was a key factor of services influencing resilience during the COVID-19 pandemic in high-performing health systems. Prompt communication and coordination among PHC, public health, and secondary care are essential [22].

At the end of 2022, COVID-19–related restrictions were lifted in China, and an epidemic wave caused by the highly transmissible Omicron SARS-CoV-2 variant placed health services in the country under extreme pressure, especially in metropolises. In Shanghai, which is the most populous metropolis in China and has a permanent population of 25 million, it is extremely difficult to deal with the spread of epidemic infections. At the end of 2022, Shanghai adopted an expansion strategy that involved allocating medical resources in secondary care institutions in a manner that favored patients with SARS-CoV-2 infection. At the same time, Shanghai, as a pilot city, was one of the first cities to promote a hierarchical diagnosis and treatment system based on PHC. As such, Shanghai provides an extremely rare opportunity to explore how the health system of a metropolis can actively respond to large-scale infections, as well as the key role of PHC in this system. In this study, we simulated the large-scale infections that occurred in Shanghai at the end of 2022 by using a simulated environment, wherein we reproduced Shanghai's response to the challenges of the fast-spreading epidemic. We then tested an alternative strategy that used a PHC-based integrated health system.

Methods

Ethical Considerations

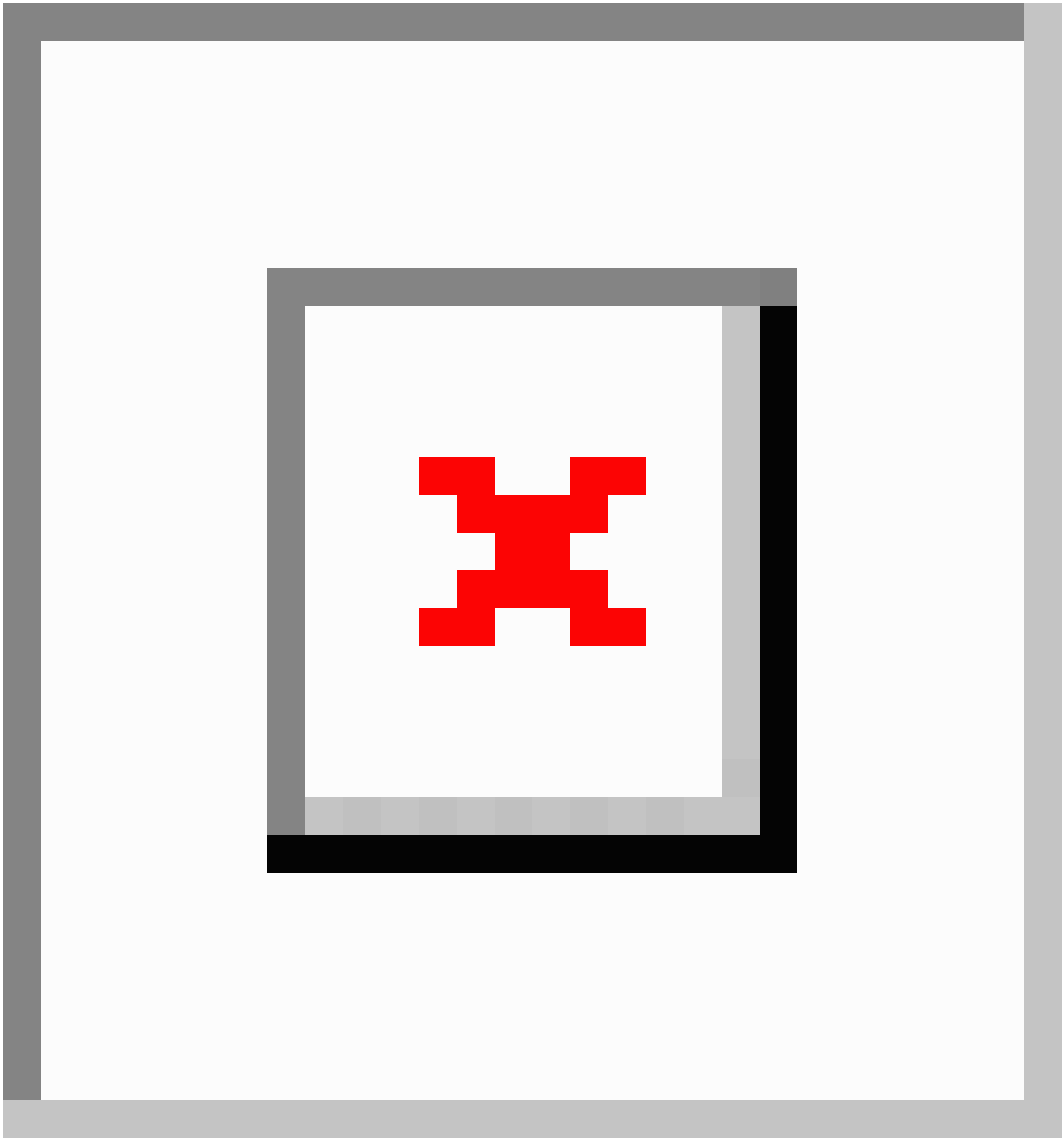
Ethics board review was not required, as this study only involved modeling and simulations. All modeling data came from public sources or published papers and did not involve ethical issues.

Study Design

System dynamics was applied in this study to simulate the mass infections and the health system performance in Shanghai. System dynamics models are established based on the feedback structures (loops) centered around the issue of concern [23,24]. The nonlinear dynamic behaviors derived from these feedback loops shed light on the underlying mechanisms that generate problematic system behaviors, which helps with understanding complex systems and finding fundamental solutions [25]. The use of system dynamics models is a suitable method for investigating public health issues that feature high-complexity systems [26,27]. In recent years, system dynamics has been widely used to model issues related to COVID-19 [28-31].

We developed a system dynamics–based model to replicate the health system in Shanghai after COVID-19–related policies were lifted. The following indicators of an overwhelmed health system were used: physician availability (the percentage of patients arriving at the hospital who could be treated) and bed availability (the percentage of patients needing hospitalization who could be admitted) in secondary hospitals. Shanghai's response to the soaring medical demands was to reallocate medical resources from other divisions to increase the supply of hospital physicians and beds for patients with COVID-19. This policy increased the capacity of secondary hospitals such that more patients could be treated and hospitalized. We also used the system dynamics model to establish a PHC-based integrated health system as an alternative option for addressing hospital overwhelm. The following three critical policy interventions were tested: first contact in PHC, identification of high-risk patients and recommendation to secondary care hospitals, and referral for a return to PHC for follow-up and recovery at the community level (Figure 1). Telemedicine services were also considered in PHC, with which more first contacts could be handled and the capacity of PHC to handle patients could be increased.

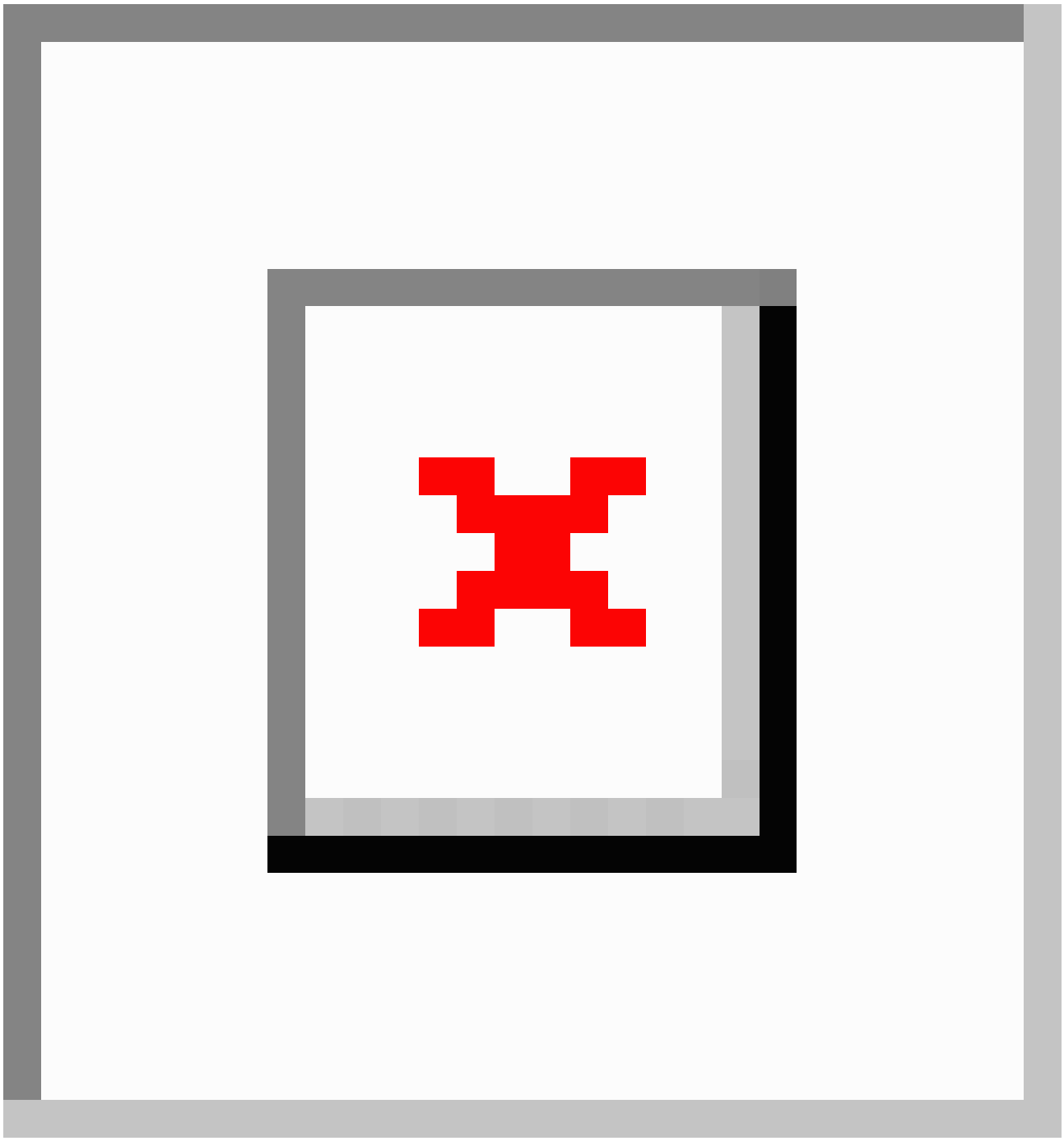
Figure 1. PHC-based health system. PHC: primary health care.



Model Structure

The Shanghai model includes the following two parts: the epidemic dynamics representing the mass infections that occurred when COVID-19 restrictions were lifted in Shanghai and the health care system response, as shown in Figure 2.

Figure 2. Shanghai system dynamics model of a PHC-based integrated system. *A*: infected population without symptoms; D: death; *E*: infected population during the incubation period; GB: getting better; GW: getting worse; HR: home recovery; *I*: infected population with symptoms; PHC: primary health care; R: recover; *RA*: population in *A* that has recovered; *S*: susceptible population without vaccination; *SV*: susceptible population with vaccination.



In Figure 2, the left part depicts an extension of the traditional Susceptible-Exposed-Infectious-Removed model, which we used to model the spread of COVID-19 in Shanghai and compute the number of symptomatic cases, of which a large proportion would require medical services. We disaggregated the total population into the following six groups.

SV and *S* represent the susceptible population with vaccination and the susceptible population without vaccination, respectively. The transformation of *SV* to *S* represents the waning effectiveness of COVID-19 vaccines, where ω is the waning effect of vaccination.

E represents the infected population during the incubation period. The transformation of *SV* to *E* and *S* to *E* represents the spread of the virus, where c is the contact rate, β is the transmission probability, and θ_1 is the effectiveness rate of vaccination against infection.

I and *A* represent the infected population with symptoms and the infected population without symptoms, respectively. α is the percentage of asymptomatic cases, and τ is the incubation period.

RA represents the population in *A* that has recovered. γ_1 is the recovery fraction among asymptomatic cases.

Patients with symptoms (ie, those from population *I*) link the left and right parts of the model. Some patients will recover at home, and others will visit a physician. Among those visiting a physician, some will first contact a PHC institution, while others will contact a secondary hospital directly. PHC institutions and secondary hospitals each have a specific capacity, and when this capacity is reached, new, excess patients cannot be treated and will have to return home. With regard to patients treated in PHC institutions and secondary hospitals, those with mild symptoms will be given prescriptions and sent home to recover. With regard to patients needing further treatment when presenting at the PHC level, general practitioners will recommend them to a secondary hospital; some patients will be hospitalized and become inpatients if hospital beds are available. Over time, some inpatients will recover, whereas others will develop severe illness and eventually recover or die. Patients who recover at home will either improve or worsen, as will untreated and treated patients from PHC institutions and secondary hospitals. The proportion of patients whose condition worsens is highest for untreated patients and lowest for treated patients. Some recovering inpatients in secondary care hospitals might recover at the community level if PHC can provide follow-up health management services. The model equations and parameter settings are detailed in sections 1 and 2 in Multimedia Appendix 1.

Data Source and Model Validation

We previously developed and validated a model of reopening in Shanghai that accounts for the epidemiological dynamics of the first Omicron wave in this metropolis during the first half of 2022 [32]. The model we established in this study was based on that previous model and was used to simulate the second Omicron wave, specifically the time when most intervention prevention control measures were lifted at the end of 2022. Data related to the spread of Omicron in Shanghai, such as the contact rate, transmission possibility, asymptomatic rate, incubation period, and recovery fraction, were obtained from previous literature about COVID-19 and, especially, Omicron (further details are reported in section 2 and Table S1 in Multimedia Appendix 1). Data related to individuals' behaviors, such as the rate of first contact in PHC and the rate of recovery at home, were set according to estimations based on our investigation of PHC, hospitals, and the community. Sensitivity tests for these parameters were conducted to check the robustness of the model (section 3.2 in Multimedia Appendix 1).

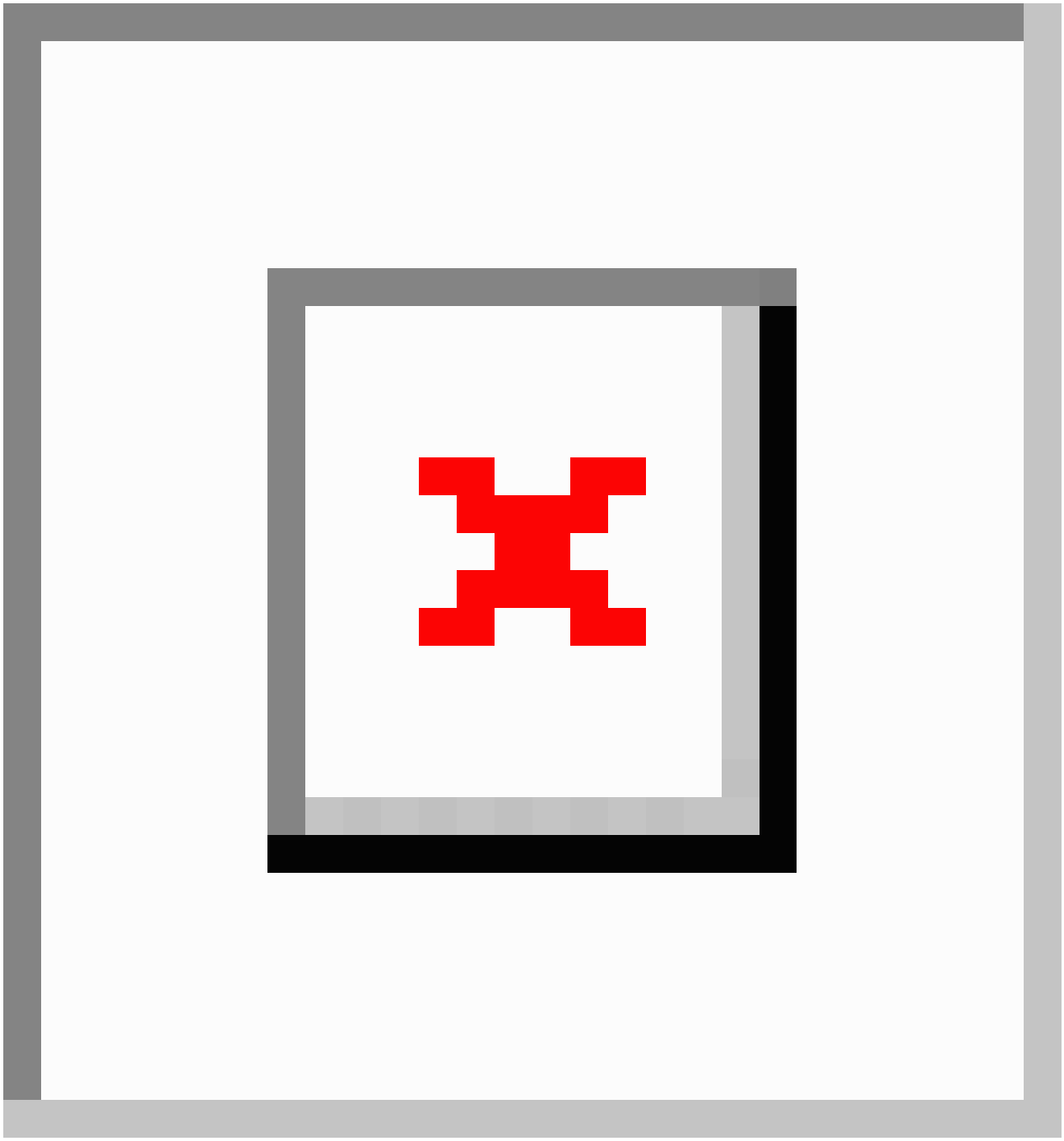
The validation of model behavior usually involves the comparison of simulation results with real-world data. Because mass COVID-19 testing was no longer required, accurate infection data were no longer available; news reports on viral infections and the level of pressure on medical resources were used as references. The model results revealed patterns that were similar to the actual situation during the second Omicron wave in Shanghai, thereby confirming the validity of the model (further details are reported in section 3 in Multimedia Appendix 1). As a result, we determined that this model could provide a simulated environment to facilitate the exploration of effective policies regarding the response to mass infections in a metropolis.

Results

Scenario 1: Medical Resource Reallocation in Secondary Care

When the strict intervention prevention controls were lifted, the policy focus changed from preventing the spread of COVID-19 to the timely treatment of patients with COVID-19. When facing massive increases in infections, physicians' availability could decline to as low as 55% if no interventions were adopted. In the case of Shanghai, a series of measures was taken to deal with the impact of large-scale infections on hospitals. When hospital physician capacity and hospital bed capacity were increased by 70% of the original capacities, the lowest physician availability and bed availability changed to approximately 85% and 70%, respectively. Moreover, bed shortages lasted approximately 8 days, which was around one-third of the bed shortage time for the scenario with no capacity extension. The peak number of severe cases decreased, as more patients could be treated promptly. Consequently, the cumulative number of deaths decreased to less than half of that for the scenario without additional resources, as shown in Figure 3.

Figure 3. Expanding capacity to meet soaring demand. Base: baseline; Ext cap 30: 30% capacity extension; Ext cap 50: 50% capacity extension; Ext cap 70: 70% capacity extension.

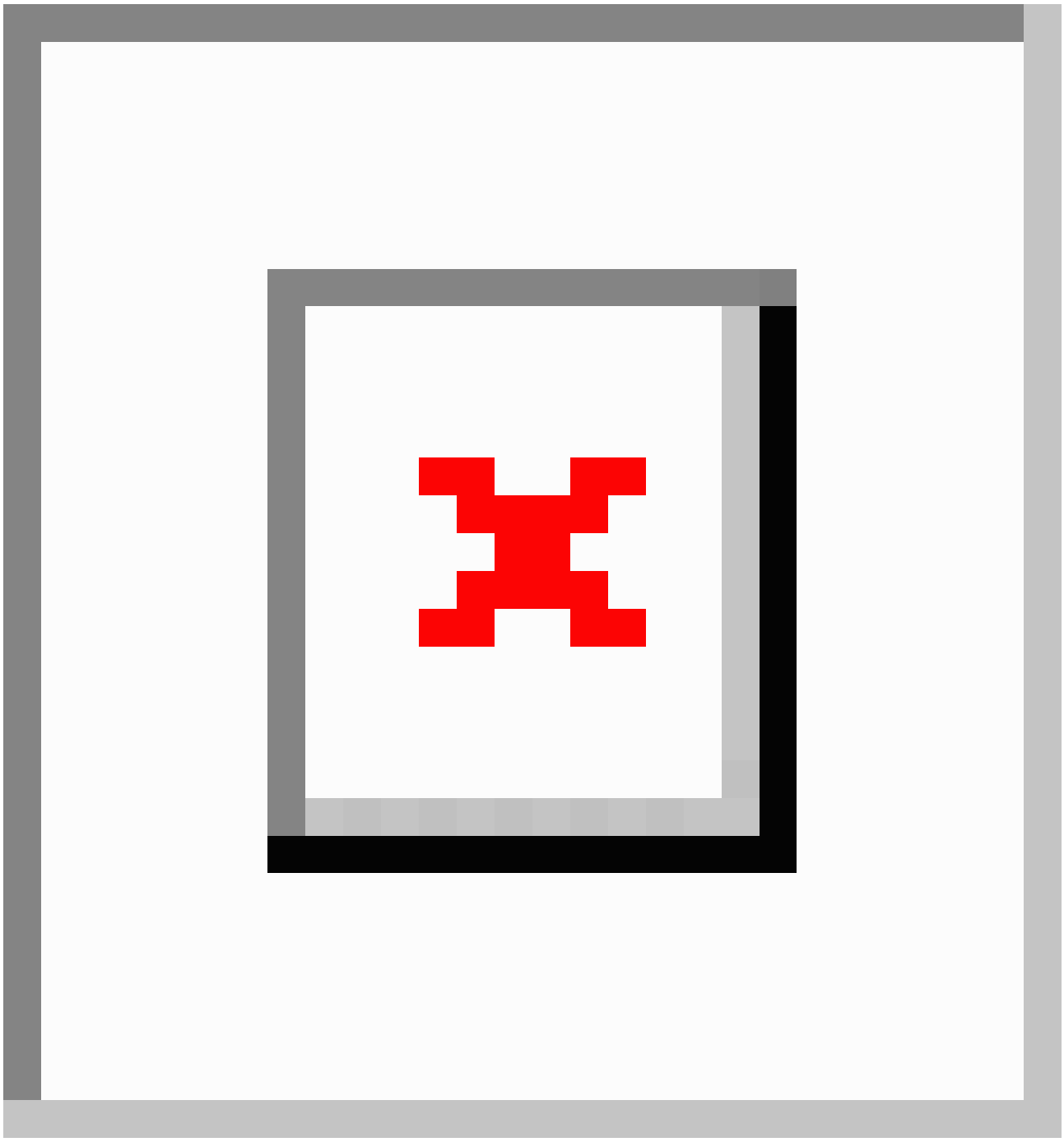


Scenario 2: PHC-Based Integrated Health Care System

Scenario 2.1: Increasing the Rate of First Contact in PHC Plus PHC Telemedicine

Facing huge increases in the number of patients, we examined ways to increase the rate of first contact in PHC, in which PHC telemedicine services were also considered. Under such circumstances, 6 scenarios were simulated, with the rate of first contact in PHC with and without telemedicine services set to 40%, 50%, and 60%. Figure 4 shows that replacing the worst scenario (40% rate of first contact in PHC without telemedicine) with the best scenario (60% rate of first contact in PHC with telemedicine) increased the lowest level of secondary hospital physician availability and that of secondary hospital bed availability by 32% (from 51% to 67%) and 111% (from 9% to 19%), respectively. Moreover, the duration of bed shortages dropped from approximately 30 days to approximately 20 days. Because more patients were promptly treated in the best scenario, the number of cumulative deaths decreased from 24,740 in the worst scenario to 15,837—a 56% decrease.

Figure 4. Scenarios with various rates of first contacts in primary health care with and without telemedicine. PHC FC 40: 40% rate of first contact in primary health care without telemedicine; PHC FC 50: 50% rate of first contact in primary health care without telemedicine; PHC FC 60: 60% rate of first contact in primary health care without telemedicine; PHC FC 40 + Telem 2: 40% rate of first contact in primary health care with telemedicine; PHC FC 50 + Telem 2: 50% rate of first contact in primary health care with telemedicine; PHC FC 60 + Telem 2: 60% rate of first contact in primary health care with telemedicine.

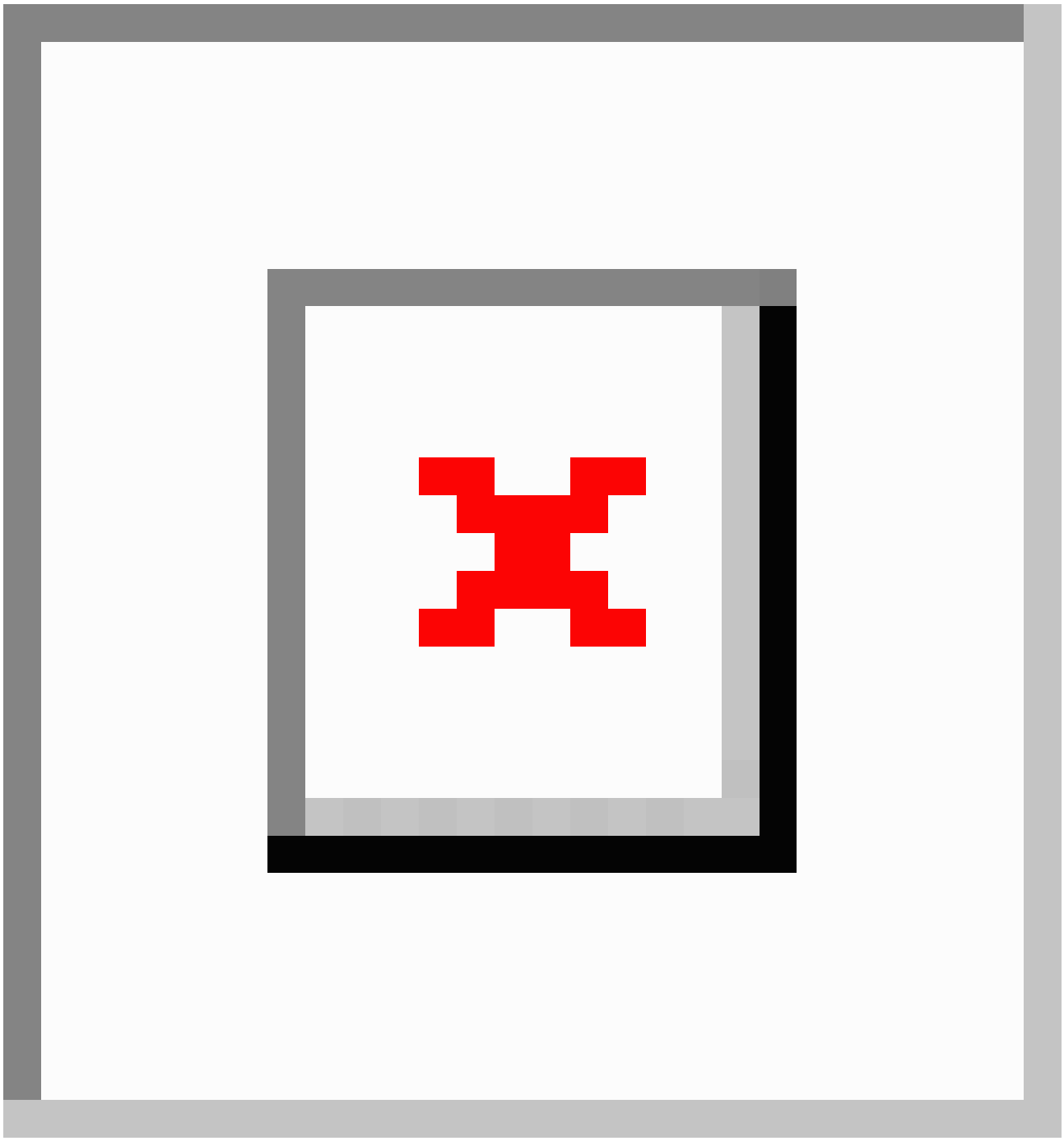


Scenario 2.2: Referral Recommendation Rate for High-Risk Patients

Two of the main tasks of general practitioners in PHC are to identify high-risk patients and provide referral recommendations to secondary care, which are related to the professional capabilities of general practitioners. We simulated the following four scenarios: (1) a recommendation rate of 80%, meaning that 20% of patients requiring advanced treatment in a hospital were not identified (ie, underrecommendation); (2) a recommendation rate of 100% without underrecommendation or overrecommendation, which is an ideal scenario; (3) a recommendation rate of 120%, meaning that 20% of patients were overreferred to secondary care; and (4) a recommendation rate of 100% but with underrecommendation and overrecommendation happening at the same time. Underrecommendation and overrecommendation, respectively, slightly increased and decreased the physician availability and bed availability. However, with underrecommendation, in which 20% of patients who needed to be treated in a hospital were not referred, some patients developed severe illness due to improper

treatment, leading to more severe cases and more cumulative deaths, as shown in Figure 5.

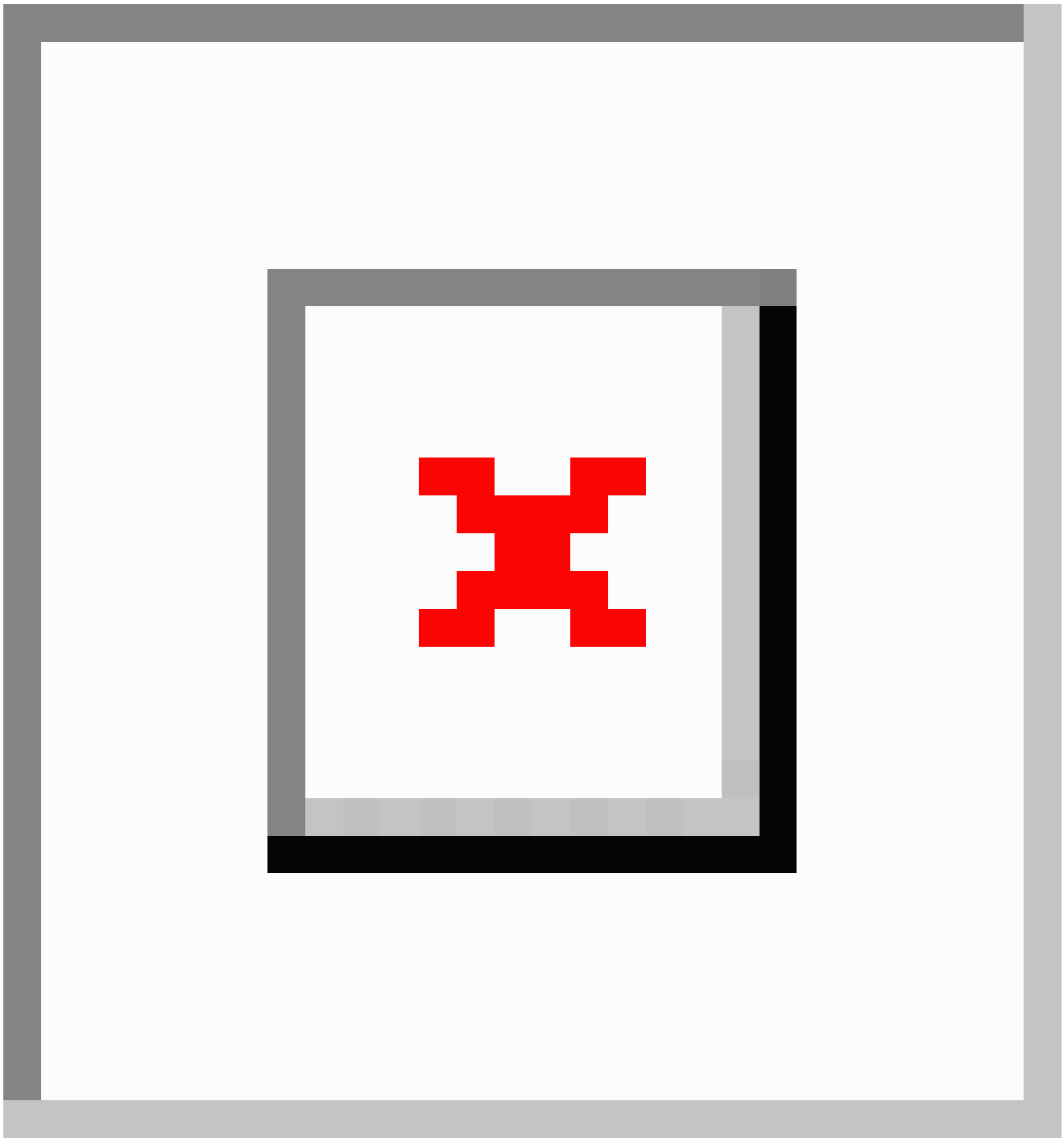
Figure 5. Scenarios involving general practitioners in primary health care with differing capabilities for identifying high-risk patients. “Ideal” refers to no underrecommendation and no overrecommendation. “Nonideal” refers to underrecommendation and overrecommendation happening at the same time.



Scenario 2.3: Follow-Up Health Management Services in PHC

When facing bed shortages, some patients with mild symptoms who are nearly recovered can be transferred to recover in the community if PHC can provide follow-up health management services. We simulated four scenarios under which 5%, 10%, 15%, and 20% of hospital inpatients were transferred back to PHC to recover. As shown in Figure 6, physician availability was not affected, but hospital bed availability improved. When 20% of inpatients could recover in the community, the lowest bed availability level reached approximately 20%, whereas this level reached 7% in the scenario where only 5% of inpatients were referred to PHC to recover. Moreover, the number of days with bed shortages was nearly halved. As a result, the peak number of severe cases decreased from 16,897 to 14,737, and the cumulative number of deaths dropped from 28,008 to 14,344.

Figure 6. Scenarios with different percentages of patients returning to primary health care for recovery. Health mngt 05: 5% rate of health management in the community; Health mngt 10: 10% rate of health management in the community; Health mngt 15: 15% rate of health management in the community; Health mngt 20: 20% rate of health management in the community.

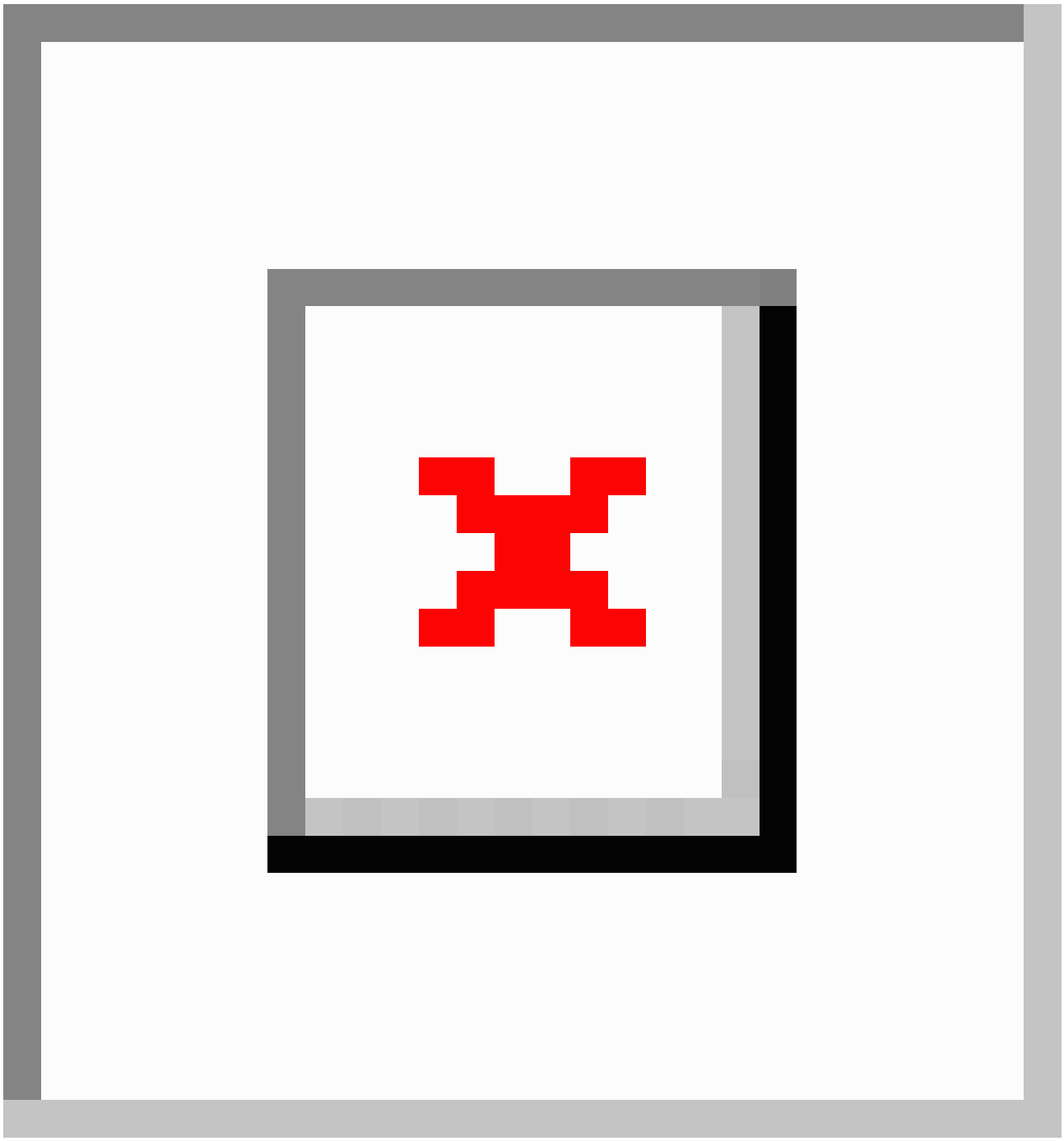


Scenario 2.4: Mixed Policy Interventions

We integrated the above three policy interventions to investigate the overall impact of a PHC-based system, as shown in Figure 7. The model simulation results showed that the lowest level of hospital physician availability ranged from 51% in the worst case (40% rate of first contact in PHC without telemedicine) to 69% in the best case (60% rate of first contact in PHC with telemedicine). The lowest level of hospital bed availability varied even more, ranging from 6% in the worst case (40% rate of first contact in PHC without telemedicine and 5% rate of health management in the community) to 51% in the best case (60% rate of first contact in PHC with telemedicine, recommendation rate of 80%, and 20% rate of health management in the community). With regard to the number of severe cases, the worst case peaked at 20,876 severe cases (40% rate of first contact in PHC without telemedicine, underrecommendation rate of 20%, and 5% rate of health management in the community), and the best case peaked at 11,984 severe cases (60% rate of first contact in PHC with telemedicine, recommendation rate of 100%, and 20% rate of health management in the community)—a decrease of approximately 75%. As for cumulative deaths, the worst case was 37,369 deaths, and the best case was only 7946 deaths—a 79% reduction. Furthermore, we identified the following relatively optimal policy intervention mix: a 60% rate of first

contact in PHC with telemedicine services, a 20% rate of referral to return to PHC, and a recommendation rate of 100% to 120%.

Figure 7. Overall impact of integrated primary health care (PHC). The red, orange, green, and blue areas represent PHC with or without telemedicine, the rate of FC in PHC, the rate of PHC recommendation to secondary care, and the rate of referral to return to PHC, respectively. The gray areas represent the biggest medical resource gaps, including hospital physician availability and hospital bed availability, and the resulting peak number of SCs and cumulative deaths. FC: first contact; SC: severe case.



Discussion

Principal Findings

The experience of hospital overwhelm during the COVID-19 pandemic highlights the need to reflect on existing health systems and search for more proactive solutions during an epidemic. In this study, we constructed a simulated policy environment to replicate Shanghai's response to the mass SARS-CoV-2 infections that occurred once restrictions were lifted. Specifically, the Shanghai Municipal Health Commission deployed medical resources in secondary care hospitals, in a manner that favored patients with SARS-CoV-2 infection, as quickly as possible. This strategy included increasing the availability of beds and reallocating more medical staff from other departments to promptly treat critically ill patients and prevent deaths. This efficient and decisive response allowed Shanghai to avoid the large-scale congestion and overwhelm of the health care system. However, Shanghai's strategy worked under the assumptions that advanced medical resources would

be available and could be deployed, and the strategy largely depended on the government's strong decision-making and coordination capabilities. At the same time, we realize that relying on PHC to alleviate congestion is an important strategy to achieve the effective allocation of medical resources, rather than only relying on temporary expansion in secondary care [33].

We proposed an alternative PHC-based strategy and tested this in a simulated policy simulation environment. We tested the rate of first contact in PHC, the rate of identifying high-risk patients for recommendation to a specialist, and the rate of return to PHC for recovery. According to the simulation results, increasing the rate of first contact in PHC could effectively alleviate the shortage of specialists in large hospitals. Additionally, telemedicine application in PHC contributed substantially to reducing congestion within hospitals engaged in COVID-19 treatment. In our model, a 60% rate of first contact in PHC with telemedicine could increase the lowest level of secondary hospital physician availability from 51% to 67% and reduce the number of cumulative deaths by 9630. The value of first contact in PHC for patients is receiving immediate medical treatment to avoid severe illness or death caused by delays in treatment, as well as reducing the shortage of medical resources in secondary hospitals. COVID-19 has accelerated the development of telemedicine. Alexander and colleagues [34] used a nationally representative audit of outpatient care to characterize primary care delivery in the United States and found that the pandemic was associated with a >25% decrease in primary care volume, which has been offset in part by increases in the delivery of telemedicine. Some believe that the boom in telemedicine during the COVID-19 pandemic could worsen health disparities [35], especially for racial and ethnic minority groups; those living in rural areas; and individuals with limited English proficiency, low literacy, or low income [36]. Nevertheless, telemedicine is an inevitable future developmental trend.

The rate of identifying high-risk patients is a crucial indicator of PHC worker capacity. We found that underidentification could result in more severe illness and more deaths, whereas overidentification could increase congestion in hospitals to some degree. For example, in the scenario with a 50% rate of first contact in PHC with telemedicine, a 120% recommendation rate reduced hospital specialist availability from 61% to 60%, whereas an 80% recommendation rate increased hospital specialist availability from 61% to 63%. A similar impact was observed on hospital bed availability. However, underrecommendation resulted in some patients (ie, those needing further treatment) failing to seek timely medical care and thus higher rates of severe illness and an increase of 3265 cumulative deaths. According to the simulation results, the effect of accurately identifying high-risk patients is limited in the existing system, possibly due to a low rate of first contact in PHC. Unlike countries in Europe and North America, China has a loose medical referral system rather than a strict referral system based on first contact in PHC [37]. China established its PHC system after the new health care reform in 2009 [38]. In October 2016, the Chinese government launched the Healthy China 2030 initiative, in which a critical component is developing a patient referral model [39]. In contrast, gatekeeping systems can ensure the efficient use of scarce medical resources in secondary care; to date, there has been no action plan to enforce the patient referral model [37]. The rate of first contact in PHC has remained at approximately 30% to 50% for the past 10 years. However, in scenarios where PHC first contact–based referral is strictly implemented, such as in the United Kingdom [40], we believe that accurate risk identification in PHC is important.

We also considered the rate of return to PHC for recovery in the community, which can accelerate bed turnover in secondary hospitals. The model simulation results showed that increasing the rate of return to PHC from 5% to 20% would increase bed availability from 6% to 16%, thereby reducing the number of cumulative deaths by approximately 13,000. According to the WHO, referral is a bidirectional process that acknowledges not only the role of the specialist but also the critical role of PHC workers in coordinating patient care over the longer term [41]. In May 2023, Shanghai issued an important document—*Implementation Plan to Further Enhance the City's Community Health Service Capabilities*—focusing on strengthening 4 functions in community health centers—PHC, health management, rehabilitation, and nursing—as well as the primary public health network [42]. COVID-19 has definitely brought challenges to PHC, but it has also provided new opportunities.

Interestingly, we also found a multiplier effect with combined policy interventions. For example, offering telemedicine services, increasing the rate of first contact in PHC from 40% to 60%, and raising the rate of referral to return to PHC from 5% to 20% respectively increased bed availability by 16.67%, 50%, and 167%. However, when combined, these policies increased the lowest bed availability level by 683%. Optimal policy intervention combinations are widely applied in health, climate change, and economics (eg, funding instruments). Policy mixing implies a focus on trade-off interactions and interdependencies among different policies, as they affect the extent to which the intended policy outcomes are achieved. It provides a window of opportunity to reconsider basic and often hidden assumptions to better deal with complex, multilevel, multiactor realities [43]. In this study, we identified a relatively optimal policy combination (ie, a 60% rate of first contact in PHC, a 110% recommendation rate, and a 20% rate of return to PHC) that could establish a strong PHC foundation and increase health system resilience by reducing medical resource gaps in responding to public health emergencies. The interplay of policies and instruments, as well as the deliberate design of policy mixes and portfolios of interventions, has received surprisingly little practical and theoretical attention so far and is vastly underrated [44].

Using the scenario of reopening in Shanghai, we built a health care system for metropolises to deal with large-scale infections and verified the role of PHC through a system simulation model. However, our study has some limitations. First, real-world data were missed, especially epidemiological data, as mass COVID-19 testing was canceled. We validated our model based on information from news reports indicating the development of the Omicron wave and web-based information. Second, data

related to individuals' behaviors, such as the rate of first contact in PHC and the rate of recovery at home, are not available. We estimated these parameters according to our investigation of PHC institutions, hospitals, and communities. Third, this study simulated a PHC-based integrated health system responding to large-scale infections (including parameters such as first-contact rate, referral rate, and recovery fraction), but our models did not tell us how the integrated system could be improved. More attention should be paid to integrated health systems in the near future by conducting more case studies or implementation research.

Conclusions

Rather than focusing on secondary care, in this study, we proposed an alternative—strengthening the health system via a bottom-up approach by using PHC as a foundation to better respond to a public health emergency. Per our PHC-based health system, an optimal PHC-based integrated strategy would be to have a 60% rate of first contact in PHC, a 110% recommendation rate, and a 20% rate of return to PHC, which could increase health system resilience during public health emergencies. A robust PHC-based integrated health system, in addition to the temporary deployment of medical resources in secondary care, could maximize the use of limited medical resources to actively respond to large-scale increases in infections. This study provides an optimal solution for constructing a PHC-based integrated health system to respond to large-scale infections. We acknowledge that there is a long way to go to achieve an integrated health system, as getting PHC to communicate and interact seamlessly with secondary care is extremely challenging globally. We advocate increasing investments in PHC to promote its overall development and conducting more research on integrated health systems in the near future.

Acknowledgments

This study was funded by the Three-Year Action Program of Shanghai Municipality for Strengthening the Construction of Public Health System (grant GWVI-11.2-YQ54), Science and Technology Committee of Shanghai Municipality Soft Science Research Plans (grant 23692113400), the National Natural Science Foundation of China (grant 72274122), and the National Social Science Foundation of China (grant 22BGL240).

Data Availability

The data used in this study are publicly available on the National Health Commission of the People's Republic of China website [45]. Our model code is available from the corresponding author on request.

Authors' Contributions

JH, HL, and YQ conceptualized this study. JH and YQ wrote the original draft and reviewed and edited the manuscript. YY was responsible for data visualization. LZ was responsible for data collection and data analysis. YQ was responsible for the methodology.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model description and definitions, parameter settings, and model validation.
[DOCX File, 691 KB - medinform_v12i1e54355_app1.docx]

References

1. WHO COVID-19 dashboard. World Health Organization. URL: https://covid19.who.int/ [accessed 2023-07-28]
2. Armocida B, Formenti B, Ussai S, Palestra F, Missoni E. The Italian health system and the COVID-19 challenge. Lancet Public Health 2020 May;5(5):e253. [doi: 10.1016/S2468-2667(20)30074-8] [Medline: 32220653]
3. da Silva SJR, Pena L. Collapse of the public health system and the emergence of new variants during the second wave of the COVID-19 pandemic in Brazil. One Health 2021 Dec;13:100287. [doi: 10.1016/j.onehlt.2021.100287] [Medline: 34222607]
4. El Bcheraoui C, Weishaar H, Pozo-Martin F, Hanefeld J. Assessing COVID-19 through the lens of health systems' preparedness: time for a change. Global Health 2020 Nov 19;16(1):112. [doi: 10.1186/s12992-020-00645-5] [Medline: 33213482]
5. Han E, Tan MMJ, Turk E, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. Lancet 2020 Nov 7;396(10261):1525-1534. [doi: 10.1016/S0140-6736(20)32007-9] [Medline: 32979936]
6. Mahendradhata Y, Andayani NLPE, Hasri ET, et al. The capacity of the Indonesian healthcare system to respond to COVID-19. Front Public Health 2021 Jul 7;9:649819. [doi: 10.3389/fpubh.2021.649819] [Medline: 34307272]
7. Legido-Quigley H, Asgari N, Teo YY, et al. Are high-performing health systems resilient against the COVID-19 epidemic? Lancet 2020 Mar 14;395(10227):848-850. [doi: 10.1016/S0140-6736(20)30551-1] [Medline: 32151326]

8. Tangcharoensathien V, Bassett MT, Meng Q, Mills A. Are overwhelmed health systems an inevitable consequence of COVID-19? experiences from China, Thailand, and New York State. BMJ 2021 Jan 22;372:n83. [doi: 10.1136/bmj.n83] [Medline: 33483336]
9. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 2;6(2):e18810. [doi: 10.2196/18810] [Medline: 32238336]
10. Lim WH, Wong WM. COVID-19: notes from the front line, Singapore's primary health care perspective. Ann Fam Med 2020 May;18(3):259-261. [doi: 10.1370/afm.2539] [Medline: 32393562]
11. Lauriola P, Martín-Olmedo P, Leonardi GS, et al. On the importance of primary and community healthcare in relation to global health and environmental threats: lessons from the COVID-19 crisis. BMJ Glob Health 2021 Mar;6(3):e004111. [doi: 10.1136/bmjgh-2020-004111] [Medline: 33692145]
12. Malaysia: a primary health care case study in the context of the COVID-19 pandemic. World Health Organization. 2023 Aug 28. URL: https://www.who.int/publications/i/item/9789240076723 [accessed 2024-02-02]
13. Frieden TR, Lee CT, Lamorde M, Nielsen M, McClelland A, Tangcharoensathien V. The road to achieving epidemic-ready primary health care. Lancet Public Health 2023 May;8(5):e383-e390. [doi: 10.1016/S2468-2667(23)00060-9] [Medline: 37120262]
14. Kupferschmidt K, Cohen J. Can China's COVID-19 strategy work elsewhere? Science 2020 Mar 6;367(6482):1061-1062. [doi: 10.1126/science.367.6482.1061] [Medline: 32139521]
15. Declaration of Astana. World Health Organization. 2018. URL: https://www.who.int/docs/default-source/primary-health/declaration/gcphc-declaration.pdf [accessed 2023-07-28]
16. Wong SYS, Tan DHY, Zhang Y, et al. A tale of 3 Asian cities: how is primary care responding to COVID-19 in Hong Kong, Singapore, and Beijing. Ann Fam Med 2021;19(1):48-54. [doi: 10.1370/afm.2635] [Medline: 33431392]
17. Kavanagh MM, Erondu NA, Tomori O, et al. Access to lifesaving medical resources for African countries: COVID-19 testing and response, ethics, and politics. Lancet 2020 May 30;395(10238):1735-1738. [doi: 10.1016/S0140-6736(20)31093-X] [Medline: 32386564]
18. Universal health coverage (UHC) in Africa: a framework for action: main report (English). : World Bank Group; 2016 URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/735071472096342073/main-report [accessed 2023-07-28]
19. Wood G. Think 168,000 ventilators is too few? Try three. The Atlantic. 2020 Apr 10. URL: https://www.theatlantic.com/ideas/archive/2020/04/why-covid-might-hit-african-nations-hardest/609760/ [accessed 2023-07-28]
20. Huang P. 10,000 cases and 500 deaths in Africa. Health officials say it's just the beginning. NPR. 2020 Apr 8. URL: https://www.npr.org/sections/coronavirus-live-updates/2020/04/08/830209940/10-000-cases-and-500-deaths-in-africa-health-officials-say-its-just-the-beginnin [accessed 2023-07-28]
21. Lal A, Erondu NA, Heymann DL, Gitahi G, Yates R. Fragmented health systems in COVID-19: rectifying the misalignment between global health security and universal health coverage. Lancet 2021 Jan 2;397(10268):61-67. [doi: 10.1016/S0140-6736(20)32228-5] [Medline: 33275906]
22. Park S, Elliott J, Berlin A, Hamer-Hunt J, Haines A. Strengthening the UK primary care response to COVID-19. BMJ 2020 Sep 25;370:m3691. [doi: 10.1136/bmj.m3691] [Medline: 32978177]
23. Forrester JW. Industrial Dynamics: MIT Press; 1961.
24. Sterman JD. Business Dynamics: Systems Thinking and Modeling for a Complex World: McGraw-Hill Education; 2000.
25. Sterman JD. Learning from evidence in a complex world. Am J Public Health 2006 Mar;96(3):505-514. [doi: 10.2105/AJPH.2005.066043] [Medline: 16449579]
26. Homer JB, Hirsch GB. System dynamics modeling for public health: background and opportunities. Am J Public Health 2006 Mar;96(3):452-458. [doi: 10.2105/AJPH.2005.062059] [Medline: 16449591]
27. Darabi N, Hosseinichimeh N. System dynamics modeling in health and medicine: a systematic literature review. Syst Dyn Rev 2020 Mar 22;36(1):29-73. [doi: 10.1002/sdr.1646]
28. Rahmandad H, Sterman J. Quantifying the COVID-19 endgame: is a new normal within reach? Syst Dyn Rev 2022 Aug 24. [doi: 10.1002/sdr.1715] [Medline: 36246868]
29. Qian Y, Xie W, Zhao J, et al. Investigating the effectiveness of re-opening policies before vaccination during a pandemic: SD modelling research based on COVID-19 in Wuhan. BMC Public Health 2021 Sep 7;21(1):1638. [doi: 10.1186/s12889-021-11631-w] [Medline: 34493226]
30. Zhao J, Jia J, Qian Y, Zhong L, Wang J, Cai Y. COVID-19 in Shanghai: IPC policy exploration in support of work resumption through system dynamics modeling. Risk Manag Healthc Policy 2020 Oct 8;13:1951-1963. [doi: 10.2147/RMHP.S265992] [Medline: 33116976]
31. Huang J, Qian Y, Shen W, et al. Optimizing national border reopening policies in the COVID-19 pandemic: a modeling study. Front Public Health 2022 Nov 30;10:979156. [doi: 10.3389/fpubh.2022.979156] [Medline: 36530669]
32. Qian Y, Cao S, Zhao L, Yan Y, Huang J. Policy choices for Shanghai responding to challenges of Omicron. Front Public Health 2022 Aug 9;10:927387. [doi: 10.3389/fpubh.2022.927387] [Medline: 36016887]

33. Huang J, Liu Y, Zhang T, et al. Can family doctor contracted services facilitate orderly visits in the referral system? a frontier policy study from Shanghai, China. Int J Health Plann Manage 2022 Jan;37(1):403-416. [doi: 10.1002/hpm.3346] [Medline: 34628680]
34. Alexander GC, Tajanlangit M, Heyward J, Mansour O, Qato DM, Stafford RS. Use and content of primary care office-based vs telemedicine care visits during the COVID-19 pandemic in the US. JAMA Netw Open 2020 Oct 1;3(10):e2021476. [doi: 10.1001/jamanetworkopen.2020.21476] [Medline: 33006622]
35. Ortega G, Rodriguez JA, Maurer LR, et al. Telemedicine, COVID-19, and disparities: policy implications. Health Policy Technol 2020 Sep;9(3):368-371. [doi: 10.1016/j.hlpt.2020.08.001] [Medline: 32837888]
36. Patton-López MM. Communities in action: pathways to health equity. J Nutr Educ Behav 2022 Jan;54(1):P94-P95. [doi: 10.1016/j.jneb.2021.09.012]
37. Xiao Y, Chen X, Li Q, Jia P, Li L, Chen Z. Towards healthy China 2030: modeling health care accessibility with patient referral. Soc Sci Med 2021 May;276:113834. [doi: 10.1016/j.socscimed.2021.113834] [Medline: 33774532]
38. Tao W, Zeng Z, Dang H, et al. Towards universal health coverage: lessons from 10 years of healthcare reform in China. BMJ Glob Health 2020 Mar 19;5(3):e002086. [doi: 10.1136/bmjgh-2019-002086] [Medline: 32257400]
39. Yang J, Siri JG, Remais JV, et al. The Tsinghua-Lancet Commission on Healthy Cities in China: unlocking the power of cities for a healthy China. Lancet 2018 May 26;391(10135):2140-2184. [doi: 10.1016/S0140-6736(18)30486-0] [Medline: 29678340]
40. Forrest CB. Primary care in the United States: primary care gatekeeping and referrals: effective filter or failed experiment? BMJ 2003 Mar 29;326(7391):692-695. [doi: 10.1136/bmj.326.7391.692] [Medline: 12663407]
41. Hort K, Gilbert K, Basnayaka P, Annear PL. Strategies to strengthen referral from primary care to secondary care in low- and middle-income countries. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/325734/9789290227090-eng.pdf?sequence=1&isAllowed=y [accessed 2023-07-28]
42. Implementation plan to further enhance the city's community health service capabilities [Article in Chinese]. Shanghai Municipal People's Government. 2023 May 10. URL: https://www.shanghai.gov.cn/nw12344/20230510/55a194b734f54655ba5fc3bc60982a5d.html [accessed 2023-07-28]
43. Flanagan K, Uyarra E, Laranja M. Reconceptualising the 'policy mix' for innovation. Res Policy 2011 Jun;40(5):702-713. [doi: 10.1016/j.respol.2011.02.005]
44. Edler J, Cunningham P, Flanagan K, Laredo P. Innovation policy mix and instrument interaction: a review. : NESTA; 2013 URL: https://research.manchester.ac.uk/en/publications/innovation-policy-mix-and-instrument-interaction-a-review [accessed 2023-07-28]
45. 疫情通报 [Article in Chinese]. National Health Commission of the People's Republic of China. URL: http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml [accessed 2024-04-22]

Abbreviations

PHC: primary health care
WHO: World Health Organization

Edited by C Lovis; submitted 07.11.23; peer-reviewed by S Kreindler, W Tao; revised version received 04.03.24; accepted 10.03.24; published 03.06.24.

Please cite as:
Huang J, Qian Y, Yan Y, Liang H, Zhao L
Addressing Hospital Overwhelm During the COVID-19 Pandemic by Using a Primary Health Care–Based Integrated Health System: Modeling Study
JMIR Med Inform 2024;12:e54355
URL: https://medinform.jmir.org/2024/1/e54355
doi:10.2196/54355

Original Paper

Value of Electronic Health Records Measured Using Financial and Clinical Outcomes: Quantitative Study

Shikha Modi[1,2], MBA, PhD; Sue S Feldman[2], RN, MEd, PhD; Eta S Berner[2], EdD; Benjamin Schooley[3], PhD; Allen Johnston[4], PhD

[1]The University of Alabama in Huntsville, Huntsville, AL, United States

[2]The University of Alabama at Birmingham, Birmingham, AL, United States

[3]Brigham Young University, Provo, UT, United States

[4]Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL, United States

Corresponding Author:
Shikha Modi, MBA, PhD
The University of Alabama in Huntsville
1610 Ben Graves Dr NW
Huntsville, AL, 35816
United States
Phone: 1 2568242437
Email: ssm0031@uah.edu

Abstract

Background: The Health Information Technology for Economic and Clinical Health Act of 2009 was legislated to reduce health care costs, improve quality, and increase patient safety. Providers and organizations were incentivized to exhibit meaningful use of certified electronic health record (EHR) systems in order to achieve this objective. EHR adoption is an expensive investment, given the resources and capital that are invested. Due to the cost of the investment, a return on the EHR adoption investment is expected.

Objective: This study performed a value analysis of EHRs. The objective of this study was to investigate the relationship between EHR adoption levels and financial and clinical outcomes by combining both financial and clinical outcomes into one conceptual model.

Methods: We examined the multivariate relationships between different levels of EHR adoption and financial and clinical outcomes, along with the time variant control variables, using moderation analysis with a longitudinal fixed effects model. Since it is unknown as to when hospitals begin experiencing improvements in financial outcomes, additional analysis was conducted using a 1- or 2-year lag for profit margin ratios.

Results: A total of 5768 hospital-year observations were analyzed over the course of 4 years. According to the results of the moderation analysis, as the readmission rate increases by 1 unit, the effect of a 1-unit increase in EHR adoption level on the operating margin decreases by 5.38%. Hospitals with higher readmission payment adjustment factors have lower penalties.

Conclusions: This study fills the gap in the literature by evaluating individual relationships between EHR adoption levels and financial and clinical outcomes, in addition to evaluating the relationship between EHR adoption level and financial outcomes, with clinical outcomes as moderators. This study provided statistically significant evidence ($P<.05$), indicating that there is a relationship between EHR adoption level and operating margins when this relationship is moderated by readmission rates, meaning hospitals that have adopted EHRs could see a reduction in their readmission rates and an increase in operating margins. This finding could further be supported by evaluating more recent data to analyze whether hospitals increasing their level of EHR adoption would decrease readmission rates, resulting in an increase in operating margins. Hospitals would incur lower penalties as a result of improved readmission rates, which would contribute toward improved operating margins.

(JMIR Med Inform 2024;12:e52524) doi:10.2196/52524

KEYWORDS

acceptance; admission; adoption; clinical outcome; cost; economic; EHR adoption; EHR; electronic health record; finance; financial outcome; financial; health outcome; health record; hospital; hospitalization; length of stay; margin; moderation analysis;

multivariate; operating margin; operating; operation; operational; profit; project management; readmission rate; readmission; total margin; value analysis; value engineering; value management

Introduction

Overview

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 was legislated to reduce health care costs, improve quality, and increase patient safety [1]. Providers and organizations were incentivized to exhibit meaningful use of certified electronic health record (EHR) systems in order to achieve this objective [1]. The HITECH Act was based on the "triple aim" of health care, which consisted of reducing costs, improving the experience of care, and improving population health, and the HITECH Act contributed to the importance of EHRs [2]. Physicians and hospitals that adopted and used certified EHRs as described in federally defined "meaningful use" criteria were awarded approximately US $27 billion in incentives [3] for eligible providers.

EHR adoption is an expensive investment, given the resources and capital that are invested [4,5]. Due to the cost of the investment, a return on the EHR adoption investment is expected. Usually, a return on adoption is measured by calculating net profit and dividing the net profit by net investment [6]. Calculating a return on investment for EHR adoption requires considering the size of the organization, the extent of the EHR adoption, and profit or improvement in terms of both the financial and clinical outcomes perspectives. Given the complex process of calculating return on investment for EHR adoption, this study evaluates return on investment in terms of how it yields value to the adopting entity. Value from the health care perspective has been described in terms of dollars (financial), productivity (clinical), effectiveness (clinical) [7], cost savings (financial) [8], improvement in clinical decisions (clinical; Rudin et al [9]), supporting triage decisions (clinical; Rudin et al [9]), supporting collaborations among the providers (clinical; Rudin et al [9]), increased productivity (financial and clinical) [9], etc. However, a gap exists in that the return on investment is not analyzed in terms of financial and clinical outcomes combined. Additionally, current literature does not review EHR adoption in terms of level of EHR adoption but rather as a binary variable of "adopted" or "not adopted." This study addresses these gaps by including a combination of both financial and clinical outcomes in a conceptual model and reviewing EHR adoption in terms of levels of EHR adoption.

The value of health IT, of which EHRs are a subset, can depend on the stakeholder and context [10-12]. Looking at value from the stakeholder perspective, for the hospital, EHR value may be reviewed in terms of improved revenue and reduced cost (outcomes); for patients, value may be to improve health and prevent illness (outcomes); for providers, value may be to reduce errors and provide efficient care (process and outcomes); and for government, it may be to improve population health through timely public health reporting and population well-being (process and outcomes) [7]. Hence, given the frequent use of different outcome categories in the literature used to measure value, this study focuses on outcomes as the main value construct and investigates value in terms of different tangible outcomes, such as financial and clinical outcomes. This study examined how EHR adoption levels are associated with value in terms of financial and clinical outcomes combined in 1 model. To address this question, this study investigated the relationship between EHR adoption levels and financial and clinical outcomes by combining both financial and clinical outcomes into 1 conceptual model.

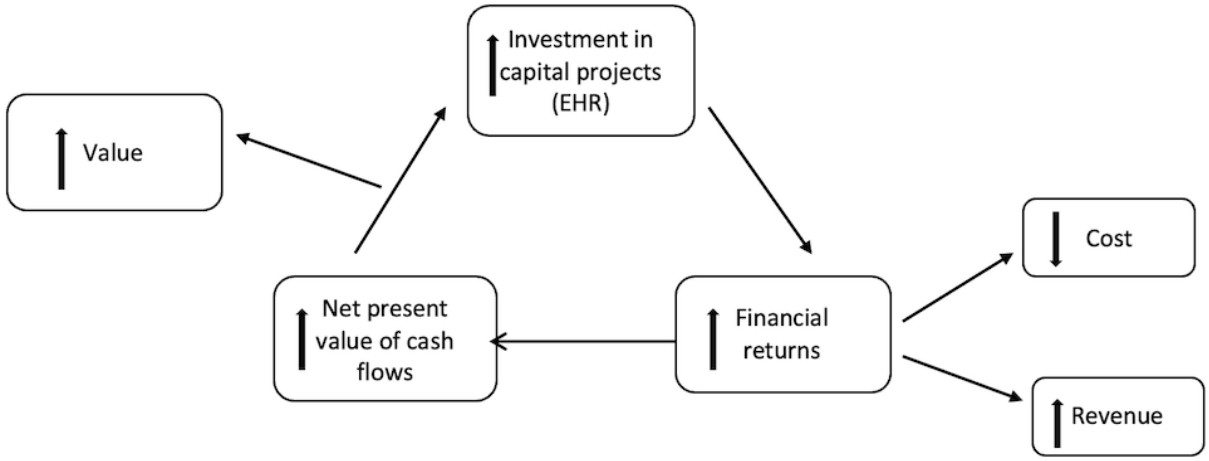
Conceptual Framework and Hypotheses

This study used the corporate financial theory of the firm [13] to guide the evaluation of the relationship between EHR adoption and financial and clinical outcomes. The corporate financial theory of the firm (Figure 1) indicates that the value of the firm, or in this case, the health care entity, is expected to be in alignment with the discounted cash flows from the investments, such as EHRs [13]. This theory indicates that a capital investment, such as EHR adoption, increases the value of the firm as it contributes toward an increase in the net present value of cash flows [13]. Multiple studies have supported the notion that EHR investments improve the value of a hospital through improved financial outcomes by way of a reduction in cost or improved revenues [4,14,15].

A study conducted by Collum et al [4] used this theory to determine an association between EHR adoption and financial outcomes (measured as profit margins and return on assets). The findings from this study indicated that financial returns depend on how long it takes for a hospital's EHR system to achieve full functionality [4], meaning it is important to consider the time variable when reviewing the outcomes of EHR adoption.

Additionally, there have been several studies that have analyzed the relationship between EHR adoption and financial outcomes without using the corporate financial theory of the firm as their guiding framework. Some of the studies from the current literature exhibited a trend that EHR adoption and financial outcomes have a nonlinear relationship [16,17], and some of the studies indicated that EHR adoption resulted in improved financial outcomes for health care organizations that adopted it over time [14,18].

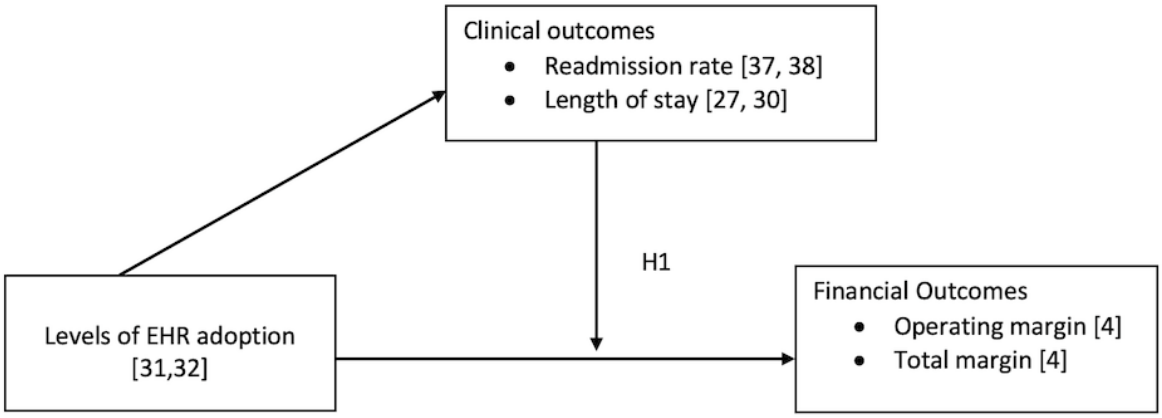
Figure 1. The corporate financial theory of the firm. EHR: electronic health record.



The literature suggests that improvement in costs and revenues is the result of improved clinical outcomes such as reduction of redundant tests [19], reduction of medication and hospital bed-related costs [20], improved ability to capture charges [15], and improved decision support systems [21]. Since this study focuses on combining both financial and clinical outcomes into 1 conceptual model, for the purpose of this study, the capital project investment (EHR adoption in this case) and improvement in financial returns (financial outcomes), tenets of the corporate financial theory of the firm, with an addition of the clinical outcomes, are integrated into a conceptual framework.

The purpose of this conceptual framework (Figure 2 [4,22-27]) is to determine if the previously stated overarching research question of "How is electronic health record adoption associated with value in terms of financial and clinical outcomes?" is supported by the following hypothesis: "The relationship between levels of EHR adoption and financial outcomes (both operating margin and total margin) is moderated by clinical outcomes (readmission rate and length of stay [LOS]) that are also affected by levels of EHR adoption (Figure 2).

Figure 2. Electronic health record (EHR) value analysis conceptual framework.



Methods

Data for this study were retrieved from multiple sources, including the Health Care Provider Cost Reporting Information System, the American Hospital Association (AHA) Annual Survey, the AHA IT Supplement Survey, and Health Care Analytics from Leavitt Partners. The study used a longitudinal design from 2014 to 2017 with 5897 hospital-year observations. Measures were divided into 2 groups: financial and clinical. Financial outcome variables were measured or operationalized using 2 variables (operating margin and total margin) that have been used in the health care literature to measure the profitability of hospitals after EHR adoption. The variables describing clinical outcomes are LOS and readmission rates, as these variables have an impact on the financial performance of the hospital [28,29]. The variables describing the financial outcomes are operating margin and total margin, as these measures include both costs and revenues described in the corporate financial theory of the firm [4,13]. The dependent variables used in this study (operating margins, total margins, LOS, and readmission rates) are not comprehensive in terms of measuring financial

and clinical outcomes for a hospital; however, for the purpose of this study, these variables are considered sufficient, given their potential association with one another.

Dependent Variables

Financial Outcome Variables

In order to gain an understanding of the financial performance of acute care hospitals, profitability ratios are the most frequently used measures [30]; hence, this study included operating margin and total margin as variables representing financial outcomes. Operating margin captures the expenses and revenues related to hospital operations. Total margin measures or captures operating and nonoperating expenses and revenues. The operating margin was calculated by dividing net patient revenues less total operating expenses by net patient revenues and multiplying the ratio by 100. The total margin variable is calculated by dividing net income by total patient revenue. The financial outcome variables are retrieved from the AHA Annual Survey (2014-2017).

Clinical Outcome Variables

Clinical outcomes were measured using LOS and readmission rates. Daniel et al [22] and Schreiber and Shaha [31] reported an intersection of financial and clinical outcomes as a result of EHR adoption and focused on LOS. These studies reported an improvement in LOS due to EHR adoption, resulting in lower plan premiums for patients [22] and costs [31]. Readmission rates are a part of the value-based purchasing program, and depending on the readmission rate, hospitals are penalized on a yearly basis, hence impacting hospital costs [28]. The readmission rates were measured for 6 conditions or procedures, as patients with these conditions are more likely to be readmitted to the hospital. These conditions are: acute myocardial infarction, chronic obstructive pulmonary disease, heart failure, pneumonia, coronary artery bypass graft surgery, and elective primary total hip arthroplasty and total knee arthroplasty [32]. LOS captures the number of days a patient spent in the hospital. Readmission rates indicate whether patients are readmitted to the hospital within 30 days of being discharged. The average LOS and readmission rates can be considered to be indicators of clinical quality outcomes by way of clinical quality measures [28]. Ben-Assuli et al [33] and Lee et al [34] have indicated improvements in average LOS and readmission rates as results of EHR adoption. To confirm these findings for the most recent data, this study analyzes how EHR adoption influences both average (LOS) and readmission rates for the selected sample.

The LOS variable is measured as the average number of days a patient stays in one hospital. The readmission rate variable is measured as the readmission rate payment adjustment factor. The full-year payment adjustment factor is based on data from the fiscal year Hospital Readmissions Reduction Program performance period (ie, July 1, 2014, to June 30, 2017). The minimum payment adjustment factor is 0.97 (ie, 3% maximum penalty). The maximum payment adjustment factor is 1 (ie, no penalty). Hospitals with higher payment adjustment factors have lower penalties [32].

Independent Variables

The level of EHR adoption is considered the major explanatory variable in this study. Hospitals are required to report the extent of adoption of each of the 28 EHR functions to the AHA IT Supplement Survey. The 28 EHR functions can be characterized into 5 different categories: clinical documentation, results viewing, computerized order entry, decision support, and bar coding. Hospitals indicate if each function is implemented in all units, 1 unit, or is in some stage of planning. A study conducted by Everson et al [23] emphasizes the reliability and validity of measuring hospital adoption of EHR with these 28 items.

In order to look at the extent of EHR adoption, Adler-Milstein et al [24] created a continuous EHR adoption measure for each hospital in each year in which they responded to the AHA IT Supplement Survey. The continuous measure was constructed as follows: for each function that was implemented in all units, a hospital received 2 points, and for each function that was implemented in at least 1 unit, a hospital received 1 point. According to the calculations, the total possible EHR adoption score ranged from 0 to 56. In order to improve interpretability, the measure was scaled by dividing each hospital's total score by 56, which yielded an EHR score ranging from 0 to 1. This strategy will be replicated in this study and applied to the EHR adoption level [24].

Control Variables

Control variables for this study include time-variant variables such as competition and payer mix. Control variables are identified based on elements that may influence the level of EHR adoption or hospital financial and clinical outcomes [4]. Since this study uses panel data, which accounts for changes in financial outcomes within hospitals due to changes in levels of EHR adoption, it is not essential to control for time-invariant hospital characteristics such as size of the hospital, ownership, system affiliation, and teaching status. For the purpose of this study, time-variant components that may change over the years, such as competition and payer mix, are considered control variables [4].

The competition construct was operationalized using the Herfindahl-Hirschman Index (HHI), which measures the concentration of an industry in a designated market. HHI was measured in terms of discharges for the health service area. Payer mix was measured using the proportion of inpatient days that were related to Medicare and Medicaid patients (Medicare percentage = total facility Medicare days/total inpatient days, and Medicaid percentage = total facility Medicaid days/total inpatient days). The AHA Annual Survey was used to collect the HHI and payer mix data.

Analysis

The unit of analysis for this study is at the hospital level. To demonstrate the appropriateness of the variables, univariate statistics and bivariate analyses were conducted. Bivariate statistics were generated for both independent and dependent variables of interest. Pairwise correlation analysis was conducted at the significance level of $P<.05$ in order to examine pairwise correlation coefficients between the continuous variables.

Multivariate relationships between different levels of EHR adoption and financial and clinical outcomes, along with the time-variant control variables, were examined using moderation analysis with a longitudinal fixed effects model [35]. Since it is unknown as to when hospitals begin experiencing improvements in financial outcomes, additional analysis was conducted using a 1- or 2-year lag for profit margin ratios [4]. Statistical significance was noted at the significance levels of $P<.10$, $P<.05$, and $P<.01$, and all statistical analyses were conducted in Stata (version 16; StataCorp).

Longitudinal Fixed Effects Moderation Analysis Model

A longitudinal fixed effects model with moderation analysis was used to analyze the multivariate relationships between different levels of EHR adoption and financial and clinical outcomes, along with the time variant control variables.

$$y_{it} = \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it1} X_{it2} + Z_{it}\lambda + \alpha_i + \mu_{it}$$

In this equation, y_{it} is the dependent variable (financial or clinical outcomes), i = hospital, and t = time. β_1 is the coefficient for the main independent variable (levels of EHR adoption), X_{it1}. β_2 is the coefficient for the moderator variable (clinical outcomes), X_{it2}. β_3 is the coefficient for the interaction of the independent variable (levels of EHR adoption) and moderator variable (clinical outcomes), $X_{it1}X_{it2}$. $Z_{it}\lambda$ represents all control variables (competition, payer mix, and years of observation). α_i is the unknown intercept for a vector of hospitals. And μ_{it} is the error term.

The hypothesis, that the relationship between EHR adoption and financial outcomes is moderated by clinical outcomes, was tested using multiple models. The models and their use are outlined in Textbox 1.

Textbox 1. Analytic models and their use.

Model 1

Determine the association between levels of electronic health record (EHR) adoption and operating margin moderated by length of stay (LOS) with the operating margins from the same year.

Model 2

Determine the association between levels of EHR adoption and operating margin moderated by readmission rates with the operating margins from the same year.

Model 3

Determine the association between levels of EHR adoption and total margin moderated by LOS with the total margins from the same year.

Model 4

Determine the association between levels of EHR adoption and total margin moderated by readmission rates with the total margins from the same year.

Model 5

Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 1-year lag in the operating margins.

Model 6

Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 2-year lag in the operating margins.

Model 7

Determine the association between levels of EHR adoption and operating margin moderated by readmission rates with a 1-year lag in the operating margins.

Model 8

Determine the association between levels of EHR adoption and operating margin moderated by readmission rates with a 2-year lag in the operating margins.

Model 9

Determine the association between levels of EHR adoption and total margin moderated by LOS with a 1-year lag in the total margins.

Model 10

Determine the association between levels of EHR adoption and total margin moderated by LOS with a 2-year lag in the total margins.

Model 11

Determine the association between levels of EHR adoption and total margin moderated by readmission rates with a 1-year lag in the total margins.

Model 12

Determine the association between levels of EHR adoption and total margin moderated by readmission rates with a 2-year lag in the total margins.

Ethical Considerations

This study was approved by the University of Alabama at Birmingham institutional review board (300003241).

Results

Overview

Descriptive statistics of acute care hospitals for the years 2014-2017 are displayed in Table 1. For acute care hospitals, average EHR adoption levels showed little variability across each observed year (approximately 0.89 for each observed year). Hospitals observed a steady decrease in average operating margin from 2014 (0.07%) to 2017 (0.057%). The average total margin across hospitals showed a decrease for 2015 (0.005%) compared with 2014 (1.014%), followed by a steady increase across years 2016 (1.136%) and 2017 (0.951%). An increase in LOS was observed for the years 2016 and 2017 (approximately 7.9 days for the year 2017 vs 3.9 days for the year 2014). Average readmission rates remained somewhat steady across all 4 observation years (approximately 0.99 for each observed year).

Table 1. Descriptive statistics of variables (N=5678 hospital-year observations).

Variables	2014 (n=1420)	2015 (n=1453)	2016 (n=1393)	2017 (n=1412)
Levels of EHR[a] adoption, mean (SD)	0.871 (0.127)	0.890 (0.121)	0.899 (0.127)	0.917 (0.102)
Operating margin, mean (SD)	0.070 (0.122)	0.065 (0.132)	0.06 (0.140)	0.057 (0.136)
Total margin, mean (SD)	1.014 (2.847)	0.005 (26.4)	1.136 (7.217)	0.951 (1.129)
Average length of stay (days), mean (SD)	3.911 (1.134)	3.881 (0.954)	7.87 (153.3)	7.945 (160.4)
Readmission rate, mean (SD)	0.998 (0.003)	0.995 (0.006)	0.995 (0.006)	0.994 (0.007)
Market competition (HHI[b]) in terms of discharges, mean (SD)	0.101 (0.199)	0.086 (0.157)	0.088 (0.172)	0.098 (0.193)
Medicare percentage, mean (SD)	0.512 (0.140)	0.518 (0.128)	0.518 (0.130)	0.521 (0.124)
Medicaid percentage, mean (SD)	0.197 (0.120)	0.202 (0.115)	0.203 (0.114)	0.204 (0.112)
Beds (n), mean (SD)	257 (231)	256 (229)	254 (232)	255 (236)
Ownership, n (%)				
Nongovernment not-for-profit	1105 (77.76)	1145 (78.8)	1177 (78.31)	1198 (78.82)
Investor-owned for-profit	294 (20.69)	295 (20.30)	311 (20.69)	305 (20.07)
Government nonfederal	22 (1.55)	13 (0.89)	15 (1)	17 (1.12)
Affiliation, n (%)				
Yes	584 (47.29)	660 (51.36)	687 (51.58)	731 (56.67)
No	651 (52.71)	625 (48.64)	645 (48.42)	559 (43.33)
Teaching status, n (%)				
Yes	560 (39.41)	569 (39.16)	595 (39.59)	599 (39.41)
No	861 (60.59)	884 (60.84)	908 (60.41)	921 (60.59)

[a]EHR: electronic health record.

[b]HHI: Herfindahl-Hirschman Index.

For time-variant control variables, the average HHI in terms of discharges across all 4 years was approximately 0.093. HHI values range from 0 to 1, where an HHI value closer to 1 means monopolistic markets, or more market share, and an HHI value closer to 0 means highly competitive markets, or less market share. For the sample used in this study, the markets appear to be highly competitive. In terms of payer mix, the Medicare percentage was similar across all 4 years (average of 0.52). Similarly, the Medicaid percentage was also similar across all 4 years (average of 0.20).

For organizational characteristics, bed size was somewhat similar across all hospitals for all observed years (approximately 255 beds per hospital). In terms of ownership status of the sample hospitals, a majority of the hospitals were nongovernment, not-for-profit hospitals (1105/1421, 78%), followed by investor-owned for-profit hospitals (294/1421, 20%) and government nonfederal hospitals (22/1421, 1.5%). In terms of system affiliation, approximately half the hospitals were affiliated with a system, and the other half were not. For teaching status, a majority of the hospitals did not hold a teaching status (861/1421, 61%).

According to the bivariate statistical analysis (Table 2), at the significance level of $P<.05$, levels of EHR adoption exhibit a positive correlation with operating margin at a magnitude of 0.0978. At the significance level of $P<.05$, readmission rate and levels of EHR adoption are negatively correlated at the magnitude of 0.0321. Even though the magnitudes are close to

0, these relationships are statistically significant at the significance level of $P<.05$.

Table 2. Bivariate analysis of variables.

Dependent variables	Independent variables: levels of EHR[a] adoption (correlation coefficients)
Operating margin	0.0978[b]
Total margin	–0.0142
Average length of stay	0.0039
Readmission rate	–0.0321[b]

[a]EHR: electronic health record.

[b]$P<.05$.

This study tested the following hypothesis that was derived from the EHR value analysis conceptual framework (Figure 2): "The relationship between EHR adoption and financial outcomes is moderated by clinical outcomes." Tables 3 and 4 provide details relative to the hypothesis.

Table 3. Fixed effects with regression analysis.

Variables	Model 1	Model 2	Model 3	Model 4
	OM[a]–LOS[b]–levels of EHR[c] adoption (Prob>F=0.0828)	OM–RR[d]–levels of EHR adoption (Prob>F=0.0116)	TM[e]–LOS–levels of EHR adoption (Prob>F=0.4532)	TM–RR–levels of EHR adoption (Prob>F=0.3388)
Levels of EHR adoption	–0.020	5.335[f]	–4.961	415.2
Dependent variables				
Average LOS	0.000	N/A[g]	–0.002	N/A
RR	N/A	4.375[h]	N/A	431.6
Levels of EHR adoption and average LOS	-0.000	N/A	0.001	N/A
Levels of EHR adoption and RR	N/A	–5.384[f]	N/A	–422.3
Control variables				
Market competition (HHI[i])	0.082	0.078	3.148	2.959
Medicare percentage	–0.009	–0.013	0.699	0.937
Medicaid percentage	–0.026	–0.026	1.211	1.343
Years				
2014	Reference	Reference	Reference	Reference
2015	–0.005	–0.007[f]	–1.001	–0.848
2016	–0.008	–0.009[f]	0.388	0.569
2017	–0.008	–0.011[j]	0.243	0.460

[a]OM: operating margin.

[b]LOS: length of stay.

[c]EHR: electronic health record.

[d]RR: readmission rate.

[e]TM: total margin.

[f]$P<.05$.

[g]N/A: not applicable.

[h]$P<.10$.

[i]HHI: Herfindahl-Hirschman Index.

[j]$P<.001$.

Table 4. Regression analysis with fixed effects for lagged variables.

Variables	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
	OM[a]–LOS[b]–levels of EHR[c] adoption with 1-year lag (Prob>F=0.0047)	OM–LOS–levels of EHR adoption with 2-year lag (Prob>F=0.0271)	OM–RR[d]–levels of EHR adoption with 1-year lag (Prob>F=0.0010)	OM–RR–levels of EHR adoption with 2-year lag (Prob>F=0.0202)	TM[e]–LOS–levels of EHR adoption with 1-year lag (Prob>F=0.6885)	TM–LOS–levels of EHR adoption with 2-year lag (Prob>F=0.6738)	TM–RR–levels of EHR adoption with 1-year lag (Prob>F=0.6492)	TM–RR– levels of EHR adoption with 2-year lag (Prob>F=0.5143)
Levels of EHR adoption	0.022	0.004	1.681	2.229	1.564	–1.547	–164.0	268.8
Average LOS	0.000	–9.46e–06	N/A[f]	N/A	0.001	–0.012	N/A	N/A
RR	N/A	N/A	0.818	2.192	N/A	N/A	–186.4	169.8
Levels of EHR adoption and average LOS	–0.000	–4.26e–06	N/A	N/A	–0.002	0.013	N/A	N/A
Levels of EHR adoption and RR	N/A	N/A	–1.663	–2.232	N/A	N/A	166.2	–271.7
Market competition (HHI[g])	0.219[h]	–0.068	0.223[h]	–0.068	3.610	–3.275	3.749	–3.103
Medicare percentage	0.063[i]	0.024	0.068[h]	0.030	–2.419	0.523	–2.416	0.783
Medicaid percentage	0.018	0.891[h]	0.018	0.092[h]	1.742	–2.822	1.741	–2.838
Years								
2014	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
2015	0.014[h]	–0.006	0.014[h]	–0.006	–1.057	–0.714	–1.178	–0.910
2016	0.011[h]	–0.008[i]	0.009	–0.008[i]	0.141	0.482	0.048	0.323
2017	0.010[h]	–0.018[j]	0.009[i]	–0.018[h]	0.008	0.508	–0.126	0.229

[a]OM: operating margin.

[b]LOS: length of stay.

[c]EHR: electronic health record.

[d]RR: readmission rate.

[e]TM: total margin.

[f]N/A: not applicable.

[g]HHI: Herfindahl-Hirschman Index.

[h]$P<.05$.

[i]$P<.10$.

[j]$P<.001$.

EHR: Length of Stay (Operating Margin and Total Margin)

Model 1 analyzed the relationship between EHR adoption levels and operating margins without any lags in operating margins, with LOS as a moderating variable for acute care hospitals. For Model 1, the prob>F was greater than 0.05, meaning this model did not provide a statistical explanation for the proposed relationship between EHR adoption levels and operating margins with LOS as a moderating variable.

Model 5 analyzed the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with LOS as the moderating variable for acute care hospitals. The prob>F was less than .05 for this model; however, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with LOS as a moderating variable. The nonsignificant results indicated a direct positive association between EHR adoption levels and operating margins and LOS; however, when LOS acts as a moderating variable, the indirect relationship between EHR adoption levels and operating margins was negative.

Model 6 analyzed the relationship between EHR adoption levels and operating margins with a 2-year lag in operating margins, with LOS as a moderating variable for acute care hospitals. Even though the prob>F was less than .05 for this model, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and

operating margins with a 1-year lag in operating margins, with LOS as a moderating variable. The nonsignificant results indicated a direct positive association between EHR adoption levels and operating margins with a 2-year lag, which was expected. Additionally, the nonsignificant results indicated a direct negative association between EHR adoption levels and LOS, which is consistent with the findings from the literature. However, when LOS is introduced as a moderating variable, the nonsignificant results indicate a negative indirect relationship between EHR adoption levels and operating margins with a 2-year lag.

Model 3 analyzed the relationship between EHR adoption levels and total margins without any lags in total margins, with LOS as a moderating variable for acute care hospitals. The prob>F was greater than 0.05, meaning the models did not provide a statistically significant explanation for the proposed relationship between EHR adoption levels and total margins without any lags in total margins, with LOS as a moderating variable.

Model 9 analyzed the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with LOS as a moderating variable for acute care hospitals. For Model 9, the prob>F was greater than 0.05, meaning this model could not accurately predict the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with LOS as a moderating variable.

Model 10 analyzed the relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with LOS as a moderating variable for acute care hospitals. For Model 10, the prob>F was greater than 0.05, which indicates that this model could not accurately predict the relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with LOS as a moderating variable.

EHR: Readmission Rate (Operating Margin and Total Margin)

Model 2 analyzed the relationship between EHR adoption levels and operating margins without any lags in operating margins, with readmission rates as a moderating variable for acute care hospitals. Hospitals with higher readmission payment adjustment factors have lower penalties [32]. This was the only model in which the results from the analysis provided statistically significant evidence to support the proposed relationship. At the significance level of $P<.05$, EHR adoption levels were positively associated with operating margins. Similarly, at the significance level of $P<.05$, readmission rates were positively associated with an increase in operating margin. However, when readmission rates are introduced as a moderating variable, the magnitude of the relationship between levels of EHR adoption and operating margins is negative.

Model 7 analyzed the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with readmission rates as a moderating variable for acute care hospitals. For Model 7, the prob>F was less than .05 for this model; however, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with readmission rates as a moderating variable. The nonsignificant results indicate a direct positive association between EHR adoption levels and operating margins with a 1-year lag and readmission rates, which is consistent with the findings from Model 2. However, when readmission rates act as a moderating variable, the nonsignificant results indicate a positive relationship between levels of EHR adoption and operating margins with a 1-year lag, which was the opposite of the results from Model 2.

Model 8 analyzed the relationship between EHR adoption levels and operating margins with a 2-year lag in operating margins, with readmission rates as a moderating variable for acute care hospitals. The prob>F was less than .05 for this model; however, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and operating margins with a 2-year lag in operating margins, with readmission rates as a moderating variable. Similar to Model 7, the nonsignificant results indicated a direct positive association between EHR adoption levels and operating margins with a 2-year lag and readmission rates, which was consistent with the findings from Model 2. However, when readmission rates act as a moderating variable, the nonsignificant results indicated a positive relationship between levels of EHR adoption and operating margins with a 2-year lag, which was the opposite of the results from Model 2.

Model 4 analyzed the relationship between EHR adoption levels and total margins without any lags in total margins, with readmission rates as a moderating variable for acute care hospitals. The prob>F was greater than 0.05, meaning this model could not provide a statistically significant explanation for the proposed relationship between EHR adoption levels and total margins without any lags in total margins, with readmission rates as a moderating variable.

Model 11 analyzed the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with readmission rates as a moderating variable for acute care hospitals. For Model 11, the prob>F was greater than 0.05, which indicates that this model could not accurately predict the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with readmission rates as a moderating variable.

Model 12 analyzed the relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with readmission rates as a moderating variable for acute care hospitals. The prob>F was greater than 0.05, meaning this model could not provide a statistically significant explanation for the proposed relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with readmission rates as a moderating variable.

Results from the regression analysis with fixed effects are displayed in Tables 3 and 4. Table 3 includes results from the regression analysis with financial and clinical outcomes from the same year. Hospitals receive their reimbursement and penalties associated with readmission rates and LOS approximately 1 to 2 years after the actual outcomes occur. In order to accommodate this situation, operating margin and total margin ratios were calculated with a 1- and 2-year lag. Table 4

presents results with lags in profit margins for acute care hospitals.

The results from Table 3 for model 2 suggest that, at the significance level of $P<.05$, a 1-unit increase in EHR adoption was associated with an increase of approximately 5.34% in the operating margin.

Table 4 displays results from the analyses with the added lag effect in operating and total margins. According to the results displayed in Table 4, it can be inferred that at the significance levels of $P<.05$, $P<.10$, or $P<.001$, there is not enough evidence to support models 5-8 from this study. For models 9-12, the models did not provide a statistical explanation for the proposed relationships. In other words, the models discussed above could not accurately predict the proposed relationships.

Discussion

Overview

The objective of this study was to determine how EHR adoption level contributes to financial and clinical outcomes for acute care hospitals.

To understand the relationship between EHR adoption level and financial outcomes, moderated by clinical outcomes, this study used a fixed effects moderation analysis model. We hypothesized that there would be a positive association between EHR adoption level and operating and total margins, with LOS and readmission rates as moderating variables.

According to the results displayed in Table 3, for models 1, 3, and 4, the prob>F was greater than .05, meaning the models did not provide a statistical explanation for the proposed relationships in these models. In other words, the models discussed above could not accurately predict the proposed relationships, and there is no evidence that EHR adoption levels have a linear relationship with or explain variance in the operating margins, total margins, and LOS.

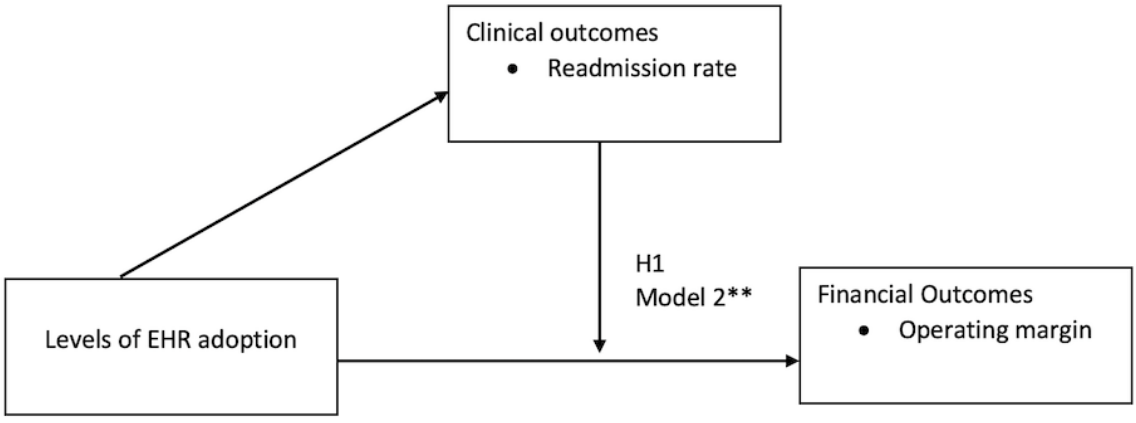
Even though the results are inverse of what was predicted in the hypothesis, these findings indicated that the relationship between EHR adoption levels and operating margins was statistically significant when it was moderated by the readmission rates variable at the significance level of $P<.05$. According to the results of the moderation analysis, as the readmission rate increases by 1 unit, the effect of a 1-unit increase in EHR adoption level on the operating margin decreases by 5.38%. In other words, when the hospital incurred lower penalties for readmissions, the operating margins increased. The minimum payment adjustment factor is 0.97 (ie, 3% maximum penalty). The maximum payment adjustment factor is 1 (ie, no penalty), and hospitals with higher payment adjustment factors have lower penalties and, in turn, larger operating margins [32].

In order to confirm any lagged effect (the timeline of hospitals receiving penalties or incentives for EHR adoption being not clear), this study included additional models that accounted for 1- and 2-year lag in the profit margin ratios (models 5-12). The results, however, did not provide any statistically significant evidence supporting a positive relationship between EHR adoption level and profit margin ratios when the lag effect was included in the model.

Findings from current literature indicate an improvement in LOS as a result of EHR adoption (not necessarily adoption level) yielding increased compensation for the loss of patient days from Center for Medicare and Medicaid Services [25]; however, for this study, none of the tested models provided a statistical explanation for the proposed relationships between EHR adoption and profit margins with LOS as moderating variables.

Even though this finding is opposite of what was proposed in the hypothesis, this finding provides statistically significant evidence that levels of EHR adoption change operating margins when readmission rates are taken into account (Figure 3). Analyzing more recent data could indicate a decrease in readmission rates as a result of increased levels of EHR adoption, yielding an increase in operating margins. The relationship between EHR adoption level and operating margins has not been previously evaluated using readmission rates as moderating variables. Hence, this finding from this study is a unique contribution to the current literature.

Figure 3. Electronic health record (EHR) value analysis framework with results. ***P*<.05.



Limitations of This Study

Regardless of the valuable contribution of the buildout of the conceptual model and the results from the analysis, this study has limitations. First, there is always a risk when using secondary data to conduct research that was not the intent when the data were collected, as this could result in inconsistency in the data collection methods due to the possibility of human error [36].

Second, this study used data from the Medicare Cost Reports to operationalize the readmission rate variable. This particular measure is reported on a 3-year rolling basis, meaning the data analyzed included a rolling average of 3 years of readmission rate data for each hospital [32]. This study operationalized the readmission rate data for specific years in order to evaluate their relationship with levels of EHR adoption and financial outcomes, which can be considered a limitation.

Conclusion

The HITECH Act has played an important role in EHRs becoming an integral part of the modern health system over the last 10 years. The goal of enacting the HITECH Act of 2009 was to reduce health care costs, improve the quality of the care provided, and increase patient safety for providers and organizations that exhibited meaningful use of certified EHR systems [1,37]. Given the cost and complexity of EHR adoption, analyzing its value from various and seemingly atypical perspectives is essential.

The current literature does a good job of providing perspectives on EHR value relative to individual financial and clinical outcomes, but it falls short in providing a collective value analysis. This study fills the gap in the literature by evaluating individual relationships between EHR adoption levels and financial and clinical outcomes, in addition to evaluating the relationship between EHR adoption level and financial outcomes, with clinical outcomes as moderators.

This study provided statistically significant evidence, indicating that there is a relationship between EHR adoption level and operating margins when this relationship is moderated by readmission rates. This finding could further be supported by evaluating more recent data to analyze whether hospitals increasing their level of EHR adoption would decrease readmission rates, resulting in an increase in operating margins. Hospitals would incur lower penalties as a result of improved readmission rates, which would contribute toward improved operating margins.

Conflicts of Interest

Not applicable.

References

1. McAlearney AS, Sieck C, Hefner J, Robbins J, Huerta TR. Facilitating ambulatory electronic health record system implementation: evidence from a qualitative study. Biomed Res Int 2013;2013:629574 [FREE Full text] [doi: 10.1155/2013/629574] [Medline: 24228257]
2. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
3. Adler-Milstein J, Green CE, Bates DW. A survey analysis suggests that electronic health records will yield revenue gains for some practices and losses for many. Health Aff (Millwood) 2013;32(3):562-570. [doi: 10.1377/hlthaff.2012.0306] [Medline: 23459736]
4. Collum TH, Menachemi N, Sen B. Does electronic health record use improve hospital financial performance? Evidence from panel data. Health Care Manage Rev 2016;41(3):267-274 [FREE Full text] [doi: 10.1097/HMR.0000000000000068] [Medline: 26052785]
5. Jang Y, Lortie MA, Sanche S. Return on investment in electronic health records in primary care practices: a mixed-methods study. JMIR Med Inform 2014 Oct 29;2(2):e25 [FREE Full text] [doi: 10.2196/medinform.3631] [Medline: 25600508]
6. Pine R, Tart K. Return on investment: benefits and challenges of baccalaureate nurse residency program. Nurs Econ 2007;25(1):13-18. [Medline: 17402673]
7. Payne TH, Bates DW, Berner ES, Bernstam EV, Covvey HD, Frisse ME, et al. Healthcare information technology and economics. J Am Med Inform Assoc 2013;20(2):212-217 [FREE Full text] [doi: 10.1136/amiajnl-2012-000821] [Medline: 22781191]
8. Peterson LT, Ford EW, Eberhardt J, Huerta TR, Menachemi N. Assessing differences between physicians' realized and anticipated gains from electronic health record adoption. J Med Syst 2011 May;35(2):151-161 [FREE Full text] [doi: 10.1007/s10916-009-9352-z] [Medline: 20703574]
9. Rudin RS, Friedberg MW, Shekelle P, Shah N, Bates DW. Getting value from electronic health records: research needed to improve practice. Ann Intern Med 2020 Jul 02;172(11 Suppl):S130-S136 [FREE Full text] [doi: 10.7326/M19-0878] [Medline: 32479182]
10. Shah GH, Leider JP, Castrucci BC, Williams KS, Luo H. Characteristics of local health departments associated with implementation of electronic health records and other informatics systems. Public Health Rep 2016;131(2):272-282 [FREE Full text] [doi: 10.1177/003335491613100211] [Medline: 26957662]

11. Feldman SS. Value proposition of health information exchange. In: Dixon BE, editor. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. Amsterdam, Netherlands: Academic Press; 2015.
12. Feldman SS. Public-Private Interorganizational Sharing of Health Data for Disability Determination. Ann Arbor, MI: ProQuest LLC; 2011.
13. Copeland TE, Weston JF, Shastri K. Financial Theory and Corporate Policy. Harlow, UK: Pearson; 2014.
14. Thouin MF, Hoffman JJ, Ford EW. The effect of information technology investment on firm-level performance in the health care industry. Health Care Manage Rev 2008;33(1):60-68 [FREE Full text] [doi: 10.1097/01.HMR.0000304491.03147.06] [Medline: 18091445]
15. Edwardson N, Kash BA, Janakiraman R. Measuring the impact of electronic health record adoption on charge capture. Med Care Res Rev 2017 Oct;74(5):582-594 [FREE Full text] [doi: 10.1177/1077558716659408] [Medline: 27416948]
16. Lim MC, Boland MV, McCannel CA, Saini A, Chiang MF, Epley KD, et al. Adoption of electronic health records and perceptions of financial and clinical outcomes among ophthalmologists in the United States. JAMA Ophthalmol 2018 Mar 01;136(2):164-170 [FREE Full text] [doi: 10.1001/jamaophthalmol.2017.5978] [Medline: 29285542]
17. Fleming NS, Becker ER, Culler SD, Cheng D, McCorkle R, da Graca B, et al. The impact of electronic health records on workflow and financial measures in primary care practices. Health Serv Res 2014 Mar;49(1 Pt 2):405-420 [FREE Full text] [doi: 10.1111/1475-6773.12133] [Medline: 24359533]
18. Wang T, Wang Y, McLeod A. Do health information technology investments impact hospital financial performance and productivity? International Journal of Accounting Information Systems 2018 Mar;28:1-13 [FREE Full text] [doi: 10.1016/j.accinf.2017.12.002]
19. Knepper MM, Castillo EM, Chan TC, Guss DA. The effect of access to electronic health records on throughput efficiency and imaging utilization in the emergency department. Health Serv Res 2018 Apr;53(2):787-802 [FREE Full text] [doi: 10.1111/1475-6773.12695] [Medline: 28376563]
20. Litzelman DK, Dittus RS, Miller ME, Tierney WM. Requiring physicians to respond to computerized reminders improves their compliance with preventive care protocols. J Gen Intern Med 1993 Jul;8(6):311-317 [FREE Full text] [doi: 10.1007/BF02600144] [Medline: 8320575]
21. Amarasingham R, Plantinga L, Diener-West M, Gaskin DJ, Powe NR. Clinical information technologies and inpatient outcomes: a multiple hospital study. Arch Intern Med 2009 Jan 26;169(2):108-114 [FREE Full text] [doi: 10.1001/archinternmed.2008.520] [Medline: 19171805]
22. Daniel GW, Ewen E, Willey VJ, Reese Iv CL, Shirazi F, Malone DC. Efficiency and economic benefits of a payer-based electronic health record in an emergency department. Acad Emerg Med 2010 Aug;17(8):824-833 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00816.x] [Medline: 20670319]
23. Everson J, Lee SYD, Friedman CP. Reliability and validity of the American Hospital Association's national longitudinal survey of health information technology adoption. J Am Med Inform Assoc 2014 Oct;21(e2):e257-e263 [FREE Full text] [doi: 10.1136/amiajnl-2013-002449] [Medline: 24623194]
24. Adler-Milstein J, Everson J, Lee SYD. EHR adoption and hospital performance: time-related effects. Health Serv Res 2015 Dec;50(6):1751-1771 [FREE Full text] [doi: 10.1111/1475-6773.12406] [Medline: 26473506]
25. Mirani R, Harpalani A. Business benefits or incentive maximization? impacts of the medicare EHR incentive program at acute care hospitals. ACM Trans Manage Inf Syst 2013 Dec;4(4):1-19 [FREE Full text] [doi: 10.1145/2543900]
26. Thirukumaran CP, Dolan JG, Reagan Webster P, Panzer RJ, Friedman B. The impact of electronic health record implementation and use on performance of the Surgical Care Improvement Project measures. Health Serv Res 2015;50(1):273-289. [doi: 10.1111/1475-6773.12191] [Medline: 24965357]
27. Wani D, Malhotra M. Does the meaningful use of electronic health records improve patient outcomes? J Oper Manag 2018;60(1):1-18. [doi: 10.1016/j.jom.2018.06.003]
28. Medicare and medicaid promoting interoperability program basics. Centers for Medicare & Medicaid Services. 2018. URL: https://www.cms.gov/medicare/regulations-guidance/promoting-interoperability-programs/medicare-medicaid-basics [accessed 2019-01-16]
29. Rojas-García A, Turner S, Pizzo E, Hudson E, Thomas J, Raine R. Impact and experiences of delayed discharge: A mixed-studies systematic review. Health Expect 2018 Mar;21(1):41-56 [FREE Full text] [doi: 10.1111/hex.12619] [Medline: 28898930]
30. Pink GH, Holmes GM, D'Alpe C, Strunk LA, McGee P, Slifkin RT. Financial indicators for critical access hospitals. J Rural Health 2006;22(3):229-236 [FREE Full text] [doi: 10.1111/j.1748-0361.2006.00037.x] [Medline: 16824167]
31. Schreiber R, Shaha SH. Computerised provider order entry adoption rates favourably impact length of stay. J Innov Health Inform 2016 May 18;23(1):166 [FREE Full text] [doi: 10.14236/jhi.v23i1.166] [Medline: 27348485]
32. Hospital readmissions reduction program (HRRP). Centers for Medicare & Medicaid Services. 2020. URL: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program [accessed 2023-11-08]
33. Ben-Assuli O, Shabtai I, Leshno M. The impact of EHR and HIE on reducing avoidable admissions: controlling main differential diagnoses. BMC Med Inform Decis Mak 2013 May 17;13(1):49 [FREE Full text] [doi: 10.1186/1472-6947-13-49] [Medline: 23594488]

34. Lee J, Kuo YF, Lin YL, Goodwin JS. The combined effect of the electronic health record and hospitalist care on length of stay. Am J Manag Care 2015 Mar 01;21(3):e215-e221 [FREE Full text] [Medline: 26014309]
35. Bailey MA. Real Econometrics The Right Tools to Answer Important Questions. New York, NY: Oxford University Press; 2017.
36. Hoffmann F, Andersohn F, Giersiepen K, Scharnetzky E, Garbe E. [Validation of secondary data. Strengths and limitations]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2008 Oct;51(10):1118-1126. [doi: 10.1007/s00103-008-0646-y] [Medline: 18985405]
37. Redd TK, Read-Brown S, Choi D, Yackel TR, Tu DC, Chiang MF. Electronic health record impact on productivity and efficiency in an academic pediatric ophthalmology practice. J AAPOS 2014 Dec;18(6):584-589 [FREE Full text] [doi: 10.1016/j.jaapos.2014.08.002] [Medline: 25456030]

Abbreviations

AHA: American Hospital Association
EHR: electronic health record
HHI: Herfindahl-Hirschman Index
HITECH: Health Information Technology for Economic and Clinical Health
LOS: length of stay

Edited by J Hefner; submitted 06.09.23; peer-reviewed by L Heryawan, A Kotlo; comments to author 23.10.23; revised version received 29.10.23; accepted 29.11.23; published 24.01.24.

Please cite as:
Modi S, Feldman SS, Berner ES, Schooley B, Johnston A
Value of Electronic Health Records Measured Using Financial and Clinical Outcomes: Quantitative Study
JMIR Med Inform 2024;12:e52524
URL: https://medinform.jmir.org/2024/1/e52524
doi:10.2196/52524
PMID:38265848

Review

Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review

Dobin Yim[1], PhD; Jiban Khuntia[2], PhD; Vijaya Parameswaran[3], RD, PhD; Arlen Meyers[2], PhD

[1]Loyola University, Maryland, MD, United States

[2]University of Colorado Denver, Denver, CO, United States

[3]Stanford University, Stanford, CA, United States

Corresponding Author:
Jiban Khuntia, PhD
University of Colorado Denver
1475 Lawrence St.
Denver, CO
United States
Phone: 1 3038548024
Email: jiban.khuntia@ucdenver.edu

Abstract

Background: Generative artificial intelligence tools and applications (GenAI) are being increasingly used in health care. Physicians, specialists, and other providers have started primarily using GenAI as an aid or tool to gather knowledge, provide information, train, or generate suggestive dialogue between physicians and patients or between physicians and patients' families or friends. However, unless the use of GenAI is oriented to be helpful in clinical service encounters that can improve the accuracy of diagnosis, treatment, and patient outcomes, the expected potential will not be achieved. As adoption continues, it is essential to validate the effectiveness of the infusion of GenAI as an intelligent technology in service encounters to understand the gap in actual clinical service use of GenAI.

Objective: This study synthesizes preliminary evidence on how GenAI assists, guides, and automates clinical service rendering and encounters in health care The review scope was limited to articles published in peer-reviewed medical journals.

Methods: We screened and selected 0.38% (161/42,459) of articles published between January 1, 2020, and May 31, 2023, identified from PubMed. We followed the protocols outlined in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to select highly relevant studies with at least 1 element on clinical use, evaluation, and validation to provide evidence of GenAI use in clinical services. The articles were classified based on their relevance to clinical service functions or activities using the descriptive and analytical information presented in the articles.

Results: Of 161 articles, 141 (87.6%) reported using GenAI to assist services through knowledge access, collation, and filtering. GenAI was used for disease detection (19/161, 11.8%), diagnosis (14/161, 8.7%), and screening processes (12/161, 7.5%) in the areas of radiology (17/161, 10.6%), cardiology (12/161, 7.5%), gastrointestinal medicine (4/161, 2.5%), and diabetes (6/161, 3.7%). The literature synthesis in this study suggests that GenAI is mainly used for diagnostic processes, improvement of diagnosis accuracy, and screening and diagnostic purposes using knowledge access. Although this solves the problem of knowledge access and may improve diagnostic accuracy, it is oriented toward higher value creation in health care.

Conclusions: GenAI informs rather than assisting or automating clinical service functions in health care. There is potential in clinical service, but it has yet to be actualized for GenAI. More clinical service–level evidence that GenAI is used to streamline some functions or provides more automated help than only information retrieval is needed. To transform health care as purported, more studies related to GenAI applications must automate and guide human-performed services and keep up with the optimism that forward-thinking health care organizations will take advantage of GenAI.

(JMIR Med Inform 2024;12:e52073) doi:10.2196/52073

KEYWORDS

generative artificial intelligence tools and applications; GenAI; service; clinical; health care; transformation; digital

Introduction

Background

Generative artificial intelligence tools and applications (GenAI) systems automatically learn patterns and structures from text, images, sounds, animation, models, or other media inputs to generate new data with similar characteristics [1]. GenAI is used to search, write, and create models, computer codes, and art forms without human assistance. GenAI has emerged significantly in the current decade to help every industry through different products such as ChatGPT, Bing Chat, Bard, LLaMA, Stable Diffusion, Midjourney, and DALL-E [2-5]. Almost all industries share an optimistic vision, with significant investment in using GenAI to transform aspects of value chains [6-10]. However, similar to many other technology hypes, whether this optimism will translate to value outcomes or be a "fad or fashion" remains to be tested over time.

The adoption of GenAI in health care is emerging. Studies point to the use of GenAI in service interactions involving breast cancer diagnoses [11], bariatric surgery [12], cardiopulmonary resuscitation [13], and breast cancer radiologic decision-making [14]. GenAI has the potential to transform by performing tasks at higher quality than humans, which may reduce errors associated with humans in expert domains such as cancer detection [15] and neurological clinical decisions [16]. The rise of GenAI is also referred to as the "second machine age" [17], whereby "instead of machines performing mechanical work they are taking on cognitive work exclusively in the human domain" [17]. Although these instances are encouraging, how exactly GenAI helps in health care processes needs to be articulated and evaluated to provide an understanding of use and value linkages [18,19]. Thus, we asked the following research questions (RQs) in this study: (1) How is GenAI used across different aspects of health care services? (RQ 1) and (2) What is the preliminary evidence of GenAI use across health care services? (RQ 2).

It is essential to explore these 2 RQs for several reasons. Exploring GenAI's use in health care services is essential for realizing its potential benefits, addressing ethical concerns, and continually improving its applications to enhance patient care and the health care ecosystem. This impact spans different areas. For instance, GenAI can help analyze data to provide personalized treatment and tailor interventions. It has shown promise in improving diagnostic accuracy, with higher levels of accuracy in the interpretation of images and scans. AI applications can enhance patient engagement by providing personalized health recommendations, reminders for medications, and real-time monitoring of vital signs. On the provider side, GenAI can save costs by streamlining administrative tasks and improving efficiency, early disease detection, and preventive care. Similarly, knowing the preliminary evidence of GenAI use across health care services is crucial for making informed decisions, ensuring regulatory compliance, building trust, guiding research initiatives, and addressing ethical considerations. This sets the stage for the responsible and effective integration of GenAI into the health care landscape.

The impact of GenAI in health care depends on various factors, including the specific application, quality of data used for training, ethical considerations, and regulatory framework in place. Continuous monitoring, evaluation, and responsible deployment are essential to maximize the positive impact and mitigate potential negative consequences. For instance, artificial intelligence (AI) assists pathologists in diagnosing diseases from pathology slides, leading to faster and more accurate diagnoses and improving patient outcomes [20]. Analysis of oncology literature, clinical trial data, and patient records can help oncologists identify personalized, evidence-based treatment options for patients with cancer, potentially improving treatment decisions [21]. AI has been applied to analyze medical images for conditions such as diabetic retinopathy, aiding in early detection and intervention [22]. AI analyzes clinical and molecular data to help physicians make more informed decisions about cancer treatment and steer them toward personalized and effective therapies [23].

Concerns about using GenAI remain because of algorithmic bias in predictive models that causes discrimination, unequal distribution of health care resources, and exacerbated health disparities [24]. Data privacy and the need for clear guidelines on AI in health care remain a gap, with reported misuse [25]. Misinterpretations or errors in algorithms can lead to incorrect diagnoses, specifically for image readings, which underscores the importance of human oversight in critical health care decisions [26]. Furthermore, implementing and maintaining AI systems can be costly, and overreliance on technology without sufficient human oversight may result in overlooking critical clinical nuances and potentially compromising patient care [27]. Therefore, it is essential to note that the impact of AI on health care is a dynamic and evolving field. Regular updates and scrutiny of the latest research and applications are necessary to understand the positive and negative aspects of GenAI in health care.

Using a literature scoping, review, and synthesis approach in this study, we evaluated the proportionate evidence of using GenAI to assist, guide, and automate clinical service functions. Technologies in general help standardize [28], provide flexibility [29], increase experience and satisfaction through relational benefits [30], induce higher switching costs [31], and enhance the overall quality [32] and value [33] of services. However, high technology may reduce personal touch, trust, and loyalty in service settings [34-38]. Complex technologies may introduce anxiety, confusion, and isolation [39] or disconnection, disruption, and passivity stressors [13] that can erode satisfaction, loyalty, and retention in service settings [28,40-42]. Given the mixed evidence in previous research on the role of technology in services [28,43,44], it is timely to assess to what extent GenAI may even have a role in shaping or disrupting health care services. Overall, the ground realities of the potential for emerging GenAI to benefit health care services rather than just being another knowledge and collation tool need to be assessed and reported to influence further research and practice activities.

Objectives

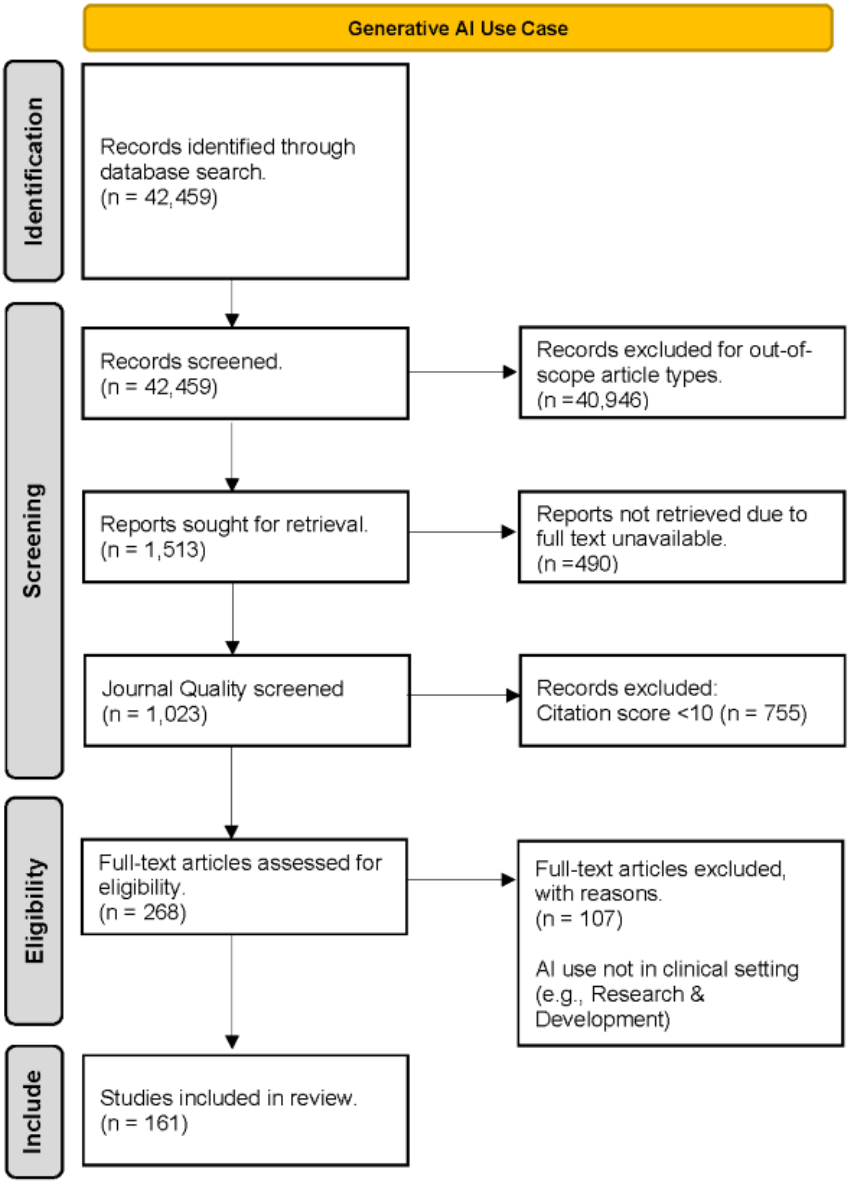
This study took a deep dive to review and synthesize preliminary evidence on how GenAI is used to assist, guide, and automate activities or functions during clinical service encounters in health care, with plausible indications for differential use. More evidence on the actual use is needed to assert that GenAI plays a considerable role in the digital transformation of health care. Therefore, this study aims to identify how GenAI is used in clinical settings by systematically reviewing preliminary evidence on its applications to assist, guide, and automate clinical activities or functions.

Methods

Article Search and Selection Strategy

This study aims to identify how physicians use GenAI in clinical settings, as evidenced in published studies. The design of this study adheres to the protocols outlined in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [45,46]. Figure 1 provides a flowchart of this study's article search and inclusion process.

Figure 1. Literature screening process for relevant articles on generative artificial intelligence (AI) tools and applications.



We focused our search exclusively on PubMed to ensure the credibility of this study's medical or clinical service settings. PubMed is part of the National Library of Medicine and a trusted national source of peer-reviewed publications on medical devices, software applications, and techniques used in the clinical setting. We performed keyword searches to retrieve relevant GenAI publications in PubMed that used "artificial intelligence" anywhere in the text of the article written in English. The sampling period of the publications was from January 1, 2020, to May 31, 2023. The search yielded 42,459 results in the first round of identification of articles for evaluation.

Within PubMed's classification system for articles, we used the "article type" that described the material presented in the article (eg, review, clinical trial, retracted publication, or letter). We

used this article type feature in the PubMed classification system to identify peer-reviewed articles and other relevant types of publications that are pertinent to our study. A total of 52.02% (22,086/42,459) of the returned articles did not have an article type assigned from the 75 article types in PubMed's classification system and were excluded from the study sample. We included clinical, multicenter, case report, news, evaluation, and validation studies. We excluded article types that were out of scope, such as uncategorized articles, government-funded studies, reviews, editorials, errata, opinion articles, nonscientific articles, retracted publications, and supplementary files. We also excluded preprint article types that were unlikely to have attracted attention. Errata or retracted publications (404/42,459, 0.95%), supplementary files (117/42,459, 0.28%), and 50 article types that had too few search returns (243/42,459, 0.57%) were also excluded.

The screening stage excluded review articles (6732/42,459, 15.86%) with an objective that was neither aligned with nor redundant to this study's goal. Opinion articles such as editorials, letters, and commentaries were excluded (2455/42,459, 5.78%). Articles whose funding came from the government or a government agency were not considered because of a conflict of interest for the researchers of the evaluated study (8936/42,459, 21.05%), and preprint articles (77/42,459, 0.2%) were excluded because of lack of availability to the public. We also considered the full text availability of the article, and 32.39% (490/1513) of the articles were excluded in the eligibility stage.

The resulting set of records included 1023 publications. To ensure the credibility of the publication source, we used CiteScore (Elsevier) [47] as a citation index to remove publication sources whose influence is limited. Any publication source whose citation index was unavailable or <10 was removed, resulting in 268 records.

In total, 2 raters, 1 author (DY) and 1 graduate assistant (BB), evaluated 161 articles. The 2 raters' agreement was 91.93%, and the expected agreement was 82.99%. The κ score was 0.5252 (SE 0.0544; *Z* score=9.66; probability>Z score=0.0000). The author and the graduate student performed manual coding by reading the paper's title, abstract, and introduction paragraph to gain a preliminary understanding of the study. After reading the abstract and introduction paragraph, each rater classified each article according to the definition of the 3 classes. For articles that were difficult to understand, the rater read the article further to gain a better understanding of the article. We defined clinical service settings to include the life cycle of physician encounters with patients for the diagnosis, prognosis, and management of health conditions. The research and development of drug discovery, for instance, was not considered. This process eliminated 107 records. The final data set of articles considered for this study was 161.

Ethical Considerations

The data collected for this study were obtained from publicly available sources. The study did not involve any interaction with users. Therefore, ethics approval was not required for this study.

Data Extraction and Categorization Process

We adopted a modified thematic synthesis approach for data analysis that involved coding the text, developing descriptive themes, and generating analytical themes [48]. Initially, each author coded each line of text extracted from the articles, assigning it to different dimensions. This line-by-line coding process facilitated identifying and capturing critical article information and concepts. Next, each author developed descriptive themes by grouping related codes and identifying common patterns or topics emerging from the coded data. These descriptive themes provided a broad overview of the various aspects of AI in the clinical service context. Building on the descriptive themes, each author generated analytical pieces to deepen the understanding and interpretation of the data. The analytical themes involved exploring relationships, connections, and implications within and across the articles, allowing for the extraction of meaningful insights.

Throughout the analysis process, all the authors engaged in extensive discussions to refine and finalize the results of the thematic synthesis. By collectively examining and interpreting the data, the research team ensured the robustness and reliability of the synthesized findings. Similar dimensions were then merged to generate the following 3 meaningful dimensions (assist, guide, and automate) and for relevance to the study objectives, as shown in Textbox 1. The researchers manually coded each article into several groups. They then tried to synthesize them into 1 of the 3 categories of *assist*, *guide*, and *automate* by looking at the title, abstract, and introduction (where applicable).

Textbox 1. Use of generative artificial intelligence tools and applications in clinical services in the reviewed articles (N=161).

Assist

- Improve diagnostic accuracy or reduce error by accessing knowledge during clinical services (141/161, 87.6%) [49-96]
- Activities:
  - Disease detection (19/161, 11.8%) [58,63,67,69,71,73,77,90,97-107]
  - Diagnosis (14/161, 8.7%) [100,108-120]
  - Screening (12/161, 7.5%) [65,86,87,93,121-128]
- Service areas:
  - Radiology (17/161, 10.6%) [49-63,65,66]
  - Cardiology (12/161, 7.5%) [67-72,74,76-79,129]
  - Gastrointestinal medicine (4/161, 2.5%) [81-84]
  - Diabetes (6/161, 3.7%) [86-91]
- Approaches and methods:
  - Deep learning (34/161, 21.1%) [49,59,60,62,63,65,68,71,79,89,100,107,108,111,115,123,125,130-145]
  - Machine learning (9/161, 5.6%) [53,55,83,91,110,146-149]
  - Image analysis (13/161, 8.1%) [68,88,104,110,111,114,116,119,133,135,138,150,151]

Guide

- Recommend treatment options, step-by-step instructions, or checklists to improve clinical services (13/161, 8.1%) [64,80,85,96,152-160]
- Personalized treatment plans (1/161, 0.6%) [64]
- Monitoring and managing (1/161, 0.6%) [96]

Automate

- Minimize or eliminate human provider involvement in clinical services or follow-ups (7/161, 4.3%) [94,95,161-165]

In addition to manual coding by human researchers, we used ChatGPT (version 3.5; OpenAI) for automatic coding. ChatGPT-3.5 was used for speed and cost. ChatGPT-4 is less accessible to users who do not have the funds to pay for its monthly subscription. ChatGPT-3.5 training used one-shot learning using the standard user interface with the "foundational" mode, and no fine-tuning was performed. Future studies may use focused data sets for fine-tuning to improve classification accuracy. However, our study demonstrates that classification accuracy is high and robust even without fine-tuning. This procedure was implemented to check for any subjective bias and demonstrate AI's potential use to complement the human coding process. The abstracts and introductions of these 161 articles were fed into ChatGPT using in-context or a few short learning processes that fine-tune a pair of domain-specific inputs and outputs to train, thereby enhancing the relevance and accuracy of ChatGPT's automated coding output [166,167].

For instance, a sample of input we used in the study was the abstract, which summarizes the article. The output is the categories identified by the experts. ChatGPT learns how to code a set of articles by repeating the pair of inputs and outputs. One-shot learning, which consists of a single pair of inputs and outputs in general, performs as well as >2 samples and zero-shot learning. The benefits of in-context learning (ICL) in ChatGPT include enhanced relevance, where the foundational model becomes better at generating content for domain-specific tasks without additional training of the full model; controlled output such as developing a single word matching the desired coding category or variable; and reduced biases inherent in manual coding. We used the definitions provided in Textbox 1 to train and restrict ChatGPT to choose only 1 of the 3 use-case categories. We further compared ChatGPT's classification with expert coding and found a high level of agreement between the 2, with a κ score of 0.94.

As mentioned previously, the manual coding process involved the raters coding and evaluating each article. After each rater coded the article, the results were compared and discussed to further refine the classification definition and derive consensus on the final assignment of the article classification. This "gold standard" classification was compared with automatic coding performed by ChatGPT (version 3.5). Automatic coding was performed by ChatGPT-3.5. Classification training was performed using one-shot ICL. ChatGPT learns how to classify articles by being fed a pair of articles and classification labels. For example, a user can feed a prompt or use control tokens to indicate an article abstract and the label associated with the article. In our context, 3 articles and labels were fed to the interface. After this initial prompt session of training on 3 classification labels, subsequent interactions of providing only

the article abstract with a prompt asking for a class label would return ChatGPT's prompt completion. Alternatively, training could involve >1 example of the article and its label, which would then be called *few-shot learning*. To summarize, 161 articles were coded by ChatGPT-3.5 based on a single instance of ICL.

Results

Findings From the Synthesis on the Use of GenAI to Assist in Different Aspects of Health Care Services

GenAI can improve clinical services in 3 ways. First, of the 161 articles, 141 (87.6%) reported using GenAI to assist services through knowledge access, collation, and filtering. The assistance of GenAI was used for disease detection (19/161, 11.8%) [58,63,67,69,71,73,77,90,97-107], diagnosis (14/161, 8.7%) [100,108-120], and screening processes (12/161, 7.5%) [65,86,87,93,121-127,168,169] in the areas of radiology (17/161, 10.6%) [49-63,65,66], cardiology (12/161, 7.5%) [67-72,74,76-79,129], gastrointestinal medicine (4/161, 2.5%) [81-84], and diabetes (6/161, 3.7%) [86-91]. Thus, although the use of GenAI has percolated across almost all disease-relevant and main service–relevant areas in health care, it is mainly for assisting through knowledge access, collation, and filtering.

The use of GenAI in disease diagnosis has long-term implications. For instance, identifying "referrable" diabetic retinopathy using routinely collected data would help in population health planning and prevention [86-90]; however, rigorous testing and validation of the applications are critical before clinical implementation [94]. Similarly, using GenAI in remote care helps improve glycemia and weight loss [95], yet challenges related to variable patient uptake and increased clinician participation necessitated by shared decision-making must be considered [96]. In radiology services, prediction models using deep learning and machine learning methods for predictive accuracy and as diagnostic aids have shown potential, and natural language processing has been used to improve readability by generating captions; however, studies report using high-quality images, highlighting the need for a future standardized pipeline for data collection and imaging detection.

In cardiology, AI analysis allows for early detection, population-level screening, and automated evaluation. It expands the reach of electrocardiography to clinical settings in which immediate interrogation of anatomy and cardiac function is needed and to locations with limited resources [67-69,71,73-75,95]. Nevertheless, there is evidence suggesting that integrating AI with patient data, including social determinants of health, enables disease prediction and early disease identification, which could lead to more precise and timely diagnoses, improving patient outcomes.

GenAI aids in diagnostic accuracy, although its focus on higher value creation in health care is limited. The articles in this review reported that they used deep learning (34/161, 21.1%) [49,59,60,62,63,65,68,71,79,89,100,107,108,111,115,123,125,130-145], machine learning (9/161, 5.6%) [53,55,83,91,110,146-149], and image analysis approaches of GenAI during the assistance process (13/161, 8.1%) [68,88,104,110,111,114,116,119,133,135,138,150,151]. Knowledge access using GenAI has the potential to enable more options and flexibility in serving patients.

Evidence of GenAI Use for Guiding or Automation Services

Only 8.1% (13/161) of the studies provided insights into how GenAI is used to guide some services by seeking recommended treatment options, step-by-step instructions, or checklists to improve clinical services [64,80,85,96,152-160]. Of the 161 studies, 1 (0.6%) study sought personalized treatment plans and discussed monitored and managed service processes using GenAI [96]. Although this use category is nascent, GenAI can help provide speed efficiency and customized solutions in health services as in other contexts [37,127,170].

Finally, only 4.3% (7/161) of the articles indicated the use of GenAI to automate any service functions that could minimize or eliminate human provider involvement. When used appropriately, automation provides a predictable, reliable, and faster experience everywhere, every time for all customers, which will be a standardized way to provide several health care services [94,95,161-165].

The use of GenAI in some instances of service automation and guidance may be in its infancy but is encouraging. Providers are trying to explore unique ways to use AI, which requires a set of steps such as understanding the current workflow and the changes needed or aspirational workflows and aligning or designing GenAI to help in the workflow. This is similar to modifying restaurant food delivery options to suit drive-in rather than sit-in options. The providers need some work to fully automate, streamline, or re-engineer the service functions using GenAI in the future.

Summary of Findings

To summarize our findings, in this study, we conducted a systematic scoping review of the literature on how GenAI is used in clinical settings by synthesizing evidence on its application to assist, guide, and automate clinical activities and functions. Of the 161 articles, 141 (87.6%) reported using GenAI to assist services through knowledge access, collation, and filtering. The assistance of GenAI was used for disease detection (19/161, 11.8%), diagnosis (14/161, 8.7%), and screening processes (12/161, 7.5%) in the areas of radiology (17/161, 10.6%), cardiology (12/161, 7.5%), gastrointestinal medicine (4/161, 2.5%), and diabetes (6/161, 3.7%). Thus, we conclude that GenAI mainly informs rather than assisting and automating service functions. Presumably, the potential in clinical service is there, but it has yet to be actualized for GenAI.

Robustness Check Using Additional Database Search

To ensure the comprehensiveness and robustness of our findings, we expanded the search to Web of Science using similar keywords and strategies (suggested by the review team). We used the same keyword, "artificial intelligence," in all text fields over the sampling period between January 1, 2020, and November 27, 2023. Our search was restricted to peer-reviewed academic journal articles written in English. We used the Web of Science–provided "Highly Cited Papers" criterion as a

filtering mechanism to follow influential papers. Given the nonclinical context of the journals in the database, we believe that filtering based on the article's importance is reasonable. Initial search results included 1958 articles from the Web of Science Core Collection. The preliminary analysis of the annual breakdown comprised 414 articles in 2023, a total of 651 articles in 2022, a total of 519 articles in 2021, and a total of 374 articles in 2020. The search results were further reduced by removing PubMed articles for redundancy, resulting in 1221 articles.

Next, Web of Science journals include medical, nonmedical, and other clinical journals. Thus, we used simple keywords for filtering nonmedical and clinical contexts. We used the keywords "medical" and "health" mentioned in the abstract, which led to 133 articles. Finally, we read the abstracts and titles to exclude survey or meta-review and nonclinical studies. This process further narrowed down the selection to 51 relevant articles. Using ChatGPT-3.5 on November 27, 2023, we applied one-shot learning by providing 3 class definitions. We asked ChatGPT-3.5 to classify the article's abstract, with 63% (32/51) in the *assist* category, 29% (15/51) in the *guide* category, and 8% (4/51) in the *automated* category. Diagnostic assistance articles dominated, similar to the results from PubMed. However, the other categories—prescriptive guidance and clinical service recommendations—were slightly higher. This difference is explained by the nonmedical and clinical nature of the journals included in the database. The "applied" nature of the journals is more likely to explore prescriptive guidance and clinical service recommendation use cases.

Discussion

Principal Findings

This study asked RQs about how GenAI is used, with evidence, to shape health care services. It showed that 11.8% (19/161) of the studies were on automation and guidance, whereas 87.6% (141/161) reflected the assistance role of GenAI. These findings are essential to discuss and distinguish between the optimism and actual use of GenAI in health care.

Study Implications

The aspiration that GenAI has the potential to change health care significantly needs a careful revisit. Health care organizations need to assess the actual ground use for GenAI and prepare for and understand the exciting possibilities with a cautious approach rather than overly high expectations. Concerns related to the cost, privacy, misuse, and regulatory aspects of implementing and using GenAI [24-26] will become more pronounced, particularly when there is a perceived overreliance without clear promising results or actual practical use [26].

The literature synthesis in this study suggests that GenAI is mainly used for screening and diagnostic purposes using knowledge access; diagnostic processes such as predicted disease outcomes, survival, or disease classification; and improvement of the accuracy of diagnosis. This solves the problem of knowledge being available and accessible in time in a well-articulated manner to provide or render the services. This could help health care providers make more accurate and timely diagnoses, leading to earlier treatment and better patient outcomes. Such knowledge distillation helps improve diagnostic accuracy through GenAI, which can provide enough knowledge to physicians during service encounters; however, this is not hugely oriented toward higher value creation in health care.

The research synthesis also suggests that there has been some use of GenAI during different steps and aspects of guiding the service delivery processes. Still, such use could be more encouraging and significant across the board. Plausibly, GenAI can analyze large amounts of disparate data from patients to suggest personalized medicine—which may help inform treatment plans for individuals. Service delivery needs some guidance or step-by-step help to be efficient and meet the duration or time requirements to render the clinical service on time, which GenAI may solve. However, we have not yet found strong evidence for such use by any health system.

Currently, the automation of service functions using GenAI has only seen minimal instances and is yet to see widespread implementation. Automation helps offset some manual activities. However, automation may help in service functions' cost, efficiency, and flexibility while maintaining some standards across similar services.

Similarly, although we did not consider this area in the synthesis as it was out of the scope of services, GenAI can also be used in drug development and clinical trial pathways—a value proposition yet to be seen in practice. However, we do not undermine that many laboratories and pharmaceutical companies have used machine learning and AI tools and techniques in drug development and clinical trials. However, reported commercial GenAI use has not come to the limelight.

Some other plausible uses of GenAI in health care include managing supply chain data, managing medical equipment assets, maintaining gadgets and equipment, and building a robust intelligent information infrastructure to support several other activities. For example, active efforts are being undertaken to incorporate GenAI, especially in administrative use cases such as the In Basket patient messaging applications. However, assessing the clinical accuracy of such tools remains a concern.

In addition, we must incorporate user-centered design and sociotechnical frameworks into designing and building GenAI for health care use cases, for instance, to explore how GenAI can prevent a common pitfall of developing models opportunistically—based on data availability or end-point labels, adopting a user-centered design framework is vital for GenAI tools [171]. Similarly, scientific or research-oriented use of GenAI for knowledge search, articulation, or synthesis is helpful [172]. However, how far that will translate to the transformative clinical health care delivery processes while creating higher-order organizational capabilities to create value remains a concern [173].

Limitations of the Study and Scope for Future Research

Several limitations and constraints affect the interpretation and generalizability of the findings of this study. Some of these limitations indicate the need for future research in relevant areas that we discuss further. First, the study's findings were

constrained by the availability of relevant and high-quality publications and the exclusion of preprints and unpublished data to limit the specifically designed scope of the study on using GenAI in health care clinical services, which influences the comprehensiveness and accuracy of the review. There also might be a tendency for studies with positive or significant results to be published, leading to a potential publication bias. In addition, harmful or neutral findings may not be adequately represented in the review, influencing the overall assessment of GenAI's effectiveness in health care. Research should focus on patient-centered outcomes, including patient satisfaction and engagement and the impact of GenAI on the patient-provider relationship. Understanding the patient perspective is crucial for successfully integrating AI technologies into health care.

Second, the field of GenAI in health care is rapidly advancing, and new technologies and applications are continuously emerging. The findings of this study might not capture the most recent developments, and the ’conclusions of this study may become outdated quickly, specifically when some technologies have the potential to be adopted beyond institutional mechanisms, such as using GenAI mobile apps to scan images for retinopathy. Furthermore, an in-depth analysis of specific GenAI applications may open newer directions, and future research should focus on specific GenAI applications to provide detailed insights into their effectiveness and limitations. This could include applications such as diagnostic tools, treatment planning algorithms, and predictive analytics. Such heterogeneity of GenAI in health care encompasses a wide range of applications, and investigating these could make it challenging to draw overarching conclusions about GenAI’s impact on clinical services.

Third, this review may not comprehensively address ethical considerations and potential biases in the use of GenAI in health care. Ethical issues related to data privacy, algorithmic bias, and the responsible deployment of AI technologies may require more in-depth exploration. Future research should systematically explore the ethical considerations associated with GenAI use in health care. This includes issues related to data privacy, consent, transparency, and the ethical deployment of AI algorithms in clinical settings. Finally, more data, papers, articles, and longitudinal developments on some applications may enrich this study and enhance its current limited generalizability. Longitudinal studies are needed to track the impact of GenAI in health care over an extended period. This will help researchers understand the sustained effects, identify potential challenges that may arise over time, and assess the scalability and adaptability of these technologies.

Future studies could undertake comparative effectiveness research to assess how GenAI compares with traditional approaches in health care. Understanding the relative advantages and disadvantages will contribute to evidence-based decision-making. In addition, it is not clear what and how to measure the GenAI applications’ effectiveness in clinical services, leading to a call for standardized study metrics that can incorporate outcome measures and evaluation frameworks. Future research should investigate how the integration of GenAI into clinical health care services affects the workflow of health care providers. This includes understanding the time savings, challenges, and potential improvements in decision-making processes. By addressing these areas, future research can contribute to a more comprehensive understanding of the role, challenges, and potential benefits of GenAI in clinical health care services.

Actionable Policy and Practice Recommendations

The proliferation of technology often outpaces the development of appropriate regulatory and policy frameworks that are necessary for guiding proper dissemination. Our call is that, given that GenAI is emerging, policy agencies and health care organizations play a role in proactively guiding the use of GenAI in health care organizations.

What are some actionable steps for stakeholders, including health care organizations and policy makers, to navigate the integration of GenAI in health care? For health care organizations, the steps may include conducting a technology assessment vis-à-vis goals to achieve outcomes from GenAI. Evaluating the existing infrastructure and technological capabilities within the health care organization to determine readiness for GenAI integration is a first step. This will provide an understanding of the current state of technology and ensure that the necessary upgrades or modifications can be implemented to support GenAI applications, thus garnering the benefits of GenAI.

The second step is to invest in staff training and education through the development of training programs to enhance the skills of health care professionals in understanding and using GenAI technologies. Well-trained staff is essential for the effective and ethical implementation of GenAI, fostering a culture of continuous learning and adaptability. Third, health care organizations need to develop and communicate clear protocols and guidelines for the use of GenAI in different health care services, outlining ethical considerations, data privacy measures, and accountability standards. Transparent protocols help ensure the responsible and standardized use of GenAI, fostering trust among health care professionals and patients.

Fourth, health care organizations need to engage in research on GenAI through collaboration with research institutions and industry partners to participate actively in studies evaluating the effectiveness and impact of GenAI applications in specific health care domains. Involvement in research contributes to the evidence base, informs best practices, and positions the organization as a leader in health care innovation. Finally, as mentioned previously, implementing the gradual integration of GenAI rather than jumping into irrational decisions is a caution. All health systems need to gradually plan and introduce GenAI technologies, starting with pilot programs in specific departments or use cases. Gradual integration allows for careful monitoring of performance, identification of potential challenges, and iterative improvement before broader implementation.

For policy makers, much work must be done at the regulatory framework level to realize GenAI better. Policy makers must establish clear and adaptive regulatory frameworks that address the unique challenges GenAI poses in health care, ensuring patient safety, data privacy, and ethical use. There is a concern

that bias in GenAI algorithms could lead to discrimination in care delivery across patients, and the role of policy guidelines in this aspect to train and use GenAI appropriately is critical. Policy frameworks must be developed to ensure less risk, safe and ethical use, and responsible effectiveness of GenAI. Policy and industry partnerships among experts to determine relevant frameworks are vital to guide the future of GenAI to help transform health care. Robust regulations will provide a foundation for the responsible and standardized integration of GenAI technologies. An underlying challenge of GenAI is integrating it across different legacy IT systems, which involves developing and adopting interoperability standards to ensure seamless communication and data exchange between different GenAI applications and existing health care systems. Interoperability enhances efficiency, reduces redundancy, and facilitates the integration of diverse GenAI solutions. In this process, creating incentives for responsible innovation for ethical considerations and the continuous improvement of GenAI applications will drive a culture of responsibility and quality improvement, aligning technological advancements with societal needs.

Policy-level efforts also need to be oriented to allocate resources to enhance health care infrastructure, including robust connectivity and data storage capabilities, to support the data-intensive nature of GenAI applications. Adequate infrastructure is crucial for the reliable and secure functioning of GenAI in health care. Many of these enhancements may require collaboration between public health care systems, private organizations, and academia to leverage collective expertise and resources for GenAI research, development, and implementation. Finally, policies that address potential biases in GenAI applications and ensure equitable access to these technologies across diverse populations are necessary to help with proactive measures to prevent the exacerbation of existing health care disparities through the adoption of GenAI.

Conclusions

GenAI is both a tool and a complex technology. Complexity is the basis for GenAI, and thus, the use of GenAI in health care creates a set of unparalleled challenges. GenAI is costly to implement and integrate across all aspects of a health system [174]. In envisioning the future of GenAI in health care, we glimpse a transformative landscape in which technology and compassion converge for the betterment of humanity. As we stand at the intersection of innovation and responsibility, the prospect of GenAI holds immense promise in revolutionizing health care, shaping a future in which personalized, efficient, and equitable clinical services are not just aspirations but tangible realities. Our vision embraces a symbiotic relationship between technology and human touch, recognizing that the power of GenAI lies not only in its computational prowess but also in its potential to amplify the capabilities of health care professionals. Picture a world in which diagnostic accuracy is elevated, treatment plans are truly personalized, and each patient's journey is marked by precision and empathy.

Crucially, this vision hinges on responsible adoption. We envisage a future in which regulatory frameworks ensure the ethical use of GenAI, safeguard patient privacy, and uphold the principles of equity. It is a future in which interdisciplinary collaboration flourishes, bridging the expertise of health care providers, policy makers, technologists, and ethicists to navigate the complexities of this evolving landscape.

In the future, the impact of AI on human lives will be profound. Patients experience a health care system that not only heals but also understands, a system in which the integration of GenAI contributes to quicker diagnoses, more effective treatments, and improved outcomes. The human experience is at the forefront—GenAI becomes a tool for health care professionals to better connect with patients and spend more time understanding their unique needs, fears, and hopes. As we embark on this journey, it is crucial to remember that the heart of health care lies in the compassion, empathy, and wisdom of its human stewards. GenAI catalyzes empowerment, freeing health care professionals from mundane tasks to engage in meaningful interactions. It fosters a health care culture in which technology serves humanity, and the collective mission is to enhance the quality of care and life.

In embracing this vision, we are not just architects of technological progress but also custodians of a future in which GenAI and human touch coalesce to redefine health care possibilities. Let our strides be guided by a commitment to responsible innovation, a dedication to inclusivity, and an unwavering focus on the well-being of those we serve. The future of GenAI in health care is not just a scientific evolution, but it is a narrative of healing; compassion; and a shared commitment to a healthier, more humane world. However, without enough evidence, we are skeptical about the current euphoria regarding GenAI in health care.

This systematic narrative review of the preliminary evidence of using GenAI in health care clinical services provides valuable insights into the evolving landscape of AI applications in health care. The existing literature synthesis reveals promising advancements and critical considerations for integrating GenAI into clinical settings. The positive evidence underscores the potential of GenAI to revolutionize health care by offering personalized treatment plans, enhancing diagnostic accuracy, and contributing to the development of innovative therapeutic solutions. The applications of GenAI in areas such as pathology assistance, oncology decision support, and medical imaging interpretation showcase its capacity to augment health care professionals' capabilities and improve patient outcomes.

However, this review also highlights several limitations and challenges that warrant careful consideration. Issues such as the quality of available data, the rapid pace of technological evolution, and the potential for algorithmic bias highlight the complexities associated with adopting GenAI in health care. Ethical concerns, data privacy considerations, and the need for transparent guidelines underscore the importance of a thoughtful and measured approach to integration.

As we navigate the preliminary evidence, it becomes evident that a collaborative effort is required among health care organizations, policy makers, researchers, and technology developers. Establishing clear regulatory frameworks, fostering interdisciplinary collaboration, and prioritizing ethical considerations are crucial steps in ensuring the responsible

deployment of GenAI. Addressing the identified limitations through targeted research initiatives, ongoing evaluation, and continuous improvement will be essential for maximizing the benefits of GenAI while mitigating potential risks.

Moving forward, it is imperative to recognize that integrating GenAI into health care is dynamic and evolving. Future research should focus on refining our understanding of the long-term impact, patient-centered outcomes, and scalability of GenAI applications. By collectively addressing the challenges outlined in this review, stakeholders can contribute to a health care landscape in which GenAI is a powerful ally in delivering personalized, efficient, and equitable clinical services.

Acknowledgments

JK expressly acknowledges the Health Administration Research Consortium at the Business School of the University of Colorado Denver for providing a platform for the stimulating discussion and insights on this topic. The authors acknowledge Mr Bhanukesh Balabhadrapatruni, graduate student fellow at the Health Administration Research Consortium, for assisting with data categorization and citation listing. AM thanks the participants from the Society of Physician Entrepreneurs for their input about artificial intelligence in health care. VP thanks Dr Ron Li at Stanford Medicine for insights and a stimulating discussion on this topic. We used the generative AI tool ChatGPT (version 3.5; OpenAI) for automatic coding and checking the accuracy of the human coding process used to categorize the articles reviewed and synthesized in this study [166,167].

Conflicts of Interest

JK is an associate editor of the Journal of Medical Internet Research.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 31 KB - medinform_v12i1e52073_app1.docx]

Multimedia Appendix 2
Conversations with ChatGPT used in the Study.
[DOCX File , 85 KB - medinform_v12i1e52073_app2.docx]

References

1. Pasick A. Artificial intelligence glossary: neural networks and other terms explained. The New York Times. 2023. URL: https://www.nytimes.com/article/ai-artificial-intelligence-glossary.html [accessed 2024-01-29]
2. Roose K. A coming-out party for generative A.I., Silicon Valley's new craze. The New York Times. 2022 Oct. URL: https://www.nytimes.com/2022/10/21/technology/generative-ai.html [accessed 2024-01-29]
3. Karpathy A, Abeel P, Brockman G, Chen P, Cheung V, Duan Y. Generative models. Open AI. 2016. URL: https://openai.com/research/generative-models [accessed 2024-01-31]
4. Metz C. OpenAI plans to up the ante in tech's A.I. race. The New York Times. 2023. URL: https://www.nytimes.com/2023/03/14/technology/openai-gpt4-chatgpt.html#:~:text=But%20in%20the%20long%20term,Brockman%20said [accessed 2024-01-29]
5. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. LaMDA: language models for dialog applications. arXiv Preprint posted online January 20, 2022 2020 [FREE Full text] [doi: 10.48550/arXiv.2201.08239]
6. Don't fear an ai-induced jobs apocalypse just yet: the west suffers from too little automation, not too much. The Economist. 2023. URL: https://www.economist.com/business/2023/03/06/dont-fear-an-ai-induced-jobs-apocalypse-just-yet [accessed 2024-01-29]
7. Harreis H, Koullias T, Roberts R, Te K. Generative AI: unlocking the future of fashion. McKinsey & Company. URL: https://www.mckinsey.com/industries/retail/our-insights/generative-ai-unlocking-the-future-of-fashion [accessed 2024-08-10]
8. Eapen TT, Venkataswamy L, Finkenstadt DJ, Folk J. How generative AI can augment human creativity. Harvard Business Review. 2023. URL: https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity [accessed 2024-01-29]
9. The race of the AI labs heats up: ChatGPT is not the only game in town. The Economist. 2023. URL: https://www.economist.com/business/2023/01/30/the-race-of-the-ai-labs-heats-up [accessed 2023-01-30]
10. Google Cloud brings generative AI to developers, businesses, and governments. Google. 2023. URL: https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-for-businesses-and-governments [accessed 2024-01-29]
11. Zheng D, He X, Jing J. Overview of artificial intelligence in breast cancer medical imaging. J Clin Med 2023 Jan 04;12(2):419 [FREE Full text] [doi: 10.3390/jcm12020419] [Medline: 36675348]
12. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023 Jun 27;33(6):1790-1796 [FREE Full text] [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]

13. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023 Apr;185:109729. [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
14. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. Preprint posted online February 7, 2023 2023 [FREE Full text] [doi: 10.1101/2023.02.02.23285399] [Medline: 36798292]
15. Cirillo D, Núñez-Carpintero I, Valencia A. Artificial intelligence in cancer research: learning at different levels of data granularity. Mol Oncol 2021 Apr;15(4):817-829 [FREE Full text] [doi: 10.1002/1878-0261.12920] [Medline: 33533192]
16. Pedersen M, Verspoor K, Jenkinson M, Law M, Abbott DF, Jackson GD. Artificial intelligence for clinical decision support in neurology. Brain Commun 2020;2(2):fcaa096 [FREE Full text] [doi: 10.1093/braincomms/fcaa096] [Medline: 33134913]
17. Brynjolfsson E, McAfee A. Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. New York, NY: WW Norton & Company; 2014.
18. Raisch S, Krakowski S. Artificial intelligence and management: the automation–augmentation paradox. Acad Manage Rev 2021 Jan 14;46(1):192-210 [FREE Full text] [doi: 10.5465/amr.2018.0072]
19. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
20. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
21. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol 2022 Jan;35(1):23-32 [FREE Full text] [doi: 10.1038/s41379-021-00919-2] [Medline: 34611303]
22. Yu SH, Kim MS, Chung HS, Hwang EC, Jung SI, Kang TW, et al. Early experience with Watson for Oncology: a clinical decision-support system for prostate cancer treatment recommendations. World J Urol 2021 Feb;39(2):407-413 [FREE Full text] [doi: 10.1007/s00345-020-03214-y] [Medline: 32335733]
23. Wang Z, Keane PA, Chiang M, Cheung CY, Wong TY, Ting DS. Artificial intelligence and deep learning in ophthalmology. In: Lidströmer N, Ashrafian H, editors. Artificial Intelligence in Medicine. Cham, Switzerland: Springer; 2022:1519-1552.
24. Osinski B, BenTaieb A, Ho I, Jones RD, Joshi RP, Westley A, et al. Artificial intelligence-augmented histopathologic review using image analysis to optimize DNA yield from formalin-fixed paraffin-embedded slides. Mod Pathol 2022 Dec;35(12):1791-1803 [FREE Full text] [doi: 10.1038/s41379-022-01161-0] [Medline: 36198869]
25. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
26. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. Med Law Rev 2023 Nov 27;31(4):501-520 [FREE Full text] [doi: 10.1093/medlaw/fwad013] [Medline: 37218368]
27. Degnan AJ, Ghobadi EH, Hardy P, Krupinski E, Scali EP, Stratchko L, et al. Perceptual and interpretive error in diagnostic radiology-causes and potential solutions. Acad Radiol 2019 Jun;26(6):833-845 [FREE Full text] [doi: 10.1016/j.acra.2018.11.006] [Medline: 30559033]
28. Khanna NN, Maindarkar MA, Viswanathan V, Fernandes JF, Paul S, Bhagawati M, et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. Healthcare (Basel) 2022 Dec 09;10(12):2493 [FREE Full text] [doi: 10.3390/healthcare10122493] [Medline: 36554017]
29. Curran JM, Meuter ML. Self-service technology adoption: comparing three technologies. J Serv Mark 2005;19(2):103-113. [doi: 10.1108/08876040510591411]
30. Choudhury V, Karahanna E. The relative advantage of electronic channels: a multidimensional view. MIS Q 2008;32(1):179. [doi: 10.2307/25148833]
31. Marzocchi GL, Zammit A. Self-scanning technologies in retail: determinants of adoption. Serv Ind J 2006 Sep;26(6):651-669 [FREE Full text] [doi: 10.1080/02642060600850790]
32. Campbell D, Frei F. Cost structure, customer profitability, and retention implications of self-service distribution channels: evidence from customer behavior in an online banking channel. Manag Sci 2010 Jan;56(1):4-24 [FREE Full text] [doi: 10.1287/mnsc.1090.1066]
33. Chen PY, Hitt LM. Measuring switching costs and the determinants of customer retention in internet-enabled businesses: a study of the online brokerage industry. Inf Syst Res 2002 Sep;13(3):255-274. [doi: 10.1287/isre.13.3.255.78]
34. Mols NP. The behavioral consequences of PC banking. Int J Bank Mark 1998;16(5):195-201 [FREE Full text] [doi: 10.1108/02652329810228190]
35. Apte UM, Vepsäläinen AP. High tech or high touch? Efficient channel strategies for delivering financial services. J Strateg Inf Syst 1993 Mar;2(1):39-54. [doi: 10.1016/0963-8687(93)90021-2]
36. Giebelhausen M, Robinson SG, Sirianni NJ, Brady MK. Touch versus tech: when technology functions as a barrier or a benefit to service encounters. J Mark 2014 Jul 01;78(4):113-124 [FREE Full text] [doi: 10.1509/jm.13.0056]
37. Selnes F, Hansen H. The potential hazard of self-service in developing customer loyalty. J Serv Res 2016 Jun 29;4(2):79-90 [FREE Full text] [doi: 10.1177/109467050142001]

38. Walker RH, Johnson LW. Why consumers use and do not use technology-enabled services. J Serv Mark 2006;20(2):125-135. [doi: 10.1108/08876040610657057]
39. Xue M, Hitt LM, Harker PT. Customer efficiency, channel usage, and firm performance in retail banking. Manuf Serv Oper Manag 2007 Oct;9(4):535-558 [FREE Full text] [doi: 10.1287/msom.1060.0135]
40. Johnson DS, Bardhi F, Dunn DT. Understanding how technology paradoxes affect customer satisfaction with self‐service technology: the role of performance ambiguity and trust in technology. Psychol Mark 2008 Apr 08;25(5):416-443 [FREE Full text] [doi: 10.1002/mar.20218]
41. Scherer A, Wünderlich NV, von Wangenheim F. The value of self-service: long-term effects of technology-based self-service usage on customer retention. MIS Q 2015 Jan 1;39(1):177-200. [doi: 10.25300/misq/2015/39.1.08]
42. Li S, Sun B, Wilcox RT. Cross-selling sequentially ordered products: an application to consumer banking services. J Mark Res 2018 Oct 10;42(2):233-239 [FREE Full text] [doi: 10.1509/jmkr.42.2.233.62288]
43. Bitner MJ, Brown SW, Meuter ML. Technology infusion in service encounters. J Acad Mark Sci 2000 Jan 01;28(1):138-149 [FREE Full text] [doi: 10.1177/0092070300281013]
44. Meuter ML, Ostrom AL, Roundtree RI, Bitner MJ. Self-service technologies: understanding customer satisfaction with technology-based service encounters. Journal of Marketing 2018 Oct 10;64(3):50-64. [doi: 10.1509/jmkg.64.3.50.18024]
45. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021:372:n160 [FREE Full text] [doi: 10.1136/bmj.n160]
46. Bitner M. Service and technology: opportunities and paradoxes. Manag Serv Qual 2001;11(6):375. [doi: 10.1108/09604520110410584]
47. Page MJ, McKenzie JA, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906 [FREE Full text] [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
48. Baker DW. Introducing CiteScore, our journal's preferred citation index: moving beyond the impact factor. Jt Comm J Qual Patient Saf 2020 Jun;46(6):309-310 [FREE Full text] [doi: 10.1016/j.jcjq.2020.03.005] [Medline: 32402761]
49. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv 2020 Jun 12;53(3):1-34 [FREE Full text] [doi: 10.1145/3386252]
50. Dong D, Fang MJ, Tang L, Shan XH, Gao JB, Giganti F, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. Ann Oncol 2020 Jul;31(7):912-920 [FREE Full text] [doi: 10.1016/j.annonc.2020.04.003] [Medline: 32304748]
51. Ruscitti P, Bruno F, Berardicurti O, Acanfora C, Pavlych V, Palumbo P, et al. Lung involvement in macrophage activation syndrome and severe COVID-19: results from a cross-sectional study to assess clinical, laboratory and artificial intelligence-radiological differences. Ann Rheum Dis 2020 Sep;79(9):1152-1155 [FREE Full text] [doi: 10.1136/annrheumdis-2020-218048] [Medline: 32719039]
52. Shao L, Yan Y, Liu Z, Ye X, Xia H, Zhu X, et al. Radiologist-like artificial intelligence for grade group prediction of radical prostatectomy for reducing upgrading and downgrading from biopsy. Theranostics 2020;10(22):10200-10212 [FREE Full text] [doi: 10.7150/thno.48706] [Medline: 32929343]
53. Liu X, Zhang D, Liu Z, Li Z, Xie P, Sun K, et al. Deep learning radiomics-based prediction of distant metastasis in patients with locally advanced rectal cancer after neoadjuvant chemoradiotherapy: a multicentre study. EBioMedicine 2021 Jul;69:103442 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103442] [Medline: 34157487]
54. Gitto S, Cuocolo R, Annovazzi A, Anelli V, Acquasanta M, Cincotta A, et al. CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas. EBioMedicine 2021 Jun;68:103407 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103407] [Medline: 34051442]
55. Zhang J, Yao K, Liu P, Liu Z, Han T, Zhao Z, et al. A radiomics model for preoperative prediction of brain invasion in meningioma non-invasively based on MRI: a multicentre study. EBioMedicine 2020 Aug;58:102933 [FREE Full text] [doi: 10.1016/j.ebiom.2020.102933] [Medline: 32739863]
56. Hindocha S, Charlton TG, Linton-Reid K, Hunter B, Chan C, Ahmed M, et al. A comparison of machine learning methods for predicting recurrence and death after curative-intent radiotherapy for non-small cell lung cancer: development and validation of multivariable clinical prediction models. EBioMedicine 2022 Mar;77:103911 [FREE Full text] [doi: 10.1016/j.ebiom.2022.103911] [Medline: 35248997]
57. Feng L, Liu Z, Li C, Li Z, Lou X, Shao L, et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. Lancet Digit Health 2022 Jan;4(1):e8-17 [FREE Full text] [doi: 10.1016/S2589-7500(21)00215-6] [Medline: 34952679]
58. Seah JC, Tang CH, Buchlak QD, Holt XG, Wardman JB, Aimoldin A, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health 2021 Aug;3(8):e496-e506 [FREE Full text] [doi: 10.1016/S2589-7500(21)00106-0] [Medline: 34219054]
59. Fontanellaz M, Ebner L, Huber A, Peters A, Löbelenz L, Hourscht C, et al. A deep-learning diagnostic support system for the detection of COVID-19 using chest radiographs: a multireader validation study. Invest Radiol 2021 Jun 01;56(6):348-356 [FREE Full text] [doi: 10.1097/RLI.0000000000000748] [Medline: 33259441]

60. Gu J, Tong T, Xu D, Cheng F, Fang C, He C, et al. Deep learning radiomics of ultrasonography for comprehensively predicting tumor and axillary lymph node status after neoadjuvant chemotherapy in breast cancer patients: A multicenter study. Cancer 2023 Feb 01;129(3):356-366. [doi: 10.1002/cncr.34540] [Medline: 36401611]
61. Jiang M, Li CL, Luo XM, Chuan ZR, Lv WZ, Li X, et al. Ultrasound-based deep learning radiomics in the assessment of pathological complete response to neoadjuvant chemotherapy in locally advanced breast cancer. Eur J Cancer 2021 Apr;147:95-105. [doi: 10.1016/j.ejca.2021.01.028] [Medline: 33639324]
62. Zhang Y, Liu M, Zhang L, Wang L, Zhao K, Hu S, et al. Comparison of chest radiograph captions based on natural language processing vs completed by radiologists. JAMA Netw Open 2023 Feb 01;6(2):e2255113 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.55113] [Medline: 36753278]
63. Yoon AP, Lee YL, Kane RL, Kuo C, Lin C, Chung KC. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open 2021 May 03;4(5):e216096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.6096] [Medline: 33956133]
64. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. JAMA Netw Open 2020 Sep 01;3(9):e2017135 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.17135] [Medline: 32970157]
65. Zhong L, Dong D, Fang X, Zhang F, Zhang N, Zhang L, et al. A deep learning-based radiomic nomogram for prognosis and treatment decision in advanced nasopharyngeal carcinoma: a multicentre study. EBioMedicine 2021 Aug;70:103522 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103522] [Medline: 34391094]
66. Lu MT, Raghu VK, Mayrhofer T, Aerts HJ, Hoffmann U. Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. Ann Intern Med 2020 Nov 03;173(9):704-713 [FREE Full text] [doi: 10.7326/M20-1868] [Medline: 32866413]
67. Ahn JS, Ebrahimian S, McDermott S, Lee S, Naccarato L, Di Capua JF, et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. JAMA Netw Open 2022 Aug 01;5(8):e2229289 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.29289] [Medline: 36044215]
68. Upton R, Mumith A, Beqiri A, Parker A, Hawkes W, Gao S, et al. Automated echocardiographic detection of severe coronary artery disease using artificial intelligence. JACC Cardiovasc Imaging 2022 May;15(5):715-727 [FREE Full text] [doi: 10.1016/j.jcmg.2021.10.013] [Medline: 34922865]
69. Kusunose K, Abe T, Haga A, Fukuda D, Yamada H, Harada M, et al. A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. JACC Cardiovasc Imaging 2020 Feb;13(2 Pt 1):374-381 [FREE Full text] [doi: 10.1016/j.jcmg.2019.02.024] [Medline: 31103590]
70. Ko WY, Siontis KC, Attia ZI, Carter RE, Kapa S, Ommen SR, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. J Am Coll Cardiol 2020 Feb 25;75(7):722-733 [FREE Full text] [doi: 10.1016/j.jacc.2019.12.030] [Medline: 32081280]
71. Vaid A, Johnson KW, Badgeley MA, Somani SS, Bicak M, Landi I, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. JACC Cardiovasc Imaging 2022 Mar;15(3):395-410 [FREE Full text] [doi: 10.1016/j.jcmg.2021.08.004] [Medline: 34656465]
72. Elias P, Poterucha TJ, Rajaram V, Moller LM, Rodriguez V, Bhave S, et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. J Am Coll Cardiol 2022 Aug 09;80(6):613-626 [FREE Full text] [doi: 10.1016/j.jacc.2022.05.029] [Medline: 35926935]
73. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021 May;27(5):815-819. [doi: 10.1038/s41591-021-01335-4] [Medline: 33958795]
74. Wu S, Chen X, Pan J, Dong W, Diao X, Zhang R, et al. An artificial intelligence system for the detection of bladder cancer via cystoscopy: a multicenter diagnostic study. J Natl Cancer Inst 2022 Feb 07;114(2):220-227 [FREE Full text] [doi: 10.1093/jnci/djab179] [Medline: 34473310]
75. Narang A, Bae R, Hong H, Thomas Y, Surette S, Cadieu C, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. JAMA Cardiol 2021 Jun 01;6(6):624-632 [FREE Full text] [doi: 10.1001/jamacardio.2021.0185] [Medline: 33599681]
76. Yuan XL, Guo LJ, Liu W, Zeng X, Mou Y, Bai S, et al. Artificial intelligence for detecting superficial esophageal squamous cell carcinoma under multiple endoscopic imaging modalities: a multicenter study. J Gastroenterol Hepatol 2022 Jan;37(1):169-178 [FREE Full text] [doi: 10.1111/jgh.15689] [Medline: 34532890]
77. Attia ZI, Kapa S, Dugan J, Pereira N, Noseworthy PA, Jimenez FL, Discover Consortium (DigitalNoninvasive Screening for COVID-19 with AI ECG Repository). Rapid exclusion of COVID infection with the artificial intelligence electrocardiogram. Mayo Clin Proc 2021 Aug;96(8):2081-2094 [FREE Full text] [doi: 10.1016/j.mayocp.2021.05.027] [Medline: 34353468]
78. Kashou AH, Medina-Inojosa JR, Noseworthy PA, Rodeheffer RJ, Lopez-Jimenez F, Attia IZ, et al. Artificial intelligence-augmented electrocardiogram detection of left ventricular systolic dysfunction in the general population. Mayo Clin Proc 2021 Oct;96(10):2576-2586 [FREE Full text] [doi: 10.1016/j.mayocp.2021.02.029] [Medline: 34120755]

79. Kwon JM, Kim KH, Medina-Inojosa J, Jeon KH, Park J, Oh BH. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. J Heart Lung Transplant 2020 Aug;39(8):805-814. [doi: 10.1016/j.healun.2020.04.009] [Medline: 32381339]
80. Asch FM, Mor-Avi V, Rubenson D, Goldstein S, Saric M, Mikati I, et al. Deep learning-based automated echocardiographic quantification of left ventricular ejection fraction: a point-of-care solution. Circ Cardiovasc Imaging 2021 Jun;14(6):e012293. [doi: 10.1161/CIRCIMAGING.120.012293] [Medline: 34126754]
81. Kashou AH, Rabinstein AA, Attia IZ, Asirvatham SJ, Gersh BJ, Friedman PA, et al. Recurrent cryptogenic stroke: a potential role for an artificial intelligence-enabled electrocardiogram? HeartRhythm Case Rep 2020 Apr;6(4):202-205 [FREE Full text] [doi: 10.1016/j.hrcr.2019.12.013] [Medline: 32322497]
82. Wu L, He X, Liu M, Xie H, An P, Zhang J, et al. Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial. Endoscopy 2021 Dec;53(12):1199-1207. [doi: 10.1055/a-1350-5583] [Medline: 33429441]
83. Yang X, Wang H, Dong Q, Xu Y, Liu H, Ma X, et al. An artificial intelligence system for distinguishing between gastrointestinal stromal tumors and leiomyomas using endoscopic ultrasonography. Endoscopy 2022 Mar;54(3):251-261. [doi: 10.1055/a-1476-8931] [Medline: 33827140]
84. Herrin J, Abraham NS, Yao X, Noseworthy PA, Inselman J, Shah ND, et al. Comparative effectiveness of machine learning approaches for predicting gastrointestinal bleeds in patients receiving antithrombotic treatment. JAMA Netw Open 2021 May 03;4(5):e2110703 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10703] [Medline: 34019087]
85. Xie X, Xiao YF, Zhao XY, Li JJ, Yang QQ, Peng X, et al. Development and validation of an artificial intelligence model for small bowel capsule endoscopy video review. JAMA Netw Open 2022 Jul 01;5(7):e2221992 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.21992] [Medline: 35834249]
86. Shung DL, Au B, Taylor RA, Tay JK, Laursen SB, Stanley AJ, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. Gastroenterology 2020 Jan;158(1):160-167 [FREE Full text] [doi: 10.1053/j.gastro.2019.09.009] [Medline: 31562847]
87. Bhuiyan A, Govindaiah A, Deobhakta A, Gupta M, Rosen R, Saleem S, et al. Development and validation of an automated diabetic retinopathy screening tool for primary care setting. Diabetes Care 2020 Oct;43(10):e147-e148 [FREE Full text] [doi: 10.2337/dc19-2133] [Medline: 32855159]
88. Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. Br J Ophthalmol 2021 May;105(5):723-728 [FREE Full text] [doi: 10.1136/bjophthalmol-2020-316594] [Medline: 32606081]
89. Olvera-Barrios A, Heeren TF, Balaskas K, Chambers R, Bolter L, Egan C, et al. Diagnostic accuracy of diabetic retinopathy grading by an artificial intelligence-enabled algorithm compared with a human standard for wide-field true-colour confocal scanning and standard digital retinal images. Br J Ophthalmol 2021 Feb;105(2):265-270. [doi: 10.1136/bjophthalmol-2019-315394] [Medline: 32376611]
90. Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun 2021 May 28;12(1):3242 [FREE Full text] [doi: 10.1038/s41467-021-23458-5] [Medline: 34050158]
91. Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD, EyeArt Study Group. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open 2021 Nov 01;4(11):e2134254 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.34254] [Medline: 34779843]
92. Ravaut M, Harish V, Sadeghi H, Leung KK, Volkovs M, Kornas K, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. JAMA Netw Open 2021 May 03;4(5):e2111315 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11315] [Medline: 34032855]
93. Bachar N, Benbassat D, Brailovsky D, Eshel Y, Glück D, Levner D, et al. An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. Am J Hematol 2021 Oct 01;96(10):1264-1274 [FREE Full text] [doi: 10.1002/ajh.26295] [Medline: 34264525]
94. Dong L, He W, Zhang R, Ge Z, Wang YX, Zhou J, et al. Artificial intelligence for screening of multiple retinal and optic nerve diseases. JAMA Netw Open 2022 May 02;5(5):e229960 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.9960] [Medline: 35503220]
95. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 2021 May;44(5):1168-1175 [FREE Full text] [doi: 10.2337/dc20-1877] [Medline: 33402366]
96. Lee Y, Kim G, Jun JE, Park H, Lee WJ, Hwang YC, et al. An integrated digital health care platform for diabetes management with ai-based dietary management: 48-week results from a randomized controlled trial. Diabetes Care 2023 May 01;46(5):959-966. [doi: 10.2337/dc22-1929] [Medline: 36821833]
97. Oikonomidi T, Ravaud P, Cosson E, Montori V, Tran VT. Evaluation of patient willingness to adopt remote digital monitoring for diabetes management. JAMA Netw Open 2021 Jan 04;4(1):e2033115 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.33115] [Medline: 33439263]

98. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020 Aug;159(2):512-20.e7. [doi: 10.1053/j.gastro.2020.04.062] [Medline: 32371116]
99. Wang P, Liu P, Glissen Brown JR, Berzin TM, Zhou G, Lei S, et al. Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study. Gastroenterology 2020 Oct;159(4):1252-61.e5. [doi: 10.1053/j.gastro.2020.06.023] [Medline: 32562721]
100. Svoboda E. Artificial intelligence is improving the detection of lung cancer. Nature 2020 Nov;587(7834):S20-S22. [doi: 10.1038/d41586-020-03157-9] [Medline: 33208974]
101. Song Z, Zou S, Zhou W, Huang Y, Shao L, Yuan J, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 2020 Aug 27;11(1):4294 [FREE Full text] [doi: 10.1038/s41467-020-18147-8] [Medline: 32855423]
102. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel AS, Naheyan T, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. Lancet Digit Health 2021 Sep;3(9):e543-e554 [FREE Full text] [doi: 10.1016/S2589-7500(21)00116-3] [Medline: 34446265]
103. Tang LY, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD. Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. Lancet Digit Health 2020 May;2(5):e259-e267 [FREE Full text] [doi: 10.1016/S2589-7500(20)30064-9] [Medline: 33328058]
104. Kim H, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020 Mar;2(3):e138-e148. [doi: 10.1016/s2589-7500(20)30003-0]
105. Wu S, Hong G, Xu A, Zeng H, Chen X, Wang Y, et al. Artificial intelligence-based model for lymph node metastases detection on whole slide images in bladder cancer: a retrospective, multicentre, diagnostic study. Lancet Oncol 2023 Apr;24(4):360-370. [doi: 10.1016/S1470-2045(23)00061-X] [Medline: 36893772]
106. Weigt J, Repici A, Antonelli G, Afifi A, Kliegis L, Correale L, et al. Performance of a new integrated computer-assisted system (CADe/CADx) for detection and characterization of colorectal neoplasia. Endoscopy 2022 Feb;54(2):180-184. [doi: 10.1055/a-1372-0419] [Medline: 33494106]
107. Homayounieh F, Digumarthy S, Ebrahimian S, Rueckel J, Hoppe BF, Sabel BO, et al. An artificial intelligence-based chest X-ray model on human nodule detection accuracy from a multicenter study. JAMA Netw Open 2021 Dec 01;4(12):e2141096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.41096] [Medline: 34964851]
108. Glissen Brown JR, Mansour NM, Wang P, Chuchuca MA, Minchenberg SB, Chandnani M, et al. Deep learning computer-aided polyp detection reduces adenoma miss rate: a united states multi-center randomized tandem colonoscopy study (CADeT-CS Trial). Clin Gastroenterol Hepatol 2022 Jul;20(7):1499-507.e4 [FREE Full text] [doi: 10.1016/j.cgh.2021.09.009] [Medline: 34530161]
109. Foersch S, Eckstein M, Wagner DC, Gach F, Woerl AC, Geiger J, et al. Deep learning for diagnosis and survival prediction in soft tissue sarcoma. Ann Oncol 2021 Sep;32(9):1178-1187 [FREE Full text] [doi: 10.1016/j.annonc.2021.06.007] [Medline: 34139273]
110. Jin EH, Lee D, Bae JH, Kang HY, Kwak M, Seo JY, et al. Improved accuracy in optical diagnosis of colorectal polyps using convolutional neural networks with visual explanations. Gastroenterology 2020 Jun;158(8):2169-79.e8. [doi: 10.1053/j.gastro.2020.02.036] [Medline: 32119927]
111. Shi Y, Wang Z, Chen P, Cheng P, Zhao K, Zhang H, Alzheimer's Disease Neuroimaging Initiative. Episodic memory-related imaging features as valuable biomarkers for the diagnosis of Alzheimer's disease: a multicenter study based on machine learning. Biol Psychiatry Cogn Neurosci Neuroimaging 2023 Feb;8(2):171-180. [doi: 10.1016/j.bpsc.2020.12.007] [Medline: 33712376]
112. Huang B, Tian S, Zhan N, Ma J, Huang Z, Zhang C, et al. Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: a retrospective multicentre study. EBioMedicine 2021 Nov;73:103631 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103631] [Medline: 34678610]
113. Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020 Oct 09;11(1):5088 [FREE Full text] [doi: 10.1038/s41467-020-18685-1] [Medline: 33037212]
114. Goh KH, Wang L, Yeow AY, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711 [FREE Full text] [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
115. Zhou Q, Zuley M, Guo Y, Yang L, Nair B, Vargo A, et al. A machine and human reader study on AI diagnosis model safety under attacks of adversarial images. Nat Commun 2021 Dec 14;12(1):7281 [FREE Full text] [doi: 10.1038/s41467-021-27577-x] [Medline: 34907229]
116. Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. Lancet Digit Health 2021 Apr;3(4):e250-e259 [FREE Full text] [doi: 10.1016/S2589-7500(21)00041-8] [Medline: 33766289]

117. Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. Lancet Digit Health 2020 Aug;2(8):e407-e416. [doi: 10.1016/s2589-7500(20)30159-x]
118. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol 2020 Feb;21(2):222-232. [doi: 10.1016/S1470-2045(19)30738-7] [Medline: 31926806]
119. Venkatesan P. Artificial intelligence and cancer diagnosis: caution needed. Lancet Oncol 2021 Oct;22(10):1364. [doi: 10.1016/S1470-2045(21)00533-7] [Medline: 34509184]
120. Gao K, Su J, Jiang Z, Zeng L, Feng Z, Shen H, et al. Dual-branch combination network (DCN): towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Med Image Anal 2021 Jan;67:101836 [FREE Full text] [doi: 10.1016/j.media.2020.101836] [Medline: 33129141]
121. Pfob A, Sidey-Gibbons C, Barr RG, Duda V, Alwafai Z, Balleyguier C, et al. Intelligent multi-modal shear wave elastography to reduce unnecessary biopsies in breast cancer diagnosis (INSPiRED 002): a retrospective, international, multicentre analysis. Eur J Cancer 2022 Dec;177:1-14. [doi: 10.1016/j.ejca.2022.09.018] [Medline: 36283244]
122. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020 Jan;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6] [Medline: 31894144]
123. Bachtiger P, Petri CF, Scott FE, Ri Park S, Kelshiker MA, Sahemey HK, et al. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. Lancet Digit Health 2022 Feb;4(2):e117-e125. [doi: 10.1016/s2589-7500(21)00256-9]
124. Kann BH, Likitlersuang J, Bontempi D, Ye Z, Aneja S, Bakst R, et al. Screening for extranodal extension in HPV-associated oropharyngeal carcinoma: evaluation of a CT-based deep learning algorithm in patient data from a multicentre, randomised de-escalation trial. Lancet Digit Health 2023 Jun;5(6):e360-e369 [FREE Full text] [doi: 10.1016/S2589-7500(23)00046-8] [Medline: 37087370]
125. Soltan AA, Kouchaki S, Zhu T, Kiyasseh D, Taylor T, Hussain ZB, et al. Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. Lancet Digit Health 2021 Feb;3(2):e78-e87 [FREE Full text] [doi: 10.1016/S2589-7500(20)30274-0] [Medline: 33509388]
126. Xie Y, Zhao L, Yang X, Wu X, Yang Y, Huang X, et al. Screening candidates for refractive surgery with corneal tomographic-based deep learning. JAMA Ophthalmol 2020 May 01;138(5):519-526 [FREE Full text] [doi: 10.1001/jamaophthalmol.2020.0507] [Medline: 32215587]
127. Abbasi J. Artificial intelligence improves breast cancer screening in study. JAMA 2020 Feb 11;323(6):499. [doi: 10.1001/jama.2020.0370] [Medline: 32044919]
128. Xu H, Tang RS, Lam TY, Zhao G, Lau JY, Liu Y, et al. Artificial intelligence-assisted colonoscopy for colorectal cancer screening: a multicenter randomized controlled trial. Clin Gastroenterol Hepatol 2023 Feb;21(2):337-46.e3 [FREE Full text] [doi: 10.1016/j.cgh.2022.07.006] [Medline: 35863686]
129. Sun Y, Zhang L, Dong D, Li X, Wang J, Yin C, et al. Application of an individualized nomogram in first-trimester screening for trisomy 21. Ultrasound Obstet Gynecol 2021 Jul;58(1):56-66 [FREE Full text] [doi: 10.1002/uog.22087] [Medline: 32438493]
130. Zeleznik R, Foldyna B, Eslami P, Weiss J, Alexander I, Taron J, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nat Commun 2021 Jan 29;12(1):715 [FREE Full text] [doi: 10.1038/s41467-021-20966-2] [Medline: 33514711]
131. Liu CM, Chang SL, Chen HH, Chen WS, Lin YJ, Lo LW, et al. The clinical application of the deep learning technique for predicting trigger origins in patients with paroxysmal atrial fibrillation with catheter ablation. Circ Arrhythm Electrophysiol 2020 Nov;13(11):e008518. [doi: 10.1161/CIRCEP.120.008518] [Medline: 33021404]
132. Qiang M, Li C, Sun Y, Sun Y, Ke L, Xie C, et al. A prognostic predictive system based on deep learning for locoregionally advanced nasopharyngeal carcinoma. J Natl Cancer Inst 2021 May 04;113(5):606-615 [FREE Full text] [doi: 10.1093/jnci/djaa149] [Medline: 32970812]
133. She Y, He B, Wang F, Zhong Y, Wang T, Liu Z, et al. Deep learning for predicting major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer: a multicentre study. EBioMedicine 2022 Dec;86:104364 [FREE Full text] [doi: 10.1016/j.ebiom.2022.104364] [Medline: 36395737]
134. Wang L, Ding L, Liu Z, Sun L, Chen L, Jia R, et al. Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. Br J Ophthalmol 2020 Mar;104(3):318-323. [doi: 10.1136/bjophthalmol-2018-313706] [Medline: 31302629]
135. Li D, Bledsoe JR, Zeng Y, Liu W, Hu Y, Bi K, et al. A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals. Nat Commun 2020 Nov 26;11(1):6004 [FREE Full text] [doi: 10.1038/s41467-020-19817-3] [Medline: 33244018]
136. Yu G, Sun K, Xu C, Shi XH, Wu C, Xie T, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nat Commun 2021 Nov 02;12(1):6311 [FREE Full text] [doi: 10.1038/s41467-021-26643-8] [Medline: 34728629]

137. Kwon JM, Cho Y, Jeon KH, Cho S, Kim KH, Baek SD, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet Digit Health 2020 Jul;2(7):e358-e367 [FREE Full text] [doi: 10.1016/S2589-7500(20)30108-4] [Medline: 33328095]
138. Lin A, Manral N, McElhinney P, Killekar A, Matsumoto H, Kwiecinski J, et al. Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study. Lancet Digit Health 2022 Apr;4(4):e256-e265 [FREE Full text] [doi: 10.1016/S2589-7500(22)00022-X] [Medline: 35337643]
139. Storelli L, Azzimonti M, Gueye M, Vizzino C, Preziosa P, Tedeschi G, et al. A deep learning approach to predicting disease progression in multiple sclerosis using magnetic resonance imaging. Invest Radiol 2022 Jul 01;57(7):423-432. [doi: 10.1097/RLI.0000000000000854] [Medline: 35093968]
140. Mao N, Zhang H, Dai Y, Li Q, Lin F, Gao J, et al. Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study. Br J Cancer 2023 Mar;128(5):793-804 [FREE Full text] [doi: 10.1038/s41416-022-02092-y] [Medline: 36522478]
141. Ueno S, Berntsen J, Ito M, Uchiyama K, Okimura T, Yabuuchi A, et al. Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: a single-center large cohort retrospective study. Fertil Steril 2021 Oct;116(4):1172-1180 [FREE Full text] [doi: 10.1016/j.fertnstert.2021.06.001] [Medline: 34246469]
142. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol 2021 Jan;22(1):132-141. [doi: 10.1016/S1470-2045(20)30535-0] [Medline: 33387492]
143. Li X, Gao H, Zhu J, Huang Y, Zhu Y, Huang W, et al. 3D deep learning model for the pretreatment evaluation of treatment response in esophageal carcinoma: a prospective study (ChiCTR2000039279. Int J Radiat Oncol Biol Phys 2021 Nov 15;111(4):926-935 [FREE Full text] [doi: 10.1016/j.ijrobp.2021.06.033] [Medline: 34229050]
144. Wu L, Ye W, Liu Y, Chen D, Wang Y, Cui Y, et al. An integrated deep learning model for the prediction of pathological complete response to neoadjuvant chemotherapy with serial ultrasonography in breast cancer patients: a multicentre, retrospective study. Breast Cancer Res 2022 Nov 21;24(1):81 [FREE Full text] [doi: 10.1186/s13058-022-01580-6] [Medline: 36414984]
145. Suri JS, Agarwal S, Saba L, Chabert GL, Carriero A, Paschè A, et al. Multicenter study on COVID-19 lung computed tomography segmentation with varying glass ground opacities using unseen deep learning artificial intelligence paradigms: COVLIAS 1.0 validation. J Med Syst 2022 Aug 21;46(10):62 [FREE Full text] [doi: 10.1007/s10916-022-01850-y] [Medline: 35988110]
146. Khurshid S, Friedman S, Pirruccello JP, Di Achille P, Diamant N, Anderson CD, et al. Deep learning to predict cardiac magnetic resonance-derived left ventricular mass and hypertrophy from 12-lead ECGs. Circ Cardiovasc Imaging 2021 Jun;14(6):e012281 [FREE Full text] [doi: 10.1161/CIRCIMAGING.120.012281] [Medline: 34126762]
147. Liu XP, Jin X, Seyed Ahmadian S, Yang X, Tian SF, Cai YX, et al. Clinical significance and molecular annotation of cellular morphometric subtypes in lower-grade gliomas discovered by machine learning. Neuro Oncol 2023 Jan 05;25(1):68-81 [FREE Full text] [doi: 10.1093/neuonc/noac154] [Medline: 35716369]
148. Akal F, Batu ED, Sonmez HE, Karadağ Ş, Demir F, Ayaz NA, et al. Diagnosing growing pains in children by using machine learning: a cross-sectional multicenter study. Med Biol Eng Comput 2022 Dec;60(12):3601-3614. [doi: 10.1007/s11517-022-02699-6] [Medline: 36264529]
149. Awada H, Durmaz A, Gurnari C, Kishtagari A, Meggendorfer M, Kerr CM, et al. Machine learning integrates genomic signatures for subclassification beyond primary and secondary acute myeloid leukemia. Blood 2021 Nov 11;138(19):1885-1895 [FREE Full text] [doi: 10.1182/blood.2020010603] [Medline: 34075412]
150. Moyer JD, Lee P, Bernard C, Henry L, Lang E, Cook F, Traumabase Group®. Machine learning-based prediction of emergency neurosurgery within 24 h after moderate to severe traumatic brain injury. World J Emerg Surg 2022 Aug 03;17(1):42 [FREE Full text] [doi: 10.1186/s13017-022-00449-5] [Medline: 35922831]
151. Hollon T, Jiang C, Chowdury A, Nasir-Moin M, Kondepudi A, Aabedi A, et al. Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. Nat Med 2023 Apr;29(4):828-832 [FREE Full text] [doi: 10.1038/s41591-023-02252-4] [Medline: 36959422]
152. Takenaka K, Ohtsuka K, Fujii T, Negi M, Suzuki K, Shimizu H, et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology 2020 Jun;158(8):2150-2157. [doi: 10.1053/j.gastro.2020.02.012] [Medline: 32060000]
153. Savage N. Why artificial intelligence needs to understand consequences. Nature (Forthcoming) 2023 Feb 24. [doi: 10.1038/d41586-023-00577-1] [Medline: 36829060]
154. -. Artificial intelligence predicts drug response. Cancer Discov 2021 Jan;11(1):4-5. [doi: 10.1158/2159-8290.CD-NB2020-109] [Medline: 33239267]
155. Wagner M, Müller-Stich BP, Kisilenko A, Tran D, Heger P, Mündermann L, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. Med Image Anal 2023 May;86:102770 [FREE Full text] [doi: 10.1016/j.media.2023.102770] [Medline: 36889206]

156. Soda P, D'Amico NC, Tessadori J, Valbusa G, Guarrasi V, Bortolotto C, et al. AIforCOVID: predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study. Med Image Anal 2021 Dec;74:102216 [FREE Full text] [doi: 10.1016/j.media.2021.102216] [Medline: 34492574]
157. Avari P, Leal Y, Herrero P, Wos M, Jugnee N, Arnoriaga-Rodríguez M, et al. Safety and feasibility of the PEPPER adaptive bolus advisor and safety system: a randomized control study. Diabetes Technol Ther 2021 Mar 01;23(3):175-186. [doi: 10.1089/dia.2020.0301] [Medline: 33048581]
158. Wathour J, Govaerts PJ, Deggouj N. From manual to artificial intelligence fitting: two cochlear implant case studies. Cochlear Implants Int 2020 Sep;21(5):299-305. [doi: 10.1080/14670100.2019.1667574] [Medline: 31530099]
159. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open 2021 Feb 01;4(2):e2037107 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
160. Eilts SK, Pfeil JM, Poschkamp B, Krohne TU, Eter N, Barth T, Comparing Alternative Ranibizumab Dosages for SafetyEfficacy in Retinopathy of Prematurity (CARE-ROP) Study Group. Assessment of retinopathy of prematurity regression and reactivation using an artificial intelligence-based vascular severity score. JAMA Netw Open 2023 Jan 03;6(1):e2251512 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.51512] [Medline: 36656578]
161. Takeda I, Yamada A, Onodera H. Artificial intelligence-assisted motion capture for medical applications: a comparative study between markerless and passive marker motion capture. Comput Methods Biomech Biomed Engin 2021 Jun;24(8):864-873. [doi: 10.1080/10255842.2020.1856372] [Medline: 33290107]
162. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, NextDREAM Consortium. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat Med 2020 Sep;26(9):1380-1384. [doi: 10.1038/s41591-020-1045-7] [Medline: 32908282]
163. Carvalho DM, Richardson PJ, Olaciregui N, Stankunaite R, Lavarino C, Molinari V, et al. Repurposing Vandetanib plus everolimus for the treatment of -mutant diffuse intrinsic pontine glioma. Cancer Discov 2022 Feb;12(2):416-431 [FREE Full text] [doi: 10.1158/2159-8290.CD-20-1201] [Medline: 34551970]
164. Sheridan C. Massive data initiatives and AI provide testbed for pandemic forecasting. Nat Biotechnol 2020 Sep;38(9):1010-1013. [doi: 10.1038/s41587-020-0671-4] [Medline: 32887968]
165. Meeuws M, Pascoal D, Janssens de Varebeke S, De Ceulaer G, Govaerts PJ. Cochlear implant telemedicine: remote fitting based on psychoacoustic self-tests and artificial intelligence. Cochlear Implants Int 2020 Sep 13;21(5):260-268. [doi: 10.1080/14670100.2020.1757840] [Medline: 32397922]
166. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45-10 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
167. Dong Q, Li L, Dai D, Zheng C, Wu Z, Chang B, et al. A survey for in-context learning. arXiv. Preprint posted online December 31, 2022 2022 [FREE Full text]
168. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, et al. Development and validation of an artificial intelligence system to optimize clinician review of patient records. JAMA Netw Open 2021 Jul 01;4(7):e2117391 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.17391] [Medline: 34297075]
169. Cowan RP, Rapoport AM, Blythe J, Rothrock J, Knievel K, Peretz AM, et al. Diagnostic accuracy of an artificial intelligence online engine in migraine: a multi-center study. Headache 2022 Jul;62(7):870-882 [FREE Full text] [doi: 10.1111/head.14324] [Medline: 35657603]
170. Curran JM, Meuter ML, Surprenant CF. Intentions to use self-service technologies: a confluence of multiple attitudes. J Serv Res 2016 Jun 29;5(3):209-224 [FREE Full text] [doi: 10.1177/1094670502238916]
171. Dabholkar PA. Consumer evaluations of new technology-based self-service options: an investigation of alternative models of service quality. Int J Res Mark 1996;13(1):29-51 [FREE Full text] [doi: 10.1016/0167-8116(95)00027-5]
172. Seneviratne MG, Li RC, Schreier M, Lopez-Martinez D, Patel BS, Yakubovich A, et al. User-centred design for machine learning in health care: a case study from care management. BMJ Health Care Inform 2022 Oct 11;29(1):e100656. [doi: 10.1136/bmjhci-2022-100656] [Medline: 36220304]
173. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Front Digit Health 2021;3:645232 [FREE Full text] [doi: 10.3389/fdgth.2021.645232] [Medline: 34713115]
174. Novak LL, Russell RG, Garvey K, Patel M, Thomas Craig KJ, Snowdon J, et al. Clinical use of artificial intelligence requires AI-capable organizations. JAMIA Open 2023 Jul;6(2):ooad028 [FREE Full text] [doi: 10.1093/jamiaopen/ooad028] [Medline: 37152469]

Abbreviations

AI: artificial intelligence
GenAI: generative artificial intelligence tools and applications
ICL: in-context learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RQ: research question

Edited by A Castonguay; submitted 21.08.23; peer-reviewed by SH Kim, Y Wang, S Pesala; comments to author 19.09.23; revised version received 12.10.23; accepted 30.01.24; published 20.03.24.

Please cite as:
Yim D, Khuntia J, Parameswaran V, Meyers A
Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review
JMIR Med Inform 2024;12:e52073
URL: https://medinform.jmir.org/2024/1/e52073
doi:10.2196/52073
PMID:38506918

Implementation Report

Maturity Assessment of District Health Information System Version 2 Implementation in Ethiopia: Current Status and Improvement Pathways

Tesfahun Melese Yilma[1*], BSc, MPH, PhD; Asefa Taddese[2*], BSc, MSc; Adane Mamuye[3*], BSc, MSc, PhD; Berhanu Fikadie Endehabtu[1*], BSc, MPH; Yibeltal Alemayehu[4], BSc, MPH, PhD; Asaye Senay[4], BSc, MSc; Dawit Daka[4], BSc, MSc; Loko Abraham[5], MD, Speciality in Pediatrics and Child Health; Rabeal Tadesse[5], BSc, MSc; Gemechis Melkamu[6], BSc, MSc; Naod Wendrad[6], BSc, MHA; Oli Kaba[6], BSc, MSc; Mesoud Mohammed[6], BSc, MSc; Wubshet Denboba[7*], BSc, MPH; Dawit Birhan[7*], BSc, MSc; Amanuel Biru[7], BSc, MSc; Binyam Tilahun[1*], BSc, MPH, MSc, PhD

[1]Department of Health Informatics, Center for Digital Health and Implementation Science, University of Gondar, Gondar, Ethiopia

[2]Department of Biostatistics and Epidemiology, Center for Digital Health and Implementation Science, University of Gondar, Gondar, Ethiopia

[3]Department of Computer Science, Center for Digital Health and Implementation Science, University of Gondar, Gondar, Ethiopia

[4]Department of Health Policy and Management, Jimma University, Jimma, Ethiopia

[5]Digital Health Activity, Addis Ababa, Ethiopia

[6]Ministry of Health, Addis Ababa, Ethiopia

[7]Data Use Partnership, Addis Ababa, Ethiopia

[*]these authors contributed equally

Corresponding Author:
Tesfahun Melese Yilma, BSc, MPH, PhD
Department of Health Informatics
Center for Digital Health and Implementation Science
University of Gondar
Chechela Street, College of Medicine and Health Sciences
University of Gondar
Gondar, 196
Ethiopia
Phone: 251 918779820
Email: tesfahun.melese@uog.edu.et

Abstract

Background: Although Ethiopia has made remarkable progress in the uptake of the District Health Information System version 2 (DHIS2) for national aggregate data reporting, there has been no comprehensive assessment of the maturity level of the system.

Objective: This study aims to assess the maturity level of DHIS2 implementation in Ethiopia and propose a road map that could guide the progress toward a higher level of maturity. We also aim to assess the current maturity status, implementation gaps, and future directions of DHIS2 implementation in Ethiopia. The assessment focused on digital health system governance, skilled human resources, information and communication technology (ICT) infrastructure, interoperability, and data quality and use.

Methods: A collaborative assessment was conducted with the engagement of key stakeholders through consultative workshops using the Stages of Continuous Improvement tool to measure maturity levels in 5 core domains, 13 components, and 39 subcomponents. A 5-point scale (1=emerging, 2=repeatable, 3=defined, 4=managed, and 5=optimized) was used to measure the DHIS2 implementation maturity level.

Results: The national DHIS2 implementation's maturity level is currently at the defined stage (score=2.81) and planned to move to the manageable stage (score=4.09) by 2025. The domain-wise maturity score indicated that except for ICT infrastructure, which is at the repeatable stage (score=2.14), the remaining 4 domains are at the defined stage (score=3). The development of a standardized and basic DHIS2 process at the national level, the development of a 10-year strategic plan to guide the implementation of digital health systems including DHIS2, and the presence of the required competencies at the facility level to accomplish

specific DHIS2-related tasks are the major strength of the Ministry of Health of Ethiopia so far. The lack of workforce competency guidelines to support the implementation of DHIS2; the unavailability of core competencies (knowledge, skills, and abilities) required to accomplish DHIS2 tasks at all levels of the health system; and ICT infrastructures such as communication network and internet connectivity at the district, zonal, and regional levels are the major hindrances to effective DHIS2 implementation in the country.

Conclusions: On the basis of the Stages of Continuous Improvement maturity model toolkit, the implementation status of DHIS2 in Ethiopia is at the defined stage, with the ICT infrastructure domain being at the lowest stage as compared to the other 4 domains. By 2025, the maturity status is planned to move from the defined stage to the managed stage by improving the identified gaps. Various action points are suggested to address the identified gaps and reach the stated maturity level. The responsible body, necessary resources, and methods of verification required to reach the specified maturity level are also listed.

(JMIR Med Inform 2024;12:e50375) doi:10.2196/50375

KEYWORDS

health information system; digital health system; District Health Information System version 2; DHIS2; maturity assessment; Stages of Continuous Improvement; Ethiopia

Introduction

Background

Health information systems (HISs) have become an essential component of evidence-based decision-making and health service delivery worldwide [1,2]. Over the past years, countries have recognized the importance of reliable health data to track key health issues and outcomes, leading to significant investments in HISs in both high- and low-income nations. Therefore, the global community has recognized that reliable health data are vital for the development of national health systems. This has been further highlighted in the 2030 Agenda of Sustainable Development Goals [3,4], which recognizes the potential of information and communication technology (ICT) to accelerate human progress, bridge the digital divide, and develop knowledge societies [5].

In low- and middle-income countries (LMICs), various national governments and some global partners, such as the World Health Organization (WHO) and United Nations International Children's Emergency Fund, have made substantial investments in strengthening their HISs [6]. In particular, Ethiopia has made remarkable achievements in implementing HIS over the years. In 2006, the Ministry of Health (MOH) undertook a Health Management Information System reform with a focus on data management, human resources, and ICT (Planning and Programming Department, unpublished data, May 2006) [7]. The MOH standardized the indicators, recording and reporting forms, procedures, and reporting channels to improve performance [8].

Moreover, the MOH has implemented different digital health initiatives, including District Health Information System version 2 (DHIS2), electronic community HIS, electronic medical record systems, electronic community-based health insurance, logistic management information system, human resource information system, master facility registries, and electronic public health emergency management system. Maturity assessment of such digital health systems is essential to guide policy makers in strategizing and prioritizing initiatives [7].

Despite the remarkable progress made in Ethiopia, a comprehensive maturity assessment is yet to be conducted to determine the maturity level of the system and to inform policy makers about potential future interventions. DHIS2 is an open-source digital health platform developed by the University of Oslo in 2006 to manage HISs. Its first implementation was in India in 2006, and its first national rollout was in Kenya in 2010 [9]. Since 2010, LMICs worldwide have adopted this software. In Ethiopia, DHIS2 has been implemented since 2018 [7]. Since then, the MOH has made significant achievements in the development and implementation of DHIS2, such as the deployment of web-based or offline instances, full ownership of DHIS2 customization and implementation, upgrading DHIS2 to version 2.30 with apps developed in-house such as those for disease and public health emergency management report, data visualization applications (scorecard, bottleneck analysis, action tracker, and custom reports), creating metadata, data set customization, incorporating reporting functionality, and new features for decision-making by integrating data quality checks and dashboards [10].

Currently, the MOH uses DHIS2 for planning, reporting, analysis, and dissemination of data for all health programs. It accurately and timely collects and aggregates data such as routine health facility data, staffing, equipment, infrastructure, population estimates, disease outbreaks, survey or audit data, patient satisfaction surveys, and longitudinal patient records. DHIS2 stores, analyzes, and evaluates both aggregate and event-based data at a health facility level [7]. Therefore, DHIS2 is accepted as a primary source of information for planning and decision-making in Ethiopia's health system. However, despite its implementation, the maturity level of DHIS2 has not yet been assessed. A maturity assessment is used to measure the current maturity status of a certain HIS to identify the strengths and improvement points and accordingly prioritize the next steps to reach higher maturity levels [11].

In Ethiopia, the Stages of Continuous Improvement (SOCI) tool was used to assess the current maturity status and improvement of the overall HIS. The assessment report indicated that the overall maturity level of the Ethiopian HIS is between the "repeatable" and "defined" maturity stages [12]. The report recommends maturity assessment of individual digital health systems (eg, DHIS2), which have a wide impact and coverage in the country.

Objective

This study aims to assess the maturity level of DHIS2 implementation in Ethiopia and propose a road map that could guide the progress toward a higher level of maturity. The assessment could help identify the current capabilities; set a baseline for measuring improvement; identify strengths, weaknesses, opportunities, and threats; identify factors influencing achievements; set new pathways for improvements; and foster a culture of excellence in Ethiopia.

Methods

Assessment Setting

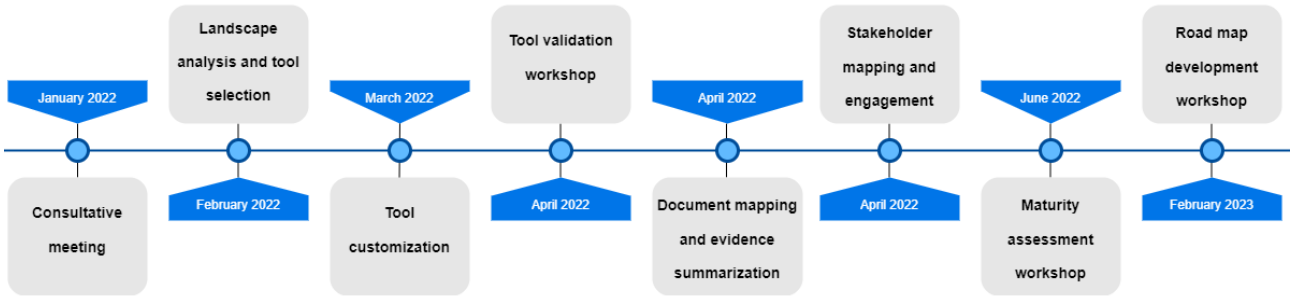
The maturity assessment of DHIS2 implementation was conducted in Ethiopia. It is a landlocked country in East Africa, with a population of 118 million and a 3-tier health care system consisting of primary health care units (health posts, health centers, and primary hospitals), general hospitals, and tertiary hospitals. The country has implemented DHIS2 since 2018. Since then, the country has made significant achievements in the customization and implementation of DHIS2 [10].

Assessment Procedure

Overview

A multimethod approach study was conducted in Ethiopia to assess the maturity level of the DHIS2 implementation. The assessment process proposed in the users' guide for the HIS interoperability maturity assessment was followed to conduct the DHIS2 maturity assessment [13]. First, we conducted various consultative meetings to develop an activity plan for the maturity assessment. Then, we conducted a landscape analysis of various maturity assessment tools to select an appropriate tool for customization. Next, tool customization, followed by a validation workshop, was conducted. The purpose of the workshop was to validate the customized tool for its clarity and appropriateness in the Ethiopian context. Then, we mapped the stakeholders and conducted a maturity assessment workshop. Finally, a road map development workshop was conducted to improve DHIS2 implementation in Ethiopia. Figure 1 shows a diagrammatic representation of the assessment procedure.

Figure 1. Diagrammatic representation of the assessment procedure.



Conducting Consultative Meetings

After the proposal to conduct a maturity assessment for DHIS2 was approved, the researchers conducted several consultative meetings with the MOH and Digital Health Activity, a partner of the MOH that supports Ethiopia in improving health care services through the implementation of digital health solutions. The first meeting was conducted on January 14, 2022, to discuss a drafted activity plan for the entire assessment procedure. The activity plan was presented in the presence of focal persons from the MOH and project facilitators from Digital Health Activity. Several useful inputs were provided to improve the activity plan. The key activities included a landscape analysis; a panel discussion among domain experts to select and customize the maturity assessment tool; a tool validation workshop involving domain experts, partners, and stakeholders to standardize the customized maturity assessment tool; a national workshop involving the MOH, partners, universities, and stakeholders; and data collection and analysis.

Landscape Analysis and Tool Selection

The researchers conducted a rigorous review of existing maturity assessment toolkits and peer reviewed publications from January 11, 2022, to February 15, 2022, to select an appropriate tool. Several tools were reviewed, including the Information Systems for Health Maturity Assessment Toolkit (IS4H) [14], the Global Digital Health Index [15], the Digital Maturity Assessment Tool [16], the HIS Interoperability Maturity Toolkit [17], the Hospital Information System Maturity Model [17], and the SOCI toolkit [18].

After reviewing each potential assessment model and conducting several discussion meetings, 2 models were shortlisted, considering their relevance, comprehensiveness, validity, credibility, and feasibility: IS4H and SOCI. IS4H is a maturity assessment model that describes the method, tool, and questions for assessing the organizational capacity of digital health systems. The model was developed by the Pan American Health Organization and WHO. In contrast, SOCI was developed by the US Centers for Disease Control and Prevention, the Health Data Collaborative Digital Health and Interoperability Working Group, and the United States Agency for International Development–funded Monitoring and Evaluation to Assess and Use Results Evaluation to help countries assess, plan, and prioritize interventions and investments to strengthen an HIS.

After rigorous deliberations regarding the strengths and limitations and comparing the shortlisted models (Multimedia Appendix 1) in a discussion among experts from the MOH, partners, and universities, SOCI was selected as the final model to assess DHIS2 implementation maturity.

Tool Customization

The assessment measures the current and desired HIS status across 5 core domains, 13 components, and 39 subcomponents and road map improvements [18]. The assessment tool measures maturity level based on 5 stages—emerging, defined, repeatable, managed, and optimized, with scores ranging from 1 to 5. Table 1 shows the definitions of each stage of the SOCI tool.

The tool was designed in 2 formats: an Excel (Microsoft Corp)–based and an app-based version of the tool available through the DHIS2 platform. In our case, we customized the Excel-based tool in March 2022 to assess the DHIS2 implementation maturity status. The customization focused on changing the generic HIS to the DHIS2 context. Throughout the Excel-based tool, we have modified the HIS concept to the DHIS2 context without altering the purpose and the core domains, components, and subcomponents of SOCI.

Table 1. The 5 stages of the Stages of Continuous Improvement tool and their definitions.

Stage	Scale	Definition
Emerging	1	Formal processes, capabilities, experience, or understanding of HIS[a] issues or activities are limited or emerging. Formal processes are not documented, and functional capabilities are at the development stage. Success depends on individual effort.
Repeatable	2	Basic processes are in place, based on previous activities or existing and accessible policies. The need for standardized processes and automated functional capabilities is known. There are efforts to document the current processes.
Defined	3	There are approved documented processes and guidelines tailored to District HIS version 2 activities. There is increased collaboration and knowledge sharing. Innovative methods and tools can be implemented and used to extend the functional capabilities.
Managed	4	Activities are under control using established processes. Requirements or goals have been developed, and a feedback process is in place to ensure that they are met. Detailed measures for processes and products are being collected.
Optimized	5	Best practices are being applied, and the system is capable of learning and adapting. The system uses experiences and feedback to rectify problems and continuously improve processes and capabilities. Future challenges are anticipated, and a plan is in place to address them through innovation and new technology. Processes are in place to ensure reviews and incorporation of relevant innovation.

[a]HIS: health information system.

Tool Validation Workshop

A validation workshop involving 18 high-level digital health experts was conducted on April 28 and 29, 2022, to standardize and validate a customized SOCI tool, with participants forming 2 groups to validate each domain, component, and subcomponent and provide constructive feedback. Following the workshop, the feedback was incorporated, which resulted in a validated and standardized SOCI tool.

Document Mapping and Evidence Summarization

The assessors used evidence from the policies and guidelines available in the Ethiopian MOH resource library to evaluate the DHIS2 implementation maturity status [19]. This supporting evidence was mapped in April 2022 based on the content and topics related to the specific component, which served as a reference for the assessors.

Stakeholder Mapping and Engagement

Potential stakeholders were identified in April 2022 after a discussion with the MOH, and those who were working on and supporting DHIS2 were selected, including national and international partners and agencies such as Ethiopian Food and Drug Administration; Ethiopian Pharmaceuticals Supply Agency; Ethiopian Public Health Institute; United Nations International Children's Emergency Fund; United States Agency for International Development; Water, Education, Economic Empowerment, Medical Care, and Alliance; Program for Appropriate Technology in Health; Bill and Melinda Gates Foundation; WHO; Population Services International; HIS Program; Institute of Chartered Accountant of Pakistan; African Medical and Research Foundation; Children's Investment Fund Foundation; Last 10 Kilometers; Data Use Partnership; Transform Primary Health Care; Project Hope; and Clinton Health Access Initiative. All directorates under the MOH, regional health bureaus, and universities were officially invited to participate in the maturity assessment, and most of the selected stakeholders agreed to participate.

Maturity Assessment Workshop

A total of 35 digital health experts were invited to participate in the maturity assessment workshop, of which only 29 (83%) participated. The workshop was conducted from June 15, 2022, to June 17, 2022, with participants representing the stakeholders identified in the mapping exercise. The workshop began with an opening remark by the Health Information Technology Director stating the workshop's objectives and the aim of the assessment, followed by a presentation on the basics, rationale, methods, and reasons for maturity assessment by the University of Gondar team. The maturity assessment tool selection process, landscape analysis, steps followed, and activities performed were presented. Participants were categorized into 2 major groups based on their expertise and experiences and oriented on how to conduct the assessment using the adopted maturity assessment toolkit, SOCI. Each team assessed DHIS2 based on the 5 domains of SOCI, with chairpersons and secretaries facilitating and documenting the assessment scores and justification for each result. Disagreements in scoring between the groups were resolved through discussions. The average score of the 39 subcomponents was summed as the total score for each component, the average score of the 13 components was

summed as the total score for each domain, and the average score of the 5 domains was considered as the total score of the current DHIS2 implementation maturity. Finally, the assessment scores, strengths, and gaps in the DHIS2 implementation were presented and discussed, and feedback was incorporated to refine the assessment results.

Road Map Development Workshop

Of the 45 digital health experts invited to the road map workshop, only 38 (84%) participated. The workshop was conducted from January 30, 2023, to February 2, 2023. The workshop aimed to improve the maturity level of DHIS2 implementation. In total, 38 high-level digital health system experts—12 from the MOH, 9 from Regional Health Bureaus, 2 from agencies, 10 donors or partners, and 5 from Capacity Building and Mentorship Program Universities—participated. The workshop was hands-on, with at least 90% of the time spent in brainstorming and discussion. This approach allowed for better experience, more retention, and more realistic evaluation of the roadmap development. The sessions were continuously linked to the strategic challenges faced by the MOH or health sector, and solutions were discussed and suggested to provide practical solutions for DHIS2 maturity challenges.

During the workshop, key issues related to the 5 domains, 13 components, and 39 subcomponents of the SOCI tool were addressed for the road mapping of DHIS2 implementation. Various discussions were conducted to uncover the strengths and gaps of DHIS2 implementation; identify specific limitations that hinder the capability and implementation of the road map of DHIS2; analyze how DHIS2 is implemented; explore various ways of future improvements on DHIS2 including alternative concepts, tools, and technologies; and scrutinize various governance documents, tools, system documentation, manuals, and other resources to assess the maturity road map of DHIS2.

Each participant integrated knowledge, cognitive strategies, and behaviors as a guide in the group work, and the road map tool was presented and discussed. A detailed road map for each domain and respective components and subcomponents was presented, and major gaps and methods to address them were discussed. A 2-year target was set, and a list of high-impact interventions was identified to address the gaps. Considerations and means of verifications were established; the primarily responsible body was identified; and finally, the road map was developed.

Ethical Considerations

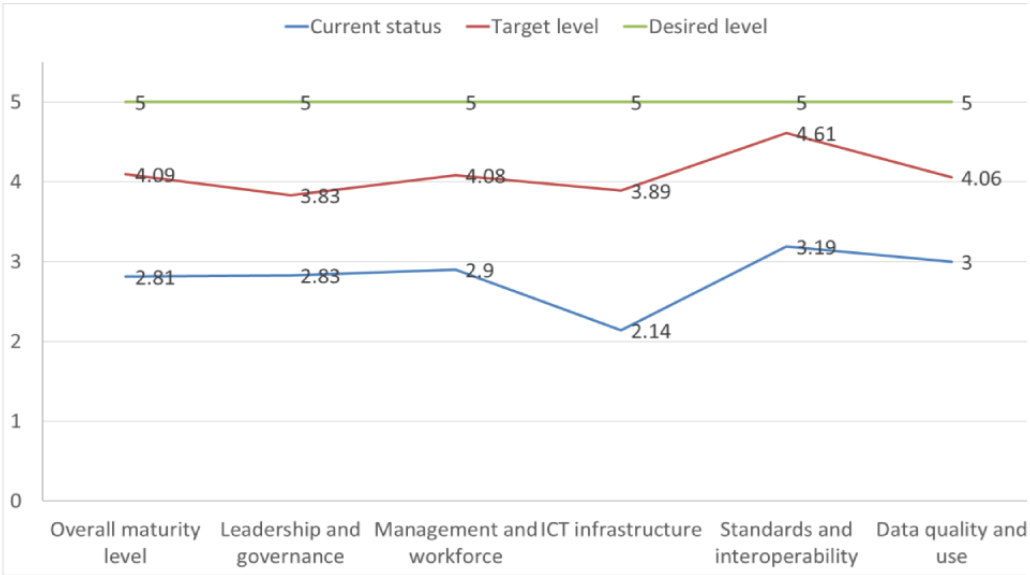
The maturity assessment was conducted in a workshop where participants completed the assessment in groups. There was no individual assessment. No personal identification was used in the assessment. Participation was voluntary. Participants were informed to express their ideas freely during the group discussion.

Implementation (Results)

Overall Current and Target Maturity Levels

The overall DHIS2 implementation in the country is currently at the “defined” stage, with a score of 2.81, and the goal is set to reach the “managed” stage of maturity, with a score of 4.09 by 2025 (Figure 2). Except for the ICT infrastructure domain, which is at the “repeatable” stage, with a score of 2.14, the other 4 domains are at the “defined” stage, with a score of approximately 3. The MOH’s strengths include the development of a standardized and basic DHIS2 process at the national level, the development of a 10-year strategic plan to guide the implementation of digital health systems, and the presence of required competencies at the facility level. However, the lack of workforce competency guidelines, the unavailability of the core competencies required to accomplish DHIS2 tasks at all levels of the health system, and ICT infrastructure challenges are the major hindrances to effectively implementing DHIS2. DHIS2 implementation is also in the “emerging” stage in terms of interoperability and data use components. Multimedia Appendix 2 provides the maturity levels of each domain, component, and subcomponent for the DHIS2 implementation.

Figure 2. A line graph showing the overall national maturity levels and targets of District Health Information System version 2 implementation. ICT: information and communication technology.



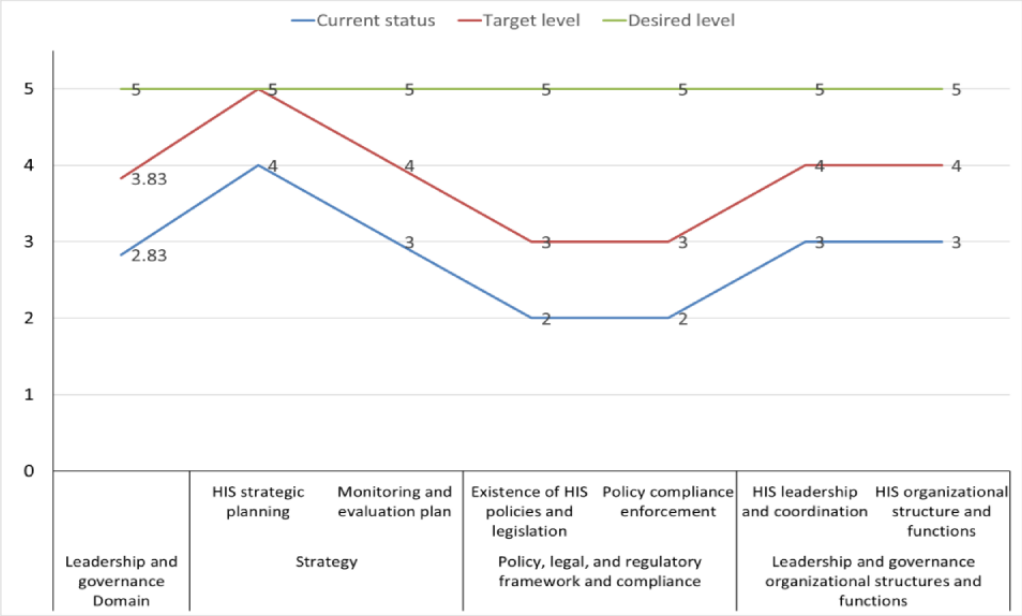
Domain-Wise Implementation Status and Improvement Road Map

Leadership and Governance

The target maturity level for the DHIS2 leadership and governance domain is to move from the defined (score=2.83) to the managed (score=3.83) stage. The "strategic planning" component is intended to move from the managed (score=3.5) to the optimized (score=4.5) maturity stage. The "policy, legal, and regulatory framework, and the compliance" component is intended to progress from the repeatable (score=2) to the defined (score=3) maturity stage, whereas the "leadership, governance, organizational structures, and functions" component is intended to move from the defined (score=3) to the managed (score=4) maturity stage (Figure 3). The MOH has a well-defined and budgeted strategy for HIS that includes DHIS2. In addition, there is a national-level cross-agency coordination group that oversees DHIS2 implementation. However, there are no well-defined mechanisms and regulatory bodies to ensure adherence to organizational policies, procedures, and best practices related to the digital health system. Multimedia Appendix 3 provides detailed information about the strengths and gaps of the ministry regarding DHIS2 implementation with respect to the "leadership and governance" domain.

The bodies responsible for performing the suggested activities (Multimedia Appendix 3) for the subcomponents are the MOH, Ethiopian Food and Drug Administration, regional health bureaus, subregional health administration authorities, and all line agencies. The resources needed to perform the operations indicated for the leadership and governance domain are adequate time, sufficient number of competent workforces, and adequate funds.

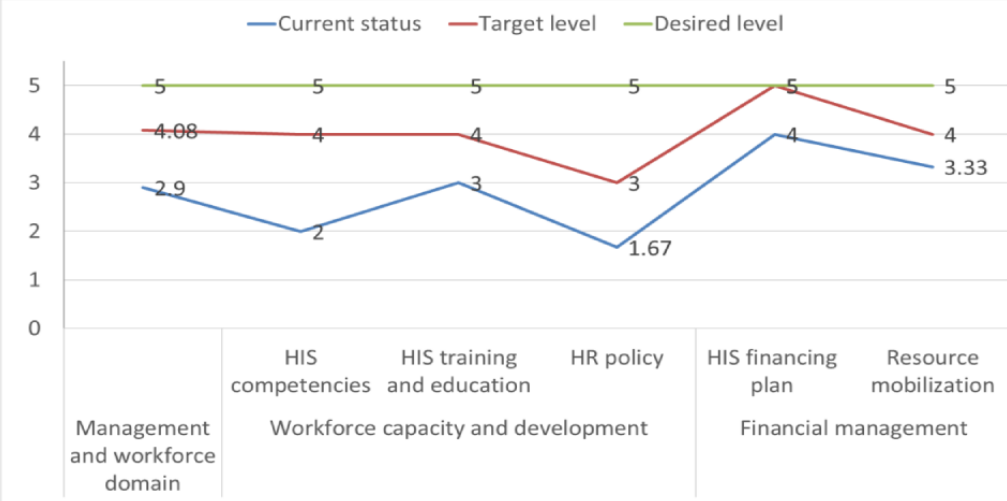
Figure 3. District Health Information System version 2 leadership and governance maturity level and targets. HIS: health information system.



Management and Workforce

The DHIS2 "management and workforce" domain is targeted to improve from the defined (score=2.9) to the managed (score=4) maturity stage. Financial management and workforce capacity and development are the 2 components that are expected to advance from the managed (score=3.6) to the optimized (score=4.5) stage and the repeatable (score=2.2) to the managed (score=3.67) stage, respectively (Figure 4). The DHIS2 academy-level training, collaboration with institutions, and customized training are identified as promising workforce competency initiatives. However, irregular DHIS2 workforce capability assessments and analyses, unclear career paths and roles, insufficient infrastructure, lack of regular feedback, and unmet staffing needs were identified as challenges (Multimedia Appendix 4). The MOH, Human Resource Directorate, Regional Health Bureau, higher institutions, associations, and civil service are responsible for preforming the planned activities. Financial and human resources, infrastructure, and connectivity are required resources. Verification methods included competency assessments, finance and monitoring and evaluation reports, and improvement at the job evaluation grading level.

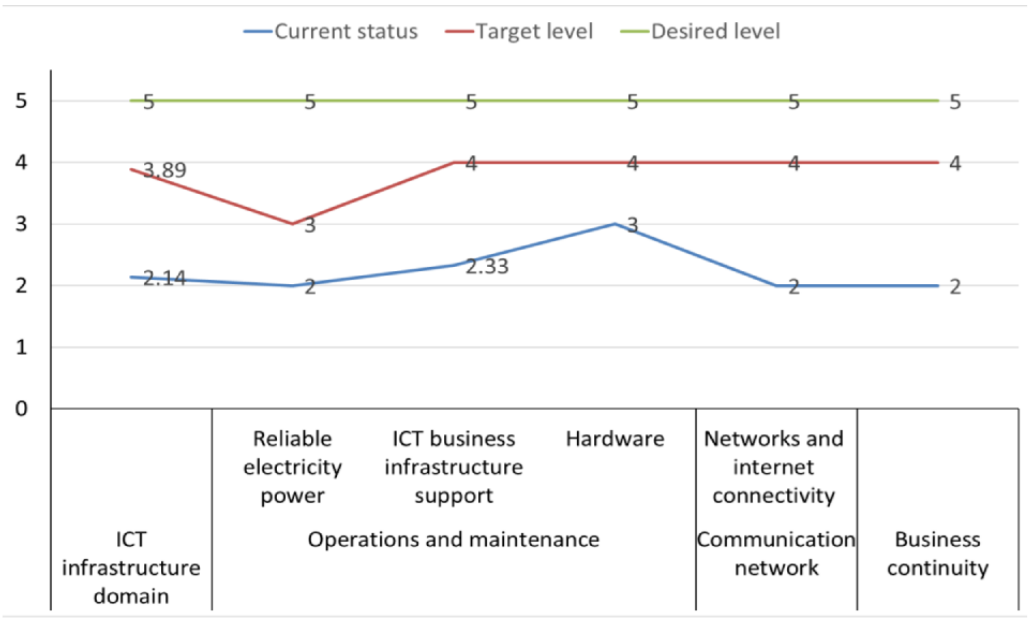
Figure 4. A line graph indicating the national District Health Information System version 2 management and workforce domain maturity level and targets. HIS: health information system; HR: human resource.



ICT Infrastructure

The information and communication domain's projected maturity levels ranged from repeatable (score=2.14) to managed (score=3.89), with a plan in place to elevate operations and maintenance from the repeatable (score=2.44) to the managed (score=3.67) level. Upgrading the local area network and wide area network components from the repeatable (score=2) to the managed (score=4) level is also planned. Similarly, the business continuity element is slated for transformation from the repeatable (score=2) to the managed (score=4) level (Figure 5). The national MOH has collected data about electricity or power access, sources, and reliability to support DHIS2 infrastructure; however, it is limited at higher levels. The MOH has also developed the facility hardware and software specifications, with some national and subnational offices having adequate hardware. Maintenance and installation of DHIS2 infrastructure are currently handled through an ad hoc mechanism (ie, there is no regular maintenance and installation). Maintenance and installation are conducted when the need arises. Simultaneously, several challenges impede the effectiveness of DHIS2 implementation under this domain, such as unstable electricity, insufficient hardware, outdated infrastructure, and limited and unstable network and internet connectivity. In addition, business continuity plans are at the emerging stage, not harmonized across DHIS2 health sector needs, and infrequently reviewed and revised (Multimedia Appendix 5). The MOH, including the chief executive officer, Regional Health Bureaus, digital health, and ICT and health infrastructure departments, is responsible for executing the planned actions in the ICT infrastructure sector. Financial resources, electrical engineers, network specialists, maintenance professionals, ICT spare parts (eg, hard drives and RAM), and other ICT accessories are all essential for achieving the predetermined goals. The methods of verification in the ICT infrastructure field include evaluation reports, network assessments, and support requests, which are measured against predetermined indicators for support from the support system.

Figure 5. A line graph indicating the national District Health Information System version 2 information and communication technology (ICT) infrastructure maturity level and targets.



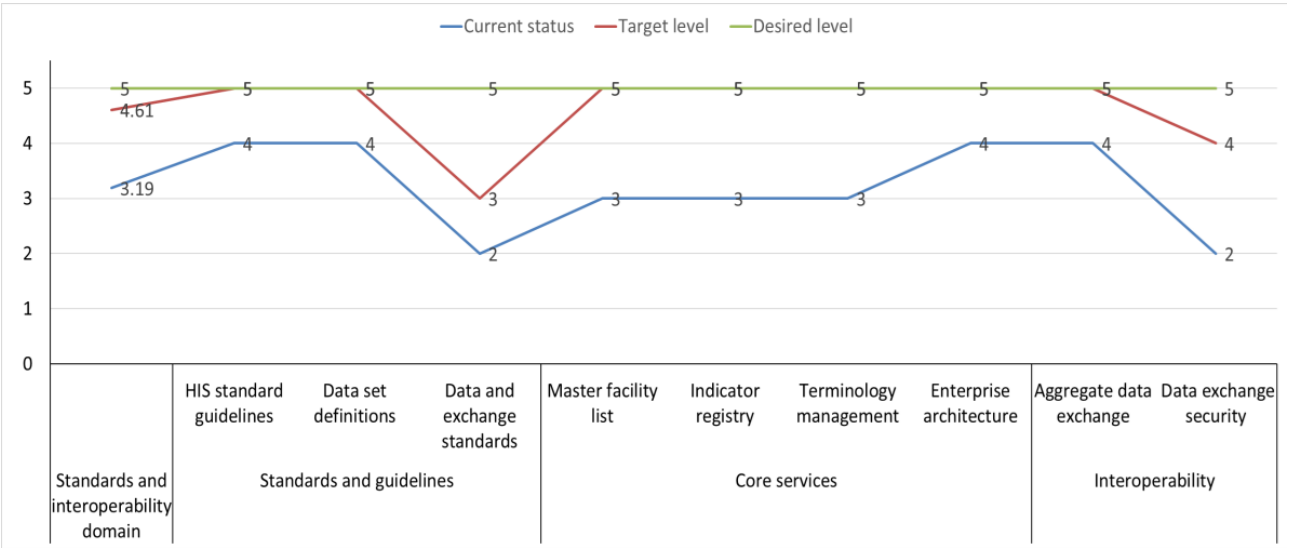
Standards and Interoperability

The "standards and interoperability" domain is an indispensable element in ensuring the success of any HIS as it facilitates the exchange of data across various systems while ensuring the accuracy of data capture and reporting. The domain's anticipated maturity levels ranged from defined (score=3.19) to optimized (score=4.61), with several components expected to move from the defined to the managed or optimized levels. The components of "standards and guidelines" are planned to have a maturity level ranging from defined (score=3.33) to managed (score=4.33), whereas the "core services" components are expected to move from the defined (score=3.25) to the optimized (score=5) level. In addition, there is a strategy to enhance the maturity level of the "interoperability and data exchange" components from defined (score=3) to optimized (score=4.5; Figure 6).

There are several strengths of the "standards and interoperability" domain, including the regular review and harmonization of indicators with international standards, the availability of different types of data exchange mechanisms, and well-documented indicator reference guides and formulas within DHIS2. The eHealth architecture includes an interoperability layer and defined shared services, and data exchange is piloted between DHIS2 and the Master Facility Registry. Vaccines, tracer drugs, and Rapid Diagnostic Test kits are available as indicators or data elements, and essential IT security measures such as virtual private networks, antivirus software, authentication, authorization, and firewalls are in place. Despite the ongoing efforts to improve standards and interoperability, there are still some gaps and areas for improvement. Although the national standard disease definitions, Health Management Information System recording and reporting guidelines, and electronic health record standards are in place, automated patient data exchange using internationally recognized standards between DHIS2 and other systems is not yet implemented. In addition, an interoperability laboratory for new exchange partners to test or onboard and a certification process do not exist yet. The DHIS2 data management and exchange standards are not integrated into the national health plan (Multimedia Appendix 6).

Technical expertise and financial resources are required to ensure the successful implementation and continued improvement in the "standards and interoperability" domain. The MOH and its partners are responsible for implementing the standards and interoperability-related activities. In addition, there must be a commitment to integrate DHIS2 data management and exchange standards into the national health plan and establish an interoperability laboratory for new exchange partners to test or onboard and a certification process.

Figure 6. A line graph indicating the national District Health Information System version 2 interoperability and standard maturity level and targets.



Data Quality and Use

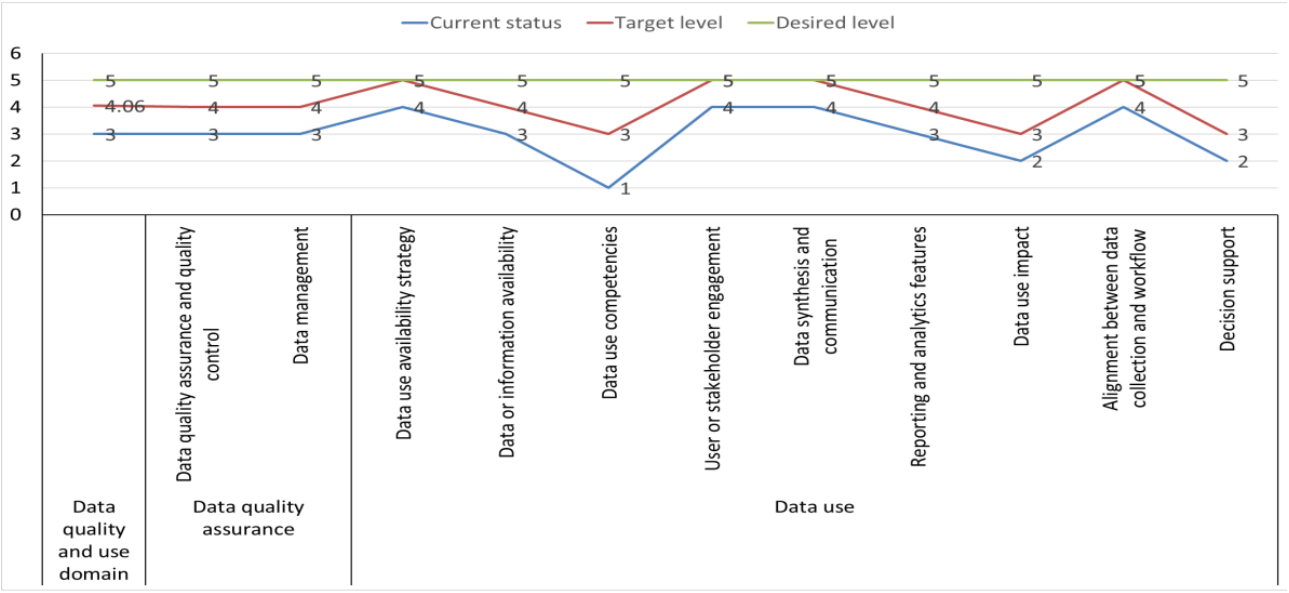
The anticipated maturity levels of the "data quality and use" domain ranged from defined (score=3) to managed (score=4.06), with planned improvements in the maturity level of data quality assurance components from the defined (score=3) to the managed (score=4) stage and data use components from the defined (score=3) to the managed (score=4.11) stage (Figure 7).

The assessment identified the strengths and gaps in this domain. The following strengths were identified: data quality assessment and auditing are performed regularly, manuals for data management and accessibility exist, and analytics and findings are shared with stakeholders on a quarterly basis. In addition, DHIS2 automated data reporting is implemented at all health facilities nationally, including sex-disaggregated data, where applicable. In contrast, there is no evidence of a national data quality governing body that meets regularly to ensure that data quality is maintained. Furthermore, the standard operating procedures and supervision for data management are inadequate. There is also a lack of common measurement metrics or indexes to monitor the progress of data or information use (Multimedia Appendix 7).

The means of verification for the "data quality and use" domain include documentation, data quality audit reports, regular technical working group minutes, practices of data exchange between systems, supervision reports, and regularly reviewed standard operating procedures. The MOH and its partners are

responsible for implementing the standards- and interoperability-related activities in this domain. Technical expertise and financial resources are required to complete the tasks mentioned in the areas of standards and interoperability.

Figure 7. A line graph showing the national District Health Information System version 2 data quality and use maturity level and targets.



Discussion

Principal Findings

Overview

The maturity assessment aims to determine the current implementation status of DHIS2 and set a road map for improvement. The results show that the DHIS2 implementation status has an average score of 2.81, indicating that it is closer to the defined (stage 3) level. This suggests that there is a standardized and basic DHIS2 process at the national level. However, the implementation of DHIS2 is limited by the lack of guidelines for relevant workforce competencies. The competencies required for DHIS2 tasks, such as knowledge and skills, are not specific to DHIS2 activities. The biggest hindrance to DHIS2 implementation in Ethiopia is the ICT infrastructure, including network and internet connectivity. This finding is consistent with the HIS maturity assessments conducted by the MOH of Ethiopia [12] and WHO [20]. It is also consistent with other African countries, such as Cape Verde, Ghana, Mali, Nigeria, and Benin, where the HIS is at the "desired" level of maturity [20].

Leadership and Governance

According to this assessment, the "leadership and governance" domain is currently closer to the defined stage of maturity, with an average score of 2.83. This indicates that the monitoring and evaluation of DHIS2 implementation lacks a clear definition as well as adherence to organizational policies, procedures, and best practices related to the digital health system. This issue is not unique to Ethiopia, as it is also prevalent in many African countries [21,22]. Situational analysis of the Africa Health Strategy 2016 to 2030 revealed that most member states have poor HIS strategies [23]. Ethiopia has just started implementing strategies to facilitate the implementation of digital health systems including DHIS2 [24]. Furthermore, various studies conducted by organizations such as WHO and the Health Metrics Network have shown that health policy is one of the weakest components of digital health and HISs in many countries, particularly LMICs [25-31]. Weak leadership and insufficient coordination are major threats to HIS implementation in LMICs [32-34]. To address these challenges, strengthening the governance and regulation of technologies, including data privacy and security and accreditation of health apps for consumers, should be prioritized for good HIS governance and leadership [35]. A robust governance framework is essential to promote HIS accountability through monitoring and regular, transparent reviews of progress and performance (MOH, unpublished data, September 2022). By prioritizing these areas, Ethiopia should work toward achieving a more effective and sustainable DHIS2 system, ultimately leading to better health care outcomes for its populations.

Management and Workforce

Regarding the "management and workforce" domain, the assessment revealed that DHIS2 implementation in Ethiopia has reached the defined stage (score=2.9). This suggests that there is a lack of regular workforce capability assessments, unclear career paths, and insufficient infrastructure, hindering the effective management and workforce practice of DHIS2 functions at the national level. The importance of effective HIS management and workforce cannot be overstated, as it is crucial for evidence-based decision-making, health service planning, and the delivery of high-quality care [36,37]. Consequently, there has been a growing interest in developing a competent HIS workforce in Africa and beyond [35,38,39]. Financing is also vital for the maintenance, development, promotion, and expansion of DHIS2, necessitating the need for an integrated system of health collaborations and programs [38,40-43]. However, studies in Ethiopia have revealed several challenges, such as poor management and governance of human resource health and weak regulatory capacity for HRH [38,44]. The MOH

has developed a 10-year HIS human resource development road map (2020-2030) and policy on eHealth architecture [45] to address these challenges. This road map provides guidance about the future need for human resources, ensuring proper national HIS support.

Regarding the "ICT infrastructure" domain, the DHIS2 implementation is currently at the repeatable stage (score=2.14). However, there are challenges such as unstandardized measurements of power outages, outdated infrastructure, and limited and unstable network and internet connectivity that hinder its effectiveness. Despite the objectives of the information revolution road map of the country to improve health care delivery through the appropriate use of ICTs, health care facilities in Ethiopia still face issues such as inadequate power sources, poor planning for replacing outdated and damaged equipment, and a lack of business continuity plan related to power supply [38,44,46]. Similar challenges have been reported in other African countries such as Kenya [47], Botswana [48], and Cameroon [49].

ICT Infrastructure

Establishing hardware and networking standards and guidelines for procurement would ensure the incorporation of related technological advancements without any hindrance to the interoperability of systems [50,51]. On the basis of the study findings, it is evident that the main determinants of DHIS2 information use are the availability of computers, communication networks, internet services, trained staff, and legislation [52]. In Africa, there is a need for an ample supply of computers, networks, internet services, and other accessories as well as training staff to boost ICT infrastructural prerequisites for the proper functioning of DHIS2 [53]. Particularly in Ethiopia, communication gaps between the internet service providers and health institutions, lack of follow-up and lack of technical support from the MOH and regional health bureaus, lack of regular network and internet connectivity assessment and reporting methods, and redundant internet or wide area network connection options are among the major challenges [44].

Standards and Interoperability

"Standards and interoperability" is another domain of DHIS2 implementation that is at the defined stage (score=3). Despite the availability of eHealth architecture that guides the interoperability of digital health systems, limited documentation has been found in Ethiopia outlining the standards for data exchange [45]. Health data exchange and harmonization rely on registry services, but in Ethiopia, there are limitations in the regular update and feedback process of the implemented core services. A client registry has not been developed, and Ethiopia does not have a national digital identification program [44]. Consistent with the call for developing standardized indicators by the World Health Assembly Resolution 63.16 [54], the Ethiopian MOH publishes health and health-related indicators annually [55], which is crucial for facilitating personal and aggregate data exchange.

Data Quality and Use

DHIS2 implementation has reached the defined stage (score=3) of maturity in terms of data quality and use. This means that there are regular data quality assessments and audits; quarterly dissemination of analytics and findings to the stakeholders; and automated data reporting using DHIS2, including sex-disaggregated data, where applicable, at all health facilities. However, there are still gaps in the national data quality governing body meeting conducted on an irregular basis to maintain data quality. Furthermore, there is a lack of common measurement metrics or indexes to monitor the progress of data or information use. Researchers agree that this lack of data quality and use is a major problem in LMICs [56,57] and needs to be addressed to solve the complex global health challenges. Thus, efforts to build a culture of data quality and use should be prioritized to ensure the effective use of data for improved health outcomes.

Limitations

This assessment has some limitations. The assessment was conducted among participants with different levels of experience and knowledge regarding DHIS2 implementation in Ethiopia. Therefore, the rating might be affected by this variation. We have attempted to minimize this limitation by letting participants have more time to discuss their differences and reach consensus. We did not collect data about DHIS2 implementation at the health facility level due to resource limitations. Therefore, the actual implementation status might not be reflected. However, we have attempted to involve participants who closely support and monitor DHIS2 implementation at the facility level.

Conclusions and Recommendations

The study used the SOCI maturity model toolkit to assess the maturity level of DHIS2 implementation in Ethiopia. The study found that DHIS2 implementation was at the defined stage of maturity in 4 of 5 domains, with a plan to move to the managed level (score=4.09) by 2025 by addressing the identified gaps.

The development of a standardized and basic DHIS2 process at the national level and the development of a 10-year strategic plan to guide the implementation of digital health systems, including DHIS2, were identified as strengths. However, the lack of workforce competency guidelines to support the implementation of DHIS2 and the gaps in knowledge and skills required to accomplish DHIS2 tasks at all levels of the health system were found to be the challenges in successfully implementing DHIS2.

The study also found that the country was in the emerging phase in terms of interoperability and data use components. Therefore, we recommend the following:

1. Clearly defined regulatory body, processes, and procedures to ensure compliance with DHIS2 standards, policies, and regulations should be established nationally. A process to review, validate, and enforce the implementation of policies, legislation, and regulations in DHIS2 should be regularized and updated as necessary to reflect the changes within the health domain. Metrics regarding DHIS2 compliance and

noncompliance should be collected, recorded, and reported regularly.

2. A platform to review DHIS2 users' competencies should be established to ensure continuous performance improvement and alignment with health care goals. This can be achieved through annual knowledge and skill evaluations and certification of the workforce. Planning, human capacity requirements, and availability related to digital health systems such as DHIS2 should be continuously improved based on the national digital health strategic plan.
3. Clear national plans and procedures for network management should be established. A dedicated network support team should be put in place at least at the district level. This will sustain the implementation of DHIS2 in the relevant facilities and offices at all levels. In addition, connectivity gaps should be documented and addressed as part of a standard process.
4. Integration of the DHIS2 data exchange and management process into the national HIS or health plan should be tracked, monitored, and reviewed through a standardized process. Moreover, it would be beneficial to have an interoperability laboratory for new exchange partners to test or for onboarding and to have a certification process.
5. Computerized alerts and reminders to managers, care providers, and patients; clinical guidelines; condition-specific order sets; focused patient data reports and summaries; documentation templates; diagnostic support; and contextually relevant reference information, among other tools, should be provided by integrating decision support apps into the DHIS2 system.

Overall, the recommendations aimed to improve the DHIS2 implementation in Ethiopia by addressing the identified gaps and enhancing the strengths. By implementing these recommendations, the country can improve the efficiency and effectiveness of its HIS, which can lead to better health outcomes for its population. However, the successful implementation of the recommendations depends on the prioritization and the capacity of the ministry and its stakeholders to allocate the necessary budget and resources.

Acknowledgments

The authors would like to thank the participants involved in the District Health Information System version 2 implementation maturity assessment process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison of features of Information Systems for Health Maturity Assessment Toolkit and Stages of Continuous Improvement for maturity assessment and improvement planning purposes.
[DOCX File , 18 KB - medinform_v12i1e50375_app1.docx]

Multimedia Appendix 2
Maturity levels of each domain, component, and subcomponent for the District Health Information System version 2 implementation.
[DOCX File , 19 KB - medinform_v12i1e50375_app2.docx]

Multimedia Appendix 3
District Health Information System version 2 road map development for the leadership and governance domain.
[DOCX File , 20 KB - medinform_v12i1e50375_app3.docx]

Multimedia Appendix 4
District Health Information System version 2 road map development for the management and workforce domain.
[DOCX File , 18 KB - medinform_v12i1e50375_app4.docx]

Multimedia Appendix 5
District Health Information System version 2 road map development for the information and communication technology infrastructure domain.
[DOCX File , 18 KB - medinform_v12i1e50375_app5.docx]

Multimedia Appendix 6
District Health Information System version 2 road map development for the standard and guideline domain.
[DOCX File , 18 KB - medinform_v12i1e50375_app6.docx]

Multimedia Appendix 7
District Health Information System version 2 road map development for the data quality and use domain.
[DOCX File , 21 KB - medinform_v12i1e50375_app7.docx]

References

1. Lippeveld T, Sauerborn R, Bodart C. Design and implementation of health information systems. World Health Organization. 2000. URL: https://www.healthdatacollaborative.org/fileadmin/uploads/hdc/_recycler_/Design_and_implementation_of_HIS.pdf [accessed 2024-03-16]
2. de Savigny D, Adam T. Systems thinking for health systems strengthening. World Health Organization. 2009. URL: https://iris.who.int/bitstream/handle/10665/44204/9789241563895_eng.pdf?sequence=1 [accessed 2024-03-16]
3. Okonjo-Iweala N, Osafo-Kwaako P. Improving health statistics in Africa. Lancet 2007 Nov;370(9598):1527-1528. [doi: 10.1016/s0140-6736(07)61644-4]
4. Pundo R, Manya AS, Mburu E, Braa J. The consistency and concurrency between the Kenya HIV/AIDS program monitoring system (KePMs) and the national reporting system (DHIS2), 2012. J Health Inform Afr 2013;1(1):2013 [FREE Full text] [doi: 10.12856/JHIA-2013-v1-i1-56]
5. United Nations General Assembly resolution 70/1. United Nations. 2015. URL: https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf [accessed 2024-03-16]
6. Health management information system (HMIS)/ monitoring and evaluation (M and E) strategic plan for Ethiopian health sector. Federal Ministry of Health, Addis Ababa. 2008. URL: https://www.cmpethiopia.org/content/download/478/2765/file/Health%20Managment%20Information%20System%20(HMIS).pdf [accessed 2024-03-16]
7. Information revolution roadmap. Minister of Health. URL: https://e-library.moh.gov.et/library/index.php/moh-resource-library/ [accessed 2024-03-16]
8. Health sector transformation plan (2008-2012 EFY) 2015/16-2019/20. Ministry of Health, Addis Ababa. 2015. URL: https://www.cmpethiopia.org/content/download/2268/9612/file/HSTP%20Final%202015-10-19.pdf [accessed 2024-03-16]
9. Health metrics network framework and standards for the development of country health information systems. World Health Organization. 2006. URL: https://paris21.org/sites/default/files/hmn_framework_1.65.pdf [accessed 2024-03-16]
10. Health information system strategic plan (2020/21-2024/25). Ministry of health - Ethiopia. URL: http://repository.iifphc.org/handle/123456789/1665?show=full [accessed 2024-03-16]
11. Proença D, Borbinha J. Maturity models for information systems - a state of the art. Procedia Comput Sci 2016;100:1042-1049 [FREE Full text] [doi: 10.1016/j.procs.2016.09.279]
12. Biru A, Birhan D, Melkamu G, Gebeyehu A, Omer AM. Pathways to improve health information systems in Ethiopia: current maturity status and implications. Health Res Policy Syst 2022 Jun 29;20(1):78 [FREE Full text] [doi: 10.1186/s12961-022-00860-z] [Medline: 35768819]
13. Health information systems interoperability maturity toolkit. MEASURE Evaluation. 2017. URL: https://www.measureevaluation.org/tools/health-information-systems-interoperability-toolkit.html [accessed 2024-03-16]
14. IS4H basics. Pan American Health Organization. 2022. URL: https://www3.paho.org/ish/index.php/en/is4h-basics [accessed 2024-03-16]
15. Digital maturity assessment tool. AARHUS University. URL: https://dbd.au.dk/ [accessed 2024-03-16]
16. Home page. MEASURE Evaluation. URL: https://www.measureevaluation.org/resources/publications/ [accessed 2024-03-16]
17. Hospital information system maturity model (HISMM). United States Agency for International Development (USAID). URL: https://www.measureevaluation.org/resources/publications/tl-17-03a.html [accessed 2024-03-16]
18. HIS stages of continuous improvement toolkit. United States Agency for International Development (USAID). 2019. URL: https://www.measureevaluation.org/his-strengthening-resource-center/his-stages-of-continuous-improvement-toolkit.html [accessed 2023-05-17]
19. Home page. Ministry of Health (MOH), Ethiopia. URL: https://e-library.moh.gov.et/library/index.php/moh-resource-library/ [accessed 2024-03-16]
20. Global digital health index. World Health Organization. URL: https://www.digitalhealthindex.org/ [accessed 2024-03-16]
21. Framework and standards for country health information systems. World Health Organization. 2008. URL: https://www.who.int/publications/i/item/9789241595940 [accessed 2024-03-16]
22. Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. World Health Organization. 2010. URL: https://iris.who.int/bitstream/handle/10665/258734/9789241564052-eng.pdf [accessed 2024-03-16]
23. Africa health strategy 2016-2030. The African Union Commission. 2030. URL: https://au.int/sites/default/files/pages/32895-file-africa_health_strategy.pdf [accessed 2024-03-16]
24. Digital health blueprint. Ministry of Health, Ethiopia. 2021. URL: https://tinyurl.com/yndskb63 [accessed 2024-03-16]
25. Gerber T, Seebreghts C. Aligning eHealth initiatives for results. Global HIT. 2016. URL: https://www.slideserve.com/joneslaura/aligning-ehealth-initiatives-for-results-study-summary-powerpoint-ppt-presentation [accessed 2024-03-16]
26. Strengthening health information systems (Resolution WHA 60.27). World Health Organization. URL: http://apps.who.int/gb/or/e/e_ss1-wha60r1.html; [accessed 2024-03-16]

27. United Nations General Assembly resolution: eHealth resolution (resolution WHA 58.28). World Health Organization. 2005. URL: http://apps.who.int/gb/or/e/e_wha58r1.html; [accessed 2024-03-16]
28. Global health initiative at a glance. U.S. Global Health Initiative. 2016. URL: http://www.ghi.gov/about/index.htm [accessed 2024-03-16]
29. Chan M, Kazatchkine M, Lob-Levyt J, Obaid T, Schweizer J, Sidibe M, et al. Meeting the demand for results and accountability: a call for action on health data from eight global health agencies. PLoS Med 2010 Jan 26;7(1):e1000223 [FREE Full text] [doi: 10.1371/journal.pmed.1000223] [Medline: 20126260]
30. Making the eHealth connection, sign on signatories. Rockefeller Foundation. 2012. URL: http://ehealth-connection.org/ehealthpetition/212 [accessed 2024-03-16]
31. Call to action: global health information forum. Global Heat Health Information Network. 2010. URL: http://www.pmaconference.mahidol.ac.th/index.php?option=com_content&view=article&id=201%3Acall-to-action-final&catid=966 %3A2010-conference&Itemid=152 [accessed 2024-03-16]
32. Alwan A, Ali M, Aly E, Badr A, Doctor H, Mandil A, et al. Strengthening national health information systems: challenges and response. East Mediterr Health J 2017 Feb 01;22(11):840-850 [FREE Full text] [doi: 10.26719/2016.22.11.840] [Medline: 28177115]
33. Glèlè Ahanhanzo Y, Ouedraogo LT, Kpozèhouen A, Coppieters Y, Makoutodé M, Wilmet-Dramaix M. Factors associated with data quality in the routine health information system of Benin. Arch Public Health 2014 Jul 28;72(1):25 [FREE Full text] [doi: 10.1186/2049-3258-72-25] [Medline: 25114792]
34. Mutale W, Chintu N, Amoroso C, Awoonor-Williams K, Phillips J, Baynes C, Population Health ImplementationTraining – Africa Health Initiative Data Collaborative. Improving health information systems for decision making across five sub-Saharan African countries: implementation strategies from the African Health Initiative. BMC Health Serv Res 2013;13 Suppl 2(Suppl 2):S9 [FREE Full text] [doi: 10.1186/1472-6963-13-S2-S9] [Medline: 23819699]
35. The Lancet Digital Health. An app a day is only a framework away. Lancet Digit Health 2019 Jun;1(2):e45 [FREE Full text] [doi: 10.1016/S2589-7500(19)30031-7] [Medline: 33323225]
36. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2024-03-16]
37. The WHO health systems framework. World Health Organization. 2013. URL: https://www.wpro.who.int/entity/health_services/health_systems_framework/en [accessed 2024-03-16]
38. National human resources for health strategic plan 2016-2025. Minister of Health, The Federal Democratic Republic of Ethiopia. 2016. URL: https://pdf.usaid.gov/pdf_docs/PA00TWMW.pdf [accessed 2024-03-16]
39. Geresu T, Shiferaw M, Mitike G, Mariam DH. Commentary: a brief review of the draft human resources for health strategic plan, Ethiopia; 2009-2020. Ethiop J Health Dev 2013;27(1) [FREE Full text]
40. Khan MH, Cruz VO, Azad A. Bangladesh's digital health journey: reflections on a decade of quiet revolution. WHO South East Asia J Public Health 2019 Sep;8(2):71-76 [FREE Full text] [doi: 10.4103/2224-3151.264849] [Medline: 31441440]
41. Australia's national digital health strategy: safe, seamless and secure: evolving health and care to meet the needs of modern Australia. Australian Government and Australian Digital Health Agency. 2018. URL: https://www.digitalhealth.gov.au/sites/default/files/2020-11/Australia%27s%20National%20Digital%20Health%20Strategy%20-%20Safe%2C%20seamless%20and%20secure.pdf [accessed 2024-03-16]
42. Understanding the influence of health information system investments on health outcomes in Côte D'ivoire: a qualitative study. MEASURE Evaluation. 2019. URL: https://www.measureevaluation.org/resources/publications/tr-19-332.html [accessed 2024-03-16]
43. WHO. The Kampala declaration and agenda for global action. World Health Organization. 2008. URL: https://www.who.int/publications/i/item/9789241596725 [accessed 2024-03-16]
44. Pathways to improve health information system in Ethiopia. Minister of Health, Ethiopia. 2021. URL: http://repository.iifphc.org/bitstream/handle/123456789/1689/Pathways-to-Improve-Health-Information-System-in-Ethiopia.pdf?sequence=1&isAllowed=y [accessed 2024-03-16]
45. Ethiopian eHealth architecture 2019. Minister of Health, Ethiopia. 2019. URL: http://repository.iifphc.org/bitstream/handle/123456789/1658/Ethiopian-Digital-Health-Blueprint.pdf?sequence=1&isAllowed=y [accessed 2024-03-16]
46. Guideline for National Health Data Access and Sharing 2021. Minister of Health, Ethiopia. URL: https://ndmc.ephi.gov.et/download/national-health-data-access-and-sharing-guideline-2021/ [accessed 2024-03-16]
47. Karuri J, Wagacha PW, Ochieng DO. Implementing a web-based routine health information system in Kenya: factors affecting acceptance and use. Int J Sci Res 2014;3(9):1-9 [FREE Full text]
48. Seloilwe E, Seitio-Kgokgwe O, Mashalla Y. Utilization of the District Health Information Software (DHIS) in Botswana: from paper to electronic based system. In: Proceedings of the 2016 IST-Africa Week Conference. 2016 Presented at: ISTAFRICA '16; May 11-13, 2016; Durban, South Africa p. 1-10 URL: https://ieeexplore.ieee.org/document/7530690 [doi: 10.1109/istafrica.2016.7530690]
49. Nah F, Sæbø JI. Analysing inhibitors of integrating and routinizing health information systems for universal health coverage: the case of Cameroon. J Health Inform 2017;4(1):114-121. [doi: 10.12856/JHIA-2017-v4-i1-176]

50. Sylva P, Abeysinghe B, James CC, Jayatilake A, Lunuwila S, Sanath D, et al. A review of eHealth policies that underpin global health care digitization. Sri Lanka J Biomed Inform 2012 Jun 07;2(4):118. [doi: 10.4038/sljbmi.v2i4.2447]
51. Submission. The Medical Software Industry Association. 2020. URL: https://www.msia.com.au/resources/submissions-correspondence/ [accessed 2024-03-11]
52. Health sector strategic and investment plan (KHSSP). Republic of Kenya Minister of Health. 2013. URL: https://extranet.who.int/countryplanningcycles/sites/default/files/planning_cycle_repository/kenya/kenya_health_strategic_plan2.pdf [accessed 2024-03-16]
53. Kuyo RO, Muiruri L, Njuguna S. Organizational factors influencing the adoption of the district health information system 2 in Uasin Gishu County, Kenya. Int J Med Res Health Sci 2018;7(10):48-57 [FREE Full text]
54. Global code of practice on the international recruitment of health personnel. World Health Organization. 2010. URL: https://cdn.who.int/media/docs/default-source/health-workforce/nri-2021.pdf?sfvrsn=326f3294_32&download=true [accessed 2024-03-16]
55. HMIS indictors referrence guide. Ministry of Health. 2022. URL: http://www.aau.edu.et/chs/?wpdmact=process&did=MzI4LmhvdGxpbms [accessed 2024-03-16]
56. Hotchkiss DR, Aqil A, Lippeveld T, Mukooyo E. Evaluation of the Performance of Routine Information System Management (PRISM) framework: evidence from Uganda. BMC Health Serv Res 2010 Jul 03;10(1):188 [FREE Full text] [doi: 10.1186/1472-6963-10-188] [Medline: 20598151]
57. Garrib A, Stoops N, McKenzie A, Dlamini L, Govender T, Rohde J, et al. An evaluation of the district health information system in rural South Africa. S Afr Med J 2008 Jul;98(7):549-552. [Medline: 18785397]

Abbreviations

DHIS2: District Health Information System version 2
HIS: health information system
ICT: information and communication technology
IS4H: Information Systems for Health Maturity Assessment Toolkit
LMICs: low- and middle-income countries
MOH: Ministry of Health
SOCI: Stages of Continuous Improvement
USAID: United States Agency for International Development
WHO: World Health Organization

Edited by C Lovis, C Perrin; submitted 15.07.23; peer-reviewed by N Stathakarou, A Babington-Ashaye, M Randriambelonoro; comments to author 25.09.23; revised version received 21.12.23; accepted 13.01.24; published 26.07.24.

Please cite as:
Yilma TM, Taddese A, Mamuye A, Endehabtu BF, Alemayehu Y, Senay A, Daka D, Abraham L, Tadesse R, Melkamu G, Wendrad N, Kaba O, Mohammed M, Denboba W, Birhan D, Biru A, Tilahun B
Maturity Assessment of District Health Information System Version 2 Implementation in Ethiopia: Current Status and Improvement Pathways
JMIR Med Inform 2024;12:e50375
URL: https://medinform.jmir.org/2024/1/e50375
doi:10.2196/50375
PMID:

Original Paper

Exploring Health Care Professionals' Perspectives on the Use of a Medication and Care Support System and Recommendations for Designing a Similar Tool for Family Caregivers: Interview Study Among Health Care Professionals

Aimerence Ashimwe[1], MSc; Nadia Davoody[1], MSc, PhD

Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Nadia Davoody, MSc, PhD
Karolinska Institutet
Tomtebodavägen 18 A
Stockholm, S-17177
Sweden
Phone: 46 (0)8 524 864 86
Email: nadia.davoody@ki.se

Abstract

Background: With the aging population on the rise, the demand for effective health care solutions to address adverse drug events is becoming increasingly urgent. Telemedicine has emerged as a promising solution for strengthening health care delivery in home care settings and mitigating drug errors. Due to the indispensable role of family caregivers in daily patient care, integrating digital health tools has the potential to streamline medication management processes and enhance the overall quality of patient care.

Objective: This study aims to explore health care professionals' perspectives on the use of a medication and care support system (MCSS) and collect recommendations for designing a similar tool for family caregivers.

Methods: Fifteen interviews with health care professionals in a home care center were conducted. Thematic analysis was used, and 5 key themes highlighting the importance of using the MCSS tool to improve medication management in home care were identified.

Results: All participants emphasized the necessity of direct communication between health care professionals and family caregivers and stated that family caregivers need comprehensive information about medication administration, patient conditions, and symptoms. Furthermore, the health care professionals recommended features and functions customized for family caregivers.

Conclusions: This study underscored the importance of clear communication between health care professionals and family caregivers and the provision of comprehensive instructions to promote safe medication practices. By equipping family caregivers with essential information via a tool similar to the MCSS, a proactive approach to preventing errors and improving outcomes is advocated.

(JMIR Med Inform 2024;12:e63456) doi:10.2196/63456

KEYWORDS

eHealth; telemedicine; mobile health; mHealth; medication management; home care; family caregivers; mobile phone

Introduction

Background

As the population ages, the demand for effective care solutions in home care intensifies, necessitating innovative approaches to medical care services [1]. According to studies [2,3], older adults often face complex health conditions and cognitive impairments; thus, drug therapy becomes more complex, and with an increasing number of new prescription drugs being approved and patients requiring more potent and sophisticated treatments, the likelihood of errors also increases. Studies show high rates of medication errors, with 23% to 92% in outpatient settings and nearly 50% after discharge, highlighting the need for effective interventions [4-6] Recently, the role of family caregivers in supporting older adults with daily activities,

including medication management, has become vital. Recognizing this, there is a growing trend to develop digital supportive systems that lighten the burden on family caregivers while enhancing the well-being of the older adults under their care. Family caregivers often struggle with the challenges of managing multiple responsibilities, navigating complex medication regimens, and ensuring adherence to promote medication without harm, all while managing their loved ones' cognitive conditions. In response to this pressing need, there is a call for proactive measures aimed at empowering family caregivers with tools and resources tailored to their unique needs [7-11].

In the realm of health care, eHealth refers to the integration of digital technologies and information and communications technologies into health care systems and processes. This integration aims to improve the delivery of health care services, enhance patient care, and streamline administrative tasks [12,13].

The adoption of eHealth solutions enables better coordination, data-driven decision-making, and personalized medicine, benefiting health care professionals, providers, patients, and family caregivers alike. Health care professionals gain improved access to patient data and tools for remote consultation, whereas providers streamline administrative tasks and enhance care delivery and patients, together with family caregivers, benefit from increased engagement, remote monitoring, and personalized interventions, leading to better health outcomes [1,14-16]. eHealth solutions empower patients to actively participate in their care, access educational resources, and engage in teleconsultations with health care professionals from the comfort of their homes. Moreover, digital health streamlines administrative tasks and improves care coordination and communication among multidisciplinary teams [17-20]. Building on these advancements, the role of eHealth solutions in home care has become increasingly important, particularly in countries such as Sweden, where a growing aging population receives care in their homes.

Home Care in Sweden and the Medication and Care Support System

Home care in Sweden is a vital component of the country's health care system, aiming to provide support and assistance to individuals who require medical or personal care in their own homes. Home care services in Sweden are organized and provided by municipal authorities, intending to promote independence, autonomy, and quality of life for individuals who may have difficulties and challenges with daily activities due to illness, disability, or aging [21,22].

Despite its well-structured and person-centered approach, several challenges persist for both patients and their family caregivers. One prominent challenge is the increasingly complex nature of medication management, particularly for older adults with multiple chronic conditions or cognitive impairments. Ensuring adherence to medication regimens, preventing medication errors, and managing potential drug interactions can pose significant hurdles. In addition, the evolving needs of individuals requiring home care demand a higher level of coordination and integration among health care providers, family caregivers, and support services, often leading to fragmented care delivery and gaps in communication [3,23,24].

Among these challenges, the implementation of a medication management system emerges as a crucial solution to enhance the efficiency, safety, and quality of care within the home care context. By leveraging technology and tailored support mechanisms, a medication and care support system (MCSS) can assist health care professionals in the process of medication management. Moreover, a medication management system can facilitate seamless communication and collaboration among health care professionals and family caregivers, fostering a more integrated and holistic approach to care delivery [25,26]. By addressing the complexities of medication management and promoting collaboration among stakeholders, an MCSS has the potential to enhance the effectiveness and sustainability of home care services, enabling individuals to age in place with dignity and independence [27,28]. One example of a digital tool that embodies this collaborative approach to medication management is the MCSS developed by Vitec Appva [29] and widely implemented in Sweden's home care services. The MCSS supports health care professionals in home care in medication administration, dosage tracking, and medication reconciliation, thereby reducing the risk of adverse drug events (ADEs) and promoting medication adherence. In addition to its core functions, the MCSS is extensively used by health care professionals across most home care settings in Sweden for task management, including inputting signing lists and tracking the completion of assigned duties. It is also used to simply see what the health care professionals need to do and sign off once they have completed it. It plays a crucial role in medication management in home care. With the widespread adoption of smartphones and tablets, health care professionals can access MCSS platforms from anywhere, allowing for real-time monitoring of medication schedules, reminders, administration logs, and communication tools [29]. The nurses have access to the MCSS through both a mobile app and a web platform (Figure 1 [29]). These tools allow them to customize and oversee various activities related to medication administration and patient care. With the mobile app, health care professionals can conveniently access the system from their smartphones or tablets, enabling them to stay connected and manage tasks on the go. Similarly, the web platform provides a robust interface accessible from computers or laptops, offering additional functionalities for detailed monitoring and analysis. This dual accessibility empowers nurses to tailor their approach to medication management, effectively monitor patient activities regardless of their location or device preference, assist in administrative tasks, and maintain the security of medical information [29].

Figure 1. Mobile and web-based medication and care support system platform. The figure was inspired by the MCCS developed by Vitec Appva [30].

However, despite the significant benefits offered by these digital solutions, challenges in medication management still persist. Studies [2,16,30,31] have shown that, even with the availability of advanced digital tools, family caregivers continue to deal with medication mismanagement in their complex caregiving responsibilities and limited communication with health care professionals. While health care professionals use the MCSS as a tool to administer medications and effectively communicate with each other, there remains a notable gap in research regarding whether a tool similar to the MCSS can also be leveraged by family caregivers to manage multiple medications, thereby mitigating medication errors in home care practices.

Studies [3,32-36] indicate that, as the population ages and drug therapy becomes more complex—with more new prescription drugs being approved and patients requiring increasingly potent and sophisticated treatments—the likelihood of errors also increases. Studies have revealed alarming statistics on medication errors and ADEs, demonstrating their profound impact on patient safety, health care providers, and health care systems. Studies [32-36] have discussed medication errors, their causes, prevention methods, and risk management strategies within the pharmacy and health care industry context. They have highlighted the importance of medication therapy as a valuable tool in treating various health conditions but also acknowledged the risks associated with its misuse. Examples of medication errors outlined in the studies include administering the wrong drug, wrong strength, or wrong dose; confusion over similar–looking or sounding drugs; incorrect routes of administration; miscalculations; and errors in prescribing and transcription. Key findings indicate that significant risk factors for medication errors include patient age of ≥60 years, inexperienced caregivers, polypharmacy (≥5 drugs), comorbidities, and multiple prescribers, which are critical to address for improving patient safety in home care [28]. In addition, a study [32] on 16,963 emergency department visits found that 3.4% were due to ADEs, with 15.1% leading to hospitalization, ADEs extending hospital stays by nearly 2 days, and medication errors adding almost 5 extra days, significantly increasing health care costs. This study emphasized the importance of implementing strategies to prevent medication errors, such as improving medication safety protocols, enhancing care delivery training, using technology for error detection and prevention, and fostering a culture of open communication and reporting within health care organizations. It underscored the need for constant caution and proactive measures to mitigate the risks associated with medication errors and ensure patient safety.

Aim of the Study

This project aimed to investigate health care professionals' perspectives on the use of the MCSS and gather recommendations for designing a similar tool for family caregivers. The focus was on identifying the necessary information and features essential for both health care professionals and family caregivers.

Methods

Study Design

In this study, an exploratory qualitative approach was adopted to investigate health care professionals' experience of using the MCSS and recommendations for designing a similar tool for family caregivers in home care settings. To gather comprehensive insights, the researchers explored the perspectives and recommendations of health care professionals actively using the MCSS in health care delivery within home care settings [25,37,38]. The research methodology used an inductive approach, focusing on deriving insights and conclusions from collected data. Thematic analysis was conducted according to the guidelines by Braun and Clarke [38], systematically identifying recurring themes and patterns in the interview data.

Study Setting and Participants

The interviews involved a total of 15 health care professionals and took place between February 2024 and May 2024 in Sweden. In this study, we involved health care professionals

working in home care in the Nora municipality. The Nora municipality is located in the northern part of Örebro County and has approximately 10,700 inhabitants. The municipality is responsible for, among other things, home care for people who need continued health care at home. Nurses, assistant nurses, and care assistants work in home health care [39]. The study participants' characteristics encompassed various factors, including demographic details, age distribution, tenure in home care, and experience with MCSS use. Factors such as diversity of perspectives, expertise, and experiences were considered.

The head of Nora Home Care was contacted and assisted with obtaining access to potential participants among the home care employees. Subsequently, the researchers personally met with these employees to provide a thorough explanation of the study's purpose. Following this, interested employees who met the study's inclusion criteria conveyed their willingness to participate via email. The participants' eligibility was determined by their expertise in medication administration, encompassing individuals such as registered nurses, nurse assistants, care assistants, and therapists employed at Nora Home Care actively using MCSS for >2 years. These criteria helped ensure that participants possessed relevant and comprehensive experiences, perspectives, and recommendations related to the research topic while maintaining the quality and validity of the study findings [36,37,40]. The researchers sent out emails to potential participants with further details about the study objectives and attached the invitation letter along with the letter of consent. In addition, a QR code linked to a Google Form was included to facilitate scheduling for interview dates and times. Participants' characteristics are presented in Table 1.

Table 1. Participant characteristics.

	Age category (y)	Occupation	Years of experience in home care	Years of MCSS[a] use
Participant 1	45-55	Registered nurse	10-20	5-10
Participant 2	45-55	Registered nurse	30-40	5-10
Participant 3	45-55	Registered nurse	20-30	5-10
Participant 4	30-40	Occupational therapist	10-15	5-10
Participant 5	20-30	Nurse assistant	5-10	5-10
Participant 6	40-50	Nurse assistant	20-40	5-10
Participant 7	55-65	Nurse assistant	40-50	5-10
Participant 8	40-50	Nurse assistant	10-15	5-10
Participant 9	40-50	Nurse assistant	10-15	5-10
Participant 10	45-55	Nurse assistant	10-20	5-10
Participant 11	25-35	Nurse assistant	5-10	5-10
Participant 12	35-45	Nurse assistant	10-15	5-10
Participant 13	40-50	Nurse assistant	5-10	5-10
Participant 14	20-30	Carre assistant	2-5	2-5
Participant 15	35-45	Care assistant	2-5	2-5

[a]MCSS: medication and care support system.

Data Collection

Fifteen interviews with health professionals administering medication in home care settings were conducted. Their responses centered on various aspects of medication management, including their own experiences, the role of family caregivers, the use of the MCSS tool, necessary information for caregivers, information needed by health care professionals in the MCSS from family caregivers, and recommended features for improving the tool and for designing a similar tool for family caregivers.

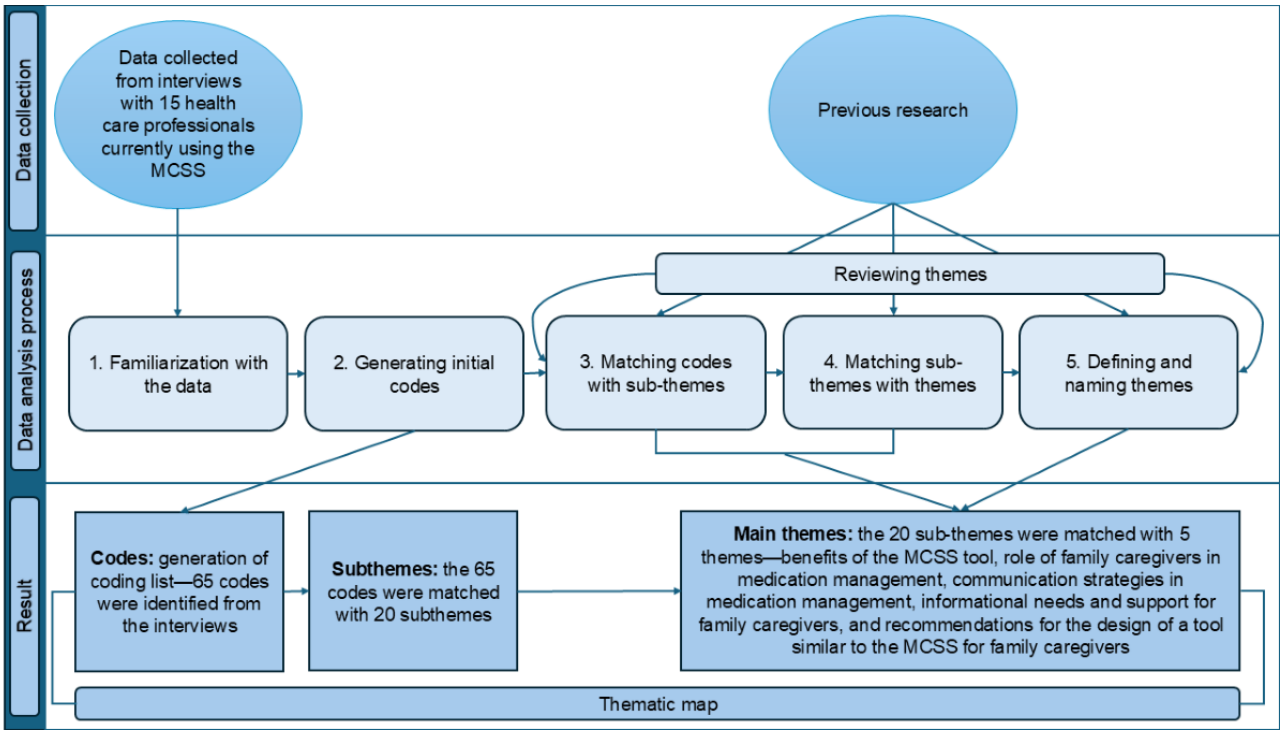
A pilot interview with a group of 3 nurses was conducted to assess the clarity of the research aim and interview questions. Feedback from the pilot test informed any necessary revisions. Each participant was required to sign an informed consent form before the interviews started. The interviews were recorded, and notes were taken to ensure the accuracy and comprehensiveness of the information gathered. Each interview lasted approximately 30 to 40 minutes. Given that our study included 15 participants, we closely monitored the data during collection and analysis to ensure that no new themes or insights were emerging. This point of saturation was reached when additional interviews did not produce any new or relevant information. The interviews were transcribed verbatim.

Data Analysis

Thematic analysis was chosen as the method to analyze the data due to its inherent flexibility and systematic framework for interpreting data from qualitative research approaches. This approach facilitated the identification of recurring patterns or themes within the dataset, allowing for a nuanced exploration of the participants' perspectives and experiences. The thematic analysis procedure, as outlined in Figure 2, inspired by the work by Matt et al [41], was guided by the step-by-step approach by Braun and Clarke [38], providing a structured methodological framework for conducting the analysis. This methodology

ensured a comprehensive examination of the data, enabling the extraction of meaningful insights that contributed to addressing the research questions effectively.

Figure 2. Data analysis process. MCSS: medication and care support system.



The analysis began with familiarization with the data, where the first author immersed herself in the data to gain a comprehensive overview. This first step involved reading through all the transcripts of the interviews multiple times to become familiar with their content. During this phase, the author took notes, highlighted sections, and looked for recurring patterns. This step was crucial for gaining insights into the data's context. The second step was to generate initial codes, which started by identifying specific segments that were relevant to the research question and objectives. The first author assigned initial codes to label them based on their content. The third step included matching codes with subthemes—during this step, both authors reviewed the initial codes and began to group them into clusters that reflected subthemes. A total of 20 identified subthemes represented a more specific aspect of the data compared to the broader themes. In this step, the authors examined relationships between codes and identified patterns that suggested subthemes. The fourth step was matching subthemes with themes, in which the authors combined related subthemes into broader themes that captured significant patterns in the data. The authors reviewed how subthemes fit together to form comprehensive themes that represented the major findings of the research. The last step was to describe what the themes represented and choose concise and descriptive names for the themes that accurately reflected their content. The data analysis was an iterative process involving creating new codes, subthemes, and themes based on emerging evidence.

Ethical Considerations

This research was carried out in Sweden. According to the Swedish Ethical Review Act, the research in our submitted manuscript does not require ethics approval as it does not handle sensitive personal information (as understood by the European General Data Protection Regulation). However, ethical requirements still apply. Consequently, prospective study participants were contacted after receiving an email invitation outlining the study's purpose. Arrangements for interviews were made via email, SMS text message, or phone call accommodating participants' preferred timing and location. Each interview session began with a comprehensive explanation of the study's aim, interview process, and participant rights. Before participation, all individuals provided written and verbal consent, with formal documentation signed by both the participant and the researchers. This consent form detailed study objectives, potential risks, confidentiality, and the right to withdraw. Participants were assured of anonymity and informed about how their data would be handled [42,43]. No compensation was provided to the participants in this study.

Results

Overview

Altogether, 5 overarching themes, 20 subthemes, and 65 codes were generated (Textbox 1 and Figures 3-7). These will be showcased alongside their corresponding summarized meaning units, with participants identified by their designated number following their quotations.

Textbox 1. An overview of the themes and subthemes.

Benefits of the medication and care support system tool

- Verification and cross-referencing of patient details and medication lists
- Ensuring adherence to prescribed schedules
- Monitoring medication intake and signing off on administration
- Electronic medication reminders and signing lists
- Digital access to medication schedules and administration records

Role of family caregivers in medication management

- Perceived benefits of family members administering medications
- Concerns about family caregivers' lack of training and experience
- Examples of medication mismanagement by family caregivers

Communication strategies in medication management

- Communication features between health care professionals and family caregivers
- Pretrip communication between health care professionals and family caregivers
- Communication during and after trips, primarily via phone calls
- Lack of direct communication during medication management periods

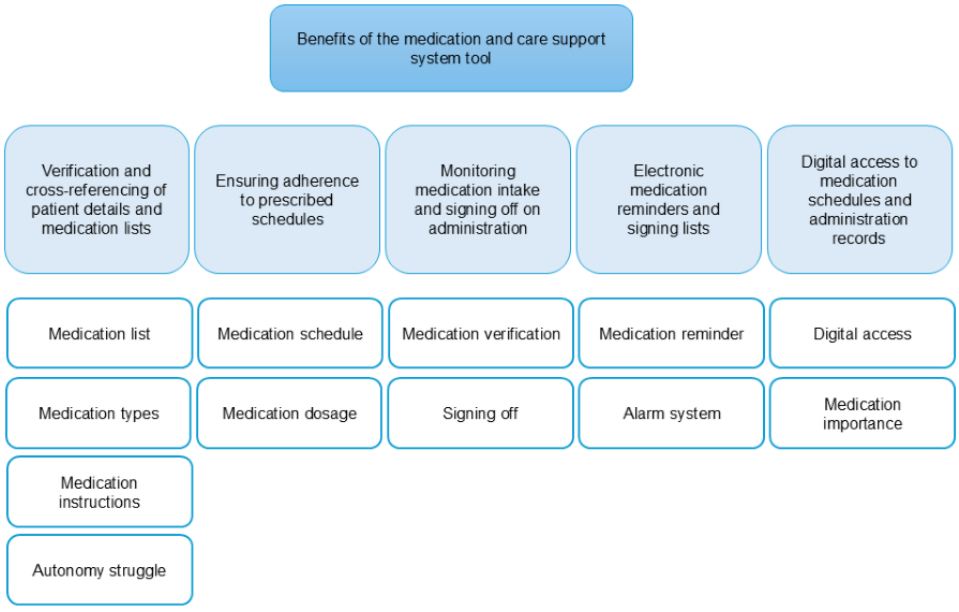
Informational needs and support for family caregivers

- Necessary information for safe medication administration
- Importance of clearly communicating medication schedules, dosages, and potential side effects
- Preferences for written information to supplement oral communication

Recommendations for the design of a tool similar to the MCSS for family caregivers

- Key features and functions of a similar MCSS for family caregivers
- Importance of a clear and simple user interface
- Privacy features to restrict access to sensitive information
- Customization options for focusing on individual patients

Figure 3. A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom.



Theme 1: Benefits of the MCSS Tool

Overview

A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom (Figure 3). The participants mentioned that the MCSS tool presents a comprehensive solution designed to streamline the medication administration process. They revealed that its functionality lies in the abilities outlined in the following sections.

Verification and Cross-Referencing of Patient Details and Medication Lists

The MCSS tool helps health care professionals verify patients' details on the app and cross-reference them with medication lists at the patients' home to ensure accuracy and consistency in medication administration:

> *We use MCCS in the municipality to ensure and know that the care recipients have received their prescribed medication. It is a way to verify that it has been done or check exactly as prescribed in the system.* [Participant 1]

> *When I arrive at the patient's [home], I open the app and go to the person who is supposed to receive medication. I verify that the name and person match, and I check the prescription list they have in their binder at home, cross-referencing it with what is displayed in the Appva MCSS app.* [Participant 6]

Ensuring Adherence to Prescribed Schedules

The participants expressed that the MCSS takes into account any specific instructions for medication timing and dosage to maintain adherence to the treatment plan. Participants discussed that this feature provides a way to track and manage medication schedules effectively for different household members. Similarly, by selecting the correct person within the app, one can view details about prescribed medications, including which medicines should be administered, the dosage instructions, and the specific times when each medication should be taken:

> *You enter the app under the right person. Then you can also check the prescription action that is available at everyone's home. It states which medicine to give, how to give it, and at what time.* [Participant 5]

The participants also described how occupational therapists and physiotherapists use the MCSS to organize training sessions and provide feedback. They mentioned that, during home training visits, the occupational therapists collaborate with physiotherapists to address patients' training needs:

> *The process, it's about when we're at home for training visits. We can say that physiotherapists are meeting someone who has a training need where staff need to be present. Then we go through what type of training program [they have], then we put in the MCSS description of the training and how often, how long approximately during the day, and then we provide feedback to both patients and staff.* [Participant 4]

Monitoring Medication Intake and Signing off on Administration

The participants discussed the MCSS's ability to monitor medication intake, and they sign off on administration on the app to confirm that the medication has been given as prescribed:

> *Once I have retrieved the medication, I observe that the person has taken or swallowed it. If I have administered the medication myself, I then sign off to confirm that it is complete.* [Participant 6]

Electronic Medication Reminders and Signing Lists

The participants appreciated the implementation of electronic medication reminders within the system that prompt health care professionals about scheduled doses. They also highlighted the importance of electronic signing lists through the MCSS to track medication administration accurately and discussed that this system promotes self-awareness by flagging instances in which medication is not taken or documented properly, ensuring optimal care:

> *An alert may come up. I can see that the patient or care recipient may not have taken their medication. Perhaps they have taken it, but it's not signed for. This is to ensure good self-awareness in medication management...* [Participant 1]

Digital Access to Medication Schedules and Administration Records

The participants emphasized the value of digital access to medication schedules and administration records, enabling health care professionals to have comprehensive and up-to-date information about medication management:

> *Is a digital platform that facilitates the management of medication and care for our patients. Through MCSS, we can schedule medication, keep track of dosages, and improve communication among healthcare professionals. Within the system, you can see whether medications have been administered or not.* [Participant 10]

Following the exploration of the benefits provided by the MCSS tool, it is equally important to examine the crucial role that family caregivers play in medication management. This next section delves into the perceived benefits and challenges faced by family caregivers, highlighting their experiences and the potential risks associated with their involvement in administering medications.

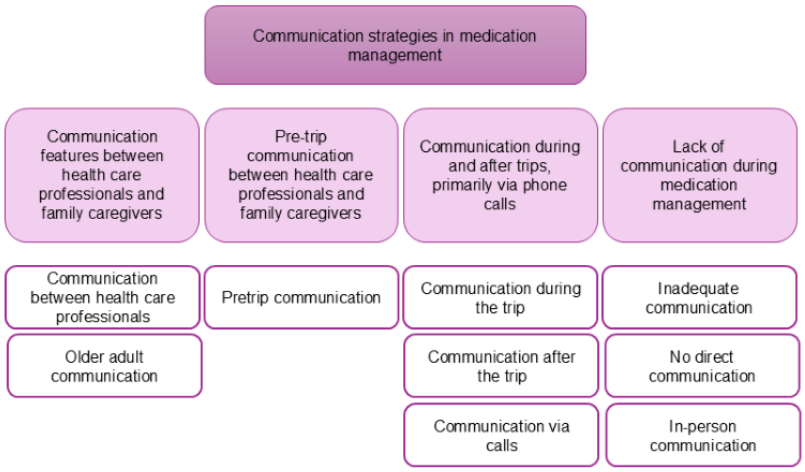
Theme 2: Role of Family Caregivers in Medication Management

Overview

A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom (Figure 4). The participants described how family members may step in to take on tasks typically performed by health care professionals and acknowledged the significant role of family members in the care of the patients. However,

the participants exposed some challenges that arose when family caregivers had the responsibility of administering medication.

Figure 4. A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom.



Perceived Benefits of Family Members Administering Medications

Participants recognized the value of family members assisting with medication administration and acknowledged that this involvement fosters a sense of collaboration and shared responsibility in the patient's care:

> *In general, it is considered good that relatives take responsibility for administering medications. It allows the patient to have autonomy and independence. That one is not so dependent on the assistance of healthcare personnel.* [Participant 1]

> *Actually, relatives take over what the staff usually do when they take care of the patient themselves. It can be temporary, for example, I am visiting my mom now at lunch and I will stay until 3 o'clock so I will give her medication at 2 o'clock.* [Participant 2]

Concerns About Family Caregivers' Lack of Training and Experience

Despite the benefits, the participants expressed concerns about the potential risks associated with family members administering medications. They noted the lack of formal training and expertise among family caregivers, which could lead to errors or omissions in medication administration:

> *Yes, it's quite a big responsibility that family members take on. It's important that they receive enough information, so they know which medications to give and at what times. There may also be a medication that shouldn't be administered if the person's condition is worse or changes in a certain way. It's quite a significant responsibility to assume...It actually carries the same risks as healthcare professionals. But the risk becomes greater because family members are inexperienced and cannot read medication lists. They don't have the training that healthcare professionals have.* [Participant 3]

Examples of Medication Mismanagement by Family Caregivers

Participants shared examples in which medication mismanagement occurred when family caregivers were responsible for administration. These examples included cases in which medications were not taken as prescribed or doses were missed altogether. They also speculated that the family caregivers may not have fully understood the patient's cognitive state or failed to double-check medication adherence. They also noted that those incidents highlighted the importance and need for clear communication and caution in medication management even when delegated to family members:

> *There was one incident that comes to mind where a woman we sent home with her medication, as her relatives had informed us, she would be going home to her daughter and taking her medication with her. But when she returned home after about a week, she hadn't taken any of her medication. She had medication three times a day...In this situation, I think that perhaps the relatives hadn't realized that the person wasn't entirely clear in her thinking. Or maybe they didn't double-check properly...but in this case, the person hadn't taken their medication. So, she brought half of the pill dispenser back home.* [Participant 6]

> *It has happened to me actually. I sent medication to a patient with dementia. And the family was supposed to take responsibility for the medication. We sent the medication box that he should take, and it was only medication for two days, but the patient took all the medication. We don't know if he took all the medications or if he threw them away...This is a bit of a problem if the family members know nothing about the medication. There are always problems with family members and patients when they are going to leave or visit.* [Participant 9]

Building on the discussion of family caregivers' roles in medication management, it is essential to consider how communication strategies impact this process. The following

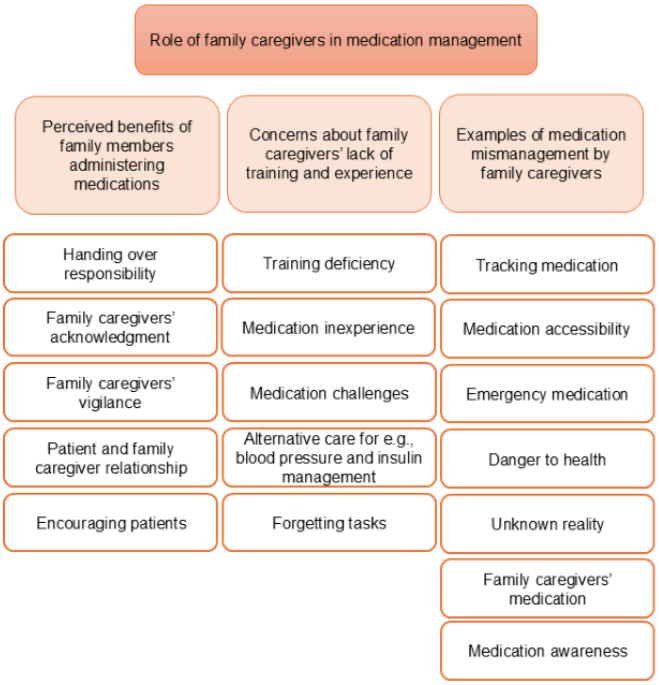
section explores the various communication features and challenges between health care professionals and family caregivers, focusing on the critical moments before, during, and after trips and the implications of inadequate direct communication during the medication management period.

Theme 3: Communication Strategies in Medication Management

Overview

A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom (Figure 5). The participants described communication strategies used between health care professionals and family caregivers typically occurring before, during, and after trips, primarily in person or through phone calls. However, the participants mentioned that they had been contacted only before the trips but never been contacted while the patient was away, whereas both family caregivers and care recipients are encouraged to communicate with health care professionals whenever they have questions if they find any changes or seek advice regarding medication management.

Figure 5. A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom.



Communication Features Between Health Care Professionals and Family Caregivers

The participants discussed the situation regarding communication between health care professionals and family caregivers:

> *Currently, we communicate with family members regarding medication administration through phone calls and in-person meetings before they travel with the patient.* [Participant 11]

Pretrip Communication Between Health Care Professionals and Family Caregivers

The participants discussed the importance of communication between health care professionals and caregivers before a trip. They mentioned that, typically, communication occurs before the trip, where caregivers inform health care professionals about their travel plans and inquire about managing medications during their absence:

> *Yes, we plan together if someone is going away for two weeks, we say, so then it's the nurse who is informed and dispenses medications for the entire period.* [Participant 3]

Communication During and After Trips, Primarily via Phone Calls

The participants highlighted a lack of ongoing communication during and after trips. They expressed that, during trips, there was no communication with family caregivers and the same happened afterward except for health care professionals possibly asking how it went. In addition, participants explained that, after trips, communication remained minimal and mentioned that this approach may not provide timely updates and not ensure that everyone involved is informed about the medication regimen while many changes may occur during the absence of the patient:

> *But during and after, we don't have any direct communication in that way. Instead, the entire responsibility is left, and it's explained beforehand. During that time, we have no communication with them, and not afterward either. Except that one can ask how it went.* [Participant 5]

Lack of Direct Communication During Medication Management Periods

The participants acknowledged challenges related to trusting family caregivers to manage medications correctly. They

mentioned the difficulty in verifying whether medications were taken properly, especially when the responsibility is entirely on the family caregiver because they cannot sign off on the medication:

> *It's a bit tricky for us as healthcare professionals; we don't receive any information about the patient, whether they have taken medication at the right time, or they haven't taken all their medications.* [Participant 9]

> *Often challenges related to them not having enough medication support. Since we are not with them, we don't know how it works for them.* [Participant 10]

The participants continued to discuss the importance of planning travels well in advance to ensure smooth coordination; otherwise, it resulted in patients not having enough medication before the trip. They mentioned the challenges regarding the fact that, in such cases, arrangements need to be made for the patient to receive the medication dispenser at the pharmacy where their family caregiver lives:

> *Also, another challenge is that when family members want to take care of the medications, they usually order upcoming medication not in good time, they always say it at the last minute, and it usually takes time to fix. It can affect the patient's illness, for example, when they are not close.* [Participant 11]

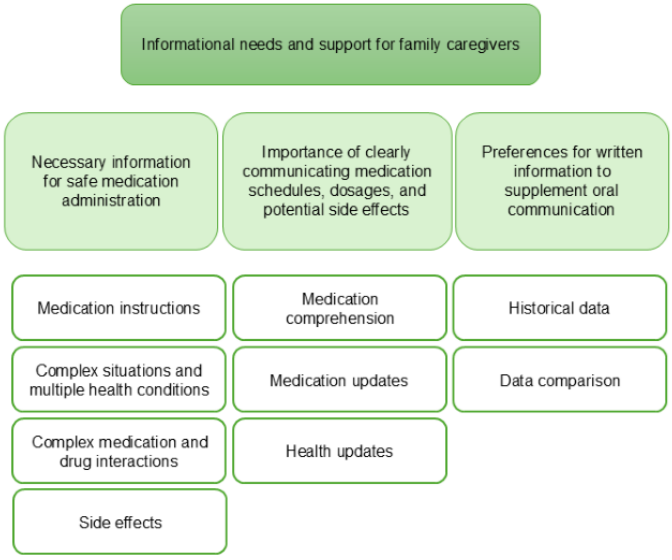
Effective communication is crucial, but ensuring that family caregivers have the right information is equally important. The next section delves into the informational needs of family caregivers, emphasizing the significance of clear instructions and understanding medication schedules and potential side effects and the preference for written materials to support oral communication.

Theme 4: Informational Needs and Support for Family Caregivers

Overview

A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom (Figure 6). The participants highlighted the need for detailed information about medication administration. They stressed the importance of providing written instructions, especially for complex situations, to ensure that family members can effectively assist with medication management.

Figure 6. A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom.



Necessary Information for Safe Medication Administration

The participants emphasized the importance of having updated information regarding any changes or updates to medication prescriptions to ensure that family caregivers and health care professionals administer the correct medications according to the latest prescriptions:

> *They can see there, for example, at nine o'clock it should be this APO on the bag, and then at twelve, it should be insulin and APO, like that.* [Participant 5]

> *They should know why they're taking the medication, which medication it is, and what illness it's for.* [Participant 12]

Participants highlighted the importance of family caregivers being updated on any changes in the medication list, such as medication names or any changes in patients' conditions that needed to be reported to health care professionals through the MCSS. They expressed that this information is crucial to health care practices to ensure that everyone is informed and involved in the patient's care:

> *Yes, one thing that can be difficult is when it says that they should give a certain medication with certain names, and then a similar product with a different name has come from the pharmacy. Then it's hard to know what it is I should give... You have to update them if there have been any changes. I always try to inform them when I can. When the patient themselves doesn't have such a good memory or understanding*

of their health condition, then I usually contact relatives and inform them. [Participant 2]

Importance of Clearly Communicating Medication Schedules, Dosages, and Potential Side Effects

The participants stressed that clear communication of medication schedules and dosages is crucial for safe administration and that the information does not need to be too much to read. They expressed that family caregivers and patients need to understand only when and how to take each medication to ensure adherence to the prescribed regimen:

> *To have information about medication, regimens, dosages, potential side effects, and what to do in case of medical emergencies. Communicating these details clearly and regularly is important to ensure safe and effective care.* [Participant 10]

They also emphasized that communicating potential side effects of medications is essential for family caregivers and patients should be aware of possible adverse reactions so that they can monitor them and seek medical attention if necessary:

> *When the patient is about to feel nauseous and vomit, maybe you understand what to do then with the medications. And sometimes that type of information needs to be written down as well, so it can be given to relatives, not just verbally; it can be a lot of information at once.* [Participant 3]
>
> *The family members have to be informed about what can sometimes happen to patients after taking such medication.* [Participant 9]

Preferences for Written Information to Supplement Oral Communication

Participants highlighted the written instructions on how to administer each medication, including dosage, route, and any special instructions or preparations that are necessary to avoid errors and make good decisions in administration:

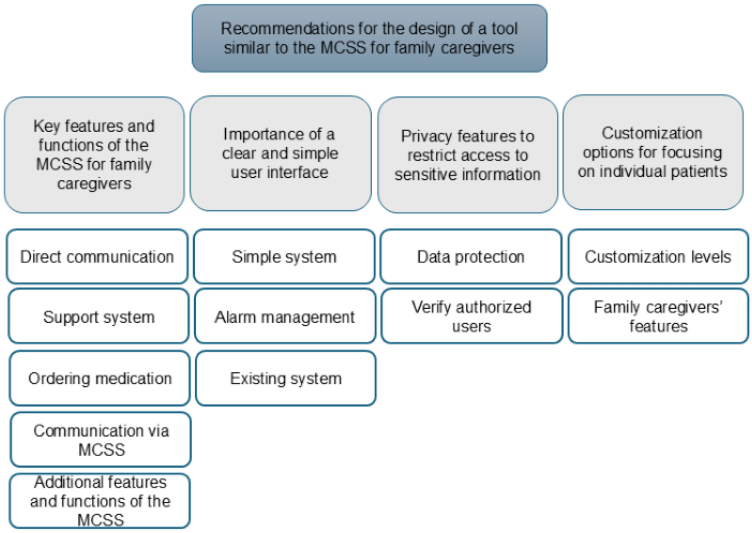
> *I recommend that all information exchanged between healthcare professionals, relatives, and patients should be in written form rather than on the phone, as it can cause misunderstandings...Relatives receive the same information about medications as the patient does, including the illness...* [Participant 13]

Building on the importance of clear communication and information, it is essential to consider how technology can support family caregivers in managing medication. The following section explores recommendations for the design of a tool similar to the MCSS, highlighting the need for key features such as a user-friendly interface, strong privacy protections, and customizable options to meet the unique needs of individual patients.

Theme 5: Recommendations for the Design of a Tool Similar to the MCSS for Family Caregivers

A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom (Figure 7). The participants recommended the design of a system similar to the MCSS for the managing of medication and health care information by family caregivers.

Figure 7. A hierarchical structure displaying the themes at the top followed by the subthemes in the middle and the corresponding codes at the bottom. MCSS: medication and care support system.



Key Features and Functions of a Similar MCSS for Family Caregivers

On the basis of the existing MCSS, participants highlighted specific features and their roles that can help family caregivers effectively manage medication.

Alarm System or Medication Reminder

Participants emphasized the importance of a reminder function within the system to prevent missed doses or overdoses. They suggested that reminders could be sent via SMS text messages or notifications to ensure timely medication administration:

> *I'm thinking that perhaps one could implement receiving a text message or something to a relative's phone when it's time for medication.* [Participant 6]

> *Electronic medication reminders are important, digital signing lists that show if medications have been administered; this gives relatives the opportunity to communicate directly with healthcare staff...These features help to increase the efficiency and safety of medication administration, as well as promote collaboration and communication between healthcare professionals and relatives.* [Participant 10]

Medication List (Dosage and Schedule)

Participants highlighted the need for easy access to a comprehensive list of medications, including both regular and as-needed (nonregular) medications. They suggested that the system should provide clear instructions regarding dosage, timing, and potential interactions to assist family caregivers in managing medications effectively:

> *If family caregivers can go in and get an overview when the care recipient has medication that you want to sign for. So, I think MCSS would help there...It's about seeing the times when medication should be given. If it's every hour or an hour before it's completely okay...Family caregivers should log in and check for as-needed medications, to see when and how often they can be given...* [Participant 1]

> *I'm thinking that as we have with MCSS, they can see there, for example, at nine o'clock it should be this APO in the bag, and then at twelve it should be insulin and APO like that.* [Participant 5]

As-Needed Medication Access

The discussion focused on the importance of medication access and patient safety. However, participants described a situation in which a family member is responsible for administering scheduled medication to a patient but access to additional medication (as needed) is restricted. If the patient suddenly needed medication during this time, the family member might not be able to provide it, leading to frustration or danger to patients:

> *...When a relative who is with a patient has received both a pharmacy bag and doses from the healthcare staff to be given at the next scheduled time, which the relatives will be responsible for and the rest is locked up in a cabinet at the patient's home and then that period begins when the patient feels unwell and needs medication, and then they don't have access to them, such as as-needed medication. So, it can be frustrating for them...In the worst cases, we lock up medications to have better control.* [Participant 2]

Participants discussed the crucial importance of medication management, particularly concerning narcotics, in home patient care. They noted that these medications are securely stored and only accessible to health care professionals due to their high vigilance requirement to avoid the potential risks associated with narcotics. This limited access can pose challenges during emergencies when family caregivers are responsible for medication administration.

Medication Instructions

Participants suggested that the system should provide detailed information about each medication, including its purpose, administration instructions, and potential side effects. They emphasized the importance of clear communication to reassure family caregivers and ensure safe medication administration:

> *Yes, perhaps relatives would need to have a description of what the medication is for in MCSS because it provides a sense of security to know what I am putting into my mouth, so that there can also be some help text provided not to give in these symptoms, for example. Or consider always giving with food or not with food and such things...Or if there's a particular pill that is important to give at an exact time, one could have extra instruction for such matters, because we don't have that now, as the staff can manage it when it's not needed in the same way now. But for relatives, it would be very good; it would reassure them in that case; it would be a completely new function.* [Participant 3]

Signing off and Medication Confirmation

Participants discussed the importance of a feature that allows family caregivers to confirm medication administration. This confirmation could serve as a record for monitoring adherence and tracking progress, providing reassurance for both caregivers and health care professionals:

> *I believe that signing list for training and rehabilitation is the best feature of relatives. Then it's also the case that we have used it, sometimes I have probably used it just to get a confirmation that it's carried out. It becomes a reassurance for us when we delegate training or rehabilitation, that we see that it's being done in the way we've agreed upon with patients and relatives.* [Participant 4]

> *And that it becomes clear for them too that they could simultaneously sign so that we still get an overview that their medication has been given at the same time so that they can also check themselves that the medication has indeed been given and that they have signed that they have given it.* [Participant 5]

Preordering Medication

Participants suggested that the family caregiver's system should include a feature for preordering medications, particularly for travel preparation. This would ensure that family caregivers have an adequate supply of medications on hand even when they are away from home:

> *Yes, the other functions, like ordering medication, because when they are not around, they have to have enough medication otherwise it's difficult to deliver medication where they are.* [Participant 2]

Importance of a Clear and Simple User Interface

Participants advocated for the importance of a system that is easy to navigate, particularly for individuals lacking technical expertise, older family caregivers, or busy working individuals.

They suggested that it should display only essential information to avoid overwhelming users:

> *It shouldn't be anything complicated, it's just that the time and date are written. It doesn't need to be anything difficult...Sometimes they may think it's too much information; it should be a bit simple. If it's just a medication bag, maybe there should just be an option to press or sign, or if you're giving insulin. That it's something simple. It can also be older people who find such things difficult.* [Participant 7]

> *And it would be a system that is easier for them because some don't have any technical training.* [Participant 11]

Privacy Features to Restrict Access to Sensitive Information

The participants expressed the importance of ensuring privacy, security, and personalized access to health care information through the system. They emphasized that the system should prevent unauthorized access; only authorized individuals such as close relatives should have access to the system and specific patient data:

> *The functions to hide are the personal identification number, or so that someone else cannot log in...It's just that only closest relatives receive all the information, for example, the personal identification number and what medication it is.* [Participant 8]

Customization Options for Focusing on Individual Patients

The participants stressed the significance of customization options for focusing on specific patients, particularly for family caregivers. They highlighted the importance of ensuring that family caregivers only have access to information relevant to their particular patient rather than being able to view data for all patients, as health care professionals can:

> *Then the system must be installed on their phones. And then they should only have access to that patient and no one else.* [Participant 2]

Discussion

Principal Findings and Comparison With Prior Work

This study delved into 15 health care professionals' perspectives regarding medication management for older adults receiving home care, with a specific focus on exploring the current digital support tool being used (MCSS). This study aimed to explore health care professionals' perspectives and gather recommendations regarding the design of a similar tool for family caregivers. The findings identified 5 key themes: benefits of the MCSS, role of family caregivers in medication management, communication strategies in medication management, informational needs and support for family caregivers, and recommendations for the design of a tool similar to the MCSS for family caregivers.

Similarly, this study found 10 key MCSS features—alarm system or medication reminder, medication list, medication dosage and schedule, as-needed medication access, medication instructions, signing off and medication confirmation, preordering medication, user-friendly interfaces, privacy features, and customization options—that are crucial for family caregivers to manage medication and have direct communication with health care professionals.

The essential features of the MCSS not only support effective medication management for family caregivers but also reflect the broader impact of the system, as highlighted by the findings, which confirm the MCSS's role in simplifying medication administration processes and facilitating direct communication with health care professionals. All participants expressed that the MCSS offers the benefits of verifying patient details, ensuring adherence to prescribed schedules, and providing electronic reminders and direct communication with health care professionals. In addition to the benefits of the MCSS tool, the crucial role of family caregivers, also known as informal caregivers, in providing unpaid support to patients cannot be overlooked. Studies [16,19,44] have shown that family caregivers offer a wide array of unpaid assistance. On the basis of these studies, family caregivers play a crucial role in medication management, particularly as health care systems increasingly rely on them to support patient care at home. These studies emphasize the growing responsibilities of family caregivers, especially in palliative care settings, where they are often tasked with managing complex medication regimens. These studies [14,19,44] highlight the fact that family caregivers are often responsible for ensuring that medication is administered correctly, managing complex medication schedules, and addressing any side effects or interactions. However, these responsibilities can be overwhelming, especially when caregivers lack formal training or adequate support. These studies highlight the need for robust support systems to ensure that family caregivers effectively handle these responsibilities. In our study, participants acknowledged the significant responsibility shouldered by family caregivers in administering medications and the need for a support system. In addition, our study highlighted insufficient communication between health care professionals and family caregivers. As the primary communication occurred before handing over medication responsibility to family caregivers during trips with patients, participants noted challenges in maintaining communication during and after the trips. The MCSS can help ease these challenges by providing tools that streamline medication management, offer clear instructions, and facilitate ongoing communication with health care professionals.

Beyond communication challenges, participants also expressed concerns regarding the lack of training and experience among family caregivers in medication administration, which led to instances of medication mismanagement. This underscored the importance of addressing educational gaps by providing an adequate support system to family caregivers to ensure safe medication practices, as was mentioned in previous studies [28,44,45]. To address this challenge identified both in these studies and in our study, the MCSS should include comprehensive training modules that help family caregivers understand how to use the system effectively. These modules could be delivered through the app or via web-based tutorials,

ensuring that family caregivers feel confident in managing medications.

Moreover, challenges related to verifying medication administration under family caregiver responsibility were expressed. The participants highlighted the importance of the MCSS, which is built with the ability to ensure proper medication administration and provide electronic reminders. They discussed that the system provides detailed medication information and sends reminders when it is time to administer medication or when time passes. In addition, previous studies have described the importance of a digital system that provides comprehensive written instructions [28,29,45] for safe medication administration and clear communication regarding medication schedules, dosages, and potential side effects. The findings of these studies align closely with our study's findings regarding the role and support needs of family caregivers and the importance of support tools that provide comprehensive and accessible information for medication management. Using a digital system among family caregivers was suggested as a crucial approach to promoting proper medication adherence and reducing the likelihood of errors associated with ADEs. These results are aligned with the results of previous studies that highlight the crucial role of digital tools in reducing medication errors and improving patient care. These studies demonstrate that telehealth interventions and telemedicine can enhance caregiver support and reduce medication management errors [10,17,20,28,46,47]. These studies emphasize the integration of telehealth for monitoring chronic conditions, reinforcing the need for a comprehensive system that supports caregivers in managing complex medication regimes. In addition, these studies discuss the prevalence of medication errors and adverse events, which our study aligns with by advocating for enhanced MCSS features to mitigate these issues.

Building on this, the participants highlighted how specific features, such as confirmation processes, could address the challenges identified in previous studies and further enhance the MCSS's effectiveness in managing medication errors and adverse events.

Our study identified that the design of an MCSS similar to that used by health care professionals is necessary for family caregivers to bridge the communication gap and improve medication management. However, while our study explored health care professionals' perspectives on this issue, more studies are needed to investigate this subject from family caregivers' and patients' perspectives. In addition, the design recommendations should be verified based on family caregivers' and patients' input.

Limitations of the Study

Some of the limitations of this study include the small sample of participants. While the in-depth nature of the interviews provided rich insights, the limited sample size may restrict the generalizability of the findings to broader populations of health care professionals working in diverse home care contexts. Future research should aim to increase the sample size to enhance the representativeness of the findings, strengthen the generalizability of the results, and ensure the broader applicability of the findings. Another notable weakness is the lack of inclusion of patient and family caregivers' perspectives in the study. By primarily focusing on health care professionals' viewpoints, this research overlooks the experiences and preferences of older adults receiving home care and of their family caregivers. By engaging patients and family caregivers in future studies, their perspectives and needs can be considered. Future research should prioritize incorporating the viewpoints of both patients and their family caregivers and the development, implementation, and evaluation of such interventions to assess their effectiveness and feasibility in real-world contexts.

Conclusions

This study delved into the complexities of medication management among older adults in home care, focusing on health care professionals' perspectives and the use of the MCSS. Through interviews, this study uncovered critical themes and recommendations, emphasizing the need to address gaps in medication management practices, particularly concerning similar support tools for family caregivers. In conclusion, incorporating the recommended design features, such as customizable interfaces and enhanced privacy controls, into the MCSS can improve usability and effectiveness for family caregivers. These enhancements are expected to facilitate better medication management, reduce errors, and strengthen communication with health care professionals. By equipping family caregivers with essential information via a tool similar to the MCSS, a proactive approach to preventing errors and improving outcomes is advocated. However, the impact of these improvements should be verified in future studies to confirm their effectiveness and address any emerging challenges.

Acknowledgments

The authors would like to thank Nora Home Care for helping with the recruitment of participants. In addition, they would like to thank all the study participants for their time and valuable insights. The use of generative artificial intelligence technology such as Grammarly (Grammarly Inc) and ChatGPT-3.5 (OpenAI) in this study was to enhance the grammatical accuracy and vocabulary richness of the data; specifically, it facilitated the incorporation of medical terminology that may not have been included in the researchers' vocabulary. This approach ensured that participants could express themselves using familiar medical terms, thereby improving the quality and authenticity of the data collected. This use of generative artificial intelligence aligns with ethical considerations as it helped maintain the integrity of participants' language while ensuring that their contributions were accurately represented in the study [42,43,48]. In addition, the tools were used in the preparation of the manuscript to harmonize the text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. The authors received no specific funding for this work.

Authors' Contributions

Both authors were involved in the design of the study. Data collection and initial data analysis were conducted by AA. The analysis was then reviewed and refined by ND. Both authors contributed to the subsequent writing and reviewing of the manuscript.

Conflicts of Interest

None declared.

References

1. Recommendations on digital interventions for health system strengthening: WHO Guideline. World Health Organization. 2019 Jun 06. URL: https://www.who.int/publications-detail-redirect/9789241550505 [accessed 2024-04-29]
2. Reinhard SC, Samis S, Levine CL. Family caregivers providing complex chronic care to their spouses. AARP Public Policy Institute. 2014. URL: https://member.aanlcp.org/wp-content/uploads/2021/03/family-caregivers-providing-complex-chronic-care-spouses-AARP-ppi-health.pdf [accessed 2024-10-14]
3. Cohen MR. Medication Errors: Causes, Prevention and Risk Management. Washington, DC: American Pharmacists Association; 2007.
4. Naseralallah L, Stewart D, Price M, Paudyal V. Prevalence, contributing factors, and interventions to reduce medication errors in outpatient and ambulatory settings: a systematic review. Int J Clin Pharm 2023 Dec;45(6):1359-1377 [FREE Full text] [doi: 10.1007/s11096-023-01626-5] [Medline: 37682400]
5. Hutton B, Kanji S, McDonald E, Yazdi F, Wolfe D, Thavorn K, et al. Incidence, causes, and consequences of preventable adverse drug events: protocol for an overview of reviews. Syst Rev 2016 Dec 05;5(1):209 [FREE Full text] [doi: 10.1186/s13643-016-0392-4] [Medline: 27919281]
6. Alqenae FA, Steinke D, Keers RN. Prevalence and nature of medication errors and medication-related harm following discharge from hospital to community settings: a systematic review. Drug Saf 2020 Jun;43(6):517-537 [FREE Full text] [doi: 10.1007/s40264-020-00918-3] [Medline: 32125666]
7. Wang J, Li X, Liu W, Yang B, Zhao Q, Lü Y, et al. The positive aspects of caregiving in dementia: a scoping review and bibliometric analysis. Front Public Health 2022 Sep 14;10:985391 [FREE Full text] [doi: 10.3389/fpubh.2022.985391] [Medline: 36187613]
8. Dementia. World Health Organization. 2023 Mar 15. URL: https://www.who.int/news-room/fact-sheets/detail/dementia [accessed 2024-04-29]
9. Lee K, Puga F, Pickering CE, Masoud SS, White CL. Transitioning into the caregiver role following a diagnosis of Alzheimer's disease or related dementia: a scoping review. Int J Nurs Stud 2019 Aug;96:119-131. [doi: 10.1016/j.ijnurstu.2019.02.007] [Medline: 30851954]
10. Graven LJ, Glueckauf RL, Regal RA, Merbitz NK, Lustria ML, James BA. Telehealth interventions for family caregivers of persons with chronic health conditions: a systematic review of randomized controlled trials. Int J Telemed Appl 2021 May 21;2021:3518050 [FREE Full text] [doi: 10.1155/2021/3518050] [Medline: 34093704]
11. Medication without harm: policy brief. World Health Organization. 2024 Mar 07. URL: https://www.who.int/publications-detail-redirect/9789240062764 [accessed 2024-04-29]
12. Global observatory for eHealth. World Health Organization. URL: https://www.who.int/observatories/global-observatory-for-ehealth [accessed 2024-04-29]
13. eHealth. World Health Organization. 2005. URL: https://iris.who.int/handle/10665/20378 [accessed 2024-04-29]
14. Ariani A, Koesoema AP, Soegijoko S. Innovative healthcare applications of ICT for developing countries. In: Qudrat-Ullah H, Tsasis P, editors. Innovative Healthcare Systems for the 21st Century. Cham, Switzerland: Springer; May 14, 2017.
15. Consolidated telemedicine implementation guide. World Health Organization. 2022 Nov 09. URL: https://www.who.int/publications-detail-redirect/9789240059184 [accessed 2024-04-29]
16. Alam S, Hannon B, Zimmermann C. Palliative care for family caregivers. J Clin Oncol 2020 Mar 20;38(9):926-936. [doi: 10.1200/JCO.19.00018] [Medline: 32023152]
17. Groom LL, McCarthy MM, Stimpfel AW, Brody AA. Telemedicine and telehealth in nursing homes: an integrative review. J Am Med Dir Assoc 2021 Sep;22(9):1784-801.e7 [FREE Full text] [doi: 10.1016/j.jamda.2021.02.037] [Medline: 33819450]
18. Alkureishi ML, Lee WW, Frankel RM. Chapter 10 - Patient-centered technology use: best practices and curricular strategies. In: Shachak A, Borycki EM, Reis SP, editors. Health Professionals' Education in the Age of Clinical Information Systems, Mobile Computing and Social Networks. Cambridge, MA: Academic Press; 2017:201-232.
19. Croteau AM, Vieru D. Telemedicine adoption by different groups of physicians. In: Proceedings of the 35th Annual Hawaii International Conference on System Sciences. 2002 Presented at: HICSS 2002; January 10, 2002; Big Island, HI. [doi: 10.1109/hicss.2002.994121]
20. Gokalp H, de Folter J, Verma V, Fursse J, Jones R, Clarke M. Integrated telehealth and telecare for monitoring frail elderly with chronic disease. Telemed J E Health 2018 Dec;24(12):940-957. [doi: 10.1089/tmj.2017.0322] [Medline: 30129884]
21. Szebehely M, Trydegård GB. Home care for older people in Sweden: a universal model in transition. Health Soc Care Community 2012 May;20(3):300-309. [doi: 10.1111/j.1365-2524.2011.01046.x] [Medline: 22141377]

22. Jarling A, Rydström I, Ernsth-Bravell M, Nyström M, Dalheim-Englund AC. Becoming a guest in your own home: home care in Sweden from the perspective of older people with multimorbidities. Int J Older People Nurs 2018 Sep;13(3):e12194. [doi: 10.1111/opn.12194] [Medline: 29603651]
23. Reinhard S. Home alone revisited: family caregivers providing complex care. Innov Aging 2019 Nov;3(Supplement_1):S747-S748 [FREE Full text] [doi: 10.1093/geroni/igz038.2740]
24. Apter AJ, Localio AR, Morales KH, Han X, Perez L, Mullen AN, et al. Home visits for uncontrolled asthma among low-income adults with patient portal access. J Allergy Clin Immunol 2019 Sep;144(3):846-53.e11. [doi: 10.1016/j.jaci.2019.05.030] [Medline: 31181221]
25. Karnehed S, Erlandsson LK, Norell Pejner M. Nurses' perspectives on an electronic medication administration record in home health care: qualitative interview study. JMIR Nurs 2022 Apr 22;5(1):e35363. [doi: 10.2196/35363] [Medline: 35452400]
26. Hughes RG, Blegen MA. Medication administration safety. In: Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Rockville, MD: Agency for Healthcare Research and Quality; 2008.
27. Marchesoni MA, Axelsson K, Lindberg I. Digital support for medication administration--a means for reaching the goal of providing good care? J Health Organ Manag 2014;28(3):327-343. [doi: 10.1108/JHOM-11-2012-0222] [Medline: 25080648]
28. Wickramasinghe N. Healthcare knowledge management: incorporating the tools, technologies, strategies, and process of knowledge management to effect superior healthcare delivery. In: Bali RK, Dwivedi AN, editors. Healthcare Knowledge Management. Health Informatics. New York, NY: Springer; 2007.
29. MCSS digital signering för HSL- och SoL-insatser. Appva. URL: https://appva.com/vi-erbjuder/appva-mcss/ [accessed 2024-01-03]
30. Committee on Family Caregiving for Older Adults, Board on Health Care Services, Health and Medicine Division, National Academies of Sciences, Engineering, and Medicine, Schulz R, Eden J. Families Caring for an Aging America. Washington, DC: National Academies Press; Nov 08, 2016.
31. Rasool MF, Rehman AU, Imran I, Abbas S, Shah S, Abbas G, et al. Risk factors associated with medication errors among patients suffering from chronic disorders. Front Public Health 2020 Nov 19;8:531038. [doi: 10.3389/fpubh.2020.531038] [Medline: 33330300]
32. Schepis TS, Klare DL, Ford JA, McCabe SE. Prescription drug misuse: taking a lifespan perspective. Subst Abuse 2020 Mar 05;14:1178221820909352. [doi: 10.1177/1178221820909352] [Medline: 32214819]
33. O'Connor MN, Gallagher P, O'Mahony D. Inappropriate prescribing: criteria, detection and prevention. Drugs Aging 2012 Jun 01;29(6):437-452. [doi: 10.2165/11632610-000000000-00000] [Medline: 22642779]
34. Kadima NJ, Nyiranteziryayo R, Umumararungu T, Adedeji AA. Use of mobile phones for patient self-reporting adverse drug reactions: a pilot study at a tertiary hospital in Rwanda. Health Technol 2020 Nov 12;11:185-191. [doi: 10.1007/s12553-020-00510-w]
35. Lo Giudice I, Mocciaro E, Giardina C, Barbieri MA, Cicala G, Gioffrè-Florio M, et al. Characterization and preventability of adverse drug events as cause of emergency department visits: a prospective 1-year observational study. BMC Pharmacol Toxicol 2019 Apr 27;20(1):21. [doi: 10.1186/s40360-019-0297-7] [Medline: 31029178]
36. Polkinghorne DE. Language and meaning: data collection in qualitative research. J Counsel Psychol 2005;52(2):137-145. [doi: 10.1037/0022-0167.52.2.137]
37. Kairuz T, Crump K, O'Brien AJ. Perspectives on qualitative research. Part 2: useful tools for data collection and analysis. Pharm J 2007 Mar;278(7445):371-377 [FREE Full text]
38. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
39. Sjuk- och tandvård för äldre. Nora Municipality. URL: https://www.nora.se/omsorgstod/stodtillaldre/sjukochtandvardforaldre.4.19b51d27126e64851c0800012060.html [accessed 2024-01-03]
40. Audulv A, Hall EO, Kneck Å, Westergren T, Fegran L, Pedersen MK, et al. Qualitative longitudinal research in health research: a method study. BMC Med Res Methodol 2022 Oct 01;22(1):255. [doi: 10.1186/s12874-022-01732-4] [Medline: 36182899]
41. Matt C, Becker M, Kolbeck A, Hess T. Continuously healthy, continuously used? – a thematic analysis of user perceptions on consumer health wearables. Pac Asia J Assoc Inf Syst 2019;11(1):108-132. [doi: 10.17705/1pais.11105]
42. Hasan N, Rana RU, Chowdhury S, Dola AJ, Khan Rony MK. Ethical considerations in research. J Nurs Res Patient Saf Pract 2021 Aug 28;1(01):1-4. [doi: 10.55529/jnrpsp.11.1.4]
43. Council for International Organizations of Medical Sciences. International Ethical Guidelines for Health-Related Research Involving Humans. Geneva, Switzerland: CIOMS; 2017.
44. Christou P. How to use artificial intelligence (AI) as a resource, methodological and analysis tool in qualitative research? Qual Rep 2023 Oct 7;28(7):1968-1980. [doi: 10.46743/2160-3715/2023.6406]
45. Choukou MA, Olatoye F, Urbanowski R, Caon M, Monnin C. Digital health technology to support health care professionals and family caregivers caring for patients with cognitive impairment: scoping review. JMIR Ment Health 2023 Jan 11;10:e40330 [FREE Full text] [doi: 10.2196/40330] [Medline: 36630174]

46. Leng M, Zhao Y, Xiao H, Li C, Wang Z. Internet-based supportive interventions for family caregivers of people with dementia: systematic review and meta-analysis. J Med Internet Res 2020 Sep 09;22(9):e19468 [FREE Full text] [doi: 10.2196/19468] [Medline: 32902388]
47. Dimant J. Medication errors and adverse drug events in nursing homes: problems, causes, regulations, and proposed solutions. J Am Med Dir Assoc 2002;3(2):S46. [doi: 10.1097/00130535-200203001-00008]
48. Sears N, Baker GR, Barnsley J, Shortt S. The incidence of adverse events among home care patients. Int J Qual Health Care 2013 Feb;25(1):16-28. [doi: 10.1093/intqhc/mzs075] [Medline: 23283731]

Abbreviations

ADE: adverse drug event
MCSS: medication and care support system

Edited by J Hefner; submitted 20.06.24; peer-reviewed by C Liu, D Patel; comments to author 14.08.24; revised version received 12.09.24; accepted 14.09.24; published 23.10.24.

Please cite as:
Ashimwe A, Davoody N
Exploring Health Care Professionals' Perspectives on the Use of a Medication and Care Support System and Recommendations for Designing a Similar Tool for Family Caregivers: Interview Study Among Health Care Professionals
JMIR Med Inform 2024;12:e63456
URL: https://medinform.jmir.org/2024/1/e63456
doi:10.2196/63456
PMID:

Enhancing Bias Assessment for Complex Term Groups in Language Embedding Models: Quantitative Comparison of Methods

Magnus Gray[1], MS; Mariofanna Milanova[2], MSc, PhD; Leihong Wu[1], PhD

1

2

Corresponding Author:
Leihong Wu, PhD

Abstract

Background: Artificial intelligence (AI) is rapidly being adopted to build products and aid in the decision-making process across industries. However, AI systems have been shown to exhibit and even amplify biases, causing a growing concern among people worldwide. Thus, investigating methods of measuring and mitigating bias within these AI-powered tools is necessary.

Objective: In natural language processing applications, the word embedding association test (WEAT) is a popular method of measuring bias in input embeddings, a common area of measure bias in AI. However, certain limitations of the WEAT have been identified (ie, their nonrobust measure of bias and their reliance on predefined and limited groups of words or sentences), which may lead to inadequate measurements and evaluations of bias. Thus, this study takes a new approach at modifying this popular measure of bias, with a focus on making it more robust and applicable in other domains.

Methods: In this study, we introduce the SD-WEAT, which is a modified version of the WEAT that uses the SD of multiple permutations of the WEATs to calculate bias in input embeddings. With the SD-WEAT, we evaluated the biases and stability of several language embedding models, including Global Vectors for Word Representation (GloVe), Word2Vec, and bidirectional encoder representations from transformers (BERT).

Results: This method produces results comparable to those of the WEAT, with strong correlations between the methods' bias scores or effect sizes (r=0.786) and P values (r=0.776), while addressing some of its largest limitations. More specifically, the SD-WEAT is more accessible, as it removes the need to predefine attribute groups, and because the SD-WEAT measures bias over multiple runs rather than one, it reduces the impact of outliers and sample size. Furthermore, the SD-WEAT was found to be more consistent and reliable than its predecessor.

Conclusions: Thus, the SD-WEAT shows promise for robustly measuring bias in the input embeddings fed to AI language models.

(JMIR Med Inform 2024;12:e60272) doi:10.2196/60272

KEYWORDS

bias; bias measurement; natural language processing; language models; artificial intelligence; input embeddings; AI; assessment; decision-making; AI-powered tool; NLP; application; AI language models

Introduction

Background

When considering the bias in artificial intelligence (AI) and how it can be mitigated, it is key to understand how to measure the bias of interest in order to properly determine the effectiveness of the mitigation technique. One common area for measuring bias in AI is with regard to the input embeddings of the AI model. Input embeddings are how the training and input data are numerically represented in order to make the data understandable to the model. Word and sentence embeddings are 2 common types of input embeddings, and these are likely to capture societal attitudes and display semantic biases [1]. For example, word embeddings may make biased associations between different genders and certain occupations (ie, nurse and female; doctor and male). Thus, in natural language processing (NLP) applications, such as large language models like ChatGPT and LLaMA, addressing bias in this area is of great importance. Several existing methods for measuring bias in input embeddings include the word embedding association test (WEAT), the Sentence Encoder Association Test (SEAT), and the Embedding Coherence Test [2-4]. The WEAT, for instance, has been rather influential, with its derivative, the SEAT, being used by several studies investigating methods of mitigating bias, including Sent-Debias and Auto-Debias [5,6]. Furthermore, the WEAT has been used to assess the stability of word embedding methods (WEMs) [7].

WEAT: The Word Embedding Association Test

The WEAT was created in 2017 to assess bias within the semantic representations of words in AI, or word embeddings [2], which represent words as a vector based on the textual context in which the word is found. This metric works by considering 2 sets of target terms (eg, science and art terms) and 2 sets of attribute terms (eg, male and female terms). The null hypothesis is that there is no difference between the sets of target words and their relative similarity to the sets of attribute words. Bias is quantified by computing the probability that a permutation of attribute words would produce the observed difference in sample means and, thus, determining the unlikelihood of the null hypothesis [2].

The WEAT was developed to be a statistical test analogous to the implicit association test (IAT), which asked participants to pair concepts or words that they implicitly associate [8]. A total of 10 WEATs were developed based on the documented human biases highlighted by the IAT. In the first WEAT study, Global Vectors for Word Representation (GloVe) word embeddings were used to numerically represent the selected words for each test and, in turn, compute bias scores (see the Methods section for more information about the WEAT's method of assessing bias). GloVe is an unsupervised learning algorithm for obtaining vector representations for words [9]. The GloVe model was trained on aggregated word-word cooccurrence statistics from a large English language corpus. The resulting word representations capture meaningful linear substructures, allowing for excellent performance on word analogy, word similarity, and named entity recognition tasks [9].

While the WEAT has become somewhat of a standard measure of bias in input embeddings, there are several limitations of using it to measure an AI language model's bias. First, the WEAT demands 2 distinct term groups for both targets and attributes, which can be challenging without prior knowledge to segregate, especially for nonbinary terms. For instance, age-related terms might include categories like infants, youth, middle-aged, and seniors, complicating differentiation. The WEAT struggles to consider the nuances across such multiple subgroups, limiting its effectiveness in scenarios where terms don't neatly divide into binary categories. Second, the current groups of terms could be incomplete, potentially introducing unwanted bias. Finally, the original bias calculation (ie, the effect size) is not that robust, and as such, it may be affected by the size or contents of the target or attribute groups. Thus, there is some room to improve this measure of bias, which leads to the focus of this study.

SEAT: The Sentence Encoder Association Test

The SEAT is a generalization of the WEAT to phrases and sentences, rather than single or compound words [3]. In more detail, this measure of bias modifies the original WEATs by inserting the words into simple sentence templates such as "This is a[n]<word > ." Furthermore, new tests were created to measure additional race- and sex-related stereotypical biases. In the SEAT study, multiple sentence encoders were evaluated, including those for popular language models, such as ELMo [10], GPT [11], and bidirectional encoder representations from transformers (BERT) [12]. In the SEAT, bias is measured the same as in the WEAT.

With sentence encoders becoming increasingly popular in NLP applications, the SEAT is a useful extension of the WEAT for measuring bias within sentence representations. Based on the SEAT study's results, sentence embeddings typically display less bias than word embeddings, and more recent sentence encoders (such as those for GPT and BERT) exhibit less bias than previous models (such as GloVe) [3]. However, the SEAT still shares the same major limitations as the WEAT, with the need to have predefined, binary sets of targets and attributes, for instance.

Applications of WEAT and SEAT

The WEAT [2] has been influential in the investigation and development of techniques for mitigating bias in AI systems, with its sentence-level extension, the SEAT [3], being used to measure and evaluate biases present in sentence presentations before and after applying debiasing techniques. More specifically, the SEAT has been used to evaluate the performance of Sent-Debias and Auto-Debias [5,6]. On the one hand, Sent-Debias [5] is a method of debiasing sentence embeddings, while on the other hand, Auto-Debias [6] is an automatic method of mitigating social biases in pretrained language models that alters the model's parameters through fine-tuning. Both methods used 6 SEAT benchmarks to measure the bias present in the sentence embeddings for various language models, both before and after their application.

The WEAT has also been used in the study of the stability of WEMs. In 2021, Borah et al [7] developed a metric for stability and evaluated a collection of WEMs, including fastText [13], GloVe [9], and Word2Vec [14], on a collection of downstream tasks, including fairness evaluation. For this task, they used the WEAT's bias score to assess the stability of the WEMs, noting a relationship between these scores and those of their developed stability metric. Thus, the WEAT can be used to determine the stability of a WEM, which is significant because stability is necessary to produce similar results across multiple experiments.

Objective

Altogether, with the WEAT's ability to assess the stability of WEMs and the SEAT's use in evaluating the performance of recent debiasing techniques such as Sent-Debias and Auto-Debias, it is evident that the WEAT has a large influence in this area of measuring bias in input embeddings. However, certain limitations of the WEAT and SEAT have been identified (ie, their nonrobust measure of bias and their reliance on predefined, binary groups of words or sentences), which may lead to inadequate measurements and evaluations of bias. Thus, this study takes a new approach at modifying this popular measure of bias, with a focus on making it more robust and applicable in other domains. Moreover, we aim to make a more flexible and user-friendly approach to measuring bias in input embeddings, where the user is not required to segregate terms to achieve a reliable bias assessment result.

Methods

WEAT Method

The WEAT's measure of bias, the effect size, is calculated similarly to Cohen *d*; with target sets *X* and *Y* and attribute sets *A* and *B*, the WEAT's effect size (*d*) is a normalized measure of how separated the 2 distributions of associations between the target and attribute are. In the formula for the effect size (Textbox 1), *s(w,A,B)* measures the association of a target word (*w*) with the attribute words (Textbox 2).

Textbox 1. Word embedding association test effect size.

$$d=\text{mean}_{x\in X} s(x,A,B)-\text{mean}_{y\in Y} s(y,A,B)\,\text{std\_dev}_{w\in X\cup Y} s(w,A,B)$$

Textbox 2. Word embedding association test association measure.

$$s(w,A,B)=\text{mean}_{a\in A}\cos(\vec{w},\vec{a})-\text{mean}_{b\in B}\cos(\vec{w},\vec{b})$$

To measure the significance of the associations between targets and attributes and determine the unlikelihood of the null hypothesis, the WEAT defines a 1-sided *P* value test with the test statistic *s*(*X*,*Y*,*A*,*B*). The test statistic (Textbox 3) measures the differential association of the 2 sets of target words with the attribute. Textbox 4 shows how the *P* value is calculated, with $(X_i,Y_i)_i$ denoting all the partitions of $X\cup Y$ into 2 sets of equal size.

Textbox 3. The word embedding association test's test statistic.

$$s(X,Y,A,B)=\sum_{x\in X} s(x,A,B)-\sum_{y\in Y} s(y,A,B)$$

Textbox 4. Word embedding association test P value.

$$\text{Pr}_i[s(X_i,Y_i,A,B)>s(X,Y,A,B)]$$

SD-WEAT Method

In this study, we introduce the SD-WEAT, a more robust and balanced method for exploring and assessing bias. Instead of using predefined word sets and using the effect size as the bias measurement score, the words are randomly replaced and the SD of multiple effect sizes is computed. This removes the need to predefine the word groups, allowing for the avoidance of biased groups of words. Furthermore, the results are more robust, as the SD is calculated over multiple runs rather than one. Together, this should allow the SD-WEAT to be a more simplistic and accessible measure of bias in input embeddings.

Multiple experiments were conducted to explore the uses of these modifications to the WEAT. In each experiment, GloVe word embeddings were used to numerically represent the words for each test, following suit from the original WEAT study. More specifically, we used the GloVe model "glove.840B.300d" from the Stanford NLP Group [15], which was trained on 840 billion tokens from Common Crawl data, has a vocab size of 2.2 million, and generates 300-dimension vectors. Moreover, other input embedding methods, including those of BERT [12], Sci-BERT [16], and BioBERT [17], have been evaluated in order to explore the differences of their biases, which may potentially come from their training datasets or techniques.

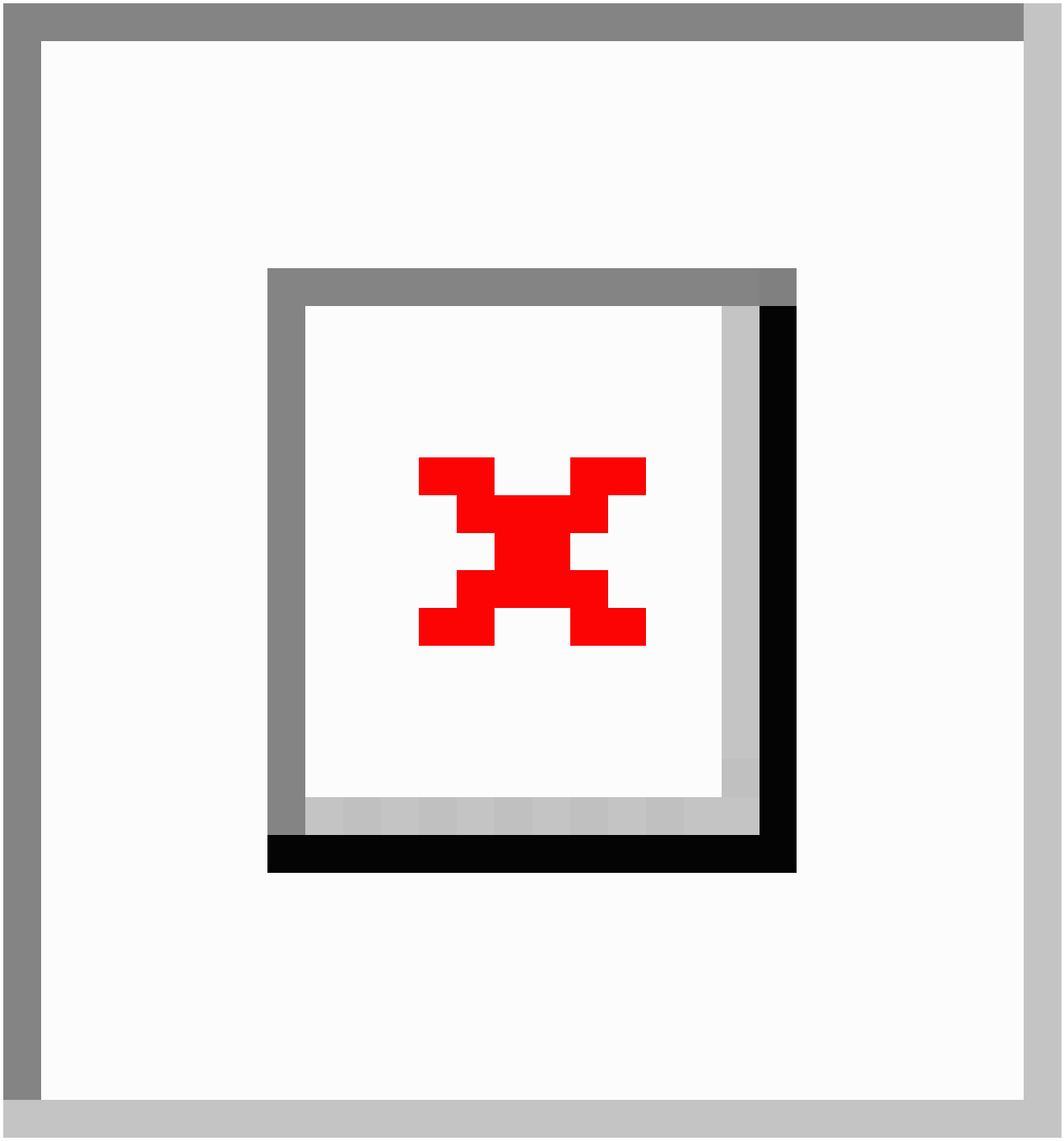
Due to the significance of the attribute sets in both the WEAT's null hypothesis and effect size calculation, it was decided to focus 2 experiments around replacing the attribute sets. The primary experiment, hereafter named "SD-WEAT," used the original 10 benchmark datasets used by the WEAT, but the attribute sets were replaced with random word sets from the combined original attribute sets. The secondary experiment, hereafter named "SD-WEAT-Negative-Control," used the original target data in the 10 benchmark datasets, but the attribute sets were replaced with random word sets derived from the GloVe dictionary.

In the SD-WEAT, for each of the 10 WEATs, the 2 sets of attribute words were pooled together into one list, and then, 100 new tests were constructed, pulling 4 words from said list to form 2 new attribute sets (of 2 words each). For example, the benchmark WEAT-7, which examines a form of gender bias, contains attribute sets with male (eg, "male," "man," "boy," etc) and female (eg, "female," "woman," "girl," etc) terms. The SD-WEAT forms 100 new tests with attribute sets containing words randomly selected from the combination of these term lists, and as a result, the new attribute sets for one of these tests may resemble the following: (1) "male" and "female"; and (2) "woman" and "boy." Assuming a normal distribution of attribute set permutations, the SD-WEAT should be capable of assessing bias between the target and attribute concepts in a more robust manner, and this should open the door for assessing bias over multiple attribute groups at once. For instance, in the case of racial bias, the SD-WEAT should be able to assess bias over multiple racial groups such as Asian, Black, or White; the different permutations could be examined at an individual level to further assess bias in this multilevel context. This concept will be further explored in future studies.

On the other hand, in the SD-WEAT-Negative-Control, a large list of words was derived from GloVe, and then, 10,000 new tests were constructed, pulling 4 words from said list to form 2 new attribute sets (of 2 words each). Because the GloVe dictionary contains multitudes more words than the original attribute sets, a larger number of tests were constructed than before, allowing for 100 groups of 100 tests to analyze the variance across groups. For each experiment, GloVe word

embeddings and a Bag-of-Words model were used to complete each test, and the SD of all the collected effect sizes was computed. Figure 1A and B illustrates the process of creating the new tests for the SD-WEAT and SD-WEAT-Negative-Control, respectively.

Figure 1. SD-WEAT experiment design. GloVe: Global Vectors for Word Representation. WEAT: word embedding association test.



Textbox 5 shows how the SD-WEAT quantifies bias, using the effect size (*d*) from the WEAT (see Textbox 1). SD was chosen over other metrics (such as average), because with a normal distribution of attribute permutations, it is expected that the average of all effect sizes is close to zero. The significance of the SD-WEAT results was also calculated. In more detail, *z* scores were calculated for each of the 10 WEAT benchmarks, using Textbox 6. Here, x is the SD-WEAT score, while μ and σ are the average and SD of the SDs for the 100 groups of 100 effect sizes in the SD-WEAT-Negative-Control, respectively. Since the WEAT uses a 1-sided, right-tailed test, *P* values were calculated from the *z* scores with the right-tailed methodology.

Textbox 5. SD-word embedding association test bias calculation.

SD_WEAT=SD(d1,d2,…d100)

Textbox 6. SD-word embedding association test significance calculation.

$z=x-\mu\sigma$

where:

x=SD_WEAT

{x}control=(SD(d1,d2,…,d100)1,…,SD(d1,d2,…,d100)100)

μ=mean(x_control)

σ=SD(x_control)

Embedding Method Comparison

To explore the differences of biases between various input embedding methods, several additional methods were evaluated, including BERT, SciBERT, and BioBERT. BERT, which stands for *bidirectional encoder representations from transformers*, is a language model that was pretrained using data from the Toronto Book Corpus and English Wikipedia with the tasks of masked language model and next sentence prediction [12]. Compared to GloVe, which was pretrained using data from Common Crawl and generates context-free representations for each word [9], BERT uses the context on either side of a word to generate that word's representation. Since the WEATs are only composed of single or compound words rather than sentences, it is expected that the representations for these words will not be influenced by context, and thus, this controls one factor that could potentially contribute to differences in bias between these 2 embedding methods. However, it is expected that differences in bias between GloVe and BERT will emerge based on the differences in their algorithms as well as their training datasets.

SciBERT and BioBERT take inspiration from BERT, using a similar architecture but modifying the model's training process. On the one hand, SciBERT was pretrained on a large corpus of scientific texts (specifically, 1.14 million papers from Semantic Scholar) rather than the more general texts used for pretraining BERT [16]. SciBERT also uses a new vocabulary based on this scientific corpus. On the other hand, BioBERT continues to pretrain BERT with a large collection of PubMed abstracts (keeping the BERT pretraining datasets), resulting in a model that performs well on biomedical tasks [17]. There is great interest in using these models in regulatory science research efforts, and thus, they were included in this analysis. The BERT, SciBERT, and BioBERT models were each obtained from the Hugging Face repository. In more detail, these 3 models are listed in Hugging Face as "bert-based-cased" [18], "allenai/scibert_scivocab_cased" [19], and "dmis-lab/biobert-v1.1" [20], respectively.

Stability of Embedding Methods

The WEAT has also been used to assess the stability of WEMs, including fastText, GloVe, and Word2Vec. In more detail, Borah et al [7] trained 3 sets of embeddings for each WEM with a Wikipedia article dataset containing approximately 46 million tokens, changing the seed for each embedding set. The 9 trained embedding sets (3 for each WEM) were then evaluated with the WEAT benchmarks, and the highest and lowest WEAT scores for each benchmark were reported for each of the 3 WEMs. The stability of the WEM is based on the range of WEAT scores; WEMs that produced more similar WEAT scores for each benchmark are more stable than those that produced more different WEAT scores. Based on their results, fastText was found to be the most stable, and Word2Vec was found to be the least stable, with GloVe somewhere in the middle [7]. This aligns with their previous findings, and thus, there is a relationship between WEAT scores and the stability of WEMs.

Our study adopts a similar, yet modified, methodology to assess the stability of the same 3 WEMs, as well as BERT, with both the WEAT and SD-WEAT. In more detail, we began by obtaining the dataset. Without access to the dataset used by the referenced study, we downloaded a Wikipedia dataset (on November 3, 2023) that was uploaded to Hugging Face, specifically the version labeled "20220301.en" [21]. The raw dataset contains over 6 million documents (articles) and 138 million sentences. To conserve computational resources and time, 3 random samples were taken from this dataset, each containing 1% of the total number of documents, or 64,587 to be exact. For the fastText, GloVe, and Word2Vec methods, these 3 sample datasets were processed similarly to those in the referenced study. More specifically, the Wikipedia articles were broken down into sentences using the "punkt" tokenizer from the Natural Language Toolkit (NLTK) package for Python. Then, the sentences were broken down into lists of words, removing stop words and converting all text to lowercase. After processing, each sample dataset contains approximately 1.4 million sentences and 19 million words (tokens). For BERT, which is a sentence embedding method, a different approach had to be taken. The BERT model required a BERT tokenizer, so instead of using a pretrained NLTK tokenizer, a collection of BERT tokenizers was trained using the raw sample datasets. The "bert-base-cased" tokenizer was used to initialize the training, and the vocabulary size was set to 32,000. Three tokenizers were trained per sample dataset, changing the seed for each, resulting in 9 unique BERT tokenizers.

Next, the models were trained. For the fastText and Word2Vec methods, the Gensim package for Python and the "FastText" [22] and "Word2Vec" [23] models were used, and for the GloVe method, the Glove-Python package [24] was used. The parameters for these models were set based on those used in the referenced study, with the vector size set to 300, the context window size set to 5, and the number of epochs set to 5. For each WEM, 3 models were trained per sample dataset, changing the seed parameter for each, resulting in 9 models per WEM. On the other hand, for BERT, the "BertForMaskedLM" model [25] from the Hugging Face package for Python was used. This model uses BERT's default vector size of 768, with the number of epochs set to 5. In total, 9 BERT models were trained, one

for each of the BERT tokenizers that were trained. Since BERT is fundamentally different than the other embedding methods, it is expected that results for these models will differ than those of the WEMs; however, the comparison between the WEAT and SD-WEAT may be useful in the examination of the SD-WEAT's strengths.

Finally, with 9 models each for fastText, GloVe, Word2Vec, and BERT (or 3 models per each of the 3 sample datasets), the stability of these embedding methods could be evaluated. Each model was assessed with the WEAT and SD-WEAT benchmarks, and the variation within these scores could be used to determine the stability of the method. Furthermore, variations within the scores for a specific benchmark can be noted to analyze the areas that have a greater impact on stability than others.

Results

SD-WEAT is Correlated With WEAT in Bias Evaluation

Figure 2A illustrates and compares the bias measurement scores for the WEAT and SD-WEAT on the 10 WEAT benchmarks with the GloVe model (see Table 1 for a more detailed breakdown of the results). As shown in Figure 2A, the effect sizes (r=0.786) and P values (r=0.776) for the 10 WEAT benchmarks between 2 approaches were highly correlated. The results demonstrate that the SD-WEAT can effectively measure bias in binary-group attribute terms, providing similar performance as the original WEAT.

Figure 2. SD-WEAT and WEAT comparison. (A) WEAT versus SD-WEAT bias scores on the 10 WEAT benchmarks. (B) WEAT versus SD-WEAT bias scores on similar benchmarks (WEAT-4 and 5). WEAT: word embedding association test.

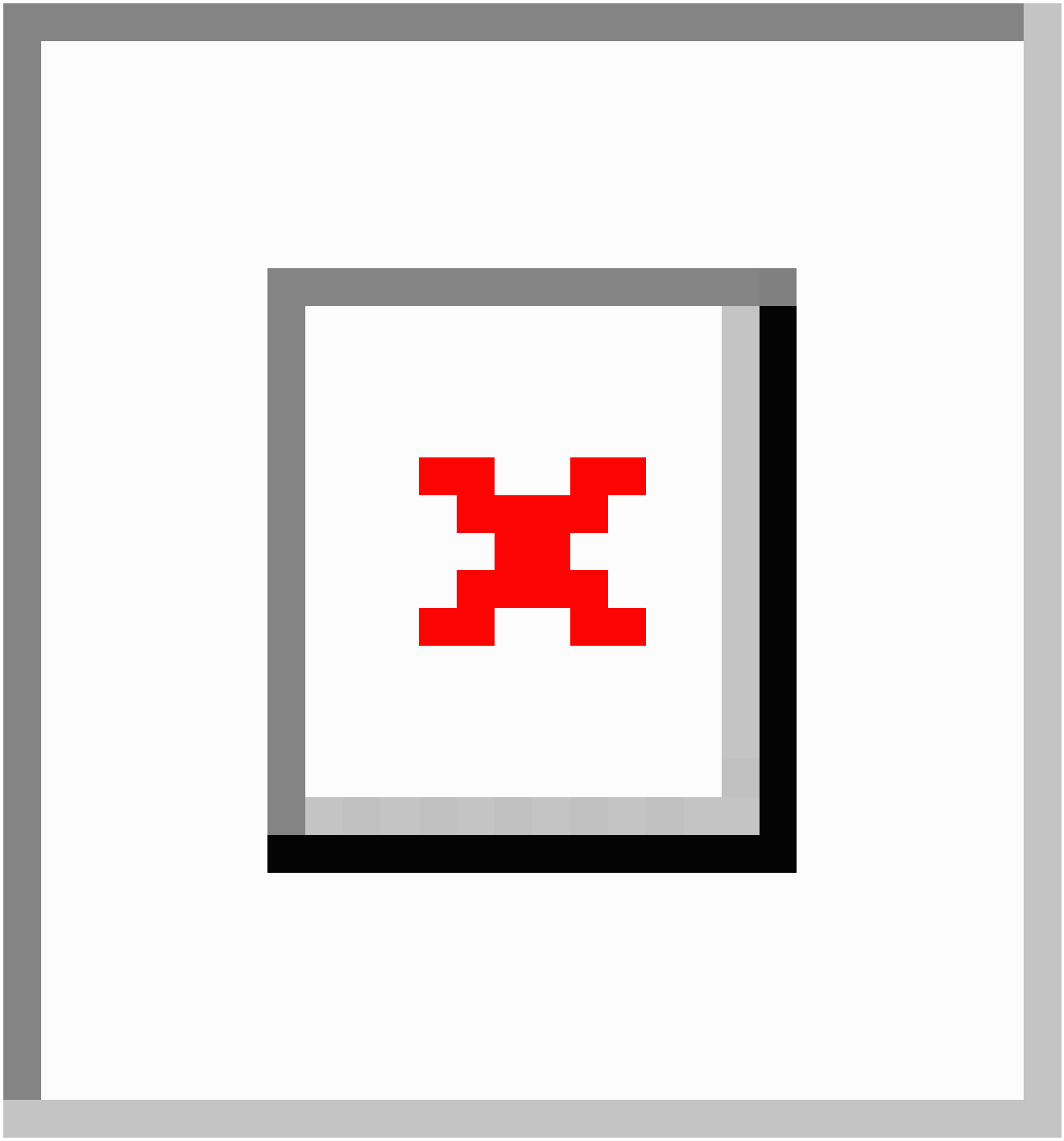


Table . SD-word embedding association test (WEAT) embedding method comparison.

Tests	GloVe[a]		BERT[b]		SciBERT		BioBERT	
	W[c]	SD-W[d]	W	SD-W	W	SD-W	W	SD-W
WEAT-1								
Bias score	1.50	1.00	0.08	0.32	−0.01	0.23	0.77	0.42
P value	<.001	<.001	.40	.55	.51	.97	.003	.97
WEAT-2								
Bias score	1.53	0.93	0.96	1.05	0.43	0.23	0.22	0.29
P value	<.001	.02	<.001	.09	.07	.81	.22	.98
WEAT-3								
Bias score	1.41	1.00	0.07	0.13	0.32	0.28	0.88	0.73
P value	<.001	<.001	.40	>.99	.10	.99	<.001	.28
WEAT-4								
Bias score	1.50	0.96	0.42	0.24	0.08	0.36	1.13	0.78
P value	<.001	.02	.12	.99	.42	.36	<.001	>.99
WEAT-5								
Bias score	1.28	0.98	0.05	0.22	0.31	0.40	.024	0.47
P value	<.001	.003	.45	>.99	.20	.01	.25	>.99
WEAT-6								
Bias score	1.81	1.35	0.02	0.17	0.40	0.45	0.23	0.27
P value	<.001	<.001	.48	>.99	.37	>.99	.33	.97
WEAT-7								
Bias score	1.06	0.69	−0.63	0.42	-0.11	0.47	−0.36	0.68
P value	.02	>.99	.90	>.99	.58	>.99	.76	.15
WEAT-8								
Bias score	1.24	0.78	−0.06	0.15	−0.31	0.51	0.03	0.58
P value	.005	.28	.56	>.99	.63	.96	.48	.93
WEAT-9								
Bias score	1.38	1.15	1.21	1.05	0.93	0.76	0.01	0.59
P value	.004	<.001	.01	.98	.052	.77	.50	<.001
WEAT-10								
Bias score	1.21	1.01	0.36	0.31	−0.31	0.85	0.10	0.41
P value	.005	.01	.26	.76	.72	.09	.43	>.99

[a]GloVe: Global Vectors for Word Representation.

[b]BERT: bidirectional encoder representations from transformers.

[c]The columns labeled “W” contain the results for the WEAT.

[d]The columns labeled “SD-W” contain the results for the SD-WEAT.

Additional patterns could be noted between the WEAT and SD-WEAT. Taking a closer look at the WEAT-4 and 5, both shared a similar focus, with the same target sets of European or African names (16 terms each) and different attribute sets formed with some combination of pleasant or unpleasant terms (25 and 8 terms each, respectively). Figure 2B depicts the WEAT and SD-WEAT scores of WEAT-4 and 5, respectively, where the SD-WEAT showed a much closer gap between 2 benchmarks. This suggested that, given the same target terms, the WEAT score was more vulnerable by the size of attribute terms, whereas the SD-WEAT can provide a more consistent effect score for bias assessment.

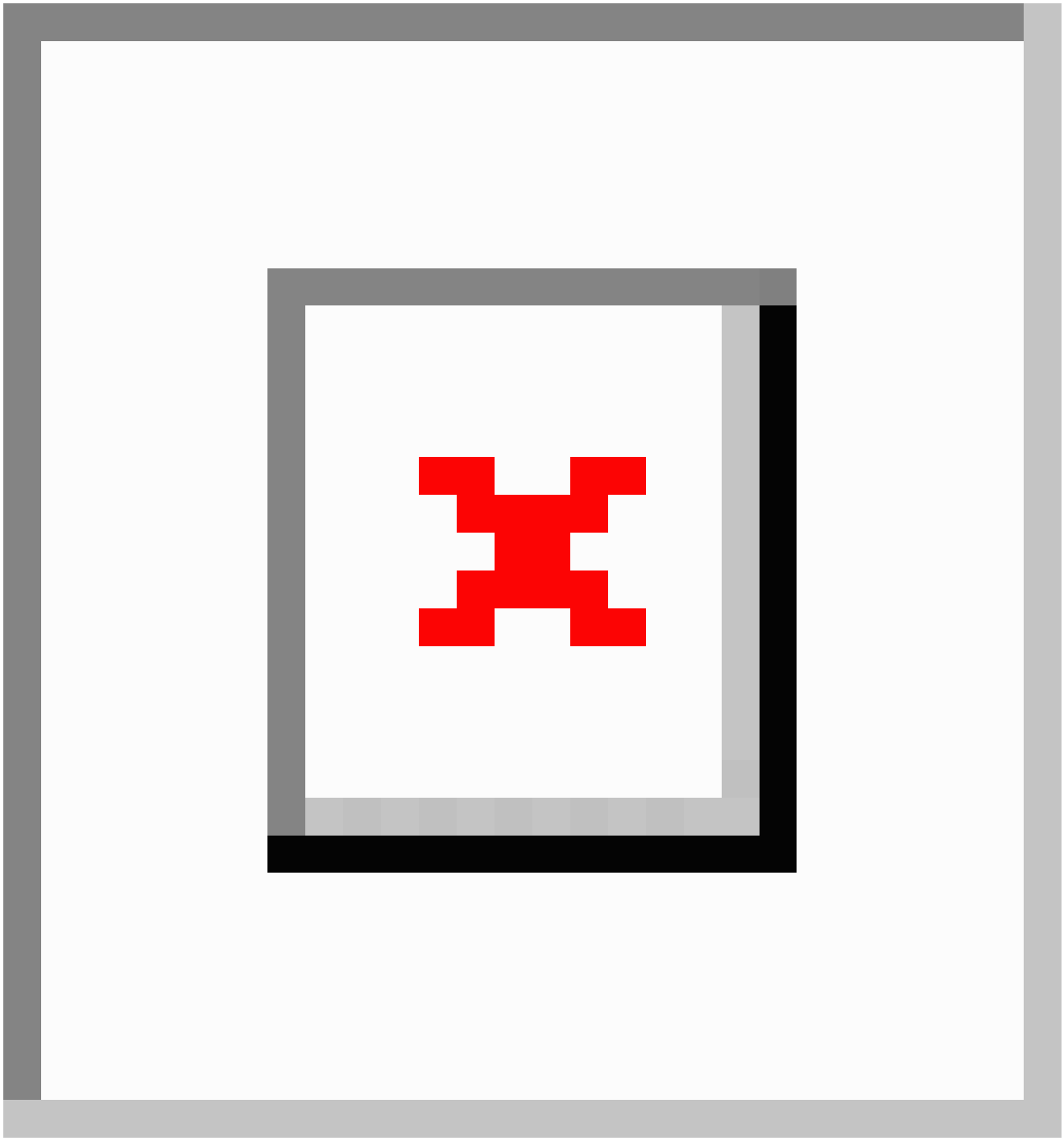
SD-WEAT’s Attribute Set Size Did Not Affect Bias Evaluation

In the primary SD-WEAT experiments, the replaced attribute sets consisted of 2 words each, which was the minimum required to measure the effect score. To examine the impact of the

attribute size, we also conducted the SD-WEAT experiments with 3- and 5-word attribute sets, keeping other parameters unchanged. Due to the original size of WEAT benchmark datasets, attribute sizes larger than 5 were not tested.

The result of this comparative analysis was illustrated in Figure 3. As shown, there was little difference among using 2-, 3-, or 5-word attribute sets. For each attribute size, the correlation between the WEAT's effect size and the SD-WEAT's SD is relatively unchanged. Thus, with the SD-WEAT methodology, using any attribute size could provide consistent results.

Figure 3. SD-WEAT attribute size analysis (WEAT vs SD-WEAT: bias score). Red, blue, and yellow represent the SD-WEAT with 2, 3, and 5 words, respectively. WEAT: word embedding association test.



Using SD-WEAT to Evaluate Various Embedding Methods

The WEAT and SD-WEAT experiments were further analyzed with BERT, SciBERT, and BioBERT, along with the original embedding method, GloVe, that the WEAT used. Table 1 contains the results. The columns labeled "W" contain the results for the WEAT, while those labeled "SD-W" contain those for the SD-WEAT. Both the bias scores (effect sizes) and *P* values are provided.

Based on the results of this analysis, the 3 BERT input embedding methods typically achieve lower WEAT and SD-WEAT scores compared to the GloVe method, indicating

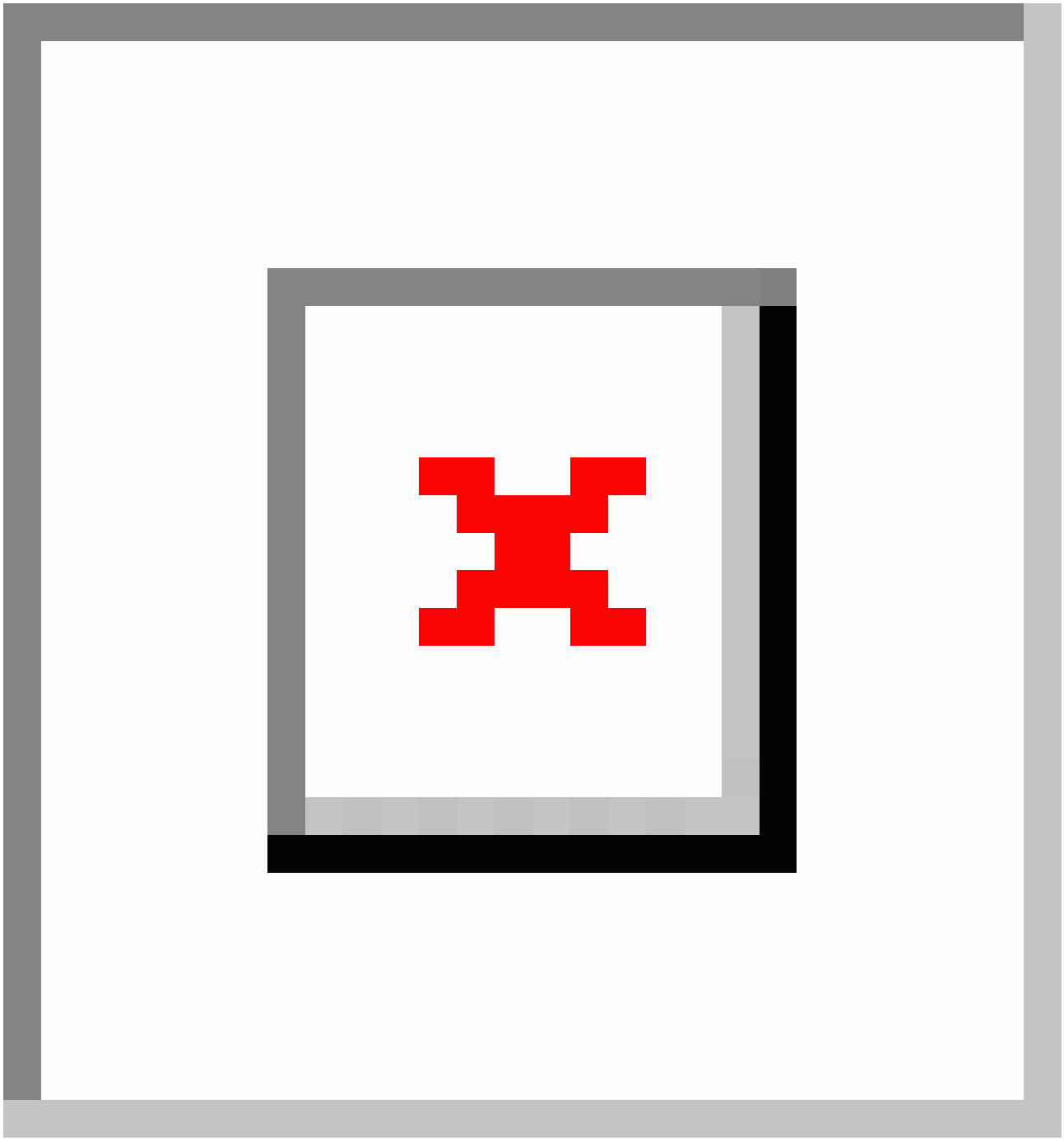
that these methods make less biased associations. Because the BERT embeddings should not be influenced by additional context with the lone words present within the WEAT benchmarks, this shows that the methods' algorithms or training datasets are likely the cause of their differences in bias. The 3 BERT models have training datasets with more objective language (ie, Wikipedia and scientific articles), which could explain why they produce lower WEAT and SD-WEAT scores than the GloVe method. Furthermore, it can be noted that the various BERT models perform differently than one another. For instance, unlike the other BERT models, BioBERT produces a significant result for WEAT-9, which focuses on the associations between mental and physical diseases and temporary and permanent terms, indicating that the attribute terms used in this specific benchmark have a significantly greater impact than random words. BioBERT was trained on biomedical data, and this could be why it performs so differently for this biomedical task.

Using SD-WEAT to Assess the Stability of Embedding Methods

The stability of fastText, GloVe, Word2Vec, and BERT was evaluated using the WEAT and SD-WEAT. Multimedia Appendix 1 shows the variability in the WEAT and SD-WEAT scores obtained for each WEAT benchmark over 9 training iterations per embedding method. The subfigures in the left column are box plots showing the spread of the scores, while the subfigures in the right column are bar charts showing the SD of these scores. At a glance, the box plots show that there is less variability in the SD-WEAT scores than the WEAT scores for each benchmark, confirmed by the lower SDs in the bar charts. Furthermore, based on the sizes of the box plots, some embedding methods appear more stable than others. For instance, fastText appears to have less variability in the WEAT and SD-WEAT scores than GloVe, Word2Vec, and BERT, indicating that this method may be the most stable of the ones evaluated.

Figure 4 compares the stability of each embedding method based on the SD of the WEAT and SD-WEAT scores over the 10 WEAT benchmarks. Recall that for each embedding method, 9 models were trained. Thus, the SD of the WEAT and SD-WEAT scores was calculated for each set of 9 models for each WEAT benchmark. Since SD is a measure of variation, lower values can be more stable. Again, fastText appears to be the most stable embedding method based on the WEAT and SD-WEAT scores. Furthermore, based on the results, there is less variability in the SD-WEAT scores than the WEAT scores.

Figure 4. Stability of word embedding methods: BERT: bidirectional encoder representations from transformers; GloVe: Global Vectors for Word Representation; WEAT versus SD-WEAT. WEAT: word embedding association test.



Overall, these findings are somewhat different than those of the referenced study, which found fastText to be the most stable and Word2Vec to be the least stable out of the 3 WEMs based on WEAT scores. In our case, fastText was indeed the most stable for the WEAT, but GloVe produced the most variable results. This difference may be a result of the difference in dataset size or contents, as we opted to use 3 smaller random samples than one larger one. This study also analyzed the stability of BERT, a sentence embedding method. However, based on the results, this method produces more variable WEAT and SD-WEAT scores than other methods, which may be a result of the fundamental differences of this model compared to the others. Moreover, some of this instability may be due to the small training sample, as the original BERT model was trained on much more data. Nonetheless, we found that all the models showed less variability with regard to the SD-WEAT than the WEAT. This shows a major strength of the SD-WEAT; this measure of bias is much more consistent and reliable than its predecessor.

Discussion

Overview

In this paper, we explored the methodology of measuring bias in the input embeddings of language embedding models, developing a novel approach, called SD-WEAT, for enhancing

bias assessment for complex term groups. This method addresses several limitations of its predecessor, the WEAT, resulting in a more robust and consistent measure of bias. Furthermore, with the SD-WEAT, it is now possible to assess bias over multilevel attribute groups, such as age, race, region, etc, in addition to binary attribute groups.

Future Directions and Limitations

In the future, new benchmarks can be established to measure bias among demographic groups or topics and a full list of attribute terms without segregation. For instance, a benchmark can be developed with target sets comprised of sex-linked medical terms and an attribute set of gender terms in order to estimate the level of bias between sex and medical conditions. In addition, by analyzing the individual trials that produced the highest bias effect, the medical terms that have the greatest association with one sex over the other can be identified. These applications may provide new insight into how these language embedding models can be applied into health care and regulatory science fields.

One limitation of the SD-WEAT is the increased computation resources due to having to generate and execute multiple runs to calculate the final bias effect score. However, the impact is rather minimal, only needing to examine the embedding method's biases once or periodically.

Advantages of SD-WEAT

The SD-WEAT not only enables the bias assessment over multilevel group terms but also enhances the bias assessment for binary group terms by avoiding unnecessary groupings of words. In some cases, it may be difficult to determine whether an intermediate word should belong to a certain grouping. Moreover, the SD-WEAT more robustly measures bias through the utilization of the SD of multiple effect sizes, and it has been found to be a more consistent and reliable measure of bias than its predecessor. As such, the SD-WEAT is a more robust and user-friendly measure of bias in input embeddings for AI language models.

Conclusions

To conclude, we introduced the SD-WEAT, a novel algorithm based on the WEAT, to enhance the bias assessment for complex term groups in language embedding models. In the future, the SD-WEAT could be applied in a regulatory science application and provide new insights to measure biases of common embedding models with regard to multilevel attribute groups, such as age, race, and region.

Acknowledgments

We wish to acknowledge that MG's contribution was made possible in part by his appointment through the Research Participation Program at the National Center for Toxicological Research, administered by the US Food and Drug Administration through the Oak Ridge Institute for Science Education.

Disclaimer

This manuscript reflects the views of the authors and does not necessarily reflect those of the Food and Drug Administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variability in word embedding association test (WEAT) and SD-WEAT scores for the 10 WEAT benchmarks over 9 training iterations.
[PNG File, 659 KB - medinform_v12i1e60272_app1.png]

References

1. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. Lang Linguist Compass 2021 Aug;15(8):e12432. [doi: 10.1111/lnc3.12432] [Medline: 35864931]
2. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science 2017 Apr 14;356(6334):183-186. [doi: 10.1126/science.aal4230] [Medline: 28408601]
3. May C, Wang A, Bordia S, Bowman SR, Rudinger R. On measuring social biases in sentence encoders. arXiv. Preprint posted online on Mar 25, 2019. [doi: 10.48550/arXiv.1903.10561]
4. Dev S, Phillips J. Attenuating bias in word vectors. Presented at: The 22nd International Conference on Artificial Intelligence and Statistics; Apr 16-18, 2019; Naha, Okinawa, Japan.
5. Liang PP, Li IM, Zheng E, Lim YC, Salakhutdinov R, Morency LP. Towards debiasing sentence representations. arXiv. Preprint posted online on Jul 16, 2020. [doi: 10.48550/arXiv.2007.08100]
6. Guo Y, Yang Y, Abbasi A. Auto-debias: debiasing masked language models with automated biased prompts. Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); May 22-27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.72]

7. Borah A, Barman MP, Awekar A. Are word embedding methods stable and should we care about it? Presented at: Proceedings of the 32nd ACM Conference on Hypertext and Social Media; Aug 30 to Sep 2, 2021; Virtual Event USA. [doi: 10.1145/3465336.3475098]
8. Greenwald AG, McGhee DE, Schwartz JL. Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 1998 Jun;74(6):1464-1480. [doi: 10.1037//0022-3514.74.6.1464] [Medline: 9654756]
9. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. Presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 25-29, 2014; Doha, Qatar. [doi: 10.3115/v1/D14-1162]
10. Sarzynska-Wawer J, Wawer A, Pawlak A, et al. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Res 2021 Oct;304:114135. [doi: 10.1016/j.psychres.2021.114135] [Medline: 34343877]
11. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. Preprint posted online on 2018 URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2024-11-05]
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
13. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. TACL 2017 Jul 1;5:135-146. [doi: 10.1162/tacl_a_00051]
14. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. Preprint posted online on Jan 16, 2013. [doi: 10.48550/arXiv.1301.3781]
15. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. Standford NLP Group. 2014. URL: https://nlp.stanford.edu/projects/glove/ [accessed 2023-12-26]
16. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv. Preprint posted online on Mar 26, 2019. [doi: 10.48550/arXiv.1903.10676]
17. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
18. Bert-base-cased. Hugging Face. URL: https://huggingface.co/bert-base-cased [accessed 2023-10-23]
19. Allenai/scibert_scivocab_cased. Hugging Face. URL: https://huggingface.co/allenai/scibert_scivocab_cased [accessed 2023-10-23]
20. Dmis-lab/biobert-v1.1. Hugging Face. URL: https://huggingface.co/dmis-lab/biobert-v1.1 [accessed 2023-10-23]
21. Datasets: wikipedia. Hugging Face. 2022. URL: https://huggingface.co/datasets/wikipedia/viewer/20220301.en [accessed 2023-03-11]
22. FastText model. Gensim. URL: https://radimrehurek.com/gensim/auto_examples/tutorials/run_fasttext.html [accessed 2023-12-20]
23. Word2Vec model. Gensim. URL: https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html [accessed 2023-12-20]
24. Maciejkula/glove-python. GitHub. 2016. URL: https://github.com/maciejkula/glove-python [accessed 2023-12-20]
25. BERT. Hugging Face. URL: https://huggingface.co/docs/transformers/v4.36.1/en/model_doc/bert#transformers.BertForMaskedLM [accessed 2023-12-20]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
GloVe: Global Vectors for Word Representation
IAT: implicit association test
NLP: natural language processing
NLTK: Natural Language Toolkit
SEAT: sentence encoder association test
WEAT: word embedding association test
WEM: word embedding method

Edited by C Lovis; submitted 06.05.24; peer-reviewed by R Huang, ST Arasteh; revised version received 24.07.24; accepted 17.08.24; published 12.11.24.

Please cite as:
Gray M, Milanova M, Wu L
Enhancing Bias Assessment for Complex Term Groups in Language Embedding Models: Quantitative Comparison of Methods
JMIR Med Inform 2024;12:e60272
URL: https://medinform.jmir.org/2024/1/e60272
doi:10.2196/60272

Semiology Extraction and Machine Learning–Based Classification of Electronic Health Records for Patients With Epilepsy: Retrospective Analysis

Yilin Xia[1,*], MD; Mengqiao He[1,*], MS; Sijia Basang[1], MD; Leihao Sha[1], MD; Zijie Huang[1], MBBS; Ling Jin[1], MBBS; Yifei Duan[1], MD; Yusha Tang[1], MD; Hua Li[1], MD; Wanlin Lai[1], MD; Lei Chen[1,2], MD

[1]

[2]

[*]these authors contributed equally

Corresponding Author:
Lei Chen, MD

Abstract

Background: Obtaining and describing semiology efficiently and classifying seizure types correctly are crucial for the diagnosis and treatment of epilepsy. Nevertheless, there exists an inadequacy in related informatics resources and decision support tools.

Objective: We developed a symptom entity extraction tool and an epilepsy semiology ontology (ESO) and used machine learning to achieve an automated binary classification of epilepsy in this study.

Methods: Using present history data of electronic health records from the Southwest Epilepsy Center in China, we constructed an ESO and a symptom-entity extraction tool to extract seizure duration, seizure symptoms, and seizure frequency from the unstructured text by combining manual annotation with natural language processing techniques. In addition, we achieved automatic classification of patients in the study cohort with high accuracy based on the extracted seizure feature data using multiple machine learning methods.

Results: Data included present history from 10,925 cases between 2010 and 2020. Six annotators labeled a total of 2500 texts to obtain 5844 words of semiology and construct an ESO with 702 terms. Based on the ontology, the extraction tool achieved an accuracy rate of 85% in symptom extraction. Furthermore, we trained a stacking ensemble learning model combining XGBoost and random forest with an F_1-score of 75.03%. The random forest model had the highest area under the curve (0.985).

Conclusions: This work demonstrated the feasibility of natural language processing–assisted structural extraction of epilepsy medical record texts and downstream tasks, providing open ontology resources for subsequent related work.

(JMIR Med Inform 2024;12:e57727) doi:10.2196/57727

KEYWORDS

epilepsy; natural language processing; machine learning; electronic health record; unstructured text; semiology; health records; retrospective analysis; diagnosis; treatment; decision support tools; symptom; ontology; China; Chinese; seizure

Introduction

Epilepsy is a major chronic neurological disorder that affects approximately 70 million people and severely reduces the quality of life of patients and their families [1]. Obtaining a correct and complete seizure semiology efficiently is essential for the diagnosis and classification of seizures. However, this process is difficult to achieve. First, the symptoms of seizures are stereotypical but variable, and the same seizure course is in fact a complex combination of multiple symptomatologic elements in time and space. Furthermore, the type of seizure an individual patient experiences can change over the course of the disease [2,3]. Second, seizures have sudden onset, resulting in a short period of time for patients or witnesses to recognize and observe them, and history taking often relies on experienced and careful questioning by epilepsy specialists rather than recording the patient's statements directly [4,5]. Finally, epilepsy specialists are scarce and unevenly distributed worldwide. Nonneurologists, medical students, caregivers, and community workers play important roles in epilepsy care but lack appropriate tools to tease out epilepsy histories and determine classifications [6-9].

In recent years, natural language processing (NLP) has been widely used in the structured processing of clinical text data and development of intelligent diagnostic tools in neurology [10]. NLP methods have been used to automatically extract details from electronic health records (EHRs) of patients with epilepsy, such as categorical diagnosis, abnormal electroencephalogram (EEG) and imaging results, and medications prescribed [11-13]. These data are also used to

accomplish tasks such as automated identification of cohorts of drug-resistant patients and long-term prognostic tracking [14,15]. However, the complexity of epilepsy symptom elements remains a challenge for entity recognition and automatic extraction classification.

Therefore, ontologies were introduced to address this complexity. The concept of ontology is derived from philosophy and is used for formal, structured, domain-specific, and human- and computer-interpretable representations of entities and relationships. It has been widely used in computers, bioinformatics, and medical informatics [16,17]. Application ontology can be used in the medical field to represent established knowledge within a domain and maintain a standardized vocabulary across multiple locations, datasets, and consortiums, allowing for automated computation and decision-making based on structured data. Application ontologies can also be combined with NLP techniques to disambiguate textual concepts and build tools for the knowledge extracted from EHRs [10,18]. This work demonstrated the feasibility of NLP-assisted structural extraction of epilepsy medical record texts and downstream tasks, providing open ontology resources for subsequent related work.

Methods

Dataset

Electronic medical record data were obtained from patients with an *International Classification of Diseases, Tenth Revision* (*ICD-10*) epilepsy diagnosis (G40 or G40.x) who were hospitalized at West China Hospital of Sichuan University and assigned an epilepsy diagnosis between 2010 and 2020. The seizure type of inpatients was determined by discharge diagnosis.

The text information of the current medical history records the details of the occurrence, evolution, diagnosis, and treatment of the patient's disease; is written in chronological order; and is divided into the following parts: onset of the disease, including the time and place of onset; antecedent symptoms; probable causes or triggers; characteristics of the main symptoms and their development and change (describing the location, nature, duration, degree, factors of relief or aggravation, and evolution of the main symptoms in sequential order); accompanying symptoms; diagnosis and treatment since the onset of the disease; and the patient's general condition since the onset of the disease.

Ethical Considerations

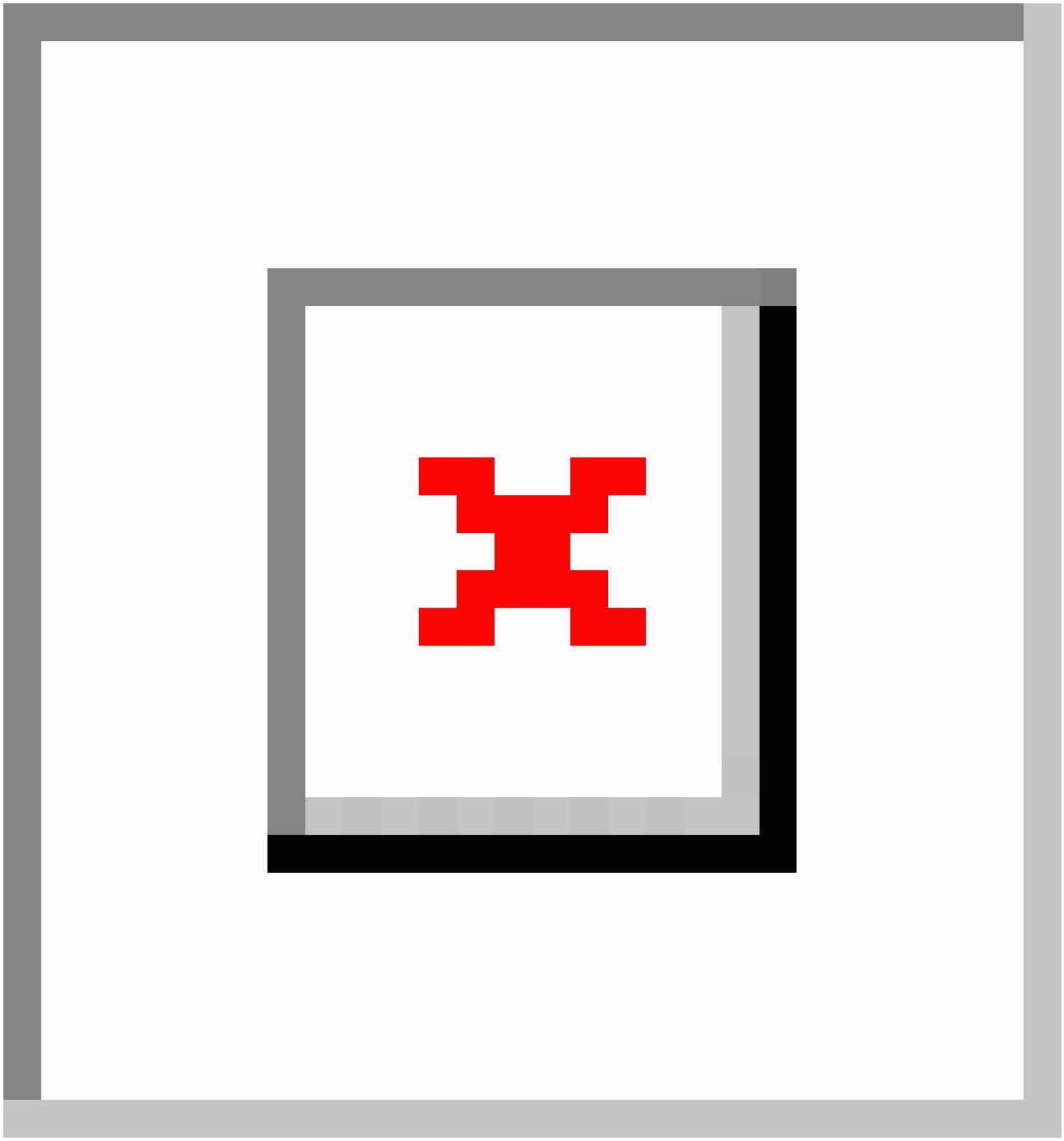
The study was reviewed and approved by the Ethics Committee of West China Hospital of Sichuan University (2022(1083)). Since the data were obtained from previous medical records, we have received approval from the ethics committee for a waiver of informed consent. The study data were deidentified, and the privacy and personal information of the subjects were protected.

Framework for Standardizing Seizure Information

We proposed a seizure extraction framework for mining and structuring important information related to seizures from the presenting medical histories of patients with epilepsy (Figure 1). The framework requires the extraction of the following information:

1. Time stamp: The important point in time at which the patient's condition has changed since today.
2. Location: Seizure site refers to the anatomical parts of the body corresponding to the symptom performance.
3. Symptom: Symptom performance refers to the symptoms and signs that appear during the seizure.
4. Duration of seizure event (episode time): Duration of epileptic events within the seizure episode.
5. Status: Occurrence state refers to the state corresponding to the symptom performance, including "with," "without," or "unknown."
6. Frequency: The frequency of seizures, for example: once a month, and so forth.

Figure 1. Example of the standardized framework.



Labeling Process

Six annotators completed the labeling process. Four of them, junior physicians (SB, LS, LJ, and YD) specializing in epilepsy or epilepsy researchers, were responsible for independently extracting seizure-related information from 2500 raw texts of presenting medical histories according to a standardized framework. Two senior physicians (HL and WL) specializing in epilepsy were responsible for discussing and formulating the framework of the annotation and the rules that should be followed during annotation to ensure reliability, providing uniform training to the annotators, and manually reviewing the final results of the annotation. Annotation rules included the following:

When a particular Chinese phrase used to describe the seizure process was a fixed collocation, the phrase was extracted as a whole without separating the verb and the object (usually a location) in it individually, in order to avoid a decrease in the specificity of the extraction.

Due to the specificity of the commonly used symptomatology phrases in the Chinese section, it is important to ensure that the symptomatic manifestations are extracted at the coarsest possible granularity, that is, descriptive phrases that include seizure state and seizure site are avoided. However, phrases should not be disassembled when they cannot be clearly recognized as symptoms, such as lip smacking (oropharyngeal automatisms) and hand rubbing (hand automatisms), and the anatomical part of the phrase should be retained. It should also be confirmed

that all seizure symptomatology is extracted from seizures and not from other symptoms accompanying epilepsy. Cognitive decline, such as memory and attention, should not be included in labeling.

Do not standardize the presentation of the extracted information and keep it as original as possible.

To assess the consistency of the annotations by the 4 annotators, 50 identical medical records were included without their knowledge. Two senior physicians provided reference standards for the annotation of the 50 medical records. We used Fleiss's κ to calculate interannotator agreement. By convention, κ value above 0.80 indicates "near-perfect" agreement.

Bilingual Ontology Construction for Seizure Semiology

Compared with other parts of the seizure information framework, epileptic semiology expression and the diversity of expression extraction tasks are more challenging, especially for Chinese EHRs of epilepsy. Therefore, we constructed a bilingual ontology to share the lexicon obtained from manual extraction and annotation. It can be further used, evaluated, and refined for future Chinese epilepsy history extraction tasks.

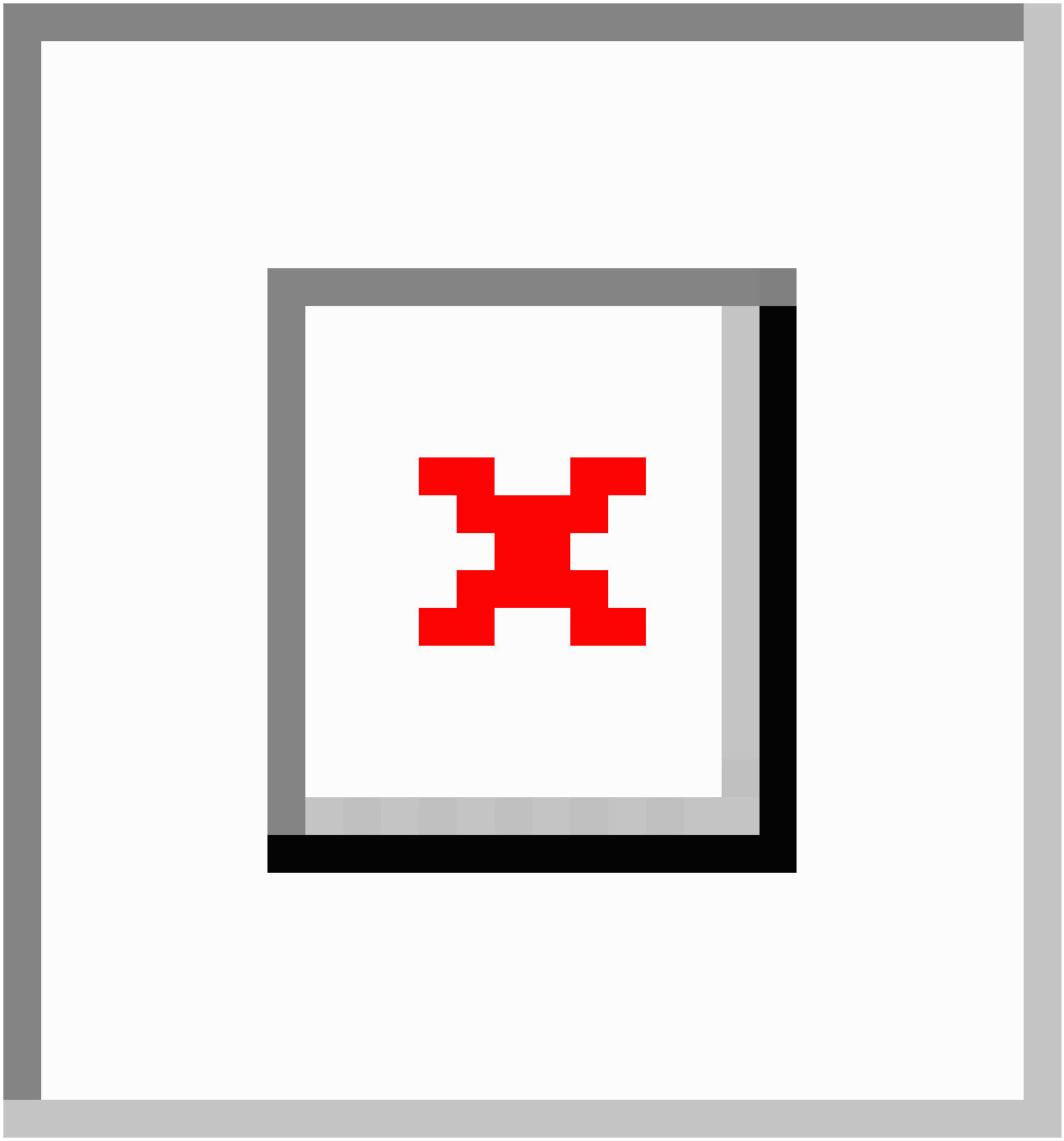
We defined the scope of this domain of ontology as epileptic semiology by reference, reused the more authoritative epilepsy-related ontologies and terminology sets as standard terminology, referred to the basic formalized ontology (BFO) as the top-level ontology, and hierarchically arranged the entities according to their domain-neutral framework. Then, we deemphasized the annotated symptoms collected in the annotation phase to eliminate redundancy and placed them into the corresponding terms as their synonymous expression properties. We used Protégé as the editor of the ontology and uploaded it in Ontology Web Language (OWL) as the first version of the world's largest ontology browser, BioPortal.

Extraction Process and Evaluation of Extraction Results

We used some NLP tools to structure the extraction of current medical history from EHRs. We imported the organized dictionaries of symptom performance, symptom nature, seizure frequency, and seizure site into the Jieba tokenizer and initialized the Part-of-Speech Tagger (Postagger) and Dependency Parser (Parser) of the pyltp [19] plug-in using existing models (pos.model, parser.model). pyltp provides a series of Chinese NLP tools, and users can use these tools for Chinese text segmentation, part-of-speech tagging, parsing, and so on.

Specifically, in the data preprocessing stage, we first imported organized dictionaries of symptom presentation, symptom type, seizure frequency, and seizure location. These dictionaries are used for subsequent segmentation and feature extraction. We used Postagger to tag the parts of speech of the tokenized results and Parser to analyze the dependency relations of the words in the current sentence or context. Next, we performed text segmentation and annotation, using Jieba Segmenter to segment the medical history text in the EHR. Jieba Segmenter is able to accurately slice and dice the text based on the imported dictionaries. Postagger was called to lexically annotate the segmentation results by identifying the lexical properties of each word. The dependencies between words are analyzed using Parser to determine the syntactic structure between words. Then, to extract symptom information, we iteratively processed the participle results by combining a list of negatives, a list of transitive or logical connectives, and a list of temporal adverbs. These normalized lists allowed us to accurately identify positive and negative symptom information. In each sentence, information such as the location, type, duration, and frequency of symptom episodes was extracted. Finally, the extracted information such as positive and negative symptoms, location, nature, duration, and frequency of episodes was structured and stored in the output dictionary according to the temporal nodes. The overall process flow is illustrated in Figure 2.

Figure 2. Extraction modeling workflow.



The software and programming languages used included Python 3.8.8, pyltp 0.2.0, pandas 1.4.2, and Jieba 0.42.1.

After the extraction was completed, we randomly selected 200 cases from all the results for manual inspection to comprehensively assess the extraction capability and obtain the accuracy for 6 aspects separately: time stamp, symptom, location, episode time, status, and frequency.

Seizure Classification Based on Machine Learning

Our work aimed to build a binary classification model capable of distinguishing between generalized and focal seizures. The analysis process, based on supervised machine learning, consisted of the following steps: data preprocessing, feature selection, algorithm selection, parameter tuning, and performance evaluation.

Data Preprocessing

Our extraction tool was used to retrieve semiology data of the patients. After preprocessing 16,587 records by *ICD* coding combined with regular expression matching, 10,098 records were excluded because they did not receive a clear classification (60%).

A total of 6489 medical history text records with a diagnosis of generalized or focal seizure were retained, including 2632 records of generalized epilepsy and 3857 records of focal epilepsy. After communication with clinicians, 103 symptom words were defined to cover the main symptoms that can occur

in patients with epilepsy. We used text-matching techniques to map the symptom descriptions in each record to these 103 symptom words. Specifically, for each record, if a symptom word was mentioned in the text, we marked the corresponding symptom word as 1; if it was not mentioned, it was marked as 0. For example, if a record mentioned "Clonic" but not "Foaming at Mouth," then the field for "Clonic" was set to 1, and the field for "Foaming at Mouth" was set to 0.

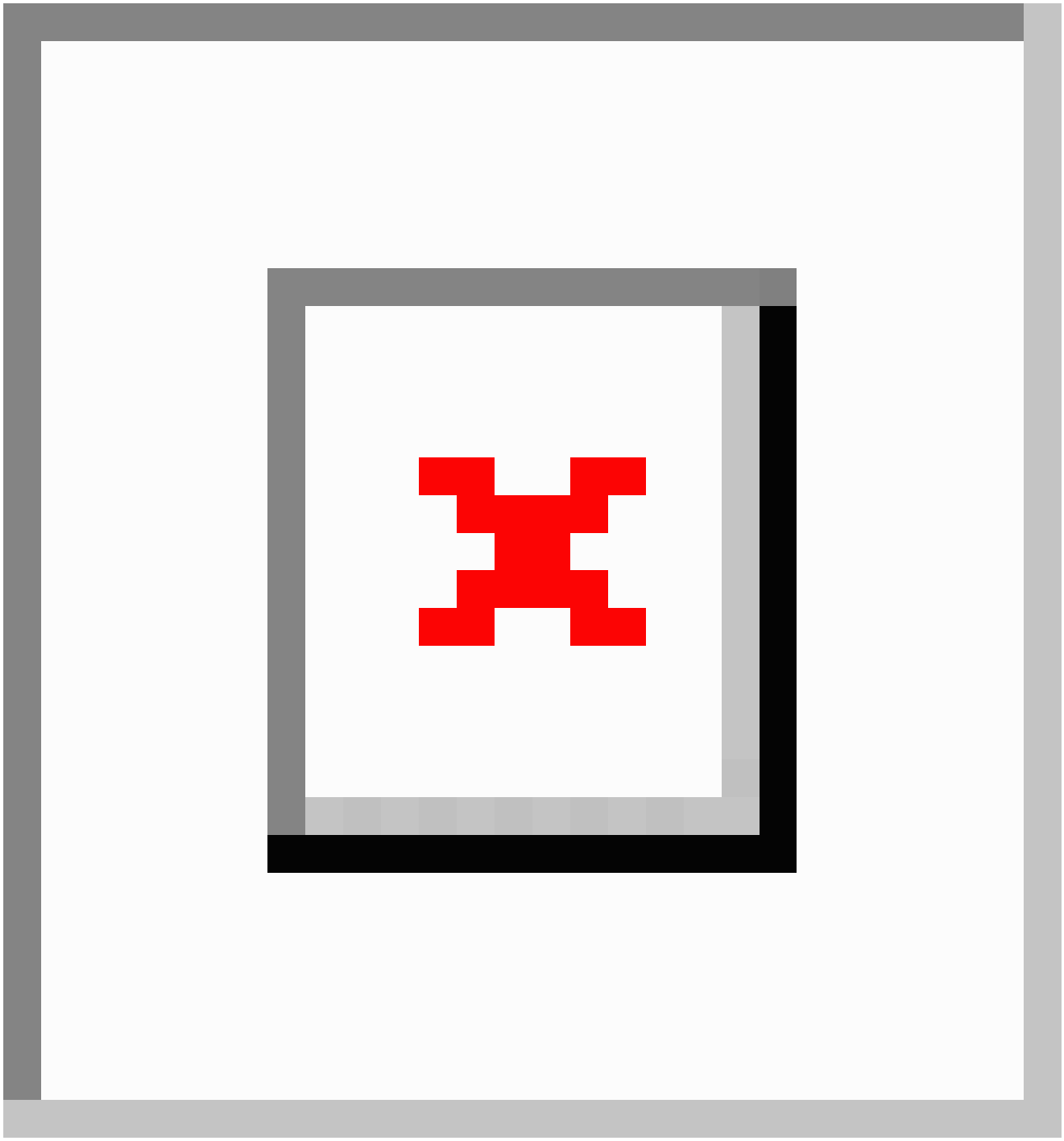
Feature Selection

We used several feature selection techniques to identify the most relevant features for the classification task. Specifically, we used recursive feature elimination, random forest–based feature importance, mutual information, and the SelectKBest method using the ANOVA *F* value. Each method was systematically applied to the feature matrix (X) and the label vector (y) to generate a reduced set of features. We varied the number of retained features (k) across multiple values to evaluate its impact on model performance. In addition, we examined the effects of different sample ratios on the model's performance.

Algorithm Selection and Parameter Tuning

Subsequently, we divided the preprocessed dataset into training and testing sets at a 7:3 ratio. We used 4 types of models as base models: decision tree [20], random forest [21], XGBoost [22], and LightGBM [23]. Using grid search algorithms and k-fold cross-validation, we optimized the hyperparameters of the models with training to enhance the model accuracy. Specific parameters are detailed in Multimedia Appendix 1. We also introduced the stacking ensemble learning method, which was conducted in 2 stages, as illustrated in Figure 3. In the first stage, we performed 5-fold cross-validation. Specifically, we divided the training dataset into 5 parts, with 4 serving as the training set for base model training and the remaining part serving as the validation set for generating new training data. Simultaneously, we predicted the entire test set (test_data) to create a new test dataset. In the second stage, we used the training and testing sets generated from the first stage as inputs for further training and prediction using the logistic regression model, resulting in the final outcome. In this study, we combined the XGBoost model with the random forest and LightGBM models for combined training and testing.

Figure 3. Stacking integration learning process. EHR: electronic health record.



Performance Evaluation

Finally, we used the test set to evaluate the precision, recall, F_1-scores, and the area under the receiver operating characteristic curve (ROC) value of the model. We designated "generalized epilepsy" as label A and "focal epilepsy" as label B. TP(*A*) represents true positives, FP(*A*) represents false positives, and FN(*A*) represents false negatives for label A, and similarly for label B.

Precision is defined by the following formula:

(1)Precision=12(TP(A)TP(A)+FP(A)+TP(B)TP(B)+FP(B))

Recall is defined by the following formula:

(2)Recall=12(TP(A)TP(A)+FN(A)+TP(B)TP(B)+FN(B))

The F_1-score (F_1) is defined by the following formula:

(3)F1=2×Precision×RecallPrecision+Recall

For the classification analysis of seizure, the following software or programming language versions were used: Python 3.8.8, NumPy 1.24.3, pandas 1.4.2, scikit-learn 1.3.2, XGBoost 2.0.1, and LightGBM 4.1.0.

Bilingual Ontology Construction for Seizure Semiology

Compared with other parts of the seizure information framework, epileptic semiology expression and the diversity of expression extraction tasks are more challenging, especially for Chinese EHRs of epilepsy. Therefore, we constructed a bilingual ontology to share the lexicon obtained from manual extraction and annotation. In developing epilepsy semiology ontology

(ESO), we followed 5 of the 7 steps of the Stanford methodology: (1) defining the domain and scope of the ontology, (2) reusing existing ontologies to the extent possible, (3) enumerating ontology terms, (4) defining classes and class hierarchies, and (5) defining class attributes (Multimedia Appendix 2).

In the first step, epileptologists and the ontology development team met biweekly to define the scope of the ontology and to ensure that the goals remained constant throughout its development. In steps 2 and 3, we standardized terminology by referring to existing, more authoritative epilepsy-related ontologies and terminology sets. In the fourth step, we adopted the BFO as the top-level ontology. In the fifth step, we de-emphasized the annotated symptoms collected in the annotation phase to eliminate redundancy and placed them into the corresponding terms as their synonymous expression properties. Finally, we rendered the ontology using the OWL in the Protégé ontology editor and uploaded it to the world's largest ontology browser, Bioportal, as a first version.

Results

Patient Cohort

The study cohort included 10,925 patients and 10,658 texts of presenting medical histories. The patient cohort included 42% (4588/10,925) females and 58% (6337/10,925) males with a mean age of 31.45 (age range: 1 - 92) years. The presenting medical history texts were independently written and completed by 117 physicians. Fifty-seven percent (6227/10,925) of the patients in the patient cohort ultimately received a definitive diagnostic classification of seizures at the time of discharge, with 32% (1992/6227) of patients having focal epilepsy and 26% (1619/6227) having generalized epilepsy.

Assessment of Labeling Quality Control Results and Extraction Capacity

In the annotation phase, we assigned 50 identical texts to the annotators without their knowledge to test the consistency of their annotations. The κ-value of the 4 annotators was 0.862, indicating a high degree of consistency.

After completing the extraction using the model, we manually inspected a random sample of 200 notes from the extraction results (which included 235 seizures) to assess the extraction performance of the model. The extraction results for the 5 dimensions are shown in Table 1.

Table . Extraction performance.

	Time stamp	Location	Symptom	Episode time	Status	Frequency
Total number of elements by reviewer annotation, "gold standard"	235	512	1325	183	1325	106
Total number of elements by algorithm report	196	516	1219	175	1302	93
Number of correct algorithm-reported elements	181	507	1126	145	1254	84
Recall, n/N (%)	181/235 (77)	507/512 (99)	1126/1325 (85)	145/183 (79)	1254/1325 (95)	84/106 (79)
Precision, n/N (%)	181/196 (92)	507/512 (98)	1126/1219 (92)	145/175 (82)	1254/1302 (96)	84/93 (90)
F_1-score	0.83	0.98	0.88	0.80	0.95	0.84

Epilepsy Semiology Ontology

The overall hierarchical structure of ESO adheres to the architecture of the top-level ontology BFO, which supports semantic interoperability between ontologies, starting from "continuant" and "occurrent" under "entity."

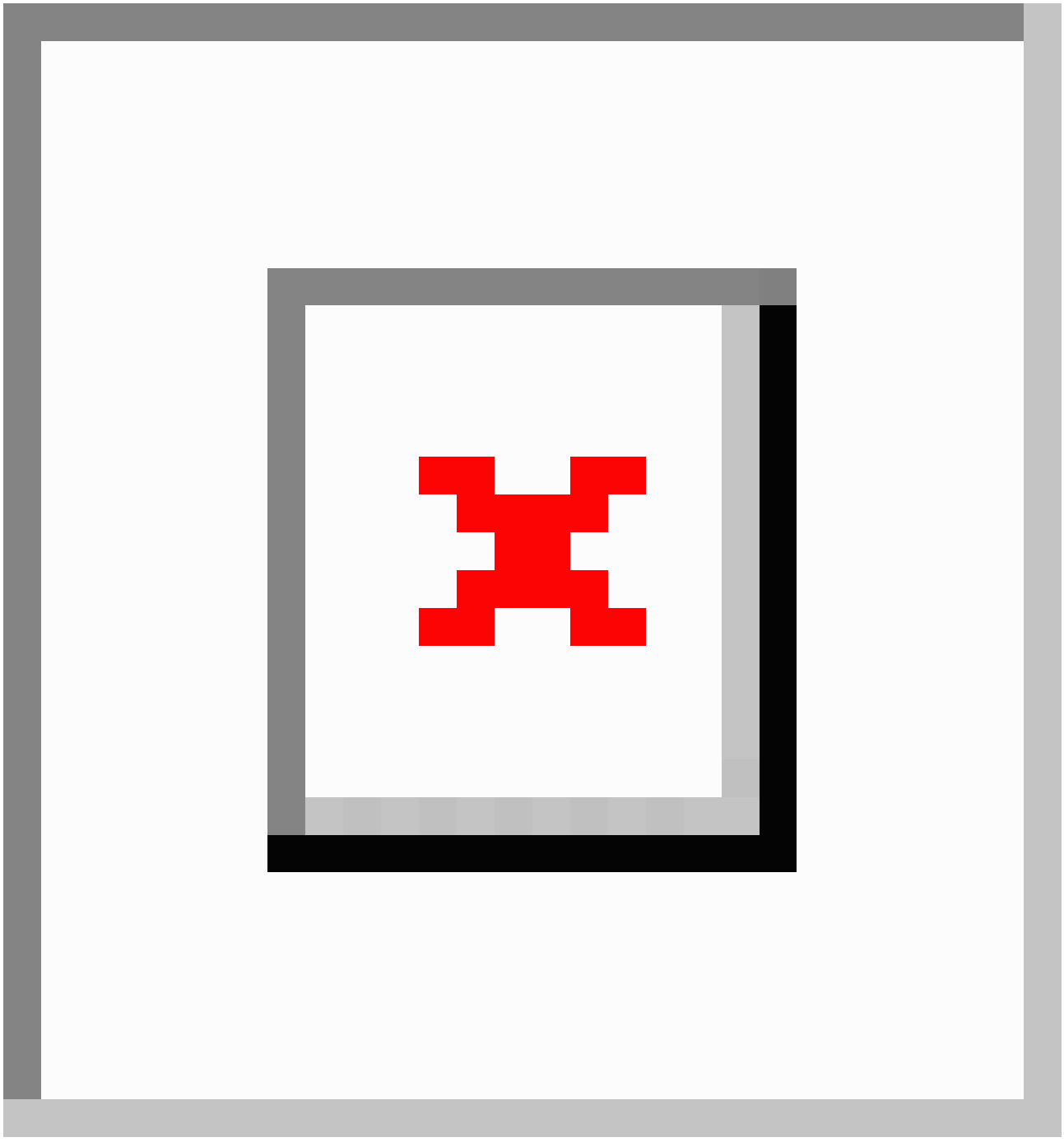
The ESO contains a total of 176 terms, most of which are based on the nominal entity "anatomical entity" and the process "physiological pathological process," with a maximum depth of 10 layers. According to the principle of ontology reuse, we partially reused and rearranged the concepts of "pathophysiological process" and its leaf nodes in epilepsy and seizure ontology (EPSO) [24] and also referred to the existing semiology terminology collection of the International League Against Epilepsy, which includes a total of 132 epilepsy semiology terms. In terms of seizure sites, we referred to the "Bodily Feature" section of Systemized Nomenclature of Medicine Clinical Terms (SNOMED CT) [25] and EPSO, which contains a total of 32 seizure-site terms. The purpose, scope, language, and users are listed in Multimedia Appendix 3.

As an important step in implementing the medical record extraction function of the application ontology, we added Chinese translations and synonyms of symptom performance as entity attributes (Multimedia Appendix 3). After annotating 2500 medical records, we obtained 5844 words of semiology. After de-emphasizing and removing nonepileptic seizure symptoms (usually abnormal general conditions and comorbid symptoms), we obtained 702 terms, 75 primary terms, and their synonyms. Among them, there were more than 30 synonyms for holding, dropping, and vocalization.

Performance of Seizure Classification

In the feature selection process, we found that choosing 103 features among the 4 feature selection methods gave the best results, and we also observed that choosing different sample ratios for training had little impact on the model performance (Multimedia Appendix 4). On this basis, we optimized the parameters and trained 4 foundational models—decision tree, random forest, XGBoost, and LightGBM—to distinguish between generalized and focal epilepsy. Figure 4A-E illustrates the contribution of each symptom feature to the predictive decisions of these models. Notably, "clonic," "tonic," "unresponsive to call," "eyes rolled up," "foaming at mouth," and "fall" are pivotal in differentiating seizure types.

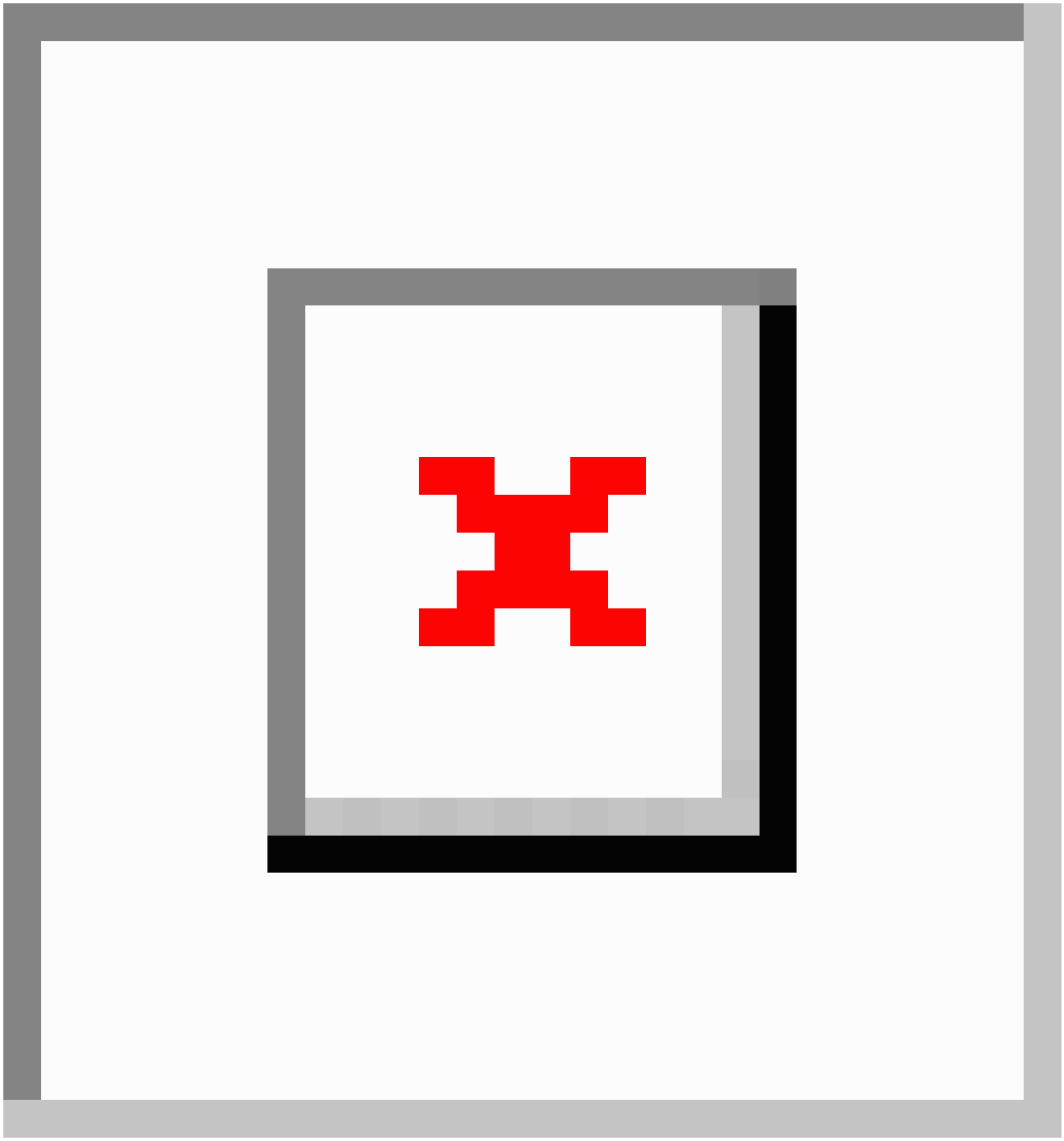
Figure 4. Distribution of important features of the base model. (A) Decision tree model important features. (B) Random forest model important features. (C) XGBoost model important features. (D) LightGBM model important features. (E) important features of the base model Wayne chart.



In addition, we trained a stacking ensemble learning model. As shown in Figure 5A-C, the stacking ensemble model outperformed the other base models in terms of precision, recall, and F_1-score. Among them, the ensemble model combining XGBoost and random forest yielded the best results, with the highest F_1-score (75.03%). We also compared the ROCs of the various models represented by different colors. Notably, the random forest model and XGBoost+random forest ensemble model outperformed the other models, as indicated by the orange and blue lines, respectively. As shown in Figure 5D, the random forest model had the highest area under the curve (AUC)—0.984—whereas the XGBoost+random forest ensemble

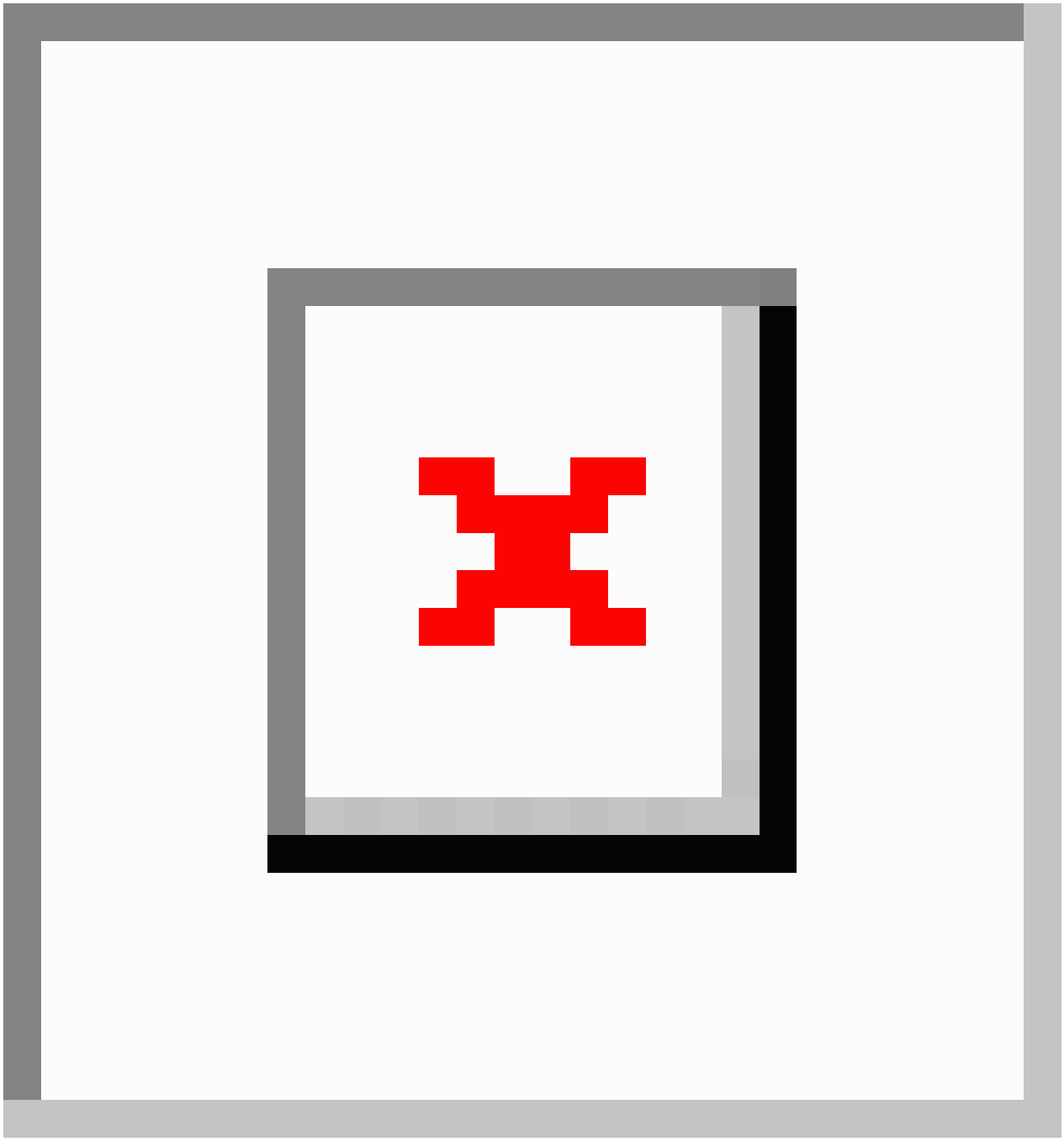
model had an AUC of 0.919, with the AUCs of the other models falling below these 2.

Figure 5. Comparison of model evaluations plotted against ROCs. (A) Comparison of precision across models. (B) Comparison of recall across models. (C) Comparison of F_1-scores across models. (D) Comparison of ROCs across models. AUC: area under the curve; ROC: receiver operating characteristic curve.



Ultimately, we selected the ensemble model combining XGBoost and random forest for predicting seizure classification and visualized its confusion matrix. As shown in Figure 6, the model has a precision of 0.68 for predicting "generalized epilepsy" and a precision of 0.80 for predicting "focal epilepsy."

Figure 6. XGBoost+random forest confusion matrix plot.



Discussion

Principal Findings

In this study, the first Chinese-English ontology of epilepsy semiology was established, the first non–English-structured extraction of epilepsy history text was achieved by combining manual annotation and NLP techniques, and automatic seizure classification was further accomplished based on the data extracted by the tool.

Comparison to Prior Work

Ninety percent of the disease burden caused by epilepsy is borne by resource-limited countries. China has more than 12% of patients with epilepsy worldwide [26,27]. The Global Burden of Disease study reported that, in 2019, China's disability-adjusted life years (DALYs) due to epilepsy accounted for 10% of the global DALYs and 94% of the DALYs in East Asia [28]. However, the development of Chinese language EHR processing tools for epilepsy has been delayed because of the lack of high-quality corpora such as relevant terminology sets. English ontologies and terminology systems, including SNOMED CT, Unified Medical Language System, and EPSO [26], are limited by the problems of diverse descriptions of Chinese medical entities, fuzzy boundaries, and the existence of nested relationships. Therefore, it is more difficult to support clinical terminology extraction from Chinese medical records after "Chinese-ization" [29]. The technical challenges of Chinese NLP lie in its complex word-splitting process, high-frequency

ambiguity phenomenon, and flexible and variable sentence construction [30]. By contrast, English NLP is relatively simple to process because of its clear separation of words by spaces, more standardized syntactic structures, and abundant processing resources. Despite these differences, the gap between Chinese and English NLP technologies is gradually narrowing as deep learning and pretrained language models continue to advance and multilingual processing capabilities are significantly enhanced. In this study, the ontology and extraction tool constructed based on the corpus of the Southwest Epilepsy Center can better serve the grassroots areas in western China, where the burden of epilepsy is high and medical resources are relatively scarce, thereby bridging the world's health disparities for people with epilepsy [26,31].

In this study, for the first time, the symptom elements of epileptic seizures were extracted at an ultrafine granularity, the accuracy of the extraction of the features reached 0.85, and the classification of generalized and focal seizures relying on the symptom features alone reached an AUC of 0.985. We also found that the key features in the classifier corresponded to the "red flag" symptoms used by human experts, yielding a list of symptoms including "clonic," "tonic," "unresponsive to call," "eyes rolled up," "foaming at mouth" and "fall," which are the same basic key features as those categorized by human experts' guidelines [2]. To the best of our knowledge, this is the first time that a present history of epilepsy has been extracted and automatically categorized with symptom element granularity [32,33]. Barbour et al [34] created regular expressions manually as well as creating false-positive filters and disambiguated them using conditional matching to extract entities such as seizure type, with internally tested F_1-values ranging from 0.86 to 0.90. Vulpius et al [35] extracted seizure epilepsy types primarily by manually constructing dictionaries.

However, these 2 studies were based only on existing unstructured diagnostic texts rather than indirect inference through medical history texts, and only automated extraction, rather than automated classification based on symptom features, was achieved. In our seizure classification task, we used a stacking integration technique to combine the XGBoost and random forest models (AUC=0.919). Despite the higher AUC of the random forest model, it may have lower precision or recall in some categories, resulting in a less favorable F_1-score than the stacking method. The stacking method, on the other hand, by combining the advantages of both random forest and XGBoost, may achieve a more balanced performance across all categories, thereby improving the F_1-score.

Although downstream tasks for seizure classification currently exist, most rely on a single-model architecture, such as support vector machine, linear model, or XGBoost [35,36]. However, by pooling multiple underlying models using stacking techniques, it is possible to improve model performance and reduce the risk of overfitting, which in turn improves the model's generalization capabilities.

Future Directions

Beyond the initial diagnosis and classification of seizure, our study has the potential to identify specific types of epilepsy. For example, the classification of adolescent myoclonic epilepsy may change over the course of a single patient's illness, with a predominance of absence and myoclonic seizures initially, followed by intensification of generalized tonic-clonic seizures in adulthood or after practice tasks [3]. This type of epilepsy is difficult to recognize because of changes from pediatric and adult neurologists. Plug-ins based on extraction and classification models can be developed to alert epileptologists to consider this particular type.

In addition, accurate extraction of seizure duration and frequency has been used in epilepsy research to help clinical researchers accurately screen retrospective cohorts in vast multicenter electronic health information databases, for example, by accelerating the speed of patient recruitment and data collection, screening of rare epilepsy cohorts [37], and screening of persistent status epilepticus in children [38]. The extracted data also enable the dynamic and automated monitoring of postmedication efficacy, epidemiological statistics, and medical economics studies on a larger scale. In the future, we will consider the use of deep learning models and the addition of multimodal features such as imaging and EEG in the seizure classification task to achieve a more accurate and dynamically changing classification capability based on the patient's journey. With further improvements in extraction and classification accuracy, automated symptom-based classification will be uniquely suited to help primary care physicians and other specialists accurately classify epilepsy and select appropriate medications. In conclusion, this work demonstrates the feasibility of NLP-assisted structured extraction of epilepsy history text and downstream tasks in Chinese and provides an open ontology resource for subsequent related work.

Limitations

This study also has some limitations. First, including the fact that the data source was only from a single center, we have not yet verified its transferability to other regions in China. Second, we have not yet applied the ontology to real clinical scenarios, such as assisting clinicians in structured and efficient registration of epilepsy history. Third, the accuracy of dependent syntax analysis is crucial to the effectiveness of information extraction, and the flexibility of Chinese grammar adds to the difficulty of the analysis. Fourth, although current deep learning techniques have gained momentum to improve the situation, they also require finer tuning and extensive contextual adaptation testing. Fifth, our ontology remains in its initial iteration. There is currently no systematic approach to quality assessment and verification. We will continue to expand and refine the ontology data. In the future, other dimensions and modalities should be added to the features, including EEG and imaging, to further improve the accuracy of classification and the completion of more downstream tasks.

Conclusions

Clinically significant seizure information was successfully extracted from Chinese medical histories using NLP. This innovative approach represents a powerful tool for clinical research, with numerous potential applications, particularly for disorders characterized by complex clinical symptoms, such as seizure disorders. During this process, we constructed a bilingual

ontology of seizure symptomatology comprising 702 terms. Furthermore, leveraging the extracted symptomatology information, we trained a binary classification model for generalized versus focal epilepsy using the stacking ensemble learning method. This demonstrates the feasibility of performing downstream tasks, such as seizure classification, based on the extracted information.

Acknowledgments

We are very grateful to Bairong Shen and Xingyun Liu from the Institute of Systems Genetics of West China Hospital for their guidance on ontology construction. This work was financially supported by TianYuan Special Funds of the National Natural Science Foundation of China (No. 12026607) and Sichuan Science and Technology Program (2023YFS0047).

Authors' Contributions

YX and LC contributed to study conception and design. YT, HL, and WL participated in data acquisition and curation. SB, LS, LJ, YD, HL, and WL participated in the data labeling process. YX and ZH contributed to ontology construction. YX, MH, SB, and LS participated in the analysis of data and extraction process. YX, MH, and LC contributed to drafting/revision of the manuscript for content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File, 574 KB - medinform_v12i1e57727_app1.docx]

Multimedia Appendix 2
Construction process of epilepsy semiology ontology.
[PNG File, 152 KB - medinform_v12i1e57727_app2.png]

Multimedia Appendix 3
Table S1. Purpose, scope, language and users of WWECA.
[XLSX File, 62 KB - medinform_v12i1e57727_app3.xlsx]

Multimedia Appendix 4
Model performance with different feature selection methods and sample ratios.
[PNG File, 1956 KB - medinform_v12i1e57727_app4.png]

References

1. Thijs RD, Surges R, O'Brien TJ, Sander JW. Epilepsy in adults. Lancet 2019 Feb 16;393(10172):689-701. [doi: 10.1016/S0140-6736(18)32596-0] [Medline: 30686584]
2. Fisher RS, Cross JH, D'Souza C, et al. Instruction manual for the ILAE 2017 operational classification of seizure types. Epilepsia 2017 Apr;58(4):531-542. [doi: 10.1111/epi.13671]
3. Cerulli Irelli E, Morano A, Orlando B, et al. Seizure outcome trajectories in a well-defined cohort of newly diagnosed juvenile myoclonic epilepsy patients. Acta Neurol Scand 2022 Mar;145(3):314-321. [doi: 10.1111/ane.13556] [Medline: 34791656]
4. Wardrope A. The promises and pitfalls of seizure phenomenology. Seizure 2023 Dec;113:48-53. [doi: 10.1016/j.seizure.2023.11.008] [Medline: 37976801]
5. Muayqil TA, Alanazy MH, Almalak HM, et al. Accuracy of seizure semiology obtained from first-time seizure witnesses. BMC Neurol 2018 Sep 1;18(1):135. [doi: 10.1186/s12883-018-1137-x] [Medline: 30172251]
6. Patterson V, Samant S, Singh MB, Jain P, Agavane V, Jain Y. Diagnosis of epileptic seizures by community health workers using a mobile app: a comparison with physicians and a neurologist. Seizure 2018 Feb;55:4-8. [doi: 10.1016/j.seizure.2017.12.006] [Medline: 29291457]
7. Goodwin M. Do epilepsy specialist nurses use a similar history-taking process as consultant neurologists in the differential diagnosis of patients presenting with a first seizure? Seizure 2011 Dec;20(10):795-800. [doi: 10.1016/j.seizure.2011.08.003] [Medline: 21920782]
8. Kakisaka Y, Jin K, Fujikawa M, Kitazawa Y, Nakasato N. Teleconference-based education of epileptic seizure semiology. Epilepsy Res 2018 Sep;145:73-76. [doi: 10.1016/j.eplepsyres.2018.06.007] [Medline: 29913406]

9. Benbir G, Demiray DY, Delil S, Yeni N. Interobserver variability of seizure semiology between two neurologist and caregivers. Seizure 2013 Sep;22(7):548-552. [doi: 10.1016/j.seizure.2013.04.001] [Medline: 23611301]
10. Ge W, Rice HJ, Sheikh IS, et al. Improving neurology clinical care with natural language processing tools. Neurology 2023 Nov 27;101(22):1010-1018. [doi: 10.1212/WNL.0000000000207853] [Medline: 37816638]
11. Cui L, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS. EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification. AMIA Annu Symp Proc 2012;2012:1191-1200. [Medline: 23304396]
12. Maldonado R, Harabagiu SM. Active deep learning for the identification of concepts and relations in electroencephalography reports. J Biomed Inform 2019 Oct;98:103265. [doi: 10.1016/j.jbi.2019.103265] [Medline: 31470094]
13. Fonferko-Shadrach B, Lacey AS, Roberts A, et al. Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system. BMJ Open 2019 Apr 1;9(4):e023232. [doi: 10.1136/bmjopen-2018-023232] [Medline: 30940752]
14. Castano VG, Spotnitz M, Waldman GJ, et al. Identification of patients with drug-resistant epilepsy in electronic medical record data using the Observational Medical Outcomes Partnership Common Data Model. Epilepsia 2022 Nov;63(11):2981-2993. [doi: 10.1111/epi.17409] [Medline: 36106377]
15. Xie K, Gallagher RS, Shinohara RT, et al. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. Epilepsia 2023 Jul;64(7):1900-1909. [doi: 10.1111/epi.17633] [Medline: 37114472]
16. Lhatoo SD, Bernasconi N, Blumcke I, et al. Big data in epilepsy: clinical and research considerations. Report from the Epilepsy Big Data Task Force of the International League Against Epilepsy. Epilepsia 2020 Sep;61(9):1869-1883. [doi: 10.1111/epi.16633] [Medline: 32767763]
17. Ong E, Wang LL, Schaub J, et al. Modelling kidney disease using ontology: insights from the Kidney Precision Medicine Project. Nat Rev Nephrol 2020 Nov;16(11):686-696. [doi: 10.1038/s41581-020-00335-w] [Medline: 32939051]
18. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018 Oct 11;379(15):1452-1462. [doi: 10.1056/NEJMra1615014] [Medline: 30304648]
19. Che W, Feng Y, Qin L, Liu T. N-LTP: an open-source neural language technology platform for Chinese. arXiv. Preprint posted online on Sep 24, 2020. [doi: 10.48550/arXiv.2009.11616]
20. Breiman L. Classification and Regression Trees, 1st edition: Routledge; 1984.
21. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, CA p. 785-794.
23. Yan J, Xu Y, Cheng Q, et al. LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol 2021 Sep 20;22(1):271. [doi: 10.1186/s13059-021-02492-y] [Medline: 34544450]
24. Sahoo SS, Lhatoo SD, Gupta DK, et al. Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care. J Am Med Inform Assoc 2014;21(1):82-89. [doi: 10.1136/amiajnl-2013-001696] [Medline: 23686934]
25. Clinical medicine. ScienceDirect. URL: https://www.sciencedirect.com/topics/medicine-and-dentistry/clinical-medicine [accessed 2024-10-15]
26. Lin Y, Hu S, Hao X, et al. Epilepsy centers in China: current status and ways forward. Epilepsia 2021 Nov;62(11):2640-2650. [doi: 10.1111/epi.17058] [Medline: 34510417]
27. Gu L, Liang B, Chen Q, et al. Prevalence of epilepsy in the People's Republic of China: a systematic review. Epilepsy Res 2013 Jul;105(1-2):195-205. [doi: 10.1016/j.eplepsyres.2013.02.002] [Medline: 23507331]
28. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222. [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
29. Rui Z, Wei C, Hao Y, Ran L, Mi-ye W. Enriching plan for Chinese synonyms in medical terms. Chin J Med Libr Inf Sci 2021;30(2):25-32. [doi: 10.3969/j.issn.1671-3982.2021.02.005]
30. de Boer HM, Mula M, Sander JW. The global burden and stigma of epilepsy. Epilepsy Behav 2008 May;12(4):540-546. [doi: 10.1016/j.yebeh.2007.12.019] [Medline: 18280210]
31. Yi H, Liu H, Wang Z, et al. The competence of village clinicians in the diagnosis and management of childhood epilepsy in Southwestern China and its determinants: a cross-sectional study. Lancet Reg Health West Pac 2020 Oct;3:100031. [doi: 10.1016/j.lanwpc.2020.100031] [Medline: 34327383]
32. Decker BM, Turco A, Xu J, et al. Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record. Seizure 2022 Oct;101:48-51. [doi: 10.1016/j.seizure.2022.07.010] [Medline: 35882104]
33. Xie K, Gallagher RS, Conrad EC, et al. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. J Am Med Inform Assoc 2022 Apr 13;29(5):873-881. [doi: 10.1093/jamia/ocac018] [Medline: 35190834]
34. Barbour K, Hesdorffer DC, Tian N, et al. Automated detection of sudden unexpected death in epilepsy risk factors in electronic medical records using natural language processing. Epilepsia 2019 Jun;60(6):1209-1220. [doi: 10.1111/epi.15966] [Medline: 31111463]

35. Vulpius SA, Werge S, Jørgensen IF, et al. Text mining of electronic health records can validate a register-based diagnosis of epilepsy and subgroup into focal and generalized epilepsy. Epilepsia 2023 Oct;64(10):2750-2760. [doi: 10.1111/epi.17734] [Medline: 37548470]
36. Fernandes M, Cardall A, Jing J, et al. Identification of patients with epilepsy using automated electronic health records phenotyping. Epilepsia 2023 Jun;64(6):1472-1481. [doi: 10.1111/epi.17589] [Medline: 36934317]
37. Barbour K, Tian N, Yozawitz EG, et al. Creating rare epilepsy cohorts using keyword search in electronic health records. Epilepsia 2023 Oct;64(10):2738-2749. [doi: 10.1111/epi.17725] [Medline: 37498137]
38. Chafjiri FMA, Reece L, Voke L, et al. Natural language processing for identification of refractory status epilepticus in children. Epilepsia 2023 Dec;64(12):3227-3237. [doi: 10.1111/epi.17789] [Medline: 37804085]

Abbreviations

AUC: area under the curve
BFO: basic formalized ontology
DALYs: disability-adjusted life years
EEG: electroencephalogram
EHR: electronic health record
EPSO: epilepsy and seizure ontology
ESO: epilepsy semiology ontology
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
NLP: natural language processing
OWL: Ontology Web Language
ROC: receiver operating characteristic curve
SNOMED CT: Systemized Nomenclature of Medicine Clinical Terms

Edited by C Lovis; submitted 25.02.24; peer-reviewed by H Lv, K Xie, P Dadheech; revised version received 23.08.24; accepted 25.08.24; published 17.10.24.

Please cite as:
Xia Y, He M, Basang S, Sha L, Huang Z, Jin L, Duan Y, Tang Y, Li H, Lai W, Chen L
Semiology Extraction and Machine Learning–Based Classification of Electronic Health Records for Patients With Epilepsy: Retrospective Analysis
JMIR Med Inform 2024;12:e57727
URL: https://medinform.jmir.org/2024/1/e57727
doi:10.2196/57727

Disambiguating Clinical Abbreviations by One-to-All Classification: Algorithm Development and Validation Study

Sheng-Feng Sung[1,2], MD, PhD; Ya-Han Hu[3], PhD; Chong-Yan Chen[3], MBA

1

2

3

Corresponding Author:
Ya-Han Hu, PhD

Abstract

Background: Electronic medical records store extensive patient data and serve as a comprehensive repository, including textual medical records like surgical and imaging reports. Their utility in clinical decision support systems is substantial, but the widespread use of ambiguous and unstandardized abbreviations in clinical documents poses challenges for natural language processing in clinical decision support systems. Efficient abbreviation disambiguation methods are needed for effective information extraction.

Objective: This study aims to enhance the one-to-all (OTA) framework for clinical abbreviation expansion, which uses a single model to predict multiple abbreviation meanings. The objective is to improve OTA by developing context-candidate pairs and optimizing word embeddings in Bidirectional Encoder Representations From Transformers (BERT), evaluating the model's efficacy in expanding clinical abbreviations using real data.

Methods: Three datasets were used: Medical Subject Headings Word Sense Disambiguation, University of Minnesota, and Chia-Yi Christian Hospital from Ditmanson Medical Foundation Chia-Yi Christian Hospital. Texts containing polysemous abbreviations were preprocessed and formatted for BERT. The study involved fine-tuning pretrained models, ClinicalBERT and BlueBERT, generating dataset pairs for training and testing based on Huang et al's method.

Results: BlueBERT achieved macro- and microaccuracies of 95.41% and 95.16%, respectively, on the Medical Subject Headings Word Sense Disambiguation dataset. It improved macroaccuracy by 0.54% - 1.53% compared to two baselines, long short-term memory and deepBioWSD with random embedding. On the University of Minnesota dataset, BlueBERT recorded macro- and microaccuracies of 98.40% and 98.22%, respectively. Against the baselines of Word2Vec + support vector machine and BioWordVec + support vector machine, BlueBERT demonstrated a macroaccuracy improvement of 2.61% - 4.13%.

Conclusions: This research preliminarily validated the effectiveness of the OTA method for abbreviation disambiguation in medical texts, demonstrating the potential to enhance both clinical staff efficiency and research effectiveness.

(JMIR Med Inform 2024;12:e56955) doi:10.2196/56955

KEYWORDS

word sense disambiguation; electronic medical records; abbreviation expansion; text mining; natural language processing

Introduction

The advent of electronic medical records (EMRs) has revolutionized data management in medical institutions by enabling the storage and collection of extensive patient data. EMRs integrate records and reports from various hospital departments, documenting diverse patient conditions and providing a comprehensive repository of information, including previous laboratory and examination reports, hospitalization and surgical procedure records, and medication histories [1-3]. EMRs contain two types of data: structured, such as physiological measurements, laboratory results, diagnostic and drug codes, and assessment scales, and unstructured, primarily consisting of textual medical records like surgical and imaging reports, pathology reports, and discharge summaries [4-10].

Recent studies have leveraged natural language processing (NLP) tools, including MetaMap, MedLEE, and Clinical Text Analysis and Knowledge Extraction System (cTAKES), to extract valuable patient information from EMRs' clinical text [11-14]. These applications range from identifying specific medical concepts to complex analyses, such as discerning relationships between medical conditions or predicting patient outcomes and disease progression [10,15-19]. However, the prevalent use of abbreviations in clinical documents poses significant challenges for NLP in clinical decision support systems, as abbreviations often have multiple meanings depending on their context, and unstandardized or local abbreviations further complicate text interpretation [20,21]. This ambiguity impedes the extraction of meaningful information, affecting clinical decision support system

performance and highlighting the need for effective methods for abbreviation disambiguation in clinical NLP applications.

Abbreviation disambiguation in NLP involves identifying the correct expansion of an abbreviation based on its context [22,23]. In this process, one-to-one (OTO) and one-to-all (OTA) approaches are two distinct strategies for resolving the meaning of abbreviations [24]. The OTO approach involves training a separate machine learning model for each specific abbreviation, learning its unique patterns and contextual cues to disambiguate its meaning. In contrast, the OTA approach uses a single machine learning model trained to disambiguate all abbreviations across various contexts.

The OTA approach in abbreviation disambiguation offers several advantages over the OTO approach. OTA is easier to scale, requiring the maintenance and updating of only a single model, whereas OTO necessitates multiple models for each abbreviation, making it less scalable. OTA is more efficient in terms of computational resources and ensures a uniform disambiguation approach, reducing inconsistencies. Additionally, OTA simplifies model management, streamlining changes and improvements. By learning general patterns and contextual cues applicable to various abbreviations, OTA enhances overall context understanding, making it suitable for applications with diverse abbreviation needs. This flexibility makes OTA particularly useful in the biomedical domain, where abbreviations can have varied meanings in different contexts.

This study aims to enhance the application of the OTA abbreviation disambiguation framework for clinical abbreviation expansion. We propose constructing an OTA disambiguation model by creating context-candidate pairs and refining word embeddings using Bidirectional Encoder Representations From Transformers (BERT) [25]. The model's effectiveness was assessed based on its predictive performance on real clinical data for the task of clinical abbreviation expansion.

Methods

Data

This study conducted experimental evaluations using 3 datasets: 2 publicly available datasets and 1 independently collected from a regional hospital in Taiwan. The first dataset, the Medical Subject Headings Word Sense Disambiguation (MSH WSD) dataset, was extracted from MEDLINE abstracts [26]. The MSH WSD dataset comprises 203 polysemous words and is divided into three sections: abbreviation set, term set, and term/abbreviation set. The abbreviation set, containing 106 ambiguous acronyms, was selected as one of our investigated datasets. The second dataset, originating from the University of Minnesota (UMN), comprises deidentified clinical text sourced from the university's hospitals [27]. The UMN dataset includes 440 frequently used abbreviations and acronyms, carefully selected from a pool of 352,267 dictated clinical notes. These two datasets are valuable resources for both NLP and medical informatics, particularly for disambiguation tasks within the health care domain [20,28,29].

Lastly, the Chia-Yi Christian Hospital (CYCH) dataset aggregates present illness data from patients at the Neurology Department of Ditmanson Medical Foundation Chia-Yi Christian Hospital. Abbreviation disambiguation results were validated by a neurologist. We narrowed the scope of abbreviations for evaluation and asked the doctor to mark the answers in advance. Specifically, we selected five frequently appearing abbreviations—ER, DM, CVA, PM, and PA—from both the UMN and CYCH datasets. We verified that the correct interpretations of abbreviations in the CYCH clinical documents matched the candidate sets for the UMN abbreviations. Due to manpower constraints, we limited our extraction to the first 1000 abbreviations for annotation by a physician. After removing one erroneous data entry and two initially overlooked abbreviations, we had a total of 998 sentences that included the five selected abbreviations. All 3 datasets were preprocessed and organized into the same format for subsequent model construction and evaluation. As shown in Table 1, the text "...He is status post a BK amputation on the right side and..." is partitioned into three parts: left, right, and target. Left denotes the text to the left of the target abbreviation, right represents the text to the right, and target indicates the target abbreviation. The remaining two fields include the correct expansion word for the target abbreviation (label) and the collection of all incorrect candidate expansion words (negs).

Table . Data schema after data preprocessing and an example sample text.

Field	Description	Example
Index	Document ID	1
Target	The target abbreviation	BK
Left	The text to the left of the target abbreviation	...He is status post a
Right	The text to the right of the target abbreviation	amputation on the right side and...
Label	The correct expansion word for the target abbreviation	below knee
Neg	The collection of all incorrect candidate expansion words. If there are multiple, separate them with commas.	BK(virus)

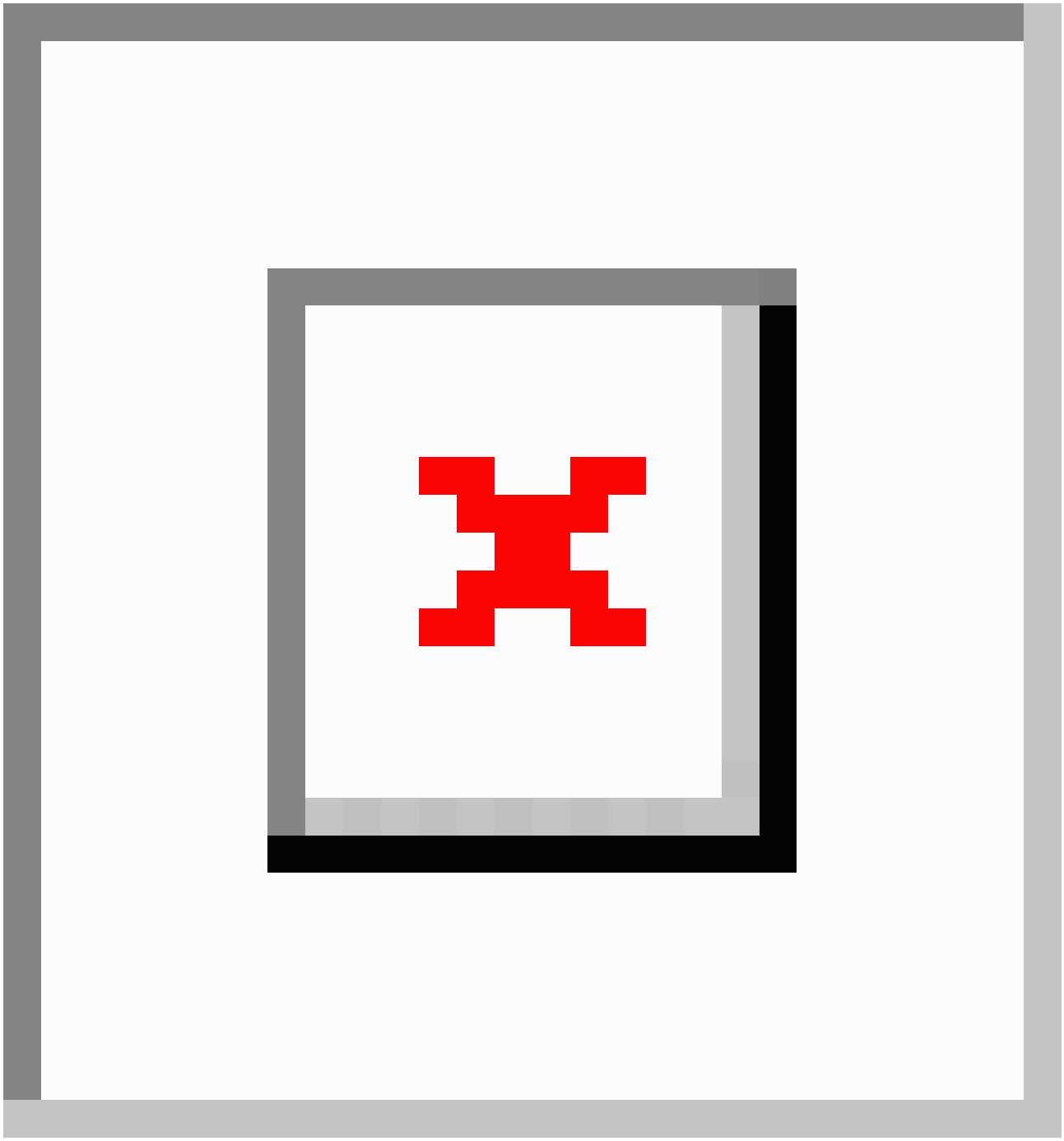
Ethical Considerations

The study protocol received formal approval from the Ditmanson Medical Foundation Chia-Yi Christian Hospital Institutional Review Board (2022074). Patient identifiers were replaced by a unique study identification number to ensure confidentiality. Informed consent was thus exempted.

The Proposed Framework

Figure 1 illustrates the proposed framework. We begin by retrieving text containing polysemous abbreviations from the 3 investigated datasets. The polysemous abbreviations are kept in their original form and marked accordingly. Subsequent steps involve common text preprocessing techniques, such as converting text to lowercase and removing certain special symbols. The preprocessed text is then adjusted to meet the input format required by BERT. Finally, the processed text is divided into training and testing sets. Three existing pretrained BERT-based models, including BERT-base-uncased [25], ClinicalBERT [30], and BlueBERT [31], are chosen and fine-tuned using these datasets, and prediction results are subsequently generated for evaluation. BERT-base-uncased is specifically used due to the common inconsistencies in capitalization within clinical texts, where lowercase letters are frequently used, even at the beginning of sentences or in abbreviations.

Figure 1. Research framework. BERT: Bidirectional Encoder Representations From Transformers; CYCH: Chia-Yi Christian Hospital; MSH WSD: Medical Subject Headings Word Sense Disambiguation; UMN: University of Minnesota.



Text Preprocessing

This study converts the text into the context-candidate pair format and adjusts the word embedding values for BERT input. Specifically, we apply GlossBERT [32] to train our model using the samples consisting of abbreviations and all of their candidate expansions. If an abbreviation has *n* candidate expansions, with only one correct answer, we produce *n* samples. This includes one sample marked as the correct expansion (indicated as 1) and *n* – 1 samples marked as incorrect expansions (indicated as 0).

Before training, we use BERT's tokenizer to convert text into WordPieces, breaking words such as "amputation" into ['amp', '##utation']. Special tokens are then added: [CLS] at the start, [SEP] to separate sentences or differentiate sections, and [PAD] to equalize sequence lengths for batch processing. For instance, when processing the sentence "He is status post a BK amputation..." with "BK" having expansions "BK(virus)" and "below knee," we generate two sequences: "[CLS] He is status post a BK amputation... [SEP] BK(virus) [SEP], 0" and "[CLS] He is status post a BK amputation... [SEP] below knee [SEP], 1."

Due to BERT's token limit of 512, sequences exceeding this are truncated. We manage sequence lengths by first converting text into WordPieces and adding necessary tokens. If the combined length of a sequence and its expansions exceeds BERT's limit, we employ a first in, first out (FIFO) strategy to ensure compliance with the token restriction.

BERT Tuning

Due to the limited dataset size, retraining a full BERT encoder was not feasible for this study. Instead, we fine-tuned existing pretrained models to assess our abbreviation disambiguation method. We selected two health care–related models, ClinicalBERT and BlueBERT, along with a generic BERT-base-uncased model as a baseline.

We adapted these models by adding a fully connected output layer. This layer consists of two linear layers and a rectified linear unit activation function, simplified as:

$$(1) y=f(\sum i=1 n w k x i+b k)$$

where wk represents the weights applied to inputs xi, and bk is the bias term.

The output layer's parameters are set as (50, 2), reflecting the size of the output from the previous layer and the number of classes (1 or 0). During prediction, the model calculates probabilities for each class. We focus primarily on the accuracy of the predictions for class 1, applying a softmax operation to enhance decision-making based on class 1's probability scores. This process optimizes our approach to evaluating the effectiveness of the trained models in context-sensitive disambiguation tasks.

Experimental Setup and Performance Measure

In our experimental evaluation, we aim to compare our proposed approach with several representative methods from prior studies on abbreviation disambiguation, focusing on model adaptability and performance across various datasets. The structure of our study is divided into two main parts.

Experiment 1 assesses the prediction performance of abbreviation expansion using both our proposed approach and baseline models (both OTO and OTA). We utilized two public datasets: MSH WSD and UMN. For MSH WSD, the OTO baselines included *k*-nearest neighbors [33], naive Bayes [26], and long short-term memory (LSTM) [34] models. For the UMN dataset, we referred to Wu et al [35] who used a combination of Word2Vec + support vector machine (SVM) as the OTO baseline. We further adapted this approach by substituting the original Word2Vec model with BioWordVec [36] (BioWordVec + SVM), which offers biomedical word embeddings via fastText, to better suit our study's focus on clinical data. Each clinical note was represented as a 200-dimensional vector. For the OTA baseline models, we implemented non–sense-based methods using BERT/XLNet, as described by Kim et al [37], which include $\text{deepBioWSD}_{\text{random embeddings}}$ and $\text{deepBioWSD}_{\text{pretrained sense embeddings}}$. Additionally, we employed sense-based methods using bidirectional LSTM, outlined by Pesaranghader et al [38], specifically masked language modeling and permutation language modeling.

Experiment 2 evaluates the prediction performance of abbreviation disambiguation within the CYCH dataset, aiming to address abbreviation ambiguity in clinical contexts. This involved training models using the UMN dataset and testing them on the CYCH dataset. The experiment was designed to test how well the fine-tuning of pretrained models could adapt to a new hospital setting, using a combination of internal and external datasets to assess accuracy in a real-world clinical environment.

We conducted experiment 2 under two distinct scenarios to assess the adaptability and effectiveness of our model in handling abbreviation disambiguation. In the first scenario, we excluded CYCH text, utilizing only the UMN dataset for training. This approach tested the model's ability to generalize from an external dataset to a new environment, applying it subsequently to 998 entries from the CYCH dataset. In the second scenario, we incorporated a small subset of CYCH text into the training process. This was designed to explore incremental learning, where the model adapts to new data while retaining previously learned information, thereby enhancing its predictive performance with minimal data and brief training periods.

Moreover, to maintain consistency and validity in our training process, it was crucial to ensure that all context-candidate pairs appeared in the training set. Consequently, the dataset was carefully screened before splitting, opting for a simple 9:1 ratio between the training and test sets instead of using cross-validation. To evaluate the model's performance, we employed metrics such as accuracy, microaccuracy, and macroaccuracy. These metrics were derived from the confusion matrix for each abbreviation, providing detailed insights into the model's efficacy across different contexts.

Results

Experiment 1

The results for experiment 1, using the MSH WSD dataset, are summarized in Table 2. For the OTO baselines, *k*-nearest neighbors achieved a macroaccuracy of 94.34%, with microaccuracy data unavailable. The naive Bayes method recorded a macroaccuracy of 93.86%, but microaccuracy was not reported. The LSTM method displayed both macro- and microaccuracy scores, which were 94.87% and 94.78%, respectively. For the OTA baselines, the sense-based deepBioWSD$_{\text{random embeddings}}$ [38] achieved macro- and microaccuracy scores of 93.88% and 93.71%, respectively. The deepBioWSD$_{\text{pretrained sense embeddings}}$ [38] improved prediction performance, with macro- and microaccuracy scores of 96.82% and 96.24%, respectively. Meanwhile, the non–sense-based methods, masked language modeling and permutation language modeling [37], recorded macroaccuracies of 95.89% and 96.83%, respectively, with microaccuracy not reported.

The BERT-base-uncased achieved macro- and microaccuracy scores of 93.64% and 93.38%, respectively. ClinicalBERT recorded a macroaccuracy of 94.77% and a microaccuracy of 94.59%. BlueBERT displayed a competitive performance with macro- and microaccuracies of 95.41% and 95.16%, respectively, on par with other evaluated methods. BlueBERT's macroaccuracy was only slightly lower than the highest performing models, deepBioWSD$_{\text{pretrained sense embeddings}}$ (96.82%) and permutation language modeling (96.83%), but higher than deepBioWSD$_{\text{random embeddings}}$ (93.88%). This demonstrates BlueBERT's robustness and effectiveness in sequence classification within this study.

The abbreviation disambiguation results for the UMN dataset, presented in Table 3, highlight the performance of various models. BlueBERT excelled, achieving macro- and microaccuracies of 98.4% and 98.22%, respectively, indicating its strong potential for disambiguation tasks. BERT-base-uncased and ClinicalBERT also showed strong performance, though slightly less than BlueBERT. In contrast, OTO-based models like Word2Vec + SVM and BioWordVec + SVM had lower accuracy scores, underscoring the advanced capabilities of the BERT models.

Table . Abbreviation disambiguation results (Medical Subject Headings Word Sense Disambiguation).

Method		Macroaccuracy (%)	Microaccuracy (%)
One-to-one			
	k-nearest neighbors [33]	94.34	—[a]
	Naive Bayes [26]	93.86	—
	Long short-term memory [34]	94.87	94.78
One-to-all			
	deepBioWSD$_{\text{random embeddings}}$ [38]	93.88	93.71
	deepBioWSD$_{\text{pretrained sense embeddings}}$ [38]	96.82	96.24
	Masked language modeling [37]	95.89	—
	Permutation language modeling [37]	96.83	—
	BERT-base-uncased	93.64	93.38
	ClinicalBERT	94.77	94.59
	BlueBERT	95.41	95.16

[a]Not applicable.

Table . Abbreviation disambiguation results (University of Minnesota).

Method (work)		Macroaccuracy (%)	Microaccuracy (%)
One-to-one			
	Word2Vec + SVM[a] [35]	95.79	—[b]
	BioWordVec + SVM	94.27	—
One-to-all			
	Masked language modeling [37]	98.39	—
	Permutation language modeling [37]	98.28	—
	BERT-base-uncased	97.59	97.27
	ClinicalBERT	98.27	98.01
	BlueBERT	98.40	98.22

[a]SVM: support vector machine.

[b]Not applicable.

Overall, the proposed OTA method, especially when implemented using the pretrained BlueBERT model, outperformed the OTO-based approaches. The OTA method's reliance on a single model, as opposed to the multiple models required by OTO methods, improves maintainability and scalability.

Experiment 2

Table 4 displays the abbreviation disambiguation results for the CYCH dataset using BlueBERT. The table compares accuracy percentages for each abbreviation when trained exclusively on external data versus including incremental amounts of CYCH data (5 and 10 samples, respectively). For example, the model recorded a 62.07% accuracy for the abbreviation DM when trained without CYCH data. With the inclusion of CYCH data, the accuracy slightly improved to 70.27% with 5 samples but then slightly decreased to 100% with 10 samples. This trend of initial improvement followed by a marginal decline was observed for other abbreviations as well. Notably, the abbreviation PA showed a substantial increase in performance; it had 0% accuracy when trained without CYCH data but reached 100% accuracy when trained with either 5 or 10 CYCH samples.

Table . Abbreviation disambiguation results of the Chia-Yi Christian Hospital (CYCH) dataset.

Abbreviation	Training without CYCH data, accuracy (%)	Training with CYCH data	
		Accuracy, includes 5 documents (%)	Accuracy, includes 10 documents (%)
ER	97.91	98.03	97.75
DM	62.07	70.27	100
CVA	96.20	98.65	100
PM	77.27	100	100
PA	0.00	100	100

Discussion

Automatic abbreviation disambiguation is crucial in clinical settings as it enhances the clarity and readability of medical records. By accurately interpreting abbreviations, it ensures that health care professionals have a precise understanding of patient information, facilitating accurate diagnoses and effective treatment plans. This automation also speeds up data processing, supports decision-making, and reduces errors, thereby improving overall health care delivery and patient safety.

Traditional OTO methods for abbreviation expansion involve constructing independent models for each abbreviation. Although this method offers high accuracy, it presents challenges in terms of maintenance and generalizability, complicating clinical applications due to the high number of models and associated maintenance costs. In contrast, this study proposes an approach that reduces the number of required models and offers better performance in clinical abbreviation restoration, thereby lowering both the operational and maintenance costs.

Compared to OTO, the OTA approach provides greater scalability, efficiency, and consistency, with a unified model that is easier to maintain and update. However, OTA approaches can be costly in terms of model retraining. Kim et al [37] highlighted that retraining the encoder necessitated high-end GPUs and substantial memory, requiring up to 14 days. Our study adopted a tuning approach using existing pretrained models, substantially cutting down training time to

approximately an hour and a half by utilizing free online resources like the K80 GPU through Kaggle Notebook. This method effectively reduces both hardware and time costs, especially beneficial in clinical settings where frequent model updates may be necessary.

This study further demonstrates the practicality of this method in various hospital scenarios, particularly addressing cross-hospital and interdepartmental issues. Our incremental learning approach has been shown to significantly improve prediction results, thereby saving considerable retraining costs.

This study has the following limitations. First, although it preliminarily validates the exceptional effectiveness of the OTA method for abbreviation disambiguation in medical texts, the evaluation is limited by the size of the datasets used. More extensive and comprehensive clinical data are required before application to further validate this method. Second, our study is constrained by the maximum sequence length restriction of the BERT model. Longer clinical notes exceeding the 512-token limit must be truncated, risking the loss of information. Analysis shows that about 16.85% of the MSH WSD dataset and only 0.03% of the UMN dataset exceed this limit. The experimental results indicate superior accuracy for the UMN dataset; however, the performance for the MSH WSD dataset is lower, likely due to significant truncation of longer texts.

Additionally, in generating context-candidate pairs, we retain all candidates and use a FIFO approach for trimming the context. If an abbreviation appears at both the beginning and end of a context and exceeds the token limit, the FIFO method may remove the initial occurrence. Conversely, a last in, first out method could remove an abbreviation appearing at the end. If the same abbreviation carries different meanings in different parts of the text, identical context-candidate pairs may be created post trimming, potentially distorting model training and leading to incorrect predictions.

This study presents an innovative approach to the disambiguation and expansion of abbreviations in clinical medical texts by utilizing context-candidate pairs and the BERT model. This method enhances the readability of medical texts, improving the efficiency of clinical staff who review EMRs and saving time for cross-disciplinary researchers analyzing clinical data, thereby increasing the effectiveness of their studies. Given that clinical medical texts are replete with abbreviations, accurate disambiguation is essential for improving text clarity and usability. Automating this process greatly assists both medical professionals and researchers. The successful application of this model on the investigated datasets underscores its effectiveness and establishes it as a valuable reference for future research in clinical abbreviation expansion.

Acknowledgments

This work was supported by the Ditmanson Medical Foundation Chia-Yi Christian Hospital Research Program under grant R112-022-1 and supported in part by the Ministry of Science and Technology of Taiwan (MOST 111-2410-H-008-026-MY2).

Data Availability

The codes used for training and evaluating the models on the Medical Subject Headings Word Sense Disambiguation and University of Minnesota datasets are available at [39] and [40] on Kaggle.

Conflicts of Interest

None declared.

References

1. Komeda Y, Handa H, Watanabe T, et al. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology 2017;93 Suppl 1:30-34. [doi: 10.1159/000481227] [Medline: 29258081]
2. Park HJ, Kim SM, La Yun B, et al. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound. Medicine (Balt) 2019 Jan;98(3):e14146. [doi: 10.1097/MD.0000000000014146] [Medline: 30653149]
3. Sato Y, Takegami Y, Asamoto T, et al. A computer-aided diagnosis system using artificial intelligence for hip fractures-multi-institutional joint development research. arXiv. Preprint posted online on Mar 11, 2020. [doi: 10.48550/arXiv.2003.12443]
4. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
5. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014;21(5):801-807. [doi: 10.1136/amiajnl-2013-001915] [Medline: 24384230]
6. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 2020 Oct 29;20(1):280. [doi: 10.1186/s12911-020-01297-6] [Medline: 33121479]

7. Hatef E, Rouhizadeh M, Nau C, et al. Development and assessment of a natural language processing model to identify residential instability in electronic health records' unstructured data: a comparison of 3 integrated healthcare delivery systems. JAMIA Open 2022 Apr;5(1):ooac006. [doi: 10.1093/jamiaopen/ooac006] [Medline: 35224458]
8. Levis M, Levy J, Dufort V, Gobbel GT, Watts BV, Shiner B. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry Res 2022 Sep;315:114703. [doi: 10.1016/j.psychres.2022.114703] [Medline: 35841702]
9. Wang M, Wei Z, Jia M, Chen L, Ji H. Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records. BMC Med Inform Decis Mak 2022 Feb 16;22(1):41. [doi: 10.1186/s12911-022-01776-y] [Medline: 35168624]
10. Sung SF, Lin CY, Hu YH. EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. IEEE J Biomed Health Inform 2020 Oct;24(10):2922-2931. [doi: 10.1109/JBHI.2020.2976931] [Medline: 32142458]
11. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21. [Medline: 11825149]
12. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236. [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
13. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. Nat Lang Eng 1995 Mar;1(1):83-108. [doi: 10.1017/S1351324900000061]
14. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513. [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
15. Garla V, Lo Re V, Dorey-Stein Z, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011;18(5):614-620. [doi: 10.1136/amiajnl-2011-000093] [Medline: 21622934]
16. Sung SF, Chen CH, Pan RC, Hu YH, Jeng JS. Natural language processing enhances prediction of functional outcome after acute ischemic stroke. J Am Heart Assoc 2021 Dec 21;10(24):e023486. [doi: 10.1161/JAHA.121.023486] [Medline: 34796719]
17. Gao J, He S, Hu J, Chen G. A hybrid system to understand the relations between assessments and plans in progress notes. J Biomed Inform 2023 May;141:104363. [doi: 10.1016/j.jbi.2023.104363] [Medline: 37054961]
18. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Med Inform Decis Mak 2020 Dec 30;20(Suppl 11):295. [doi: 10.1186/s12911-020-01318-4] [Medline: 33380338]
19. Sung SF, Chen K, Wu DP, Hung LC, Su YH, Hu YH. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study. Int J Med Inform 2018 Apr;112:149-157. [doi: 10.1016/j.ijmedinf.2018.02.005] [Medline: 29500013]
20. Moon S, Pakhomov S, Melton GB. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. AMIA Annu Symp Proc 2012;2012:1310-1319. [Medline: 23304410]
21. Wu Y, Tang B, Jiang M, Moon S, Denny JC, Xu H. Clinical acronym/abbreviation normalization using a hybrid approach. In: Forner P, Navigli R, Tufis D, Ferro N, editors. Working Notes for (CLEF) 2013 Conference, Valencia, Spain, September 23-26, 2013: CEUR-WS.org; 2013. URL: https://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-WuEt2013.pdf [accessed 2024-09-18]
22. Moon S, Berster BT, Xu H, Cohen T. Word sense disambiguation of clinical abbreviations with hyperdimensional computing. AMIA Annu Symp Proc 2013 Nov 16;2013:1007-1016. [Medline: 24551390]
23. Wu Y, Denny JC, Trent Rosenbloom S, et al. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). J Am Med Inform Assoc 2017 Apr 1;24(e1):e79-e86. [doi: 10.1093/jamia/ocw109] [Medline: 27539197]
24. Li Y, Wang H, Li X, Deng S, Su T, Zhang W. Disambiguation of medical abbreviations for knowledge organization. Inf Processing Manage 2023 Sep;60(5):103441. [doi: 10.1016/j.ipm.2023.103441]
25. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
26. Jimeno-Yepes AJ, McInnes BT, Aronson AR. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics 2011 Jun 2;12(1):1-14. [doi: 10.1186/1471-2105-12-223] [Medline: 21635749]
27. Moon S, Pakhomov S, Melton G. Clinical abbreviation sense inventory. University of Minnesota: University Digital Conservancy. 2012 Oct 31. URL: https://conservancy.umn.edu/items/6651323b-444a-479e-a41a-abca58c2e721 [accessed 2024-09-18]
28. Finley GP, Pakhomov SVS, McEwan R, Melton GB. Towards comprehensive clinical abbreviation disambiguation using machine-labeled training data. AMIA Annu Symp Proc 2016 Feb 10;2016:560-569. [Medline: 28269852]
29. Grossman Liu L, Grossman RH, Mitchell EG, et al. A deep database of medical abbreviations and acronyms for natural language processing. Sci Data 2021 Jun 2;8(1):149. [doi: 10.1038/s41597-021-00929-4] [Medline: 34078918]

30. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv. Preprint posted online on Apr 6, 2019. [doi: 10.18653/v1/W19-1909]
31. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv. Preprint posted online on Jun 13, 2019. [doi: 10.48550/arXiv.1906.05474]
32. Huang L, Sun C, Qiu X, Huang X. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In: Inui K, Jiang J, Ng V, Wan W, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Association for Computational Linguistics; 2019:3509-3514. [doi: 10.18653/v1/D19-1355]
33. Sabbir A, Jimeno-Yepes A, Kavuluru R. Knowledge-based biomedical word sense disambiguation with neural concept embeddings. Proc IEEE Int Symp Bioinformatics Bioeng 2017 Oct;2017:163-170. [doi: 10.1109/BIBE.2017.00-61] [Medline: 29399672]
34. Jimeno Yepes A. Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. J Biomed Inform 2017 Sep;73:137-147. [doi: 10.1016/j.jbi.2017.08.001] [Medline: 28797709]
35. Wu Y, Denny JC, Rosenbloom ST, et al. A preliminary study of clinical abbreviation disambiguation in real time. Appl Clin Inform 2015 Jun 3;6(2):364-374. [doi: 10.4338/ACI-2014-10-RA-0088] [Medline: 26171081]
36. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52. [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
37. Kim J, Gong L, Khim J, Weiss JC, Ravikumar P. Improved clinical abbreviation expansion via non-sense-based approaches. Proc Mach Learn Res 2020 May 10;136:161-178. [doi: 10.1038/s41597-019-0055-0]
38. Pesaranghader A, Matwin S, Sokolova M, Pesaranghader A. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. J Am Med Inform Assoc 2019 May 1;26(5):438-446. [doi: 10.1093/jamia/ocy189] [Medline: 30811548]
39. Chen S. MSH_paper_bert. Kaggle. 2022 Feb. URL: https://www.kaggle.com/code/dsaddicter/msh-paper-bert [accessed 2024-09-12]
40. Chen S. UMN_paper_bert. Kaggle. 2022 Feb. URL: https://www.kaggle.com/code/dsaddicter/umn-paper-bert [accessed 2024-09-12]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
cTAKES: Clinical Text Analysis and Knowledge Extraction System
CYCH: Chia-Yi Christian Hospital
EMR: electronic medical record
FIFO: first in, first out
LSTM: long short-term memory
MSH WSD: Medical Subject Headings Word Sense Disambiguation
NLP: natural language processing
OTA: one-to-all
OTO: one-to-one
SVM: support vector machine
UMN: University of Minnesota

Edited by C Lovis; submitted 31.01.24; peer-reviewed by J Zaghir, P Han, S Setia; revised version received 29.08.24; accepted 01.09.24; published 01.10.24.

Please cite as:
Sung SF, Hu YH, Chen CY
Disambiguating Clinical Abbreviations by One-to-All Classification: Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e56955
URL: https://medinform.jmir.org/2024/1/e56955
doi:10.2196/56955

Natural Language Processing Versus Diagnosis Code–Based Methods for Postherpetic Neuralgia Identification: Algorithm Development and Validation

Chengyi Zheng[1], MS, PhD; Bradley Ackerson[2], MD; Sijia Qiu[1], MS; Lina S Sy[1], MPH; Leticia I Vega Daily[1], MSW; Jeannie Song[1], MPH; Lei Qian[1], PhD; Yi Luo[1], PhD; Jennifer H Ku[1], MPH, PhD; Yanjun Cheng[1], MS; Jun Wu[1], MS, MD; Hung Fu Tseng[1,3], MPH, PhD

1

2

3

Corresponding Author:
Chengyi Zheng, MS, PhD

Abstract

Background: Diagnosis codes and prescription data are used in algorithms to identify postherpetic neuralgia (PHN), a debilitating complication of herpes zoster (HZ). Because of the questionable accuracy of codes and prescription data, manual chart review is sometimes used to identify PHN in electronic health records (EHRs), which can be costly and time-consuming.

Objective: This study aims to develop and validate a natural language processing (NLP) algorithm for automatically identifying PHN from unstructured EHR data and to compare its performance with that of code-based methods.

Methods: This retrospective study used EHR data from Kaiser Permanente Southern California, a large integrated health care system that serves over 4.8 million members. The source population included members aged ≥50 years who received an incident HZ diagnosis and accompanying antiviral prescription between 2018 and 2020 and had ≥1 encounter within 90 - 180 days of the incident HZ diagnosis. The study team manually reviewed the EHR and identified PHN cases. For NLP development and validation, 500 and 800 random samples from the source population were selected, respectively. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-score, and Matthews correlation coefficient (MCC) of NLP and the code-based methods were evaluated using chart-reviewed results as the reference standard.

Results: The NLP algorithm identified PHN cases with a 90.9% sensitivity, 98.5% specificity, 82% PPV, and 99.3% NPV. The composite scores of the NLP algorithm were 0.89 (F-score) and 0.85 (MCC). The prevalences of PHN in the validation data were 6.9% (reference standard), 7.6% (NLP), and 5.4% - 13.1% (code-based). The code-based methods achieved a 52.7% - 61.8% sensitivity, 89.8% - 98.4% specificity, 27.6% - 72.1% PPV, and 96.3% - 97.1% NPV. The F-scores and MCCs ranged between 0.45 and 0.59 and between 0.32 and 0.61, respectively.

Conclusions: The automated NLP-based approach identified PHN cases from the EHR with good accuracy. This method could be useful in population-based PHN research.

(JMIR Med Inform 2024;12:e57949) doi:10.2196/57949

KEYWORDS

postherpetic neuralgia; herpes zoster; natural language processing; electronic health record; real-world data; artificial intelligence; development; validation; diagnosis; EHR; algorithm; EHR data; sensitivity; specificity; validation data; neuralgia; recombinant zoster vaccine

Introduction

Herpes zoster (HZ) or shingles is a painful dermatomal vesicular disease that results from the reactivation of the latent varicella-zoster virus in the nerve ganglia [1]. Nearly all adults have the varicella-zoster virus dormant in their nervous system [2], and the estimated lifetime risk of HZ was approximately 30% prior to the availability of the zoster vaccine [3]. HZ usually begins with a prodromal stage of discomfort, followed by a painful, itchy rash on one unilateral dermatome that lasts 2 to 4 weeks [4]. Patients with HZ may develop postherpetic neuralgia (PHN)—dermatomal pain persisting at least 90 days after the appearance of the acute HZ rash [3,5]. PHN is the most common complication of HZ and greatly lowers patients' quality of life [3].

Population-based studies using real-world data are cost-effective ways to address many questions about PHN [3]. However, accurately identifying PHN is difficult. Clinical trials rely on

predetermined follow-up visits, which are difficult to replicate in real-world settings [6,7]. Due to time and resource constraints, prospective studies have mainly been limited to hundreds of patients with HZ and smaller numbers of PHN cases [3]. Retrospective studies of PHN have relied heavily on diagnosis codes [8-13], which lack accuracy [3,14], or manual chart review [14-16], which is costly and time-consuming. Moreover, despite the widespread use of code-based algorithms, only a few publications included PHN algorithm validation results [8,10].

Natural language processing (NLP), a subfield of artificial intelligence, has been used to identify and extract information from unstructured clinical data. We previously developed NLP methods to identify HZ ophthalmicus and HZ ophthalmicus with eye involvement, which are also common HZ complications [17,18]. In this study, we developed and validated an NLP algorithm to identify PHN. Using manual chart-reviewed results as a reference standard, we compared the performance of the NLP algorithm with that of 5 previously published code-based algorithms.

Methods

Setting

This study was conducted at Kaiser Permanente Southern California (KPSC), an integrated health care system with 16 hospitals and 197 medical offices that serves over 4.8 million members. The prepaid health plan incentivizes members to use services at KPSC facilities. The electronic health record (EHR) system at KPSC stores all aspects of member care, including sociodemographic characteristics, medical encounters, diagnoses, laboratory tests, pharmacy use, immunization records, membership history, and billing and claims.

PHN Case Definition

PHN was defined as pain or discomfort consistent with the HZ episode ≥90 days after the initial HZ diagnosis; the symptoms were at the location of the initial HZ rash and were not due to other obvious causes [19-21].

Data Sets

This study used EHR data of patients aged ≥50 years who each had an incident HZ diagnosis and associated antiviral prescription between 2018 and 2020 at KPSC. All patients had to have at least 1 year of membership prior to the index (incident HZ diagnosis) date so that comorbidities and health care use could be ascertained. Among patients with ≥1 encounter during the 90 - 180 days after the incident HZ diagnosis, trained research associates reviewed their EHRs based on the PHN abstraction instructions (Multimedia Appendix 1). An infectious disease physician (BKA) reviewed all possible or unclear cases. From these reviewed cases, we randomly selected 500 cases for NLP development and 800 cases for NLP validation. Because the NLP work was done concurrently with the manual review, the development data set was collected at an earlier stage, when the reviewed cohort had a greater proportion of Asian and recombinant zoster vaccine–vaccinated patients.

Reference Standard

Among the 800 cases in the validation data set, BKA reviewed 37 HZ cases that research associates had identified as unclear PHN cases. Because reviewers sometimes missed positive mentions of PHN, BKA rereviewed cases in the validation set where NLP results differed from reviewer results. Nine cases were corrected from negative to positive PHN. These manually reviewed results served as the reference standard for assessing the performance of PHN identification algorithms.

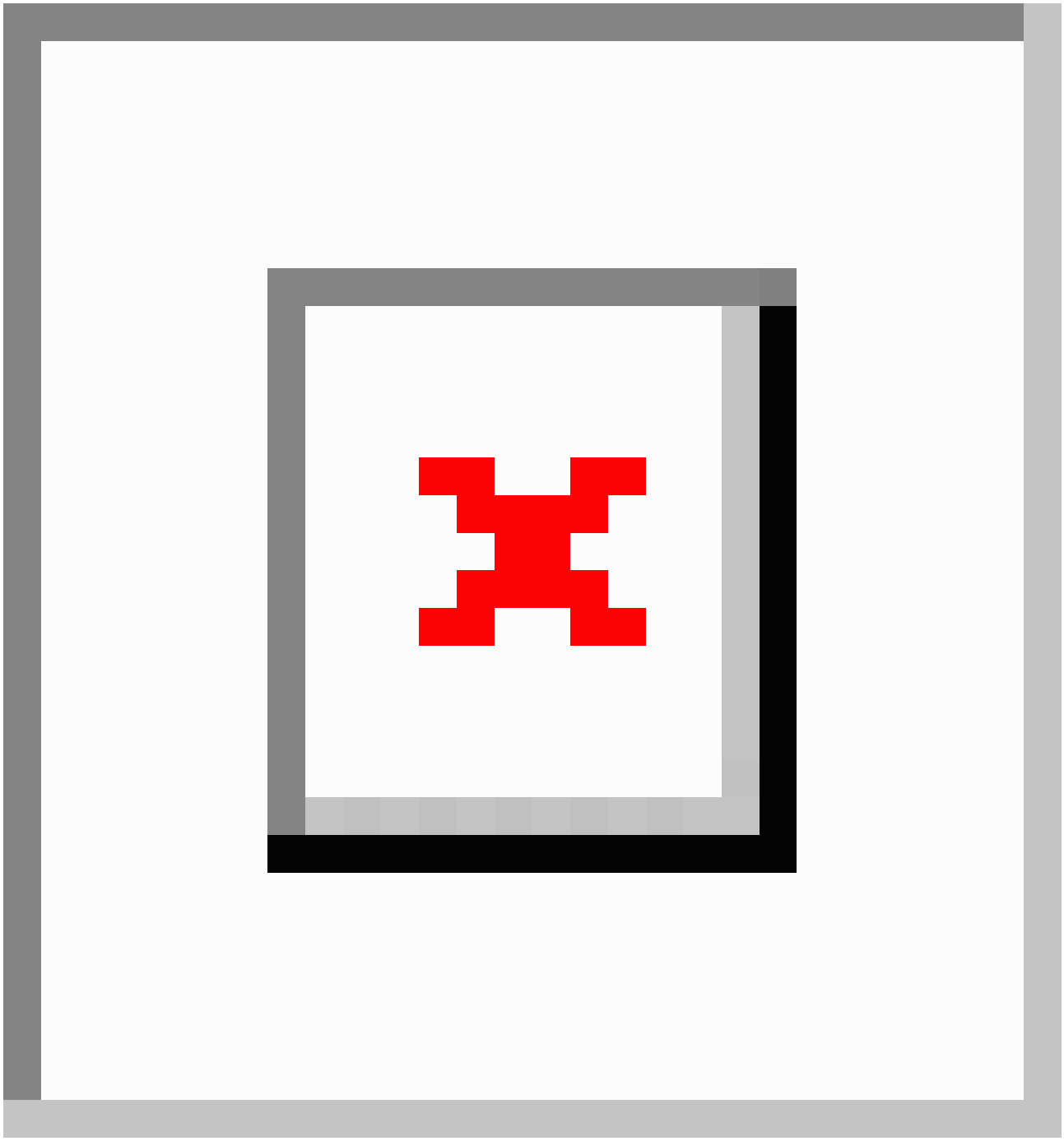
NLP Algorithm Development

We developed the NLP algorithm based on our previous work [17,18,22-26]. Multimedia Appendix 2 describes the steps for preprocessing text and generating nomenclature. We created the rule-based NLP algorithm using the Linguamatics I2E software (Linguamatics, an IQVIA company). Each note was searched at different levels: section (eg, “Physical Exam,” “Assessment/Plan”), cross-sentence, intrasentence, and phrase. A distance-based relationship algorithm was applied to identify related terms based on the number of words or sentences between them. The relationship search identified the words or phrases (eg, negated, uncertain, and hypothetical statements) that modified the concepts of interest.

Figure 1 depicts an overview of the NLP algorithm. We separated the extracted clinical texts into 3 time periods: index (acute HZ) period (−7 to 21 d from incident HZ diagnosis date), transitional (subacute HZ) period (22 to 89 d), and risk (defined PHN) period (90 to 180 d). We developed search queries to identify the HZ anatomic locations in the index episode and PHN-related evidence in the transitional and risk periods. Supporting evidence of PHN included explicit mention of ongoing PHN, symptom location and causality, and PHN listed in the assessment and plan section. Counterevidence of PHN included differential diagnoses, recurrent HZ, and resolved PHN. We excluded sections and statements that may have been copied forward as historical information.

The PHN decision algorithm was implemented in Python language, which incorporated the evidence from the NLP search queries and classified each case based on decision rules. To exclude the copy-pasted results, the NLP program ran search queries on both the transitional and risk periods and compared the results to locate identical sequences of text. The algorithm considered the time sequences of identified evidence. The symptom location during the risk period was compared with the index HZ location. Because adjacent dermatomes might be difficult to distinguish clinically, symptom location during the index and risk periods had to occur in the same or surrounding dermatomes (eg, face and neck). Based on the development data set, we tested and updated the algorithm.

Figure 1. Diagram of NLP algorithm. HZ: herpes zoster; NLP: natural language processing; PHN: postherpetic neuralgia.



Implementation of Published PHN Identification Algorithms

We selected and implemented 5 code-based PHN identification algorithms based on the variety of their algorithms, the journal category and impact factor, the publication year, the total citations, and the size of the study (Table 1). The first code-based method (C1: Yanni et al [27]) exclusively used PHN-related diagnosis codes (Multimedia Appendix 3). The remaining 4 algorithms (C2: Klompas et al [8]; C3: Klein et al [10]; C4: Forbes et al [9]; C5: Munoz-Quiles et al [11]) used additional structured data, such as diagnosis codes for HZ, neuralgia, and chronic pain; prescriptions for analgesics, antidepressants, and anticonvulsants; and clinical visit data.

Table . List of sources for selected code-based methods.

Method	Herpes zoster cases, n	Journal category	Journal (IF[a])	Year	TC[b]
C1 [27]	21,146	General medicine	BMJ Open (2.9)	2018	67
C2 [8]	2089	General medicine	Mayo Clinic Proceeding (8.9)	2011	125
C3 [10]	62,205	Immunology	Vaccine (5.5)	2019	32
C4 [9]	119,413	Neurology	Neurology (10.1)	2016	155
C5 [11]	87,086	Infectious diseases	Journal of Infection (28.2)	2018	38

[a]IF: impact factor based on Journal Citation Report released in 2023.

[b]TC: total citations based on Google Scholar as of July 1, 2024.

Validation and Analysis

The results generated from the various algorithms were evaluated against the chart-reviewed reference standard validation data set. We counted the numbers of true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) cases to calculate the performance metrics: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F-score [28], and Matthews correlation coefficient (MCC) [29].

The F-score is a combination metric in machine learning and NLP research. It is defined as a weighted harmonic mean of sensitivity and PPV, where the parameter β represents the relative importance of sensitivity versus PPV.

$$F-score=(\beta 2+1)*PPV*sensitivity\beta 2*PPV+sensitivity$$

Since a minority of patients with HZ will develop PHN and FNs and sensitivity are more important than PPV, we chose β=2 to favor sensitivity over PPV. The F-score's value ranges from 0 to 1, with higher values suggesting better prediction. However, because the F-score does not include TN in its formula, MCC has been proposed as a better overall measurement than the F-score as well as the area under the receiver operating characteristic curve in binary classification [29,30]. The MCC formula considers all 4 confusion matrix categories, with values between −1 to 1, where ±1 denotes perfect agreement or disagreement between actuals and predictions, and 0 indicates randomness.

$$MCC=TP*TN-FP*FN(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)$$

The prevalence proportion of PHN was calculated as the number of identified PHN cases per 100 cases of HZ.

Ethical Considerations

The KPSC institutional review board approved this study (institutional review board number: 12270). A waiver of informed consent was granted for this study because this was a data-only minimal-risk study.

Results

Study Population

The characteristics of the study population are presented in Table 2. The mean (SD) ages of the development and validation data sets were 69.5 (9.1) and 70.0 (9.8) years, respectively, with 329 (65.8%) and 542 (67.8%) being female. The development data set had a higher proportion of Asian (177/500, 35.4% vs 110/796, 13.8%) and recombinant zoster vaccine–vaccinated patients (61/500, 12.2% vs 28/800, 3.5%). There were no significant differences between the development and validation data sets in terms of clinical visits and comorbidities prior to the index date. In the development and validation data sets, approximately one-third of the patients had diabetes (159/500, 31.8% and 262/796, 32.8%, respectively), while less than one-quarter had chronic pulmonary disease (88/500, 17.6% and 163/800, 20.4%) and depression (103/500, 20.6% and 189/800, 23.6%). The development data had a higher proportion of patients with cancer as compared with the validation data (52/500, 10.4% vs 54/800, 6.8%).

Table . Characteristics of patients in the development and validation data sets.

Characteristics		Patients		*P* value[a]
		Development (n=500)	Validation (n=800)	
Age (year), mean (SD)		69.5 (9.1)	70.0 (9.8)	.47
Age group (years), n (%)				.52
	50 - 59	70 (14.0)	103 (12.9)	
	60 - 69	172 (34.4)	282 (35.3)	
	70 - 79	187 (37.4)	280 (35.0)	
	≥80	71 (14.2)	135 (16.9)	
Sex, n (%)				.47
	Female	329 (65.8)	542 (67.8)	
	Male	171 (34.2)	258 (32.3)	
Race or ethnicity, n (%)				<.01
	Non-Hispanic White	198 (39.6)	424 (53.0)	
	Hispanic	102 (20.4)	213 (26.6)	
	Asian/Pacific Islanders	177 (35.4)	110 (13.8)	
	Non-Hispanic Black	18 (3.6)	41 (5.1)	
	Other/Multiple/Unknown	5 (1.0)	12 (1.5)	
Number of outpatient/digital visits 6 months before HZ [b] diagnosis date, n (%)				.91
	0 - 1	68 (13.6)	106 (13.3)	
	2 - 5	168 (33.6)	262 (32.8)	
	≥6	264 (52.8)	432 (54.0)	
Number of emergency department visits 6 months before HZ diagnosis date, n (%)				.08
	0	420 (84.0)	641 (80.1)	
	≥1	80 (16.0)	159 (19.9)	
Number of hospitalizations 6 months before HZ diagnosis date, n (%)				.45
	0	475 (95.0)	752 (94.0)	
	≥1	25 (5.0)	48 (6.0)	
Comorbidity 1 year before HZ diagnosis date, n (%)				
	Allergic rhinitis	38 (7.6)	51 (6.4)	.39
	Asthma	41 (8.2)	82 (10.3)	.22
	Atopic dermatitis	5 (1.0)	9 (1.1)	.83
	Cancer	52 (10.4)	54 (6.8)	.02
	Chronic pulmonary disease	88 (17.6)	163 (20.4)	.22
	Depression	103 (20.6)	189 (23.6)	.20
	Diabetes	159 (31.8)	262 (32.8)	.72
	Epilepsy and recurrent seizures	8 (1.6)	8 (1.0)	.34
	Heart failure	24 (4.8)	55 (6.9)	.13
	Rheumatoid arthritis	25 (5.0)	38 (4.8)	.84
	Systemic lupus erythematosus	6 (1.2)	6 (0.8)	.41
Recombinant zoster vaccine, n (%)				<.01
	Unvaccinated	439 (87.8)	772 (96.5)	

Characteristics		Patients		*P* value[a]
		Development (n=500)	Validation (n=800)	
	1-dose vaccinated	32 (6.4)	9 (1.1)	
	Fully (2-dose) vaccinated	29 (5.8)	19 (2.4)	

[a]χ^2 test was used for categorical variables, and Wilcoxon test was used for continuous variables.

[b]HZ: herpes zoster.

Validation Data Set

In the validation data set, the numbers of clinical notes in the index, transitional, and risk periods were 12,158, 14,446, and 18,895, respectively. The percentages of HZ- or PHN-relevant notes were 26.2%, 8.2%, and 3.2%, respectively for the index, transitional, and risk periods. Most of the HZ index visits occurred in primary care, urgent care, emergency departments, and hospital settings (Multimedia Appendix 4). After the index period, HZ-related mentions were much less frequently documented in urgent care visit notes, but more frequently documented in specialist visit notes (41 specialties).

Application of NLP on Validation Data Set

Out of the 800 patients in the validation data set, the NLP algorithm identified 796 patients with HZ who had at least 1 note with HZ- or PHN-related terms in the index period. Among the 4 remaining patients, 2 patients had their index HZ diagnosed outside KPSC and had no follow-up visits in the index period. For the remaining 2 patients, HZ-related symptoms were documented, but no mention of HZ or PHN was made in the clinical notes. Among these 796 patients, the NLP algorithm identified the HZ anatomic location for 751 (94.3%) patients, and among them, 611 (81.3%) had laterality information (Multimedia Appendix 5). In the transitional and risk periods, the NLP algorithm identified positive mentions of any pain or discomfort in 370 (46.3%) and 425 (53.1%) patients, respectively.

Validation Results

In the validation data set, the NLP algorithm achieved a 90.9% sensitivity, 98.5% specificity, 82% PPV, and 99.3% NPV (Table 3). The composite scores of the NLP algorithm were 0.89 (F-score) and 0.85 (MCC). Of the 800 patients in the validation data set, 55 (6.9%) were chart-confirmed as PHN. The prevalence proportion of PHN identified by the NLP algorithm was 7.6%.

Table . Performance characteristics of natural language processing and code-based methods for identifying postherpetic neuralgia as compared with chart-confirmed reference standard.

Method	PHN[a] (%)	TP[b]	TN[c]	FN[d]	FP[e]	Sensitivity (%)	Specificity (%)	PPV[f] (%)	NPV[g] (%)	F-score	MCC[h]
NLP[i]	7.6	50	734	5	11	90.9	98.5	82.0	99.3	0.89	0.85
C1	5.4	31	733	24	12	56.4	98.4	72.1	96.8	0.59	0.61
C2	10.9	34	692	21	53	61.8	92.9	39.1	97.1	0.55	0.44
C3	5.5	31	732	24	13	56.4	98.3	70.5	96.8	0.59	0.61
C4	13.1	29	669	26	76	52.7	89.8	27.6	96.3	0.45	0.32
C5	9.3	31	702	24	43	56.4	94.2	41.9	96.7	0.53	0.44

[a]PHN: postherpetic neuralgia.

[b]TP: true-positive.

[c]TN: true-negative.

[d]FN: false-negative.

[e]FP: false-positive.

[f]PPV: positive predictive value.

[g]NPV: negative predictive value.

[h]MCC: Matthews correlation coefficient.

[i]NLP: natural language processing.

Error Analysis of NLP Validation Results

Error analysis of the FN and FP cases is presented in Table 4. Some of the NLP-related errors were caused by the selection of data sources. For 2 FN cases, NLP incorrectly classified them as PHN negative when statements were found indicating HZ-associated pain had resolved even though additional evidence showed the patients still had other PHN-related symptoms. The FP cases were caused by copied-and-pasted text, incorrect causality attribution of symptoms, misclassified recurrent HZ cases as PHN, and unclear clinical documentation.

Table . Error analysis of natural language processing false-negatives and false-positives.

Type of NLP[a] error		Description	Number of cases
False-negative			5
	EHR[b] data source	• Case 1: We did not include one free text table (formatted messages) from the Epic EHR. • Case 2: PHN[c] was mentioned in a clinical note from the hematology department, which was excluded from NLP processing.	2
	Unclear documentation	• HZ[d] or PHN was not stated in the clinical note, which was required by NLP to reduce false-positive hits.	1
	Symptom	• While the patient stated that HZ-associated pain had resolved, documents also indicated that the patient still had other PHN-related symptoms (prickling sensation and itchy).	2
False-positive			11
	EHR data source	• We included Epic's SmartData elements, which lacked specificity for PHN identification.	2
	Unclear documentation	• In 2 cases, the text was copied from the clinical notes in the index period. The NLP copy-and-paste detection algorithm was only applied to the clinical notes in the transitional period. • In another 2 cases, PHN and PHN-related medications were listed in the assessment and plan sections. However, it was unclear whether the patient had ongoing symptoms.	4
	Acute HZ	• NLP misclassified 2 acute HZ cases that occurred in the risk period as PHN.	2
	Causality	• Case 1: Pain thought to be due to chalazion based on information in follow-up visits. • Case 2: PHN was listed in the assessment section and tramadol and gabapentin were listed in the plan section. However, the medications were likely for lumbosacral radiculopathy.	2
	Symptom	• The patient reported generalized symptoms (nausea) since HZ, but there was no mention of concomitant sensory changes such as pain, thus the case did not meet our PHN definition.	1

[a]NLP: natural language processing.

[b]EHR: electronic health record.

[c]PHN: postherpetic neuralgia.

[d]HZ: herpes zoster.

Code-Based Methods

The prevalence proportions of PHN identified by code-based methods ranged from 5.4% to 13.1%. The code-based methods achieved a 52.7% - 61.8% sensitivity, 89.8% - 98.4% specificity, 27.6% - 72.1% PPV, and 96.3% - 97.1% NPV. The F-scores and MCCs ranged between 0.45 and 0.59 and between 0.32 and 0.61, respectively. The more sophisticated algorithms were no better than the PHN diagnosis code–only method as measured by the F-score or MCC. Although each component of the code-based methods identified PHN cases, most of them did not contribute to identifying additional true PHN cases beyond those identified by PHN diagnosis codes, and those that did have much lower PPVs (C4.3: 10.3%, C4.2: 26.1% and C2.2: 37.7%) than the PHN diagnosis code–only method (C1, PPV 72.1%) (Table 5). We re-reviewed all FP cases from code-based methods C1 and C3 and randomly sampled the remaining FP cases from approaches C2, C4, and C5. Among the 20 reviewed FP cases, we found that none were true PHN cases.

Table . Postherpetic neuralgia cases identified by code-based methods.

Method			PHN[a] diagnosis code used	PHN, n (%)	TP[b] (PPV[c]), n (%)	Supplementary TP[d]
C1[e]			✓	43 (5.4)[f]	31 (72.1)[f]	—[g]
C2			✓	87 (10.9)[f]	34 (39.1)[f]	3[f]
	C2.1		✓	43 (5.4)	31 (72.1)	—
	C2.2			69 (8.6)	26 (37.7)	3
	C2.3			4 (0.5)	1 (25.0)	—
C3			✓	44 (5.5)[f]	31 (70.5)[f]	—
	C3.1		✓	24 (3)	18 (75.0)	—
	C3.2		✓	7 (0.9)	7 (100.0)	—
	C3.3		✓	41 (5.1)	29 (70.7)	—
	C3.4		✓	23 (2.9)	17 (73.9)	—
C4			✓	105 (13.1)[f]	29 (27.6)[f]	2[f]
	C4.1		✓	36 (4.5)	26 (72.2)	—
	C4.2			69 (8.6)	18 (26.1)	2
		C4.2.1		7 (0.9)	6 (85.7)	—
		C4.2.2		2 (0.3)	1 (50.0)	—
		C4.2.3		64 (8.0)	16 (25.0)	1
		C4.2.4		2 (0.3)	1 (50.0)	1
	C4.3			29 (3.6)	3 (10.3)	2
		C4.3.1		25 (3.1)	1 (4.0)	1
		C4.3.2		2 (0.3)	1 (50.0)	1
		C4.3.3		2 (0.3)	1 (50.0)	—
C5			✓	74 (9.3)[f]	31 (41.9)[f]	—
	C5.1		✓	43 (5.4)	31 (72.1)	—
	C5.2			18 (2.3)	14 (77.8)	—
	C5.3			34 (4.3)	2 (5.9)	—

[a]PHN: postherpetic neuralgia.

[b]TP: true-positive.

[c]PPV: positive predictive value.

[d]Supplementary contributions to the number of correctly identified positive cases, apart from method C1.

[e]Method C1 only used PHN diagnosis codes.

[f]Overall performance.

[g]Not applicable.

Discussion

Principal Findings

We developed and validated NLP algorithms to identify PHN using various clinical data sources from EHRs. Compared with the chart-reviewed reference standard, the NLP algorithms showed high accuracy. This study demonstrates the feasibility of population-based PHN studies using EHR data with an automated method.

Using manual review to identify PHN cases is often infeasible for population-based research because a large volume of clinical notes would need to be reviewed. In contrast, the size of the study population and length of follow-up have little impact on running the NLP algorithm. Moreover, our NLP algorithm can readily capture PHN at varied time intervals, providing an efficient method to assess the long-term impact of PHN and compare results with studies using different PHN risk windows. Furthermore, studies can use NLP alone or with manual review confirmation. For example, a manual review of the NLP-positive cases (n=61) could increase the specificity and PPV to 100% and improve the F-score from 0.89 to 0.93 and MCC from 0.85 to 0.95; this is more efficient than a manual review of all 800 HZ cases.

Implementing NLP on EHR data presents challenges. In this study, data sources accounted for one-quarter of NLP errors (2 FNs and 2 FPs). First, clinical data were stored in a variety of locations within our institution's complex EHR system, which contains over 900,000 database tables. It is often difficult to locate the database table storing the data displayed in the EHR user interface. One FN case resulted from not including a previously unknown table. Second, selecting data sources for NLP processing is often a tradeoff. One FN and 2 FP cases resulted from including or excluding certain data sources. EHRs have also made it easy to create lengthy and bloated notes [31,32]. According to recent research, over half of clinical note content is duplicated or copied from earlier notes [32-34]. Clinicians may copy from prior visit notes to improve recall and clinical reasoning [35]. However, these replicated contents may lack temporal or contextual information, making them difficult to identify manually and challenging for NLP.

Because PHN-related symptoms such as pain and discomfort are common in a variety of medical conditions with numerous plausible causes, identifying PHN necessitates integrating the NLP-identified PHN symptoms with their associated anatomic location, temporality, and causality. These elements, however, are not always explicitly stated in clinical documents. About half of the NLP FP cases were from incorrectly attributing the complaint or treatment to PHN. These FP cases were partially explained by the NLP algorithm's preference for sensitivity over specificity.

Another popular method of PHN identification is using coded data from administrative claims or EHR, which could include a large sample size at a low cost. However, many of the code-based PHN identification algorithms have not been validated [3]. We implemented and validated 5 code-based algorithms, including 1 that solely uses PHN diagnosis codes (C1) and 2 that had previously been validated (C2 and C3). To maximize their sensitivity, algorithms C2-C5 used the "OR" statement to combine various criteria. The downside of using the "OR" logic is the loss of PPV. Algorithms C2-C5 all had worse PPV than the diagnosis codes–only algorithm (C1). However, in our study, the sensitivity of these algorithms ranged from 53% to 62%, with only C2 outperforming C1 (62% vs 56%). Algorithms C2-C5 had lower PPVs (28% - 71%) than C1 (72%). With such limited sensitivities, these algorithms may miss roughly half of the PHN cases. In our study, aside from the PHN diagnosis codes, the other diagnosis codes and prescription data had little impact on true case identification, instead adding complexity and increasing FPs.

Studies have used the similarity of the PHN proportions to construct the validity of their case-finding algorithms [8-11]. Administrative database studies reported PHN (pain persisting for ≥90 days) prevalences of 3% - 14% (Multimedia Appendix 6) [3], which are comparable to the 5.4% - 13.1% prevalences of the code-based approaches in our study. The broad range of prevalences identified in previous code-based studies could be caused by variations in study design, population, and data source [3]. However, the code-based approaches in this study had the same population and data source. Only the variation in algorithms could cause such a wide disparity.

We expanded the validations conducted for the 2 previously validated algorithms, which were performed on EHR data. The C2 (Klompas et al [8]) algorithm was only validated with the 30-day definition in the original study, and it had 86% sensitivity and 78% PPV. In our study, algorithm C2 with the 90-day definition had notably lower sensitivity (62%) and PPV (39%). One main contributor to the variability in performance is the difference in the temporal criteria. According to Yawn [36], up to 75% of pain present at 30 days disappears at 90 days, and the prevalence of PHN decreased by sixfold when the definition was changed from 30 days to 90 days. As prevalence decreases, so do the sensitivity and PPV [37,38]. The same trends were also reported in the original C2 paper; the PPVs for different PHN search criteria using the 30-day definition (29% - 95%) were nearly double that of using the 90-day definition (15% - 52%). The discrepancy in C2 algorithm performance between the original study and this study could be further explained by the differences in case definition. Our case definition for PHN is based on persistent PHN-related symptoms and causal attribution, not diagnosis code or medication. Algorithm C2 used ongoing symptoms or renewal of medication for HZ. The use of medications to identify PHN has some drawbacks, as PHN-related medications have a wide range of indications. For example, gabapentin, a first-line therapy for PHN, has over 20 approved and off-label uses [39]. Furthermore, prescriptions can be refilled in the absence of active PHN symptoms for various non-PHN disorders.

The original C3 (Klein et al [10]) algorithm was only validated on potential PHN cases identified by its 4 component criteria, rather than randomly selected HZ cases; only PPVs were reported. In this study, the 4 criteria of the C3 algorithm had PPVs ranging from 71% to 100%, which is consistent with the previous study's findings (PPVs ranging from 73% to 96%). The C3 algorithm was one of the best-performing code-based algorithms based on F-score and MCC. However, its low sensitivity (56%) and PPV (71%) indicate considerable misclassification. The lower overall PPV is partly due to the "OR" logic of the 4 criteria. Because Klein et al [10] did not describe the case definition or chart review rules, we were unable to assess their impact on the performance differences between the original C3 study and this study.

The substantial misclassification of coded methods as observed in this study could have a substantial impact on measuring incidence, identifying risk factors, and assessing vaccine effectiveness. Code-based method studies (C4 and C5) had identified depression, diabetes mellitus, heart failure, and chronic obstructive pulmonary disease as risk factors for PHN. It is conceivable that the link between depression and PHN is caused by using anticonvulsants and tricyclic antidepressants to identify PHN. The inclusion of prescriptions for pain medications and chronic pain codes may contribute to the association of diabetes mellitus [40], heart failure [41], and chronic obstructive pulmonary disease [42] with PHN.

Study Strengths and Limitations

This study was conducted within a large integrated health care system with comprehensive EHRs. Because the health plan provides strong incentives for members to use its facilities,

clinical documentation is expected to be more detailed. We developed NLP algorithms to identify PHN from various unstructured data sources within EHRs, such as clinical notes, which contain a wealth of information but differ greatly in structure, content, and quality. The algorithms were highly accurate, as evidenced by our validation. Compared with studies based on self-reported pain scores collected through surveys, EHR-based studies measure the health care burden of PHN, which is more clinically relevant. This study also has limitations. The reference standard relied on the review of EHRs which could be erroneous and incomplete [14]. Moreover, rereviewing cases in the validation set where NLP results differed from research associates' results may result in bias in favor of higher performance of the NLP algorithm. On the other hand, reconciling discrepant results improved the quality of the reference standard. Additionally, diagnosis codes, prescriptions, clinical documentation language, and style can differ between institutions and physicians. Our NLP method may perform differently in other test data sets.

Conclusions

PHN-related diagnosis codes have low sensitivity for identifying PHN cases. Additional diagnosis codes and prescription data did little to improve sensitivity while significantly lowering the PPV. Using clinical text from the EHR, the NLP-based method identified PHN cases with high accuracy. Our NLP method can be used in EHR-based studies to identify PHN risk factors and evaluate the effectiveness of vaccinations and treatments against PHN.

Acknowledgments

A part of this work was previously presented at IDWeek 2023, which was held in Boston, Massachusetts on October 12, 2023. This work was supported by Kaiser Permanente Southern California internal research funds.

Conflicts of Interest

HFT received research funding from GSK for a shingles vaccine–related study. He also received funding from Moderna for studies unrelated to this publication.

Multimedia Appendix 1
Postherpetic neuralgia abstraction decision rules.
[DOCX File, 30 KB - medinform_v12i1e57949_app1.docx]

Multimedia Appendix 2
Additional details on natural language processing algorithm development.
[DOCX File, 30 KB - medinform_v12i1e57949_app2.docx]

Multimedia Appendix 3
Code-based methods.
[DOCX File, 38 KB - medinform_v12i1e57949_app3.docx]

Multimedia Appendix 4
The proportion of herpes zoster– or postherpetic neuralgia–related notes by department/specialty.
[DOCX File, 81 KB - medinform_v12i1e57949_app4.docx]

Multimedia Appendix 5
Number and percentage of herpes zoster locations as identified by natural language processing.
[DOCX File, 32 KB - medinform_v12i1e57949_app5.docx]

Multimedia Appendix 6
Reported postherpetic neuralgia rates in previous administrative database studies.
[DOCX File, 32 KB - medinform_v12i1e57949_app6.docx]

References

1. Gershon AA, Breuer J, Cohen JI, et al. Varicella zoster virus infection. Nat Rev Dis Primers 2015 Jul 2;1:15016. [doi: 10.1038/nrdp.2015.16] [Medline: 27188665]
2. Gnann JW, Whitley RJ. Clinical practice. herpes zoster. N Engl J Med 2002 Aug 1;347(5):340-346. [doi: 10.1056/NEJMcp013211] [Medline: 12151472]
3. Kawai K, Gebremeskel BG, Acosta CJ. Systematic review of incidence and complications of herpes zoster: towards a global perspective. BMJ Open 2014 Jun 10;4(6):e004833. [doi: 10.1136/bmjopen-2014-004833] [Medline: 24916088]

4. Johnson RW, Alvarez-Pasquin MJ, Bijl M, et al. Herpes zoster epidemiology, management, and disease and economic burden in Europe: a multidisciplinary perspective. Ther Adv Vaccines 2015 Jul;3(4):109-120. [doi: 10.1177/2051013615599151] [Medline: 26478818]
5. Johnson RW, Rice ASC. Postherpetic neuralgia. N Engl J Med 2014 Oct 16;371(16):1526-1533. [doi: 10.1056/NEJMcp1403062]
6. Rowbotham M, Harden N, Stacey B, Bernstein P, Magnus-Miller L. Gabapentin for the treatment of postherpetic neuralgia: a randomized controlled trial. JAMA 1998 Dec 2;280(21):1837-1842. [doi: 10.1001/jama.280.21.1837] [Medline: 9846778]
7. Lal H, Cunningham AL, Godeaux O, et al. Efficacy of an adjuvanted herpes zoster subunit vaccine in older adults. N Engl J Med 2015 May 28;372(22):2087-2096. [doi: 10.1056/NEJMoa1501184] [Medline: 25916341]
8. Klompas M, Kulldorff M, Vilk Y, Bialek SR, Harpaz R. Herpes zoster and postherpetic neuralgia surveillance using structured electronic data. Mayo Clin Proc 2011 Dec;86(12):1146-1153. [doi: 10.4065/mcp.2011.0305]
9. Forbes HJ, Bhaskaran K, Thomas SL, et al. Quantification of risk factors for postherpetic neuralgia in herpes zoster patients: a cohort study. Neurology (ECronicon) 2016 Jul 5;87(1):94-102. [doi: 10.1212/WNL.0000000000002808] [Medline: 27287218]
10. Klein NP, Bartlett J, Fireman B, et al. Long-term effectiveness of zoster vaccine live for postherpetic neuralgia prevention. Vaccine (Auckl) 2019 Aug 23;37(36):5422-5427. [doi: 10.1016/j.vaccine.2019.07.004] [Medline: 31301920]
11. Muñoz-Quiles C, López-Lacort M, Orrico-Sánchez A, Díez-Domingo J. Impact of postherpetic neuralgia: a six year population-based analysis on people aged 50 years or older. J Infect 2018 Aug;77(2):131-136. [doi: 10.1016/j.jinf.2018.04.004] [Medline: 29742472]
12. Suaya JA, Chen SY, Li Q, Burstin SJ, Levin MJ. Incidence of herpes zoster and persistent post-zoster pain in adults with or without diabetes in the United States. Open Forum Infect Dis 2014 Sep;1(2):ofu049. [doi: 10.1093/ofid/ofu049] [Medline: 25734121]
13. Hillebrand K, Bricout H, Schulze-Rath R, Schink T, Garbe E. Incidence of herpes zoster and its complications in Germany, 2005-2009. J Infect 2015 Feb;70(2):178-186. [doi: 10.1016/j.jinf.2014.08.018] [Medline: 25230396]
14. Yawn BP, Wollan P, St Sauver J. Comparing shingles incidence and complication rates from medical record review and administrative database estimates: how close are they? Am J Epidemiol 2011 Nov 1;174(9):1054-1061. [doi: 10.1093/aje/kwr206] [Medline: 21920944]
15. Tanenbaum HC, Lawless A, Sy LS, et al. Differences in estimates of post-herpetic neuralgia between medical chart review and self-report. J Pain Res 2020;13:1757-1762. [doi: 10.2147/JPR.S255238] [Medline: 32765050]
16. Tseng HF, Lewin B, Hales CM, et al. Zoster vaccine and the risk of postherpetic neuralgia in patients who developed herpes zoster despite having received the zoster vaccine. J Infect Dis 2015 Oct 15;212(8):1222-1231. [doi: 10.1093/infdis/jiv244] [Medline: 26038400]
17. Zheng C, Luo Y, Mercado C, et al. Using natural language processing for identification of herpes zoster ophthalmicus cases to support population-based study. Clin Exp Ophthalmol 2019 Jan;47(1):7-14. [doi: 10.1111/ceo.13340] [Medline: 29920898]
18. Zheng C, Sy LS, Tanenbaum H, et al. Text-based identification of herpes zoster ophthalmicus with ocular involvement in the electronic health record: a population-based study. Open Forum Infect Dis 2021 Feb;8(2):ofaa652. [doi: 10.1093/ofid/ofaa652] [Medline: 33575426]
19. Delaney A, Colvin LA, Fallon MT, Dalziel RG, Mitchell R, Fleetwood-Walker SM. Postherpetic neuralgia: from preclinical models to the clinic. Neurotherapeutics 2009 Oct;6(4):630-637. [doi: 10.1016/j.nurt.2009.07.005] [Medline: 19789068]
20. Forbes HJ, Thomas SL, Smeeth L, et al. A systematic review and meta-analysis of risk factors for postherpetic neuralgia. Pain 2016 Jan;157(1):30-54. [doi: 10.1097/j.pain.0000000000000307] [Medline: 26218719]
21. Coplan PM, Schmader K, Nikas A, et al. Development of a measure of the burden of pain due to herpes zoster and postherpetic neuralgia for prevention trials: adaptation of the brief pain inventory. J Pain 2004 Aug;5(6):344-356. [doi: 10.1016/j.jpain.2004.06.001] [Medline: 15336639]
22. Zheng C, Rashid N, Wu YL, et al. Using natural language processing and machine learning to identify gout flares from electronic clinical notes. Arthritis Care Res (Hoboken) 2014 Nov;66(11):1740-1748. [doi: 10.1002/acr.22324] [Medline: 24664671]
23. Zheng C, Sun BC, Wu YL, et al. Automated identification and extraction of exercise treadmill test results. J Am Heart Assoc 2020 Mar 3;9(5):e014940. [doi: 10.1161/JAHA.119.014940] [Medline: 32079480]
24. Zheng C, Yu W, Xie F, et al. The use of natural language processing to identify Tdap-related local reactions at five health care systems in the Vaccine Safety Datalink. Int J Med Inform 2019 Jul;127:27-34. [doi: 10.1016/j.ijmedinf.2019.04.009] [Medline: 31128829]
25. Zheng C, Duffy J, Liu ILA, et al. Identifying cases of shoulder injury related to vaccine administration (SIRVA) in the United States: development and validation of a natural language processing method. JMIR Public Health Surveill 2022 May 24;8(5):e30426. [doi: 10.2196/30426] [Medline: 35608886]
26. Zheng C, Rashid N, Koblick R, An J. Medication extraction from electronic clinical notes in an integrated health system: a study on aspirin use in patients with nonvalvular atrial fibrillation. Clin Ther 2015 Sep;37(9):2048-2058. [doi: 10.1016/j.clinthera.2015.07.002] [Medline: 26233471]

27. Yanni EA, Ferreira G, Guennec M, et al. Burden of herpes zoster in 16 selected immunocompromised populations in England: a cohort study in the Clinical Practice Research Datalink 2000-2012. BMJ Open 2018 Jun 7;8(6):e020528. [doi: 10.1136/bmjopen-2017-020528] [Medline: 29880565]
28. Derczynski L. Complementarity, F-score, and NLP evaluation. Presented at: Tenth International Conference on Language Resources and Evaluation (LREC'16); May 23-28, 2016; Portorož, Slovenia.
29. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020 Jan 2;21(1):6. [doi: 10.1186/s12864-019-6413-7] [Medline: 31898477]
30. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min 2023 Feb 17;16(1):4. [doi: 10.1186/s13040-023-00322-4] [Medline: 36800973]
31. Downing NL, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: are we ignoring the real cause? Ann Intern Med 2018 Jul 3;169(1):50-51. [doi: 10.7326/M18-0139] [Medline: 29801050]
32. Rule A, Bedrick S, Chiang MF, Hribar MR. Length and redundancy of outpatient progress notes across a decade at an academic medical center. JAMA Netw Open 2021 Jul 1;4(7):e2115334. [doi: 10.1001/jamanetworkopen.2021.15334] [Medline: 34279650]
33. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and sources of duplicate information in the electronic medical record. JAMA Netw Open 2022 Sep 1;5(9):e2233348. [doi: 10.1001/jamanetworkopen.2022.33348] [Medline: 36156143]
34. Wang MD, Khanna R, Najafi N. Characterizing the source of text in electronic health record progress notes. JAMA Intern Med 2017 Aug 1;177(8):1212-1213. [doi: 10.1001/jamainternmed.2017.1548] [Medline: 28558106]
35. Zheng C, Lee MS, Bansal N, et al. Identification of recurrent atrial fibrillation using natural language processing applied to electronic health records. Eur Heart J Qual Care Clin Outcomes 2024 Jan 12;10(1):77-88. [doi: 10.1093/ehjqcco/qcad021] [Medline: 36997334]
36. Yawn BP. Post-shingles neuralgia by any definition is painful, but is it PHN? Mayo Clin Proc 2011 Dec;86(12):1141-1142. [doi: 10.4065/mcp.2011.0724] [Medline: 22134931]
37. Murad MH, Lin L, Chu H, et al. The association of sensitivity and specificity with disease prevalence: analysis of 6909 studies of diagnostic test accuracy. CMAJ 2023 Jul 17;195(27):E925-E931. [doi: 10.1503/cmaj.221802] [Medline: 37460126]
38. Tenny S, Hoffman MR. Prevalence. In: StatPearls [Internet]: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK430867/ [accessed 2024-09-04]
39. Yasaei R, Katta S, Patel P, Saadabadi A. Gabapentin. In: StatPearls [Internet]: StatPearls Publishing; 2024. URL: https://www.ncbi.nlm.nih.gov/books/NBK493228/ [accessed 2024-09-04]
40. Dyck PJ, Kratz KM, Karnes JL, et al. The prevalence by staged severity of various types of diabetic neuropathy, retinopathy, and nephropathy in a population-based cohort: the Rochester Diabetic Neuropathy Study. Neurology (ECronicon) 1993 Apr;43(4):817-824. [doi: 10.1212/wnl.43.4.817] [Medline: 8469345]
41. Alemzadeh-Ansari MJ, Ansari-Ramandi MM, Naderi N. Chronic pain in chronic heart failure: a review article. J Tehran Heart Cent 2017 Apr;12(2):49-56. [Medline: 28828019]
42. Andenæs R, Momyr A, Brekke I. Reporting of pain by people with chronic obstructive pulmonary disease (COPD): comparative results from the HUNT3 population-based survey. BMC Public Health 2018 Jan 25;18(1):181. [doi: 10.1186/s12889-018-5094-5] [Medline: 29370850]

Abbreviations

EHR: electronic health record
FN: false-negative
FP: false-positive
HZ: herpes zoster
KPSC: Kaiser Permanente Southern California
MCC: Matthews correlation coefficient
NLP: natural language processing
NPV: negative predictive value
PHN: postherpetic neuralgia
PPV: positive predictive value
TN: true-negative
TP: true-positive

Edited by Q Chen; submitted 01.03.24; peer-reviewed by A Hosny, K Kawai; revised version received 02.07.24; accepted 08.07.24; published 10.09.24.

Please cite as:
Zheng C, Ackerson B, Qiu S, Sy LS, Daily LIV, Song J, Qian L, Luo Y, Ku JH, Cheng Y, Wu J, Tseng HF
Natural Language Processing Versus Diagnosis Code–Based Methods for Postherpetic Neuralgia Identification: Algorithm Development and Validation
JMIR Med Inform 2024;12:e57949
URL: https://medinform.jmir.org/2024/1/e57949
doi:10.2196/57949

The Impact of Collaborative Documentation on Person-Centered Care: Textual Analysis of Clinical Notes

Victoria Stanhope[1], MSW, PhD; Nari Yoo[1], MA; Elizabeth Matthews[2], MSW, PhD; Daniel Baslock[3], MSW, PhD; Yuanyuan Hu[4], MSW, PhD

1

2

3

4

Corresponding Author:
Victoria Stanhope, MSW, PhD

Abstract

Background: Collaborative documentation (CD) is a behavioral health practice involving shared writing of clinic visit notes by providers and consumers. Despite widespread dissemination of CD, research on its effectiveness or impact on person-centered care (PCC) has been limited. Principles of PCC planning, a recovery-based approach to service planning that operationalizes PCC, can inform the measurement of person-centeredness within clinical documentation.

Objective: This study aims to use the clinical informatics approach of natural language processing (NLP) to examine the impact of CD on person-centeredness in clinic visit notes. Using a dictionary-based approach, this study conducts a textual analysis of clinic notes from a community mental health center before and after staff were trained in CD.

Methods: This study used visit notes (n=1981) from 10 providers in a community mental health center 6 months before and after training in CD. LIWC-22 was used to assess all notes using the Linguistic Inquiry and Word Count (LIWC) dictionary, which categorizes over 5000 linguistic and psychological words. Twelve LIWC categories were selected and mapped onto PCC planning principles through the consensus of 3 domain experts. The LIWC-22 contextualizer was used to extract sentence fragments from notes corresponding to LIWC categories. Then, fixed-effects modeling was used to identify differences in notes before and after CD training while accounting for nesting within the provider.

Results: Sentence fragments identified by the contextualizing process illustrated how visit notes demonstrated PCC. The fixed effects analysis found a significant positive shift toward person-centeredness; this was observed in 6 of the selected LIWC categories post CD. Specifically, there was a notable increase in words associated with achievement (β=.774, *P*<.001), power (β=.831, *P*<.001), money (β=.204, *P*<.001), physical health (β=.427, *P*=.03), while leisure words decreased (β=−.166, *P*=.002).

Conclusions: By using a dictionary-based approach, the study identified how CD might influence the integration of PCC principles within clinical notes. Although the results were mixed, the findings highlight the potential effectiveness of CD in enhancing person-centeredness in clinic notes. By leveraging NLP techniques, this research illuminated the value of narrative clinical notes in assessing the quality of care in behavioral health contexts. These findings underscore the promise of NLP for quality assurance in health care settings and emphasize the need for refining algorithms to more accurately measure PCC.

(JMIR Med Inform 2024;12:e52678) doi:10.2196/52678

KEYWORDS

person-centered care; collaborative documentation; natural language processing; concurrent documentation; clinical documentations; visit notes; community; health center; mental health center; textual analysis; clinical informatics; behavioral health; mental health; linguistic; linguistic inquiry; dictionary-based; sentence fragment; psychology; psychological; clinical information; decision-making; mental health services; clinical notes; NLP

Introduction

Collaborative documentation (CD) is a specified behavioral health practice where clinicians complete visit notes jointly with consumers during the session [1]. Through deliberate clinical strategies, such as sharing the computer screen, reading visit notes aloud, and actively seeking consumer's input into the content of the session note [2], CD is a person-centered strategy that aims to engage and empower individuals and facilitate a mutual agreement on treatment progress, service goals, and session activities. Both as a means to promote person-centered care (PCC) and make health information more accessible and transparent to consumers, the practice of CD is being widely

disseminated through formal and informal training for community mental health providers [3].

PCC, also referred to as patient-centered care, is a paradigm shift in health care that is defined by the Institute of Medicine as care that is responsive to individual client preferences, needs, and values [4]. A key part of behavioral health reform, PCC moves away from disease-centered treatment to a more holistic approach that engages individuals as active, empowered partners in their care. Person-centered care planning (PCCP) is a recovery-oriented practice that integrates principles of PCC into the service planning process. PCCP orients service planning and documentation to the unique personal life goals of the consumer [5], and provides a framework for operationalizing PCC in practice through six core principles, which are as follows: (1) PCC is based on the person's own unique life goals and aspirations; (2) PCC is oriented toward promoting recovery rather than only minimizing illness and symptoms; (3) PCC articulates the person's own role and the role of both paid practitioners and natural supports in assisting the person to achieve his or her own goals; (4) PCC focuses and builds on the person's capacities, strengths, and interests; (5) PCC emphasizes the use of natural community settings rather than segregated program settings; and (6) PCC anticipates and allow for uncertainty, setbacks, and disagreements as inevitable steps on the path to greater self-determination [6].

While CD has emerged as a recognized person-centered practice strategy by fully engaging the consumer in decisions about their care, there remains a very limited evidence base demonstrating its clinical effectiveness, including its impact on PCC. Existing research demonstrates that CD is aligned with consumer preferences; a recent scoping review found that the ability to read their visit notes improved consumers' experience in mental health care, including their ability to remember their plans of care, understand their treatment, and trust decisions made with providers [7]. In community mental health, the use of CD specifically has been found to strengthen the therapeutic alliance [8], and improve service engagement, both in terms of visit attendance and medication adherence [1].

Despite this preliminary support for CD, more work is needed to examine its impact on quality of care. Adding urgency to this knowledge gap, regulatory changes to the 21st Century Cures Act now mandate organizations to make electronic health and mental health information, including many types of visit notes, accessible online to service users [9]. As a consequence, best practices for using visit notes to support the provision of high-quality PCC are needed.

Clinical informatics, which provides highly efficient ways to mine data within the electronic health record (EHR), is a promising methodological approach to examining the impact of CD on clinical quality. Although behavioral health has lagged behind medical settings in the adoption of EHRs, now the majority of behavioral health settings document visit notes via the EHR [10]. The shift from paper records to EHRs provides an unprecedented opportunity for clinical informatics to inform quality improvement in behavioral health care. Visit notes include nuanced information about care processes, session content, provider perspective, and the consumer experience that are not captured in other more structured fields of the EHR, offering valuable insight into clinical quality that has not yet been systematically targeted in mental health services research [11]. Researchers have applied manual content analysis to visit notes to evaluate dimensions of PCC [12,13], but these efforts are inevitably limited in scope and dependent on the interpretive lens of the researcher [14].

One clinical informatics strategy that can parse large volumes of unstructured narrative information into quantitative data is natural language processing (NLP) [15]. While some text mining approaches use words as the unit of analysis, NLP is able to capture the complexity of unstructured narrative using underlying metadata, which examines how words relate to each other in a sentence and the semantic context of a sentence [16]. The method involves syntactic processing, information extraction, and capturing meaning and relationships across concepts. NLP has been predominantly used for detecting pathology and predicting behavior [17] including measuring the following: alcohol misuse in trauma patients [18], suicidal behavior [19], adverse childhood experiences among VA patients [20], smoking status [21], and sentiment at discharge [22]. Recent studies have used this method to measure quality and safety in nursing care [23], identify integrated care elements within primary care [24], and detect changes in clinical documentation after opening notes to service users [25], however, there have been fewer studies that have used NLP to analyze indicators of mental health care quality, including PCC. By providing a framework to systematically categorize and compare the contents of clinical notes, NLP is well poised to provide novel insight into how CD affects clinical quality and PCC.

The dictionary-based approach in NLP is a method that uses a predefined lexicon to identify and extract certain types of words or phrases within a given text. This approach is often used in tasks such as part-of-speech tagging, named entity recognition, and sentiment analysis. In the dictionary-based approach, the dictionary consists of a list of words or phrases along with their associated tags or labels. For example, in part-of-speech tagging, the dictionary might contain a list of common nouns, verbs, and adjectives, each with a corresponding tag that indicates the word's part of speech. One of the benefits of the dictionary-based approach is its simplicity and ease of implementation, as it targets only a predefined lexicon, or set of words, and does not require the use of complex algorithms or machine learning models. One of the most dictionary-based approaches is LIWC (Linguistic Inquiry and Word Count), and LIWC analysis has been found to be particularly useful for identifying and analyzing the emotional state of individuals in mental health–related text data [26]. Despite its potential, such as its application in oncology settings to examine the changes in clinical notes after patient access [27], the dictionary-based approach has not yet been applied to clinical notes in behavioral health care settings for its change after CD.

This study examines the effect of CD on the person-centeredness of documentation within a community mental health setting using a quasi-experimental pretest-posttest design. The study adapted a well-established dictionary to conduct sentiment

analysis of provider clinic visit notes before and after providers were trained in CD.

Methods

Ethical Considerations

This study (IRB-FY2022-5838) was granted an exemption by the New York University Institutional Review Board.

Data Source

The study setting was a community mental health center, which provided a range of services to people with severe mental illnesses including outpatient therapy, assertive community treatment, community support programs, and psychiatric rehabilitation. The data set, or corpus, is visit notes completed by providers trained in CD. Providers were trained in CD by MTM Services [28], a leader in CD training. Their training consisted of tailored in-person workshops, technical assistance, and a "train the trainer" series designed for practice sustainability. The clinic training consisted of 8 hours of web-based training and customized consultation support for implementation.

The visit notes were clinical narrative documents completed when the provider had an in-person contact with a service user. Visit notes document the following: (1) sessions focused on developing or revising the service plan, a narrative clinical document completed every 6 months with the service user detailing goals, strengths and barriers, short-term objectives, supports, professional/ billable services, and natural support and self-directed actions, and (2) a visit focused on progress made towards completing the steps in the service plan. Inclusion criteria for providers were that they were: full-time employees of the clinic; provided services to adults with severe mental illnesses; trained in CD; and had been employed in their position a year prior and the year after being trained.

This study sampled all visit notes completed by participating providers 6 months prior to CD training and all visit notes completed during the sixth month following training (with a month lag for implementation time) between July 1, 2015, and March 10, 2020. Based on anticipated documentation rates of a clinic note being generated by each visit, we recruited 10 providers, generating a total of 1981 visit notes. On average, 198.1 notes were included per provider with a standard deviation of 37.8. Sampled notes were deidentified but linked to providers through unique identifiers. In addition to the session narrative, each note was comprised of the following sections: (1) Provider ID, (2) Time and Date of Service, (3) Therapy Modality, and (4) Length of Stay.

Analytic Strategy

We used LIWC-22 to compute the scores for PCC-related sentiment and linguistic categories [29] in each sampled note. The LIWC method is a text analysis tool that uses linguistic algorithms to identify and categorize words in a text according to their psychological properties. The method is based on the LIWC dictionary, which contains over 5000 words and word stems organized into linguistic and psychological categories. In this study, 3 domain experts used a consensus approach to select LIWC categories that mapped onto the 6 principles of PCC [6] with several of the categories mapping onto more than one domain. This iterative process involved 8 rounds of independent coding by 3 raters using Excel, followed by discussions to resolve discrepancies and achieve consensus on the final coding of all data points. According to the Cohen κ measure, raters had an average interrater agreement of 88% across the LIWC categories [30]. Out of a possible 107 LIWC categories, we selected 12 categories. The selected LIWC domains and their associated PCC principles are summarized in Tables 1 and 2 [6,30].

Table . Mapping PCCP[a] principles onto LIWC[b] categories.

PCCP principles	LIWC categories
PCCP is based on the person's own unique life goals and aspirations.	Lifestyle (leisure, home, work, money, religion), social referents (family, friend)
PCCP is oriented toward promoting recovery rather than only minimizing illness and symptoms.	Health (physical, wellness)
PCCP articulates the person's own role and the role of both paid practitioners and natural supports in assisting the person to achieve his or her own goals.	Social referents (family, friend), physical (physical, wellness)
PCCP focuses and builds on the person's capacities, strengths, and interests.	Drives (achievement, affiliation, power), lifestyle (leisure, home, work, money, religion)
PCCP emphasizes the use of natural community settings rather than segregated program settings.	Lifestyle (leisure, home, work, money, religion)
PCCP anticipates and allows for uncertainty, setbacks, and disagreements as inevitable steps on the path to greater self-determination.	Drives (affiliation, achievement, power)

[a]PCCP: person-centered care planning.

[b]LIWC: Linguistic Inquiry and Word Count.

Table . Sample fragments from clinic notes generated by contextualizing LIWC[a] subcategories.

LIWC subcategories		Clinic note excerpt
Drives		
	Achievement	• Was able to identify her strengths, abilities, and self-identified progress in therapy reports that she has been engaging more socially with friends and that she has been trying to express herself
	Affiliation	• Unexpected death of her pet was huge stressor that triggers increases in intensity of depression improvement with community resources and social networking due to becoming more integrated with his new community and within his daughter's school district
	Power	• The client reports with a positive outlook that she feels more in control and is excited to receive praise for using her therapy learning • Had no outbursts or over reactions recently and feels proud of her assertive but in control manner
Lifestyle		
	Home	• Highlighting that he doesn't like the apartment "being so quiet" when his son is gone… • The client has increased productivity at home with baking and wrapping presents.
	Leisure	• She is also making self-care more of a priority, "I scheduled a cruise and it's just my sister and I going" • Progress is that he has begun basketball
	Money	• She admits that she has no savings of her own but she knows that she will get alimony • He has figured out a plan to pay for housing
	Religion	• Topics of no control include people's religious actions and beliefs, elements within his own church and community, as well as the political culture • She reported that she has been supported by her church and increased her faith significantly
	Work	• He continues to apply for jobs and is now working with workforce development. • Not working currently, sent about a couple of job applications, continues with college course work
Physical		
	Physical	• Strengthen his tongue and swallowing skills, will occur to help reduce his concern regarding health issues • Barriers to maintaining treatment plans goals, because her varying blood sugars have caused severe mood swings

LIWC subcategories		Clinic note excerpt
	Wellness	• Emily is making progress in her goals to increase positive self-worth and applying healthy coping skills • Trouble with motivation at times and needing to clear his mind, discussed option of yoga and mindfulness
Social referents		
	Family	• Has been spending time with her family and working to express her feelings and needs when appropriate • Progress is that she has figured out how to resolve some of her parenting issues
	Friend	• States that she feels she can "cut loose and have fun" with her friends • Regards to her recent trip to [place] which she really enjoyed she made new friends

[a]LIWC: Linguistic Inquiry and Word Count.

Analysis

Using the LIWC categories described above, we compared notes pre and post CD training to examine differences in PCC. The complete visit note was used as the unit of analysis. For data preprocessing, we cleaned the data and converted the information into a structured format that made it amenable to identifying patterns in the data. The LIWC-22 dictionary is case-insensitive and allows for matching 2-word phrases. The LIWC-22 software removes extra whitespace characters by default. While irrelevant words are not explicitly defined in the dictionary, we removed section headings (eg, Location) from the clinical notes before processing the text to eliminate some irrelevant words. Negated phrases (eg, "not happy") are not treated differently from nonnegated phrases in the standard scoring. To address this limitation, we included an additional analysis in Multimedia Appendix 1 that controls for the negations score.

To validate the team's selection of LIWC categories, we first used the Contextualizer function of LIWC-22 to generate sentence fragments containing words related to each LIWC domain included in the analysis. We then analyzed changes in the clinical note before and after CD training, using the complete visit note as the unit of analysis. To calculate changes in the content characteristics of clinical notes before and after CD training, we calculated frequency scores of each LIWC category for every clinical note since this study focuses on examining the presence of words from the LIWC dictionary.

Instead of using LIWC scores based on percentages, we used frequency scores. This decision was made based on previous research indicating that CD enhances the length of clinical notes in terms of word and character count [31]. Additionally, we used frequencies to assess the presence of PCC-related language in clinical notes. Our primary interest was to capture whether clinicians used PCC-related words in their documentation, even if these words did not constitute a large proportion of the total text. By focusing on word frequencies, we aimed to mitigate the potential impact of note length inflation due to CD practices, such as copy-and-pasting or using templates [32,33]. The LIWC frequency scores are calculated by the following steps. First, we calculated LIWC scores, which are determined by the percentage of words in a text that belong to specific linguistic categories. Then, to find the frequency of each category within a clinical note, the respective LIWC percentage is multiplied by the total word count of the note. For example, an LIWC value of 1.02 for achievement indicates that the note contains 1.02 words related to achievement per the LIWC dictionary.

We used a fixed effects model that included the provider as a categorical variable. The changes in the notes before and after CD training were calculated while accounting for nesting within the therapist using individual fixed-effects models. This approach allowed us to examine whether the changes in PCC language use before and after the CD training varied across the 10 providers in our sample. To further investigate these differences, we calculated the intraclass correlation coefficient (ICC) and conducted paired sample *t* tests for each provider (Multimedia Appendix 2). The LIWC version 22 was used for dictionary-based sentiment analysis and STATA (version 17.0; Statacorp) was used for statistical analyses.

Results

Contextualizing LIWC Categories

The sentences generated by the Contextualizer function illustrated how the LIWC categories mapped onto the PCCP principles (see Table 2). The drives subcategories (achievement, affiliation, and power) reflected a strength-based approach by describing the positive changes made by the client; greater self-determination by feeling more in control; and interests by capturing how a client feels connected to community, people, and pets. The lifestyle subcategories (home, leisure, money, religion, and work) captured the unique details of the person's life that are needed to individualize treatment, including their beliefs, values, and preferences which inform their personal life goals. Examples included going on a cruise as part of self-care and seeking employment. Lifestyle categories also illustrated

people's interests such as playing sports and describing their life in the community such as attending church. Physical categories (physical and wellness) demonstrated a more holistic approach to the client by paying attention to how physical health affects mental health and also to a focus on recovery by including activities that promote wellness such as yoga and health coping skills. The social referents category (family and friend) demonstrated the role family and friends play as natural supports such as going on vacation with your sister or having fun with your friends.

Changes in LIWC Categories

Overall, there was a significant positive change in 4 of the selected LIWC categories indicating person-centeredness after the providers had been trained in CD. As shown in Table 3, the fixed effects regression analysis found the following among the 12 selected characteristics: an increased use in 4 categories, decreased use in 4 categories, and no change in use in 4 categories, while controlling for length of sessions at the therapist level. Within the drives category, we observed a significant increase in words associated with achievement (β=.774, *P*<.001, ICC=0.146) and power (β=.831, *P*<.001, ICC=0.072), with the ICC values indicating that 14.6% and 7.2% of the variance in these word categories, respectively, could be attributed to differences between therapists. In the lifestyle category, there was an increase in the use of words related to home (β=.047, *P*=.35, ICC=0.060) and work (β=.047, *P*=.12, ICC=0.285), but these changes are not statistically significant. The ICC values suggest that 6.0 and 28.5% of the variance in these word categories, respectively, could be attributed to differences between therapists. On the other hand, leisure and religion-associated words showed a significant decrease (β=−.166, *P*=.002, ICC=0.033; β=−.105, *P*<.001, ICC=0.031, respectively), while words associated with money displayed substantial increases (β=.204, *P*<.001, ICC=0.035). The ICC values for these categories indicate that 3.3%, 3.1%, and 3.5% of the variance, respectively, could be attributed to differences between therapists. In the health category, there was a notable increase in the use of physical health–related words (β=.427, *P*=.03, ICC=0.159), with 15.9% of the variance attributable to differences between therapists. In contrast, wellness-related words decreased significantly (β=−.427, *P*<.001, ICC=0.211), with 21.1% of the variance attributable to differences between therapists. In the social referents category, the use of family-related words did not show any significant change (β=−.016, *P*=.89, ICC=0.085), with 8.5% of the variance attributable to differences between therapists. However, the frequency of friend-related words decreased significantly (β=−.084, *P*=.005, ICC=0.028), with only 2.8% of the variance attributable to differences between therapists.

Table . Person-centeredness before and after collaborative documentation (CD). Coefficients were reported. Standard errors are in parentheses. β denotes coefficients of fixed-effects models. Fixed-effects estimates were based on models from the STATA module "xtreg" commands, clustered by therapist and with controls for length of session (minutes).

Category		Sample words	Frequency mean		Fixed effects		
			Before CD	After CD	ICC[a]	β (SE)	*P* value
Drives							
	Achievement	work, better, best, working	4.04	5.09	0.146	.774 (0.129)	<.001
	Affiliation	we, our, us, help	4.28	4.24	0.212	.086 (0.138)	.53
	Power	wn, order, allow, power	1.75	2.66	0.072	.831 (0.094)	<.001
Lifestyle							
	Home	home, house, room, bed	0.66	0.77	0.060	.047 (0.051)	.35
	Leisure	game, fun, play, party	0.78	0.62	0.033	−.166 (0.053)	.002
	Money	business, pay, price, market	0.20	0.41	0.035	.204 (0.035)	<.001
	Religion	god, hell, christmas, church	0.21	0.10	0.031	−.105 (0.029)	<.001
	Work	work, school, working, class	6.20	6.68	0.285	.234 (0.152)	.12
Health							
	Physical	medic, food, patients, eye	5.43	5.90	0.159	.427 (0.193)	.03
	Wellness	healthy, gym, supported, diet	0.93	0.59	0.211	−.416 (0.057)	<.001
Social referents							
	Family	parent, mother, father, baby	2.24	2.09	0.085	−.016 (0.116)	.89
	Friend	friend, boyfriend, girlfriend, dude	0.26	0.18	0.028	−.084 (0.03)	.005

[a]ICC: intraclass correlation coefficient.

Discussion

Principal Findings

The contextualizing analysis provided insight into the documentation content reflecting the selected LIWC categories and demonstrated how person-centered principles can be integrated into clinical documentation. Using LIWC categories illustrated how providers described their clients in ways that gave a sense of their lives beyond their mental health. These details included what they care about and how that related to their personal goals (getting a job, financial situation, going on a cruise, or attending church), their life beyond the clinic in the community (their home life, family and friends, and their community). The sentences also showed when clinicians used a strengths-based approach, the nature and content of their clinic notes changed in ways that moved beyond symptoms [34].

The quantitative results indicating whether there was an increase in person-centeredness of clinical documentation as indicated by relevant LIWC categories were mixed, with a significant increase in half the subcategories. The most pronounced positive increase was within the drives category, with words associated with power and achievement increasing. In terms of PCC, this indicates providers made more reference to self-determination, including how the client has made progress, and their strengths. Lifestyle categories, which include words related to hobbies and other social activities were more mixed, showing that providers were not consistent in increasing their focus on personal life goals or taking a holistic view of the client. In health, there was a significant increase in references to physical health but a decrease in references to wellness. This may reflect the increasing efforts to integrate health into their clinical interventions [35], but does not indicate a more recovery-oriented focus. Finally, in terms of social referents, there was no change in family references, which may be due to the fact that family inclusion is a common best practice of PCC [36,37], and a decrease in references to friends, often considered a source of natural supports within PCC.

Existing work has suggested that CD can improve important indicators of PCC, including service engagement and the quality of the working alliance [1,8], but research has yet not illuminated how this practice impacts care processes, including how person-centered principles are integrated into clinical interventions. Through analysis of session notes, this study found an increase in strengths-based approaches to clinical documentation following CD training, which may also reflect a shift towards interventions that emphasize self-determination. In addition to expanding the limited evidence base around the impact of CD on clinical quality, this study uniquely describes the mechanisms through which CD supports alliance building and engagement in care.

To meet the much-documented challenges of measuring PCC [38,39], a core component of health care reform, this study sought to harness the richness of clinical narrative data using a machine learning algorithm. While some studies have used NLP to study psychotherapy sessions [25], this is one of the first to mine narrative psychosocial documentation in behavioral health settings. Overall, this methodological approach has considerable potential to sample large quantities of documentation to measure different aspects of care quality, including PCC. NLP can be used both by researchers to examine how quality of care predicts clinical outcomes and by clinics to promote and document quality improvement.

Despite this potential, there are considerable challenges in calibrating algorithms so that they can accurately capture more nuanced aspects of care, such as person-centeredness. This study chose to use a well-validated dictionary designed to capture psychological concepts within narrative data. Although the study team was able to map existing LIWC categories onto established principles of PCC, the algorithm was not explicitly designed to measure these constructs, which may have contributed to the lack of positive findings within certain subcategories. Furthermore, the LIWC dictionary's focus on single words limits the algorithm's ability to capture more nuanced meanings that occur when words are evaluated within the larger context of surrounding phrases or sentences. This suggests the need to develop an algorithm focused specifically on PCC.

This study showed the potential for using NLP techniques to measure the quality of care within behavioral health settings. As more care standards demand that clinics demonstrate PCC [40] within mental health, there is a pressing need for feasible methods to capture this quality dimension. Being able to use at the aggregate level note data that reflects more nuanced and individualized aspects of care would help clinics document and report PCC.

In addition, our study highlights the value of clinical notes for research in behavioral health settings. Clinical notes have been harnessed for research purposes but have mainly been confined to hospital settings due to the scarcity of publicly available data [41]. The nature of psychosocial documentation differs from other clinical notes, which require different analytics and models and there is a need for publicly available data sets for analyzing psychotherapy notes in the United States. Furthermore, mental health notes often contain identifiable and sensitive information different from other clinical notes for physical illnesses, so an in-depth discussion on ethics, privacy, and deidentification and the development of techniques such as word embedding models to improve the privacy of clinical notes [42,43] for mental health notes is required.

Limitations

The study was limited to the scope of the LIWC categories rather than an algorithm developed specifically to capture PCC and therefore, was not able to measure the concept in its entirety. Furthermore, the analysis limited itself to categories with a positive valence rather than also measuring the inverse of PCC. Although well validated, the LIWC can still fail to capture the meaning of words and mis-categorize them but as it is used more, the algorithm will continue to be trained and improved. Overall, while documentation is an important indicator of PCC, it does not directly capture the interpersonal interactions between the provider and the clinician which shape a client's experience of PCC.

While our fixed effects model controlled for Provider ID and Date of Service, we were unable to account for potential differences in therapy modality or length of stay. In our study setting, the therapy modality primarily consisted of individual psychotherapy sessions excluding 2 notes (1 group therapy and 1 family therapy) resulting in insufficient variation to control for this variable. Additionally, our current data set did not include sufficient information on length of stay to include this variable in the model. Future research could benefit from examining the influence of therapy modality and longitudinal factors on PCC language use in clinical documentation.

It is important to acknowledge that while our study demonstrates an increase in person-centered content within clinic visit notes following the implementation of the CD, we did not directly assess whether the increased PCC in the clinic visit notes was associated with improved PCC practices. Future research should investigate the relationship between the presence of PCC in clinical notes and its impact on PCC practices and outcomes.

Conclusion

This study is an important first step in using NLP to measure the quality of care through narrative clinical notes in behavioral health settings. We were able to identify key PCC principles within the notes using a dictionary-based approach and examine whether CD changes the way providers document with respect to PCC. This demonstrates the potential for NLP to be used by both researchers and clinics as a quality improvement tool and the importance of further developing algorithms that can capture the nuances of PCC.

Acknowledgments

This work was funded by the Constance and Martin Silver Center on Data Science and Social Equity at New York University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Person-centeredness before and after collaborative documentation: control for negation words.
[DOCX File, 21 KB - medinform_v12i1e52678_app1.docx]

Multimedia Appendix 2
Paired sample *t* test.
[DOCX File, 33 KB - medinform_v12i1e52678_app2.docx]

References

1. Stanhope V, Ingoglia C, Schmelter B, Marcus SC. Impact of person-centered planning and collaborative documentation on treatment adherence. Psychiatr Serv 2013 Jan;64(1):76-79. [doi: 10.1176/appi.ps.201100489] [Medline: 23280459]
2. Matthews EB. Integrating the electronic health record into behavioral health encounters: strategies, barriers, and implications for practice. Adm Policy Ment Health 2017 Jul;44(4):512-523. [doi: 10.1007/s10488-015-0676-3]
3. DiCarlo R, Garcia YE. Chapter 4 - Electronic record keeping and psychotherapy alliance: the role of concurrent collaborative documentation. In: Tettegah SY, Garcia YE, editors. Emotions, Technology, and Health: Academic Press; 2016:63-82.
4. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press; 2001.
5. Tondora J, Miller R, Slade M, Davidson L. Partnering for Recovery in Mental Health: A Practical Guide to Person-Centered Planning: John Wiley & Sons; 2014.
6. Tondora J, Pocklington S, Osher D, Davidson L. Implementation of person-centered care and planning: from policy to practice to evaluation.: Substance Abuse and Mental Health Services Administration; 2005. URL: https://www.hsri.org/files/uploads/publications/ImplementationOfPersonCenteredCareandPlanning.pdf [accessed 2024-09-13]
7. Schwarz J, Bärkås A, Blease C, et al. Sharing clinical notes and electronic health records with people affected by mental health conditions: scoping review. JMIR Ment Health 2021 Dec 14;8(12):e34170. [doi: 10.2196/34170] [Medline: 34904956]
8. Matthews EB. Computer use in mental health treatment: understanding collaborative documentation and its effect on the therapeutic alliance. Psychotherapy (Chic) 2020 Jun;57(2):119-128. [doi: 10.1037/pst0000254] [Medline: 31599638]
9. O'Neill S, Blease C, Delbanco T. Open notes become law: a challenge for mental health practice. Psychiatr Serv Am Psychiatric Assoc 2021 Jul 1;72(7):750-751. [doi: 10.1176/appi.ps.202000782] [Medline: 33971748]
10. Agency for Planning and Evaluation Certified Community Behavioral Health Clinics Demonstration Program: Report to Congress, 2019.: Washington, DC: Department of Health and Human Services Sep 2019 URL: https://aspe.hhs.gov/reports/certified-community-behavioral-health-clinics-demonstration-program-report-congress-2019 [accessed 2022-10-13]
11. Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. Comput Inform Nurs 2009 Jul;27(4):215-223. [doi: 10.1097/NCN.0b013e3181a91b58] [Medline: 19574746]
12. Haselden M, Dixon LB, Overley A, et al. Giving back to families: evidence and predictors of persons with serious mental illness contributing help and support to families. Community Ment Health J 2018 May;54(4):383-394. [doi: 10.1007/s10597-017-0172-1] [Medline: 29022227]
13. Butler JM, Gibson B, Patterson OV, et al. Clinician documentation of patient centered care in the electronic health record. BMC Med Inform Decis Mak 2022 Mar 12;22(1):65. [doi: 10.1186/s12911-022-01794-w] [Medline: 35279157]
14. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res 2016 Jun;25(2):86-100. [doi: 10.1002/mpr.1481]
15. Edgcomb JB, Zima B. Machine learning, natural language processing, and the electronic health record: innovations in mental health services research. Psychiatr Serv 2019 Apr;70(4):346-349. [doi: 10.1176/appi.ps.201800401] [Medline: 30784377]
16. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019 May;125:37-46. [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]
17. Mascio A, Kraljevic Z, Bean D, et al. Comparative analysis of text classification approaches in electronic health records. 2020 Presented at: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing p. 86-94. [doi: 10.18653/v1/2020.bionlp-1.9]
18. Afshar M, Phillips A, Karnik N, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. J Am Med Inform Assoc 2019 Mar 1;26(3):254-261. [doi: 10.1093/jamia/ocy166] [Medline: 30602031]
19. Carson NJ, Mullin B, Sanchez MJ, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLoS ONE 2019 Feb;14(2):e0211116. [doi: 10.1371/journal.pone.0211116] [Medline: 30779800]

20. Hammond KW, Ben‐Ari AY, Laundry RJ, Boyko EJ, Samore MH. The feasibility of using large‐scale text mining to detect adverse childhood experiences in a VA‐treated population. J Trauma Stress 2015 Dec;28(6):505-514. [doi: 10.1002/jts.22058] [Medline: 26579624]
21. Rajendran S, Topaloglu U. Extracting smoking status from electronic health records using NLP and deep learning. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:507-516. [Medline: 32477672]
22. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. PLoS ONE 2015 Aug 24;10(8):e0136341. [doi: 10.1371/journal.pone.0136341] [Medline: 26302085]
23. Zerden LDS, Lombardi BM, Richman EL, Fraher EP, Shoenbill KA. Harnessing the electronic health record to advance integrated care. Fam Syst Health 2021 Mar;39(1):77-88. [doi: 10.1037/fsh0000584] [Medline: 34014732]
24. Rahimian M, Warner JL, Jain SK, Davis RB, Zerillo JA, Joyce RM. Significant and distinctiven-grams in oncology notes: a text-mining method to analyze the effect of opennotes on clinical documentation. JCO Clin Cancer Inform 2019 Dec;3(3):1-9. [doi: 10.1200/CCI.19.00012] [Medline: 31184919]
25. Atkins DC, Steyvers M, Imel ZE, Smyth P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Impl Sci 2014 Dec;9(1). [doi: 10.1186/1748-5908-9-49] [Medline: 24758152]
26. Blease C, Torous J, Hägglund M. Does patient access to clinical notes change documentation? Front Public Health 2020 Nov 27;8:577896. [Medline: 33330320]
27. Alpert JM, Morris BB, Thomson MD, Matin K, Sabo RT, Brown RF. Patient access to clinical notes in oncology: a mixed method analysis of oncologists' attitudes and linguistic characteristics towards notes. Patient Educ Couns 2019 Oct;102(10):1917-1924. [doi: 10.1016/j.pec.2019.05.008] [Medline: 31109771]
28. MTM Services. 2023. URL: https://www.mtmservices.org/ [accessed 2023-08-31]
29. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The development and psychometric properties of LIWC-22. 2022 Feb. [doi: 10.13140/RG.2.2.23890.43205]
30. Klein D. Implementing a general framework for assessing interrater agreement in stata. Stata J 2018 Dec;18(4):871-901. [doi: 10.1177/1536867X1801800408]
31. Yoo N, Matthews E, Baslock D, Stanhope V. Impact of collaborative documentation on completeness and length of clinical notes in behavioral health settings. Psychiatr Serv 2023 Aug 2;75(2):186-190. [doi: 10.1176/appi.ps.20230118] [Medline: 37528697]
32. Thornton JD, Schold JD, Venkateshaiah L, Lander B. Prevalence of copied information by attendings and residents in critical care progress notes. Crit Care Med 2013 Feb;41(2):382-388. [doi: 10.1097/CCM.0b013e3182711a1c] [Medline: 23263617]
33. Wang MD, Khanna R, Najafi N. Characterizing the source of text in electronic health record progress notes. JAMA Intern Med 2017 Aug 1;177(8):1212. [doi: 10.1001/jamainternmed.2017.1548]
34. Braun MJ, Dunn W, Tomchek SD. A pilot study on professional documentation: do we write from a strengths perspective? Am J Speech Lang Pathol 2017 Aug 15;26(3):972-981. [doi: 10.1044/2017_AJSLP-16-0117] [Medline: 28637055]
35. Scharf DM, Eberhart NK, Schmidt N, et al. Integrating primary care into community behavioral health settings: programs and early implementation experiences. Psychiatr Serv 2013 Jul;64:660-665. [doi: 10.1176/appi.ps.201200269]
36. Frampton SB, Giuliano M. Patient-centered care: the North Star to guide us during uncertainty into a better day. Int J Qual Health Care 2023 Aug 9;35(3). [doi: 10.1093/intqhc/mzad061] [Medline: 37556113]
37. Boise L, White D. The family's role in person-centered care: practice considerations. J Psychosoc Nurs Ment Health Serv 2004 May;42(5):12-20. [doi: 10.3928/02793695-20040501-04]
38. Stanhope V, Baslock D, Tondora J, Jessell L, Ross AM, Marcus SC. Developing a tool to measure person-centered care in service planning. Front Psychiatry 2021 Aug 2;12:681597. [doi: 10.3389/fpsyt.2021.681597] [Medline: 34408678]
39. Burgers JS, van der Weijden T, Bischoff E. Challenges of research on person-centered care in general practice: a scoping review. Front Med 2021 Jun 24;8:1-9. [doi: 10.3389/fmed.2021.669491]
40. Person-centered service planning guidelines for Medicaid managed care organizations, local departments of social services, and health homes. New York State Department of Health. 2022. URL: https://www.health.ny.gov/health_care/managed_care/plans/pcsp_guidelines.htm [accessed 2023-08-30]
41. Zhou N, Wu Q, Wu Z, Marino S, Dinov ID. DataSifterText: partially synthetic text generation for sensitive clinical notes. J Med Syst 2022 Nov 16;46(12):96. [doi: 10.1007/s10916-022-01880-6] [Medline: 36380246]
42. Abdalla M, Abdalla M, Rudzicz F, Hirst G. Using word embeddings to improve the privacy of clinical notes. J Am Med Inform Assoc 2020 Jun 1;27(6):901-907. [doi: 10.1093/jamia/ocaa038] [Medline: 32388549]
43. Abdalla M, Abdalla M, Hirst G, Rudzicz F. Exploring the privacy-preserving properties of word embeddings: algorithmic validation study. J Med Internet Res 2020 Jul 15;22(7):e18055. [doi: 10.2196/18055] [Medline: 32673230]

Abbreviations

CD: collaborative documentation
EHR: electronic health record

ICC: intraclass correlation coefficient
LIWC: Linguistic Inquiry and Word Count
NLP: natural language processing
PCC: person-centered care
PCCP: person-centered care planning

Edited by J Hefner; submitted 12.09.23; peer-reviewed by J Abbas, K Woo, LDS Zerden; revised version received 07.06.24; accepted 26.06.24; published 20.09.24.

Please cite as:
Stanhope V, Yoo N, Matthews E, Baslock D, Hu Y
The Impact of Collaborative Documentation on Person-Centered Care: Textual Analysis of Clinical Notes
JMIR Med Inform 2024;12:e52678
URL: https://medinform.jmir.org/2024/1/e52678
doi:10.2196/52678

Data Set and Benchmark (MedGPTEval) to Evaluate Responses From Large Language Models in Medicine: Evaluation Development and Validation

Jie Xu[1], DHM; Lu Lu[1], MA; Xinwei Peng[1], MM; Jiali Pang[1], MS; Jinru Ding[1], MEng; Lingrui Yang[2], MSc; Huan Song[3,4], PhD; Kang Li[3,4], PhD; Xin Sun[2], MD; Shaoting Zhang[1], PhD

1

2

3

4

Corresponding Author:
Shaoting Zhang, PhD

Abstract

Background: Large language models (LLMs) have achieved great progress in natural language processing tasks and demonstrated the potential for use in clinical applications. Despite their capabilities, LLMs in the medical domain are prone to generating hallucinations (not fully reliable responses). Hallucinations in LLMs' responses create substantial risks, potentially threatening patients' physical safety. Thus, to perceive and prevent this safety risk, it is essential to evaluate LLMs in the medical domain and build a systematic evaluation.

Objective: We developed a comprehensive evaluation system, MedGPTEval, composed of criteria, medical data sets in Chinese, and publicly available benchmarks.

Methods: First, a set of evaluation criteria was designed based on a comprehensive literature review. Second, existing candidate criteria were optimized by using a Delphi method with 5 experts in medicine and engineering. Third, 3 clinical experts designed medical data sets to interact with LLMs. Finally, benchmarking experiments were conducted on the data sets. The responses generated by chatbots based on LLMs were recorded for blind evaluations by 5 licensed medical experts. The evaluation criteria that were obtained covered medical professional capabilities, social comprehensive capabilities, contextual capabilities, and computational robustness, with 16 detailed indicators. The medical data sets include 27 medical dialogues and 7 case reports in Chinese. Three chatbots were evaluated: ChatGPT by OpenAI; ERNIE Bot by Baidu, Inc; and Doctor PuJiang (Dr PJ) by Shanghai Artificial Intelligence Laboratory.

Results: Dr PJ outperformed ChatGPT and ERNIE Bot in the multiple-turn medical dialogues and case report scenarios. Dr PJ also outperformed ChatGPT in the semantic consistency rate and complete error rate category, indicating better robustness. However, Dr PJ had slightly lower scores in medical professional capabilities compared with ChatGPT in the multiple-turn dialogue scenario.

Conclusions: MedGPTEval provides comprehensive criteria to evaluate chatbots by LLMs in the medical domain, open-source data sets, and benchmarks assessing 3 LLMs. Experimental results demonstrate that Dr PJ outperforms ChatGPT and ERNIE Bot in social and professional contexts. Therefore, such an assessment system can be easily adopted by researchers in this community to augment an open-source data set.

(JMIR Med Inform 2024;12:e57674) doi:10.2196/57674

KEYWORDS

ChatGPT; LLM; assessment; data set; benchmark; medicine

Introduction

The development of large language models (LLMs) has revolutionized natural language processing, raising significant interest in LLMs as a solution for complex tasks such as instruction execution and elaborate question-answering in various domains [1]. Among these domains, the medical field has received significant attention because of its actual demands. Recently, progress has been achieved in medical education [2], patient care management [3], medical exams [4], and other medical applications.

Despite their capabilities, LLMs are prone to generating hallucinations (not fully reliable responses) [5,6]. Hallucinations in LLMs' responses create substantial risks, potentially

threatening patient's physical safety and leading to serious medical malpractice. Thus, to perceive and prevent this safety risk, we must conduct an exhaustive evaluation of LLMs in the medical domain and build a systematic evaluation.

However, conducting an exhaustive evaluation for LLMs is not trivial. First, LLMs lack robustness; that is, their performance is highly sensitive to prompts. White et al [7] showed that a meticulously crafted and thoroughly tested prompt could greatly improve performance and produce superior results. Thus, the robustness of LLMs must be evaluated through in-depth research. Second, the evaluation criteria of LLMs are critical. Recent evaluations have been mainly based on automatic metrics [8-10] (eg, bilingual evaluation understudy, Recall-Oriented Understudy for Gisting Evaluation, and accuracy) in popular applications such as machine translation and text summarization. Despite their high efficiency, these automatic metrics are insufficient for using LLMs in real-world medical scenarios. Other factors such as the logical coherence of responses, social characteristics like tone, and the ability to understand contextual information are essential influential factors [6,11-17].

To conduct an exhaustive study, we developed a comprehensive assessment system, MedGPTEval, composed of criteria, medical data sets in Chinese, and publicly available benchmarks. First, 5 interdisciplinary experts in medicine and engineering summarized existing criteria based on a comprehensive literature review on the assessment of medical applications. The experts have rich research experience in artificial intelligence (AI) or big data, but specific subdisciplines and majors may vary, including AI and health care management, AI and clinical medicine, AI and medical imaging, clinical medicine and big data, AI, medical imaging, and computer vision. Second, these candidate criteria were optimized using a Delphi method. In the realms of health care [18,19] and the foresight of interdisciplinary future-built environments [20], the Delphi method has emerged as an efficacious instrument for amalgamating the insights of experts across diverse domains, fostering consensus, and refining standards. This approach serves to harmonize the interests of all pivotal stakeholders, thereby amplifying the efficacy and transparency of value-based outcomes [19]. The obtained evaluation criteria cover medical professional capabilities, social comprehensive capabilities, contextual capabilities, and computational robustness, with 16 detailed indicators. Third, 3 clinical experts designed medical data sets to interact with LLMs, including 27 medical dialogues and 7 case reports in Chinese. The case data set is adapted and constructed based on real clinical cases. We have adopted multiple rounds of internal review and expert review processes, and have conducted verification consistent with actual clinical scenarios to ensure the accuracy and practicality of the data. Finally, benchmarking experiments were conducted on the data sets. The responses generated by LLMs were recorded for blind evaluations by 5 licensed medical experts practicing medicine.

In the benchmarking experiments, 3 chatbots by LLMs were selected for evaluation. First, ChatGPT, an LLM created by OpenAI, has gained global popularity owing to its exceptional language capabilities [2]. However, ChatGPT has not been specifically trained for the medical domain [21]. Second, ERNIE Bot is an LLM developed by Baidu, Inc, a Chinese computer technology company [22]. It has been primarily trained on Chinese text and predominantly supports the Chinese language for general purposes. Third, Doctor PuJiang (Dr PJ) is an LLM created by the medical research group of the Shanghai Artificial Intelligence Laboratory. Dr PJ has been trained based on massive Chinese medical corpora and supports various application scenarios, such as diagnosis, triage, and medical question-answering. Note that ChatGPT and ERNIE Bot are general-purpose conversational AI systems, while Dr PJ is an LLM fine-tuned specifically for medical use. To promote research on the evaluation of medical LLMs, we conducted benchmarking experiments on the proposed medical data sets in Chinese. Experimental results show that Dr PJ outperformed ChatGPT and ERNIE Bot in both the multi-turn medical dialogues (scores of 13.95 vs 13.41 vs 12.56 out of 16) and the case report scenarios (scores of 10.14 vs 8.71 vs 8.0 out of 13).

The scale of the data set remains limited. We urge researchers in this community to join this open project via email (xujie@pjlab.org.cn). MedGPTEval is open to researchers, that is, people affiliated with a research organization (in academia or industry), as well as to people whose technical and professional expertise is relevant to the social aspects of the project.

The contribution of this work is 2-fold:

1. By conducting a thorough study of LLMs used in the medical context and collaborating with domain experts, we established comprehensive evaluation criteria to assess the medical responses of LLMs.
2. Based on the criteria, we released a set of open-source data sets for the evaluation of medical responses in Chinese and conducted benchmark experiments on 3 chatbots, including ChatGPT.

Methods

Evaluation Criteria

The evaluation criteria for assessing the LLMs were summarized by a thorough literature review. The evaluation criteria were then optimized using the Delphi method [23]. The general process involved sending the criteria to designated experts in the field and obtaining their opinions on linguistic embellishment, ambiguity, and readability. After generalizing and corrections, we provided anonymous feedback to each expert. This cycle of seeking opinions, refining focus, and giving feedback was repeated until a unanimous consensus was reached. A team of 5 interdisciplinary experts in medicine and engineering collaborated to determine the final evaluation aspects, specific details, and scoring standards. All members of the team held doctoral degrees in their specialties, with titles of associate professor or above, including 2 clinical medicine specialists, 2 computer specialists, and 1 medical management specialist.

Medical Data Sets in Chinese

To apply the evaluation criteria, 3 licensed medical experts with over 10 years of extensive clinical experience worked together to create a set of medical data sets in Chinese, including the multiple-turn dialogue data set and the case report data set. The

case report data set necessitated a singular round of questioning and encompasses an elaborate medical record of the patient, including age, gender, medical history (personal and familial), symptoms, medication history, and other relevant information. In addition, the medical problem consulted had to be clearly described. In contrast, the data set with multiple-turn dialogue was derived through an iterative process comprising four rounds. The initial round was initiated with the patient's symptoms, followed by supplementary descriptions of medication, examination, or other symptom-related queries. The data set with multiple-turn dialogue required careful consideration to assess contextual relevance.

Benchmark

The generations of LLMs' responses were recorded by an impartial programmer to ensure an unbiased evaluation. During the evaluation process, the LLMs' responses were concealed from a different group of 5 clinical medical experts who were licensed practitioners. They have similar years of clinical experience, and we have unified training on assessment processes and criteria to account for the impact of differences in clinical practice on the assessment process. The clinical fundamental response performances of 3 LLMs (ChatGPT, ERNIE Bot, and Dr PJ) were then compared based on the assessment criteria outlined above and on the proposed medical data sets. The data sets proposed by 5 clinical medical experts based on actual clinical experience and clinical confusion, and determined through peer review and discussion were used to evaluate the medical and social capabilities of the LLMs, while the multiple-turn dialogue data set was used to additionally assess their contextual abilities. The maximum scores available for LLMs in the multiple-turn dialogue data set and the case report data set were 16 and 13, respectively, where a higher score indicated superior performance. Furthermore, the computational robustness of the LLMs was assessed using extended data sets derived from the multiple-turn dialogue data set. Lastly, a subset of the case reports was randomly selected and comprehensively reviewed by five medical experts. The benchmark assessment methods are summarized in Table 1.

Table . Summary of benchmark assessment.

Data sets and assessment aspects		Assessment approaches
Medical dialogue		
	Medical professional capabilities, social comprehensive capabilities, contextual capabilities	Maximum score of 16
	Computational robustness	Percentage
Case report		
	Medical professional capabilities, social comprehensive capabilities	Maximum score of 13
	Computational robustness	Percentage
	Comprehensive review	Comments

Ethical Considerations

This study does not include human participants (ie, no human subject experimentation or intervention was conducted) and does not require institutional review board approval.

Results

Comprehensive Assessment Criteria

The draft evaluation criteria for assessing the LLMs were summarized by a thorough literature review [6,7,11-14,16,17,24] from 4 aspects: medical professional capabilities, social comprehensive capabilities, contextual capabilities, and computational robustness. All 5 interdisciplinary experts made suggestions for fine-tuning the assessment method, and they reached a consensus using the Delphi method to make it more scientifically rigorous and easier to read [23].

Medical Professional Capabilities

The professional comprehensive capabilities of LLMs' answers were evaluated using 6 indicators [7,12,17]: (1) accuracy, requiring that there are no medical errors in the answers and that the answers do not provide any harmful information to patients (accuracy can also include the evaluation of safety); (2) informativeness, where a 3-point Likert scale was used to evaluate the informativeness of the answers (0: incomplete; 1: adequate; 2: comprehensive); (3) expansiveness, meaning that the answers contain useful information besides the medical knowledge included in the question; (4) logic, with a 3-point Likert scale (0: the answer is irrelevant to the topic; 1: off topic, the answer does not directly address the topic but is still relevant; 2: on topic, the answer addresses the topic directly and positively); (5) prohibitiveness, where the LLMs correctly identify medical vocabulary or prohibited vocabulary; and (6) sensitivity, ensuring that LLMs' answers do not contain any politically sensitive expressions. Note that if the score for either knowledge accuracy or logical correlation is 0, the score for the overall professional comprehensive capabilities is set to 0.

Social Comprehensive Capabilities

We conducted an overall evaluation of the social comprehensive performances using 4 indicators [6,11,12,14]: (1) comprehension, where a binary scale is used to evaluate the readability of the answers (0: awkward sounding—all answers are professional and not explanatory; 1: understandable—intuitive and easy to understand); (2) tone, which pertains to the appropriate use of mood/tone in the generated responses by the LLMs, including the use of mood

particles, symbols, emotional rhythm, and emotional intensity; (3) empathy, where the accuracy of the scenario analysis is considered, including emotional understanding and reasoning; and (4) social decorum, using a 3-point Likert scale to evaluate the social decorum (0: rude–not matching any friendly social keywords or displaying malicious language attacks; 1: general—matching 1-2 keywords; 2: graceful—matching 3 or more keywords).

Contextual Capabilities

Three indicators were used to access the contextual capabilities [13,24] only in the multiple-turn dialogue data set, as follows: (1) repeated answer, which means that no duplicate answers should appear in the responses generated by LLMs; (2) anaphora matching, which involves correctly identifying and matching the abbreviations or aliases of medical professional terms used in the dialogue; and (3) key information, where LLMs can recognize and include all relevant information from the question in their response, particularly those that have been repeated 2 or more times in the questions. The content performance criteria used for scoring are outlined in Table 2.

Table . Summary of evaluation aspects, indicators, criteria, and data sets.

Evaluation aspects		Data sets	Evaluation criteria	Score
Medical professional capabilities		Both		
	Accuracy[a]		No medical knowledge errors are present in the answer	1
	Informativeness		Comprehensive: answers include additional information beyond the expectations	2
	Expansiveness		Answers include content from aspects other than medical knowledge included in the question	1
	Logic[a]		On topic: the answers address the topic directly and positively	2
	Prohibitiveness		The model can correctly identify medical or prohibited terms	1
	Sensitivity		There is no political sensitivity expressed in the answers of chatbots by LLM[b]	1
Social comprehensive capabilities		Both		
	Comprehension		Understandable: the answers are intuitive and easy to understand	1
	Tone		The answers use correct modal particles and symbols	1
	Empathy		The answers can accurately empathize with the patient	1
	Social decorum		Appropriate: matching 3 or more keywords	2
Contextual capabilities		Multiple-turn dialogue		
	Repeated answer		The model has no duplicate answers	1
	Anaphora matching		The model can identify medical professional abbreviations and aliases	1
	Key information		The model can identify key information that appears 2 or more times	1

[a]Highest priority. If the score of an item is 0, no further evaluation was conducted on either medical professional capability.

[b]LLM: large language model.

Computational Robustness

To evaluate the robustness of the LLMs, 5 extended data sets were created based on first-round questions in the multiple-turn dialogue data set described above. Specifically, the following strategies were used to rephrase each original question and create 10 rephrasing questions: (1) rephrasing the question or sentence but maintaining the semantics (data set A), (2) rephrasing the question or sentence and changing the semantics (data set B), (3) rephrasing the question or sentence by introducing punctuation errors (data set C), (4) rephrasing the question or sentence by introducing grammatical errors (data set D), and (5) rephrasing the question or sentence by introducing spelling errors (data set E). Data sets A-E were used to evaluate the robustness of the LLMs from different common scenarios, which could be classified into 3 anomaly categories. Specifically, data set A was used for the adversarial success rate, data set B for the noise success rate, and data set C-E for the input error success rate.

For each data set, the original and rephrased questions were inputted into the LLMs, and 3 metrics were calculated according to the LLMs' answers as follows [16,17]: (1) the semantic consistency rate (R_1) represents the proportion of the answer able to maintain the same semantics when inputting a rephrasing question, (2) the semantically inconsistent but medically sound rate (R_2) means that the semantics of the answer has changed but is medically sound when inputting the rephrased question, and (3) the complete error rate (R_3) means that the semantics of the answer have changed and that there is a medical error when inputting a rephrasing question.

Medical Data Sets in Chinese

Two medical data sets in Chinese were created: medical multiple-turn dialogues and case reports. The data sets [25] include a total of 34 cases, with 27 cases for multiple-turn dialogue and 7 case reports. Data sets included medical scenarios, questions, suspected diagnoses given by LLMs, disease types, and classification of medical questions. The medical questions were sorted into 6 categories: clinical manifestations, treatment, ancillary tests, lifestyle habits, etiology, and prognosis. Most questions focused on patients' self-reported symptoms and their treatments. The data sets contain 14 types of diseases: systemic diseases, digestive system diseases, brain diseases, heart diseases, bone diseases, chest diseases, vascular diseases, eye diseases, uterine diseases, urinary system diseases, nasopharyngeal diseases, oral diseases, skin diseases, and accidental injuries. Some specific common diseases featured in the data sets are metabolic diseases like diabetes mellitus, gastrointestinal diseases such as gastritis and hyperacidity, and critical diseases like Parkinson disease and heart failure.

Benchmarks Based on ChatGPT, ERNIE Bot, and Dr PJ

Analysis of the Results in Two Medical Scenarios

As shown in Table 3, three assessment aspects were covered in the multiple-turn dialogue evaluation: medical professional capabilities, social comprehensive capabilities, and contextual capabilities. Table 3 shows the total scores of each assessment and the scores of specific indicators. Dr PJ outperformed ChatGPT and ERNIE Bot, with total scores of 13.95, 13.41, and 12.56, respectively. ChatGPT achieved a slightly higher score of 6.30 in medical professional capabilities, compared to 6.25 for Dr PJ and 5.63 for ERNIE Bot. Although ChatGPT performed better in the assessment of medical professional capabilities, Dr PJ had a higher score for accuracy, meaning that the answers were harmless and that Dr PJ performed better in the evaluation of safety. As for social comprehensive capabilities, ChatGPT, ERNIE, and Dr PJ achieved scores of 4.26, 4.33, and 4.70, respectively. Dr PJ achieved a score of 3.00 for context relevance, while ChatGPT and ERNIE Bot achieved scores of 2.85 and 2.59, respectively.

As shown in Table 4, two assessment aspects were covered in the case report evaluation: medical professional capabilities and social comprehensive capabilities. Dr PJ outperformed ChatGPT and ERNIE Bot, with total scores of 10.14, 8.71, and 8.00, respectively. As for medical professional capabilities, Dr PJ achieved 6.86, higher than that of ChatGPT (6.43) and ERNIE Bot (5.71). Similarly, Dr PJ had the highest score (1.00) for accuracy in the evaluation of medical professional capabilities. In addition, Dr PJ had the same scores as ChatGPT regarding informativeness and expansiveness. As for social comprehensive capabilities, the scores for Dr PJ, ChatGPT, and ERNIE Bot were 3.29, 2.29, and 2.29, respectively. Specific scores for each indicator can be found in Table 4.

Table . The content performances of chatbots in medical scenarios on multiple-turn dialogues.

Evaluation indicators	Chatbots		
	ChatGPT	ERNIE Bot	Doctor PuJiang
Total score (maximum score: 16)	13.41	12.56	13.95
Medical professional capabilities (maximum score: 8)	6.30	5.63	6.25
Accuracy	0.91	0.79	0.94
Informativeness	1.40	1.22	1.31
Expansiveness	0.19	0.12	0.17
Logic	1.81	1.50	1.84
Prohibitiveness	1.00	1.00	1.00
Sensitivity	1.00	1.00	1.00
Social comprehensive capabilities (maximum score: 5)	4.26	4.33	4.70
Comprehension	0.96	0.96	0.96
Tone	0.96	1.00	1.00
Empathy	0.70	0.70	0.85
Social decorum	1.63	1.67	1.89
Contextual capabilities (maximum score: 3)	2.85	2.59	3.00
Repeated answer	0.96	0.81	1.00
Anaphora matching	0.96	0.85	1.00
Key information	0.93	0.93	1.00

Table . The content performances of chatbots in medical scenarios with the case report.

Evaluation indicators	Chatbots		
	ChatGPT	ERNIE bot	Doctor PuJiang
Total score (maximum score: 13)	8.71	8.00	10.14
Medical professional capabilities (maximum score: 8)	6.43	5.71	6.86
Accuracy	0.86	0.71	1.00
Informativeness	1.43	1.14	1.43
Expansiveness	0.43	0.43	0.43
Logic	1.71	1.43	2.00
Prohibitiveness	1.00	1.00	1.00
Sensitivity	1.00	1.00	1.00
Social comprehensive capabilities (maximum score: 5)	2.29	2.29	3.29
Comprehension	1.00	1.00	1.00
Tone	0.29	0.14	0.71
Empathy	0.00	0.14	0.29
Social decorum	1.00	1.00	1.29

Comprehensive Review of Detailed Case Reports

The comments of 2 case reports by 5 medical experts are shown in Multimedia Appendix 1. Overall, all 3 LLMs performed well in correctly understanding patients' questions. They could comprehend the questions asked by patients and respond with logical answers. However, Dr PJ outperformed the others in terms of sociality. Additionally, Dr PJ answered the questions in an orderly manner, with clear and intuitive serial numbers listed.

Computational Robustness Performance

The results in Table 5 show that Dr PJ outperformed ChatGPT and ERNIE Bot in the semantic consistency rate, with a higher adversarial success rate, noise success rate, and input error success rate. This indicates that Dr PJ was the best at

maintaining the same semantics of the model answers when questions were paraphrased. Furthermore, in the complete error rate category, both Dr PJ and ERNIE Bot had lower error rates than ChatGPT, suggesting that the semantics of the answer changed when the question was altered. Dr PJ also had a low probability of medical errors.

Table . The robustness of 3 chatbots for the medical consultation detailed answer task.

Chatbots, anomaly category, and data set		R_1[a] (%)	R_2[b] (%)	R_3[c] (%)
ChatGPT				
ASR[d]				
	Data set A	15	65	20
NSR[e]				
	Data set B	15	55	30
IESR[f]				
	Data set C	0	100	0
	Data set D	30	40	30
	Data set E	20	80	0
ERNIE Bot				
ASR				
	Data set A	10	85	5
NSR				
	Data set B	0	100	0
IESR				
	Data set C	0	100	0
	Data set D	20	80	0
	Data set E	20	80	0
Doctor PuJiang				
ASR				
	Data set A	15	80	5
NSR				
	Data set B	35	65	0
IESR				
	Data set C	60	40	0
	Data set D	50	40	10
	Data set E	80	20	0

[a]R_1: semantic consistency rate.

[b]R_2: semantically inconsistent but medically sound.

[c]R_3: complete error rate.

[d]ASR: adversarial success rate.

[e]NSR: noise success rate.

[f]IESR: input error success rate.

Discussion

Principal Findings

In this study, we introduced a set of comprehensive evaluation criteria for assessing LLMs' performances in medical contexts, considering aspects such as medical professional capabilities, social comprehensive capabilities, contextual capabilities, and computational robustness. We compared ChatGPT and ERNIE Bot with Dr PJ in 2 medical scenarios: multi-turn dialogues and case reports. Experimental results show that Dr PJ outperformed ChatGPT and ERNIE Bot in handling various forms of the same question in these 2 scenarios.

Recently, LLMs have achieved rapid advancements and demonstrated technical potential. However, only a few

question-and-answer evaluation methods have been developed for nonmedical fields or accuracy aspects. Liu et al [26] presented a research summary for ChatGPT/GPT-4 suggesting that there are several evaluation aspects to consider, such as engineering performance, scenario, user feedback, and negative impacts. Similarly, West [17] evaluated the accuracy of ChatGPT-3.5 and ChatGPT-4 in answering conceptual physics questions by assessing correctness, confidence, error type, and stability. Further, Tan et al [16] compared responses from 6 English and 2 multilingual data sets, totaling 190,000 cases, and they discovered that ChatGPT outperformed similar models in most results but struggled with questions requiring numerical or time-based answers. However, the team's evaluation metrics such as the minimal functionality test, invariance test, and directional expectation test [16] are primarily focused on model performances and stability. Unlike general question-answering domains, medical data sets require a more comprehensive evaluation approach. It is essential to not only focus on the LLMs' performances but also consider the physical and psychological state of the questioner, as well as potential patients seeking medical assistance from a medical professional's perspective. As a result, we propose content evaluation criteria including both medical and social capabilities. Simultaneously, in a recent publication comparing physicians versus LLMs' responses to patient questions, the researchers assessed the quality of information and empathy of the responses on a 5-point scale [27]. Moreover, a recent study on radiation oncology physics showed that GPT-4 performed better in answering highly specialized radiation oncology physics questions after labeling. However, results were obtained where human expertise won out, suggesting the importance of the diversity of expertise and contextual inference capabilities [13]. Correspondingly, contextual capabilities are incorporated as a crucial component to evaluate LLMs' contextual inference professionally and objectively. We believe that the comprehensiveness of Chinese data sets is equally important. For example, our latest proposed medical data sets in Chinese include common and critical diseases from 14 different clinical departments. Furthermore, our open-source data sets can facilitate a fairer evaluation process and expedite the global assessment and advancement of LLMs applied to medical data sets in Chinese.

Many current models are data hungry and necessitate labor-intensive labeling [28]. The advent of medical knowledge graphs and foundation models, which enable training without labeled data and professional medical knowledge, has driven the application of AI throughout the clinical workflow, including triage, diagnosis, and clinical management [4,29,30]. Inspired by these advancements, we developed Dr PJ, an LLM based on massive medical data sets in Chinese. Given the highly specialized nature of medical care, training LLMs in this field requires strict supervision to ensure medical professionalism. Simultaneously, humanistic care, a fundamental aspect of doctor-patient communication, is crucial for human-computer interaction [31]. Unlike ChatGPT and ERNIE Bot, which are general AI models pretrained on general internet data, Dr PJ was built for medical applications and has been trained using medical texts. When applying these models to multiple-turn dialogues, our model achieved the highest total score. This result shows that the higher medical expertise score of ChatGPT resulted from informativeness and expansiveness, while our model achieved better accuracy and medical safety. Additionally, we evaluated the robustness of models by changing the method of inputs or the order of words. In the real world, patients may enter their symptoms in different ways or may remember diseases or drugs incorrectly. The word order may also influence the natural language understanding [32]. Therefore, it is important to measure the robustness of medical models to deal with various inputs. Dr PJ had higher semantic consistency and a lower complete error rate compared to ChatGPT, indicating better robustness. Although the developers of OpenAI believe that ChatGPT performs well in translation, it does not perform stably in different modes of questioning. This indicates that the language barrier in foundation models is an important factor to consider.

Limitations

Limitations remain in the evaluation system and LLM development. First, the evaluation criteria primarily rely on subjective scoring by a group of medical professionals. Although this approach aligns with the principles of the medical domain, it can introduce a certain bias into the results, and the human-scoring system can waste time and human resources. Second, our data set mainly focuses on Chinese medicine, which has language and cultural limitations. This may have some impact on the generalizability of the findings. Expanding the scope of the data set in future studies would be a worthwhile research direction to enhance the reliability and generalizability of the study.

Future Directions

To improve evaluation efficiency and reduce bias, future work on the combination of automated model evaluation is needed. Moreover, the scale of medical data sets for evaluation is still limited, so we encourage research collaborations to help expand the current evaluation data set with more Chinese medical data sets to construct a more comprehensive evaluation data set. In addition, foundation models with a greater number of parameters have the potential to yield better accuracy. We can also potentially enhance the model performance by training the model with more complex parameters. Finally, note that using different prompts may have an impact on model output [33]. Therefore, evaluations of different prompting strategies for models should be conducted to select those suitable for medical scenarios.

Conclusions

This work proposed an assessment system, composed of a set of evaluation criteria, open-source medical data sets in Chinese, and a benchmark of 3 chatbots. Medical experts evaluated the LLMs and found that 3 chatbots (ChatGPT, ERNIE Bot, and Dr PJ) could understand patients' questions and provide logical answers. Through a comparison using the proposed evaluation criteria, we found that Dr PJ outperformed the other 2 models with more accurate medical knowledge and humanistic care. Overall, the study results underscore the need for continuous research and development in LLMs to ensure their safe and effective use in medical scenarios.

Acknowledgments

This research was supported by the Shanghai Artificial Intelligence Laboratory.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comments of detailed case reports.
[DOC File, 56 KB - medinform_v12i1e57674_app1.doc]

References

1. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR Med Inform 2022 Feb 10;10(2):e32875. [doi: 10.2196/32875] [Medline: 35142635]
2. Anders BA. Why ChatGPT is such a big deal for education. C2C Digital Magazine 2023;1(18) [FREE Full text]
3. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/index/chatgpt/ [accessed 2023-09-05]
4. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. medRxiv. Preprint posted online on Feb 1, 2023. [doi: 10.1101/2023.01.30.23285067] [Medline: 36778449]
5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
6. Hagendorff T, Fabi S, Kosinski M. Machine intuition: uncovering human-like intuitive decision-making in GPT-3.5. arXiv. Preprint posted online on Dec 10, 2022. [doi: 10.48550/arXiv.2212.05206]
7. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. Preprint posted online on Feb 21, 2023. [doi: 10.48550/arXiv.2302.11382]
8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023 Feb 9;2(2):e0000198. [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Balas M, Ing EB. Conversational AI models for ophthalmic diagnosis: comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. JFO Open Ophthalmol 2023 Mar;1:100005. [doi: 10.1016/j.jfop.2023.100005]
10. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. Preprint posted online on Feb 28, 2023. [doi: 10.21203/rs.3.rs-2566942/v1] [Medline: 36909565]
11. Hu T, Xu A, Liu Z, et al. Touch your heart: a tone-aware chatbot for customer care on social media. In: CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems: Association for Computing Machinery; 2018:1-12. [doi: 10.1145/3173574.3173989]
12. Liang H, Li H. Towards standard criteria for human evaluation of chatbots: a survey. arXiv. Preprint posted online on May 24, 2021. [doi: 10.48550/arXiv.2105.11197]
13. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023 Jul 17;13:1219326. [doi: 10.3389/fonc.2023.1219326] [Medline: 37529688]
14. Chaves AP, Gerosa MA. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. Int J Hum Comput 2021 May 9;37(8):729-758. [doi: 10.1080/10447318.2020.1841438]
15. Yang X, Li Y, Zhang X, Chen H, Cheng W. Exploring the limits of ChatGPT for query or aspect-based text summarization. arXiv. Preprint posted online on Feb 16, 2023. [doi: 10.48550/arXiv.2302.08081]
16. Tan Y, Min D, Li Y, et al. Evaluation of ChatGPT as a question answering system for answering complex questions. arXiv. Preprint posted online on Mar 14, 2023. [doi: 10.48550/arXiv.2303.07992]
17. West CG. AI and the FCI: can ChatGPT project an understanding of introductory physics? arXiv. Preprint posted online on Mar 26, 2023. [doi: 10.48550/arXiv.2303.01067]
18. Taylor E. We agree, don't we? The Delphi method for health environments research. HERD 2020 Jan;13(1):11-23. [doi: 10.1177/1937586719887709] [Medline: 31887097]
19. Swart ECS, Parekh N, Daw J, Manolis C, Good CB, Neilson LM. Using the Delphi method to identify meaningful and feasible outcomes for pharmaceutical value-based contracting. J Manag Care Spec Pharm 2020 Nov;26(11):1385-1389. [doi: 10.18553/jmcp.2020.26.11.1385] [Medline: 33119437]
20. Sala Benites H, Osmond P, Prasad D. A future-proof built environment through regenerative and circular lenses—Delphi approach for criteria selection. Sustainability 2022 Dec 29;15(1):616. [doi: 10.3390/su15010616]
21. King MR. The future of AI in medicine: a perspective from a chatbot. Ann Biomed Eng 2023 Feb;51(2):291-295. [doi: 10.1007/s10439-022-03121-w] [Medline: 36572824]

22. Sun Y, Wang S, Feng S, et al. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. arXiv. Preprint posted online on Dec 29, 2021. [doi: 10.48550/arXiv.2107.02137]
23. Côrte-Real N, Ruivo P, Oliveira T, Popovič A. Unlocking the drivers of big data analytics value in firms. J Bus Res 2019 Apr;97:160-173. [doi: 10.1016/j.jbusres.2018.12.072]
24. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv. Preprint posted online on Jan 10, 2023. [doi: 10.48550/arXiv.2201.11903]
25. Open-source question database for MedGPTEval. Google Docs. 2023. URL: https://qr02.cn/DBeS9U [accessed 2024-06-03]
26. Liu Y, Han T, Ma S, et al. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. arXiv. Preprint posted online on Aug 22, 2023. [doi: 10.48550/arXiv.2304.01852]
27. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
28. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:191-200. [Medline: 32477638]
29. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021 Jun 3;4(1):93. [doi: 10.1038/s41746-021-00464-x] [Medline: 34083689]
30. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv. Preprint posted online on Feb 26, 2023. [doi: 10.1101/2023.02.21.23285886] [Medline: 36865204]
31. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. JAMA 2018 Jan 2;319(1):19-20. [doi: 10.1001/jama.2017.19198] [Medline: 29261830]
32. Pham TM, Bui T, Mai L, Nguyen A. Out of order: how important is the sequential order of words in a sentence in natural language understanding tasks? arXiv. Preprint posted online on Jul 26, 2021. [doi: 10.48550/arXiv.2012.15180]
33. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023 Sep 30;55(9):1-35. [doi: 10.1145/3560815]

Abbreviations

AI: artificial intelligence
Dr PJ: Doctor PuJiang
LLM: large language model

Edited by C Lovis; submitted 25.02.24; peer-reviewed by K Chen, PF Chen; revised version received 03.04.24; accepted 04.05.24; published 28.06.24.

Please cite as:
Xu J, Lu L, Peng X, Pang J, Ding J, Yang L, Song H, Li K, Sun X, Zhang S
Data Set and Benchmark (MedGPTEval) to Evaluate Responses From Large Language Models in Medicine: Evaluation Development and Validation
JMIR Med Inform 2024;12:e57674
URL: https://medinform.jmir.org/2024/1/e57674
doi:10.2196/57674

Impact of Translation on Biomedical Information Extraction: Experiment on Real-Life Clinical Notes

Christel Gérardin[1], MSc, MD; Yuhan Xiong[1,2], MS; Perceval Wajsbürt[3], PhD; Fabrice Carrat[1,4], MD, PhD; Xavier Tannier[5], PhD

1

2

3

4

5

Corresponding Author:
Christel Gérardin, MSc, MD

Abstract

Background: Biomedical natural language processing tasks are best performed with English models, and translation tools have undergone major improvements. On the other hand, building annotated biomedical data sets remains a challenge.

Objective: The aim of our study is to determine whether the use of English tools to extract and normalize French medical concepts based on translations provides comparable performance to that of French models trained on a set of annotated French clinical notes.

Methods: We compared 2 methods: 1 involving French-language models and 1 involving English-language models. For the native French method, the named entity recognition and normalization steps were performed separately. For the translated English method, after the first translation step, we compared a 2-step method and a terminology-oriented method that performs extraction and normalization at the same time. We used French, English, and bilingual annotated data sets to evaluate all stages (named entity recognition, normalization, and translation) of our algorithms.

Results: The native French method outperformed the translated English method, with an overall F_1-score of 0.51 (95% CI 0.47-0.55), compared with 0.39 (95% CI 0.34-0.44) and 0.38 (95% CI 0.36-0.40) for the 2 English methods tested.

Conclusions: Despite recent improvements in translation models, there is a significant difference in performance between the 2 approaches in favor of the native French method, which is more effective on French medical texts, even with few annotated documents.

(JMIR Med Inform 2024;12:e49607) doi:10.2196/49607

KEYWORDS

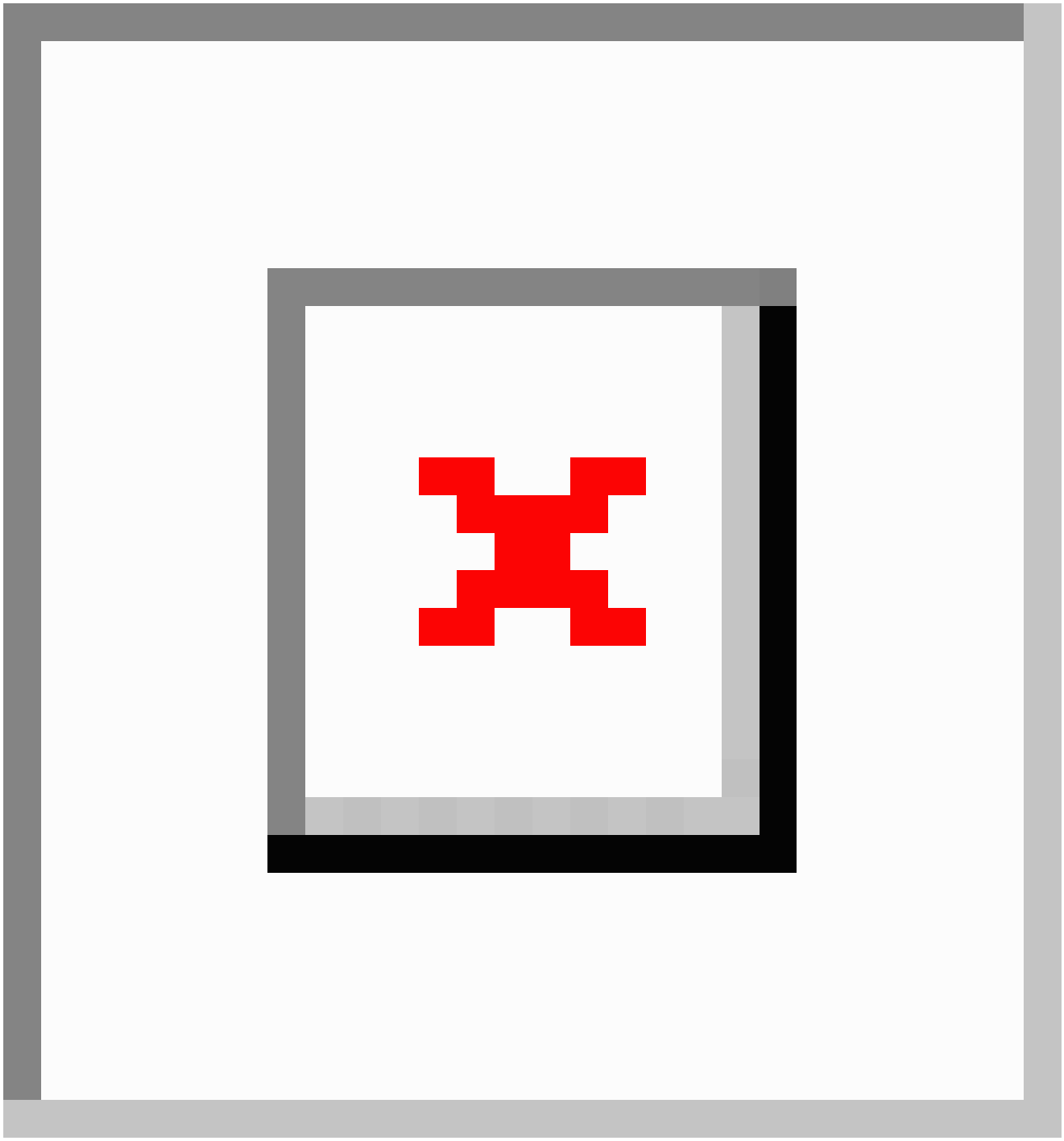
concept normalization; named entity recognition; natural language processing; translation; translational tool; biomedical data set; bilingual language model

Introduction

Named entity recognition (NER) and term normalization are important steps in biomedical natural language processing (NLP). NER is used to extract key information from textual medical reports, and normalization consists of matching a specific term to its formal reference in a shared terminology such as the United Medical Language System (UMLS) Metathesaurus [1]. Major improvements have been made recently in these areas, particularly for English, as a huge amount of data is available in the literature and resources. Modern automatic language processing relies heavily on pretrained language models, which enable efficient semantic representation of texts. The development of algorithms such as transformers [2,3] has led to significant progress in this field.

In Figure 1, the term “mention level” indicates that the analysis is carried out at the level of a word or small group of words: first at the NER stage (in blue) and then during normalization (in green); finally, all mentions with normalized concept unique identifiers (CUIs) are aggregated at the “document level” (orange part). The sets of aggregated CUIs per document predicted by the native French and translated English approaches are then compared to the manually annotated gold standard.

Figure 1. Overall objective of the method: translating plain text to the CUI codes of the UMLS Metathesaurus, document by document. CHEM: Chemicals & Drugs; CUI: concept unique identifier; DISO: Disorders; PROC: Procedures; UMLS: United Medical Language System.



In many languages other than English, efforts remain to be made to obtain such results, notably due to a much smaller quantity of accessible data [4]. In this context, our work explores the relevance of a translation step for the recognition and normalization of medical concepts in French biomedical documents. We compared 2 methods: (1) a native French approach where only annotated documents and resources in French are used and (2) a translation-based approach where documents are translated into English, in order to take advantage of existing tools and resources for this language that would allow the extraction of concepts mentioned in unpublished French texts without new training data (zero-shot), as proposed in van Mulligen et al [5].

We evaluated and discussed the results on several French biomedical corpora, including a new set of 42 annotated hospitalization reports with 4 entity groups. We evaluated the normalization task at the document level, in order to avoid a cross-language alignment step at evaluation time, which would add a potential level of error and thus make the results more difficult to interpret (see word alignment in Gao and Vogel [6] and Vogel et al [7]). This normalization was carried out by mapping all terms to their CUI in the UMLS Metathesaurus [1]. Figure 1 summarizes these various stages, from the raw French text and the translated English text to the aggregation and comparison of CUIs at the document level. Our code is available on GitHub [8].

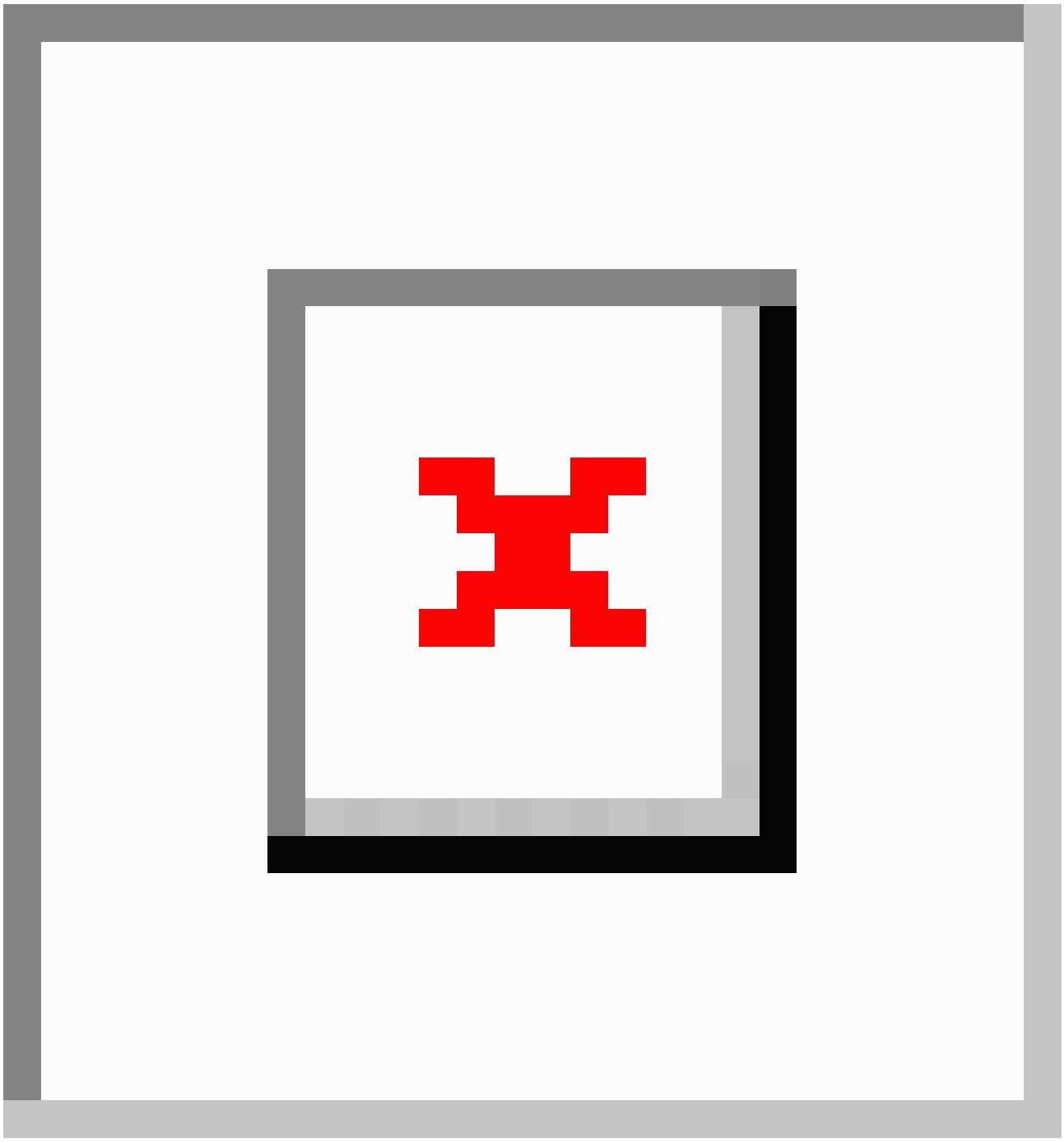
The various stages of our algorithms rely heavily on transformers language models [2]. These models currently represent the state of the art for many NLP tasks, such as machine translation, NER, classification, and text normalization (also known as entity binding). Once trained, these models can represent any specific language, such as biomedical or legal. The power of these models comes from their neural architecture but also largely depends on the amount of data they are trained on. In the biomedical field, 2 main types of data are available: public articles (eg PubMed) and clinical electronic medical record databases (eg MIMIC-III [9]), and the most powerful models are, for example, BioBERT [10], which has been trained on the whole of PubMed in English, and ClinicalBERT [11], which has been trained on PubMed and MIMIC-III. In French, the variety of models is less extensive, with CamemBERT [12] and FlauBERT [13] for the general domain and no specific model available for the biomedical domain.

In Figure 2, axis 1 (green axis on the left) corresponds to the native French branch with a NER step based on a FastText model trained from scratch on French clinical notes and a CamemBERT model. A multilingual Bidirectional Encoder Representations From Transformers (BERT) model was then used for the normalization step, with 2 models tested: a deep multilingual normalization model [14] and CODER [15] with the full version. Axes 2.1 and 2.2 (the 2 purple axes on the right) correspond to the translated English branches, with a first translation step performed by the OPUS-MT-FR-EN model [16] for both. Axis 2.1 (left) was conducted with decoupled NER and normalization steps: FastText trained from PubMed and MIMIC-III [17] for NER, and deep multilingual normalization [14] or CODER [15] with the English version for normalization. Axis 2.2 (right) used a single system for the NER and normalization stages: MedCAT [18].

In addition to particularly powerful English-language pretrained models, universal biomedical terminologies (ie, the UMLS Metathesaurus) also contain many more English terms than other languages. For example, the UMLS Metathesaurus [1] contains at least 10 times more English terms than French terms, which may enable rule-based models to perform better in English. As mentioned above, each reference concept in the UMLS Metathesaurus [1] is assigned a CUI, associated with a set of synonyms, possibly in several languages, and a semantic group, such as *Disorders*, *Chemicals & Drugs*, *Procedure*, *Anatomy*, etc.

In parallel, the performance of machine translation has also improved thanks to the same type of transformer-based language models, and recent years have seen the emergence of high-quality machine translations, such as OPUS-MT developed by Tiedemann et al [16], Google Translate, and others. These 2 observations have led several research teams to add a translation step in order to analyze medical texts, for example, to extract relevant mentions in ultrasound reports [19,20] or in the case of the standardization of medical concepts [14,15,21]. Work in the general (nonmedical) domain has also focused on alignment between named entities in parallel bilingual texts [22,23].

Figure 2. Diagram of different experiments comparing French and English language models without and with intermediate translation steps. CHEM: Chemicals & Drugs; CUI: concept unique identifier; DEVI: Devices; DISO: Disorders; EHR: electronic health record; EN: English; FR: French; FT: fine-tuned; PROC: Procedures; UMLS: United Medical Language System.



Methods

Approaches

Overview

Figure 2 shows the main approaches and models used in our study. We explored 1 "native French approach axis" (axis 1 in Figure 2), based on French linguistic models learned from and applied to French annotated data, and 2 "translated English approach axes" (axes 2.1 and 2.2), based on a translation step and concept extraction tools in English. We compared the performance of all axes with the average of the document-level CUI prediction precisions for all documents.

Native French Approach

Axis 1 consisted of 2 stages: a NER stage and a normalization stage. For the NER stage, we used the nested NER algorithm. Next, a normalization step was performed by 2 different algorithms: a deep multilingual normalization model [14] and CODER [15] with the *CODER all* version.

Translated-English Approach

First, axes 2.1 and 2.2 consisted of a translation step, performed by the state-of-the-art OPUS-MT-FR-EN [16] or Google

Translate algorithm. Second, similar to axis 1, axis 2.1 was based on a NER step and a normalization step. The NER step was performed by the same algorithm but trained on the National NLP Clinical Challenges (N2C2) 2019 data set [24] without manual annotation realignment; for the normalization step, we used the same deep multilingual algorithm [14] and the English version of CODER [15] based on a BioBERT model [10]. This axis allows us to compare 2 methods whose difference lies solely in the translation step.

Axis 2.2 was based on the MedCAT [18] algorithm, which performs NER and normalization simultaneously. In this case, we compared the native French method with a state-of-the-art, ready-to-use English system, which is not available in French.

Data Sets

For all our experiments, we chose to focus on 4 semantic groups of the UMLS Metathesaurus [1]: *Chemical & Drugs* ("CHEM"); *Devices* ("DEVI"), corresponding to medical devices such as pacemakers, catheters, etc; *Disorders* ("DISO"), corresponding to all signs, symptoms, results (eg, positive or negative results of biological tests), and diseases; and *Procedures* ("PROC"), corresponding to all diagnostic and therapeutic procedures such as imaging, biological tests, operative procedures, etc, as well as the corresponding number of documents.

Table 1 shows the data sets used for all our experiments and the corresponding number of documents. First, 2 French data sets were used for the final evaluation, as well as for training the axis-1 models. QUAERO is a freely available corpus [25] based on pharmacological notes with 2 subcorpora: MEDLINE (short sentences from PubMed abstracts) and EMEA (drug package inserts). We also annotated a new data set of real-life clinical notes from the Assistance Publique Hôpitaux de Paris data warehouse, described in Section S1 in Multimedia Appendix 1.

Table . Overview of all data sets used. When a data set is used for both training and testing, 80% of the data set is used for training and 20% is used for testing. Thus, for the EMEA data set, 30 documents were used for training and 8 for testing, 34 French notes were used for training and 8 for testing, and so on.

Variables	Languages and data sets						
	French			English		English and French	
	QUAERO [25]		French notes	N2C2[a] 2019 [24]	Mantra [26]	WMT[b] 2016 [27]	WMT 2019 [28]
	EMEA	MEDLINE					
Type	Drug notices	MEDLINE titles	French notes	English notes	Drug notices and MEDLINE titles	PubMed abstracts	PubMed abstracts
Size (documents), n	38	2514	42	100	200	>600,000 sent	6542
Use							
Train NER[c]	✓	✓	✓	✓			
Test NER	✓	✓	✓	✓			
Normalization	✓	✓	✓	✓			
Test MedCAT				✓	✓		
Translation (fine-tuning)						✓	✓
Translation (test)						✓	

[a]N2C2: National Natural Language Processing Clinical Challenges.

[b]WMT: Workshop on Machine Translation.

[c]NER: named entity recognition.

Second, we used the N2C2 2019 corpus [24] with annotated CUIs, on which we automatically added semantic group information from the UMLS Metathesaurus [1], to train the axis-2.1 system and evaluate the NER and English normalization algorithms. We also used the Mantra data set [26], a multilingual reference corpus for biomedical concept recognition.

Finally, we refined and tested the translation algorithms on the Workshop on Machine Translation biomedical corpora of 2016 [27] and 2019 [28]. A detailed description of the number of respective entities in the data sets can be found in Table S1 in Multimedia Appendix 1.

The annotation methods for the French corpus are detailed in Section S1 and Figure S1 in Multimedia Appendix 1. The distribution of entities for this annotation is detailed in Table S1 in Multimedia Appendix 1.

Translation

We used and compared 2 main algorithms for the translation step: the OPUS-MT-FR-EN model [16], which we tested without and with *fine-tuning* on the 2 biomedical translation corpora of 2016 and 2019 [27,28], and Google Translate as a comparison model.

NER Algorithm

For this step, we used the algorithm of Wajsbürt [29] described in Gérardin et al [30]. This model is based on the representation of a BERT transformer [3] and calculates the scores of all possible concepts to be predicted in the text. The extracted concepts are delimited by 3 values: start, end, and label. More precisely, the encoding of the text corresponds to the last 4 layers of BERT, FastText integration, and a max-pool Char-CNN [31] representation of the word. The decoding step is then performed by a 3-layer long short-term memory [32] with learning weights [33], similar to the method in Yu et al [34]. A sigmoid function was added to the vertex. Values (start, end, and label) with a score greater than 0.5 were retained for prediction. The loss function was a binary cross-entropy, and we used the Adam optimizer [35].

In our experiments, for the native French axis (axis 1 in Figure 2), the pretrained embeddings used to train the model were based on a FastText model [36], trained from scratch on 5 gigabytes of clinical text, and a CamemBERT-large model [12] *fine-tuned* on this same data set. For English axis 2.1, the pretrained models were BioWordVec [17] and ClinicalBERT [11].

Normalization Algorithms

Overview

This stage of our experiments was essential for comparing a method in native French and one translated into English, and it consisted of matching each mention extracted from the text to its associated CUI in the UMLS Metathesaurus [1]. We compared 3 models for this step, described below: the deep multilingual normalization algorithm developed by Wajsbürt et al [14]; CODER [15]; and the MedCAT [18] model, which performs both NER and normalization.

These 3 models require no training data set other than the UMLS Metathesaurus.

Deep Multilingual Normalization

This algorithm by Wajsbürt et al [14] considers the normalization task as a highly multiclass classification problem with cosine similarity and a softmax function as the last layer. The model is based on contextual integration, using the pretrained multilingual BERT model [3], and works in 2 steps. In the first step, the BERT model is fine-tuned and the French UMLS terms and their corresponding English synonyms are learned. Then, in the second step, the BERT model is frozen and the representation of all English-only terms (ie, those present only in English in the UMLS Metathesaurus [1]) is learned. The same training is used for the native French and translated English approaches. This model was trained with the 2021 version of the UMLS Metathesaurus [1], corresponding to the version used for annotating the French corpus. The model was thus trained on over 4 million concepts corresponding to 2 million CUIs.

CODER

The CODER algorithm [15] was developed by contrastive learning on the basis of the medical knowledge graph of the UMLS Metathesaurus [1], with concept similarities being calculated from the representation of terms and relations in this knowledge graph. Contrastive learning is used to learn embeddings through multisimilarity loss [37]. The authors have developed 2 versions: a multilingual version based on the multilingual BERT [3] and an English version based on the pretrained BioBERT model [10]. We used the multilingual version for axis 1 (native French approach) and the English version for axis 2.1. Both types of this model (*CODER all* and *CODER en*) were trained with the 2020 version of UMLS (publicly available models). *CODER all* [15] was trained on over 4 million concepts corresponding to 2 million CUIs, and *CODER en* was trained on over 3 million terms and 2 million CUIs.

For the deep multilingual model and the CODER model, in order to improve performance in terms of accuracy, we chose to add semantic group information (ie, *Chemical & Drugs*, *Devices*, *Disorders*, and *Procedures*) to the model output: that is, from the first *k* CUIs chosen from a mention, we selected the first from the corresponding group.

The MedCAT algorithm is described in detail in Section S1 in Multimedia Appendix 1.

Ethical Considerations

The study and its experimental protocol were approved by the Assistance Publique Hôpitaux de Paris Scientific and Ethical Committee (IRB00011591, decision CSE 20-0093). Patients were informed that their electronic health record information could be reused after an anonymization process, and those who objected to the reuse of their data were excluded. All methods were applied in accordance with the relevant guidelines (*Commission nationale de l'informatique et des libertés* reference methodology MR-004 [38]).

Results

The sections below present the performance results for each stage. The N2C2 2019 challenge corpus [24] enabled us to evaluate the performance of our English models on clinical data, and the Biomedical Translation 2016 shared task [27] allowed us to evaluate our translation performance on biomedical data with a BLEU score [39].

NER Performances

To be able to compare our approaches in native French and translated English, we used the same NER model, trained and tested on each of the data sets described above. Table 2 shows the corresponding results. Overall F_1-scores were similar across data sets: from 0.72 to 0.77.

Table . Named entity recognition (NER) performance for each model. For all experiments, we used the same NER algorithm but with different pretrained models. The best performance values are italicized.

Groups	Data sets and models								
	EMEA test, with FastText*[a] and Camem-BERT-FT [12]			French notes, with FastText* and Camem-BERT-FT			N2C2[b] 2019 test, with BioWordVec [17] and ClinicalBERT [11]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CHEM[c]	0.80	0.83	0.82	0.84	0.88	0.86	0.87	0.85	0.86
DEVI[d]	0.42	0.81	0.55	0.00	0.00	0.00	0.58	0.51	0.54
DISO[e]	0.54	0.63	0.59	0.67	0.65	0.66	0.74	0.72	0.73
PROC[f]	0.73	0.78	0.74	0.78	0.72	0.75	0.80	0.78	0.79
Overall	*0.71*	*0.77*	*0.74*	*0.73*	*0.71*	*0.72*	*0.78*	*0.76*	*0.77*

[a]FastText* corresponds to a FastText model [36] trained from scratch on our clinical data set.

[b]N2C2: National Natural Language Processing Clinical Challenges.

[c]CHEM: Chemical & Drugs.

[d]DEVI: Devices.

[e]DISO: Disorders.

[f]PROC: Procedures.

Normalization Performances

This section presents only the normalization performance based on the gold standard's entity mentions, without the intermediate steps. The results are summarized in Table 3. The deep multilingual algorithm performed better for all corpora tested, with an improvement in F_1-score from +0.6 to +0.11. By way of comparison, the winning team of the 2019 N2C2 had achieved an accuracy of 0.85 using the N2C2 data set directly to train their algorithm [24]. In our context of comparing algorithms between 2 languages, the normalization algorithms were not trained on data other than the UMLS Metathesaurus. MedCAT's performance (shown in Table S2 in Multimedia Appendix 1) cannot be directly compared with that of other models, as this method performed both NER and normalization in a single step. However, we note that this algorithm performed as well as axis 2.1 in terms of overall performance, as shown in Table 4.

Table . Performance of the normalization step. Model results were calculated from the annotated data sets, focusing on the 4 semantic groups of interest: *Chemical & Drugs*, *Devices*, *Disorders*, and *Procedures*. The best performance values are italicized.

Algorithms	Data set models		
	EMEA test	French notes	N2C2[a] 2019 test
Deep multilingual normalization	*0.65*	*0.57*	*0.74*
CODER all	0.58	0.51	—[b]
CODER en	—	—	0.63

[a]N2C2: National Natural Language Processing Clinical Challenges.

[b]Not applicable.

Table . Overall performances. The normalization step was performed by the deep multilingual model and the translation was performed by the OPUS-MT-FR-EN FT model. The best performance values are italicized.

Methods	EMEA test			French notes		
	Precision	Recall	F_1-score (95% CI)	Precision	Recall	F_1-score (95% CI)
Axis 1 (French NER[a]+normalization)	0.63	0.60	*0.61 (0.53-0.65)*	0.49	0.53	*0.51 (0.47-0.55)*
Axis 2.1 (Translation+NER+normalization)	0.53	0.40	0.45 (0.38-0.51)	0.41	0.38	0.39 (0.34-0.44)
Axis 2.2 (Translation+MedCAT [18])	0.53	0.46	0.49 (0.38-0.54)	0.38	0.38	0.38 (0.36-0.40)

[a]NER: named entity recognition.

Translation Performances

For both translation models, the respective BLEU scores [39] were calculated on the shared 2016 Biomedical Translation Task [27]. The chosen BLEU algorithm was the weighted geometric mean of the n-gram precisions per sentence.

A fine-tuned version of OPUS-MT-FR-EN [16] was also tested on the 2016 and 2019 Biomedical Translation shared tasks. For fine-tuning, we used the following hyperparameters: a maximum sequence length of 128 (mainly for computational memory reasons), a learning rate of 2×10^{-5}, and a weight decay of 0.01, and we varied the number of epochs up to 15 epochs (the error function curve stops decaying after 10 epochs). The Google Translate model could not be used for our clinical score experiments for reasons of confidentiality.

Table 5 presents the BLEU scores for the 3 models, showing that fine-tuning the OPUS-MT-FR-EN model [16] on biomedical data sets gave the best results, with a BLEU score [39] of 0.51. This was the model used to calculate the overall performance of axes 2.1 and 2.2.

Table . Translation performances: BLEU scores of the translation models. The best performance value is italicized.

Models	WMT[a] Biomed 2016 test
Google Translate	0.42
OPUS-MT-FR-EN	0.31
OPUS-MT-FR-EN FT[b]	*0.51*

[a]WMT: Workshop on Machine Translation.

[b]OPUS-MT-FR-EN FT corresponds to the OPUS-MT-FR-EN model [16] *fine-tuned* on biomedical translated corpus from the WMT Biomedical Translation Tasks in 2016 [27] and 2019 [28].

Overall Performances From Raw Text to CUI Predictions

This section presents the overall performance of the 3 axes, in an end-to-end pipeline. For axis 2, the results are those obtained with the best normalization algorithm (presented in Table 3). The model used for translation is the OPUS-MT-FR-EN [16] fine-tuned model. The results are presented in Table 4, with the best results obtained by the native French approach on the EMEA corpus [25] and French clinical notes. The 95% CIs were calculated using the empirical bootstrap method [40].

Discussion

Principal Findings

In this paper, we compared 2 approaches for extracting medical concepts from clinical notes: a French approach based on a French language model and a translated English approach, where we compared 2 state-of-the-art English biomedical language models, after a translation step. The main advantages of our experiment are that it is reproducible and that we were able to analyze the performance of each step of the algorithm: NER, normalization, and translation, and to test several models for each step.

The Quality of the Translation Is Not Sufficient

We showed that the native French approach outperformed the 2 translated English approaches, even with a small French training data set. This analysis confirms that, where possible, an annotated data set improves feature extraction. The evaluation of each intermediate step showed that the performance of each module was similar in French and English. We can therefore conclude that it is rather the translation phase itself that is of insufficient quality to allow the use of English as a proxy without a loss of performance. This is confirmed by the translation performance calculations, where the calculated BLEU scores were relatively low, although improved by a fine-tuning step.

In conclusion, although translation is commonly used for entity extraction or term normalization in languages other than English

[5,20,41-43], due to the availability of turnkey models that do not require additional annotation by a clinician, we showed that this induces a significant performance loss.

Commercial application programming interface–based translation services could not be used for our task due to data confidentiality issues. However, the OPUS-MT model is considered state of the art, it is adjustable to domain-specific data, and the translation results presented in Table 5 confirm the absence of performance difference between this model and the Google Translate model.

Although our experiments were carried out on a single language, the French-English pair is one of the best performers in recent translation benchmarks [16]. Other languages are unlikely to produce significantly better results.

Error Analysis

In these experiments, the overall results may appear low, but the task is still complex, especially because the UMLS Metathesaurus [1] contains many synonyms with different CUIs. To better understand this, we performed an error analysis on the normalization task only, as shown in Table S3 in Multimedia Appendix 1, with a physician's evaluation, on a sample of 100 errors for both models. We calculated that 24% (24/100) and 39% (39/100) of the terms found by the deep normalization algorithm [14] and CODER [15], respectively, were in fact synonyms but had 2 different UMLS CUIs. This highlights the difficulty of achieving normalization on the UMLS Metathesaurus. The UMLS Metathesaurus indeed groups together numerous terminologies whose mapping between terms is often imperfect, implying that certain synonyms, as shown here, do not have the same CUI, as pointed out by Cimino [44] and Jiménez-Ruiz et al [45]. For example, "cardiac ultrasound" has the CUI of C1655737, whereas "echocardiography" has another CUI of C0013516; similarly, "H/O: thromboembolism" has a CUI of C0455533, whereas "history of thromboembolism" has a CUI of C1997787, and so on.

Moreover, to be more precise, each axis had its own errors: overall, the errors in axis 2 were essentially due to the loss of information in translation. One notable error was literal translation: for example, "dispersed lupus erythematous" instead of "systemic lupus erythematosus," or "crepitant" instead of "crackles." This loss of translation led to more errors in the extraction of named entities.

In addition to the loss of translation information, axis 2.1 was also penalized by the NER step, due to the difference between the training set (N2C2 notes) and the test set (the translated French notes; the aim being to compare the performance of English-language turnkey models with the performance of French-language models from an annotated set). Axis 2.1, for example, omitted the names of certain drugs more often.

Finally, both axes were penalized by abbreviations. These were often badly translated (for example, the abbreviation "MFIU" for "mort foetale in utero," meaning "intrauterine fetal death," was not translated), which penalized axis 2. Nevertheless, if they were indeed extracted by NER steps in axis 1, they were not correctly normalized due to the absence of a corresponding CUI in the UMLS Metathesaurus.

Limitations

This work has several limitations. First, the actual French clinical notes contained very few terms in the *Devices* semantic group, which prevented the NER algorithm from finding them in the test data set. However, this drawback, which penalized the native French approach, still allowed us to draw a conclusion for the results. Furthermore, in this study, we did not take into account attributes of the extracted terms such as negation, hypothetical attribute, or belonging to a person other than the patient for comparison purposes, as the QUAERO [25] and N2C2 2019 [24] data sets did not have this labeled information.

Acknowledgments

The authors would like to thank the Assistance Publique Hôpitaux de Paris (AP-HP) data warehouse, which provided the data and the computing power to carry out this study under good conditions. We wish to thank all the medical colleges, including internal medicine, rheumatology, dermatology, nephrology, pneumology, hepato-gastroenterology, hematology, endocrinology, gynecology, infectiology, cardiology, oncology, emergency, and intensive care units, that gave their permission for the use of the clinical data.

Data Availability

The data sets analyzed as part of this study are not accessible to the public due to the confidentiality of data from patient files, even after deidentification. However, access to raw data from the Assistance Publique Hôpitaux de Paris (AP-HP) data warehouse can be granted by following the procedure described on its website [46]: by contacting the ethical and scientific committee at secretariat.cse@aphp.fr. Prior validation of access by the local institutional review committee is required. In the case of non-APHP researchers, a collaboration contract must also be signed.

Authors' Contributions

CG contributed to conceptualization, data curation, formal analysis, investigation, methodology, software, validation, original drafting, writing—original version, and writing—revision and editing the manuscript. YX contributed to investigation, methodology, software, and validation. PW contributed to investigation, software, and revision of the manuscript. FC contributed to conceptualization, methodology, project administration, supervision, writing—original version, and writing—revision and editing

of the manuscript. XT contributed to conceptualization, formal analysis, methodology, writing—original version, and writing—revision and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of the data sets, an example of the clinical notes annotation, French corpus annotation, MedCAT performances, and error analysis.
[DOCX File, 154 KB - medinform_v12i1e49607_app1.docx]

References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(suppl 1):D267-D270. [doi: 10.1093/nar/gkh061] [Medline: 14681409]
2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017) 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 2024-03-15]
3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers): Association for Computational Linguistics; 2019:4171-4186. [doi: 10.18653/v1/N19-1423]
4. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12. [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
5. van Mulligen EM, Afzal Z, Akhondi SA, Vo D, Kors JA. Erasmus MC at CLEF Ehealth 2016: concept recognition and coding in French texts. In: Balog K, Cappellato L, Ferro N, Macdonald C, editors. Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum CEUR Workshop Proceedings, Vol 1609: CEUR-WS.org; 2016:171-178 URL: https://ceur-ws.org/Vol-1609/16090171.pdf [accessed 2024-03-15]
6. Gao Q, Vogel S. Parallel Implementations of word alignment tool. In: Cohen KB, Carpenter B, editors. SETQA-NLP '08: Software Engineering, Testing, and Quality Assurance for Natural Language Processing: Association for Computational Linguistics; 2008:49-57. [doi: 10.5555/1622110.1622119]
7. Vogel S, Ney H, Tillmann C. HMM-based word alignment in statistical translation. In: COLING '96: Proceedings of the 16th Conference on Computational Linguistics - Volume 2: Association for Computational Linguistics; 1996:836-841. [doi: 10.3115/993268.993313]
8. ChristelDG/biomed_translation. GitHub. URL: https://github.com/ChristelDG/biomed_translation [accessed 2024-03-15]
9. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
10. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
11. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
12. Martin L, Muller B, Ortiz Suárez PJ, et al. CamemBERT: a tasty French language model. In: Kurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2020:7203-7219. [doi: 10.18653/v1/2020.acl-main.645]
13. Le H, Vial L, Frej J, et al. FlauBERT: unsupervised language model pre-training for French. In: Calzolari N, Béchet F, Blanche P, et al, editors. Proceedings of the Twelfth Language Resources and Evaluation Conference: European Language Resources Association; 2020:2479-2490 URL: https://aclanthology.org/2020.lrec-1.302 [accessed 2024-03-15]
14. Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform 2021 Feb;114:103684. [doi: 10.1016/j.jbi.2021.103684] [Medline: 33450387]
15. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform 2022 Feb;126:103983. [doi: 10.1016/j.jbi.2021.103983] [Medline: 34990838]
16. Tiedemann J, Thottingal S. OPUS-MT - building open translation services for the world. In: Martins A, Moniz H, Fumega S, et al, editors. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation: European Association for Machine Translation; 2020:479-480 URL: https://aclanthology.org/2020.eamt-1.61 [accessed 2024-03-15]
17. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52. [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
18. Kraljevic Z, Bean D, Mascio A, et al. MedCAT -- medical concept annotation tool. arXiv. Preprint posted online on Dec 18, 2019. [doi: 10.48550/arXiv.1912.10166]

19. Campos L, Pedro V, Couto F. Impact of translation on named-entity recognition in radiology texts. Database (Oxford) 2017 Jan 1;2017(2017):bax064. [doi: 10.1093/database/bax064] [Medline: 29220455]
20. Suarez-Paniagua V, Dong H, Casey A. A multi-BERT hybrid system for named entity recognition in Spanish radiology reports. In: Faggioli G, Ferro N, Joly A, Maistro M, Piroi F, editors. Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, Vol 2936: CEUR-WS.org; 2021:846-856 URL: https://ceur-ws.org/Vol-2936/paper-70.pdf [accessed 2024-03-15]
21. Perez-Miguel N, Cuadros M, Rigau G. Biomedical term normalization of EHRs with UMLS. In: Calzolari N, Choukri K, Cieri C, et al, editors. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018): European Language Resources Association (ELRA); 2018:2045-2051 URL: https://aclanthology.org/L18-1322 [accessed 2024-03-15]
22. Chen Y, Zong C, Su KYS. On jointly recognizing and aligning bilingual named entities. In: Hajič J, Carberry S, Clark S, Nivre J, editors. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2010:631-639 URL: https://aclanthology.org/P10-1065 [accessed 2024-03-15]
23. Chen Y, Zong C, Su KYS. A joint model to identify and align bilingual named entities. Comput Linguist 2013 Jun 1;39(2):229-266. [doi: 10.1162/COLI_a_00122]
24. Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural Language Processing (NLP) Clinical Challenges (N2C2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc 2020 Oct 1;27(10):1529-1537. [doi: 10.1093/jamia/ocaa106] [Medline: 32968800]
25. Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French medical corpus: a resource for medical entity recognition and normalization. Presented at: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing - BioTextM2014; May 26-31, 2014; Reykjavik, Iceland p. 24-30 URL: https://perso.limsi.fr/pz/FTPapiers/Neveol_BIOTEXTM2014.pdf [accessed 2024-03-15]
26. Kors JA, Clematide S, Akhondi SA, van Mulligen EM, Rebholz-Schuhmann D. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. J Am Med Inform Assoc 2015 Sep;22(5):948-956. [doi: 10.1093/jamia/ocv037] [Medline: 25948699]
27. Bojar O, Chatterjee R, Federmann C. Findings of the 2016 Conference on Machine Translation. In: Bojar O, Buck C, Chatterjee R, et al, editors. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers: Association for Computational Linguistics; 2016:131-198. [doi: 10.18653/v1/W16-2301]
28. Bawden R, Bretonnel Cohen K, Grozea C, et al. Findings of the WMT 2019 Biomedical Translation Shared Task: evaluation for MEDLINE abstracts and biomedical terminologies. In: Bojar O, Chatterjee R, Federmann C, et al, editors. Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2): Association for Computational Linguistics; 2019:29-53. [doi: 10.18653/v1/W19-5403]
29. Wajsbürt P. Extraction and Normalization of Simple and Structured Entities in Medical Documents [thesis].: Sorbonne Université; 2021 Dec URL: https://theses.hal.science/THESES-SU/tel-03624928v1 [accessed 2024-03-15]
30. Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artif Intell Med 2022 Jun;128:102311. [doi: 10.1016/j.artmed.2022.102311] [Medline: 35534148]
31. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Knight K, Nenkova A, Rambow O, editors. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Association for Computational Linguistics; 2016:260-270. [doi: 10.18653/v1/N16-1030]
32. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
33. Kim J, El-Khamy M, Lee J. Residual LSTM: design of a deep recurrent architecture for distant speech recognition. 2017 Presented at: Interspeech 2017; Aug 20-24, 2017; Stockholm, Sweden p. 1591-1595. [doi: 10.21437/Interspeech.2017-477]
34. Yu J, Bohnet B, Poesio M. Named entity recognition as dependency parsing. In: Jurafsky D, Chai J, Schulter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2020:6470-6476. [doi: 10.18653/v1/2020.acl-main.577]
35. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Dec 22, 2014. [doi: 10.48550/arXiv.1412.6980]
36. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017 Dec 1;5:135-146. [doi: 10.1162/tacl_a_00051]
37. Wang X, Han X, Huang W, Dong D, Scott MR. Multi-similarity loss with general pair weighting for deep metric learning. 2019 Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 15-20, 2019; Long Beach, CA p. 5017-5025. [doi: 10.1109/CVPR.2019.00516]
38. CNIL (Commission Nationale de l'Informatique et des Libertés). URL: https://www.cnil.fr/en/home [accessed 2024-03-15]
39. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics: Association for Computational Linguistics; 2002:311-318. [doi: 10.3115/1073083.1073135]

40. Dekking FM, Kraaikamp C, Lopuhaa HP, Meester LE. A Modern Introduction to Probability and Statistics: Understanding Why and How: Springer Nature; 2007.
41. Cotik V, Rodríguez H, Vivaldi J. Spanish named entity recognition in the biomedical domain. In: Lossio-Ventura J, Muñante D, Alatrista-Salas H, editors. Information Management and Big Data. SIMBig 2018. Communications in Computer and Information Science, vol 898: Springer:233-248. [doi: 10.1007/978-3-030-11680-4]
42. Hellrich J, Hahn U. Enhancing multilingual biomedical terminologies via machine translation from parallel corpora. In: Métais E, Roche M, Teisseire M, editors. Natural Language Processing and Information Systems. NLDB 2014. Lecture Notes in Computer Science, vol 8455: Springer; 2014:9-20. [doi: 10.1007/978-3-319-07983-7_2]
43. Attardi G, Buzzelli A, Sartiano D. Machine translation for entity recognition across languages in BIOMEDICAL documents. In: Forner P, Navigli R, Tufis D, Ferro N, editors. Working Notes for CLEF 2013 Conference. CEUR Workshop Proceedings, Vol 1179: CEUR-WS.org; 2013. URL: https://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFER-AttardiEt2013.pdf [accessed 2024-03-15]
44. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51. [doi: 10.1136/jamia.1998.0050041] [Medline: 9452984]
45. Jiménez-Ruiz E, Grau BC, Horrocks I, Berlanga R. Logic-based assessment of the compatibility of UMLS ontology sources. J Biomed Semantics 2011 Mar 7;2 Suppl 1(Suppl 1):S2. [doi: 10.1186/2041-1480-2-S1-S2] [Medline: 21388571]
46. Assistance Publique Hôpitaux de Paris. URL: www.eds.aphp.fr [accessed 2024-03-18]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
CUI: concept unique identifier
N2C2: National Natural Language Processing Clinical Challenges
NER: named entity recognition
NLP: natural language processing
UMLS: United Medical Language System

Edited by C Lovis; submitted 03.06.23; peer-reviewed by L Modersohn, M Torii; revised version received 07.01.24; accepted 10.01.24; published 04.04.24.

Please cite as:
Gérardin C, Xiong Y, Wajsbürt P, Carrat F, Tannier X
Impact of Translation on Biomedical Information Extraction: Experiment on Real-Life Clinical Notes
JMIR Med Inform 2024;12:e49607
URL: https://medinform.jmir.org/2024/1/e49607
doi:10.2196/49607

Original Paper

Additional Value From Free-Text Diagnoses in Electronic Health Records: Hybrid Dictionary and Machine Learning Classification Study

Tarun Mehra[1], MD; Tobias Wekhof[2], PhD; Dagmar Iris Keller[3,4], MD

[1]Department for Medical Oncology and Hematology, University Hospital of Zurich, Zurich, Switzerland

[2]Center of Economic Research, ETH Zurich, Zurich, Switzerland

[3]Faculty of Medicine, University of Zurich, Zurich, Switzerland

[4]Emergency Department, University Hospital of Zurich, Zurich, Switzerland

Corresponding Author:
Tarun Mehra, MD
Department for Medical Oncology and Hematology
University Hospital of Zurich
Rämistrasse 100
Zurich, 8091
Switzerland
Phone: 41 44255 ext 1111
Email: tarun.mehra@usz.ch

Abstract

Background: Physicians are hesitant to forgo the opportunity of entering unstructured clinical notes for structured data entry in electronic health records. Does free text increase informational value in comparison with structured data?

Objective: This study aims to compare information from unstructured text-based chief complaints harvested and processed by a natural language processing (NLP) algorithm with clinician-entered structured diagnoses in terms of their potential utility for automated improvement of patient workflows.

Methods: Electronic health records of 293,298 patient visits at the emergency department of a Swiss university hospital from January 2014 to October 2021 were analyzed. Using emergency department overcrowding as a case in point, we compared supervised NLP-based keyword dictionaries of symptom clusters from unstructured clinical notes and clinician-entered chief complaints from a structured drop-down menu with the following 2 outcomes: hospitalization and high Emergency Severity Index (ESI) score.

Results: Of 12 symptom clusters, the NLP cluster was substantial in predicting hospitalization in 11 (92%) clusters; 8 (67%) clusters remained significant even after controlling for the cluster of clinician-determined chief complaints in the model. All 12 NLP symptom clusters were significant in predicting a low ESI score, of which 9 (75%) remained significant when controlling for clinician-determined chief complaints. The correlation between NLP clusters and chief complaints was low (r=−0.04 to 0.6), indicating complementarity of information.

Conclusions: The NLP-derived features and clinicians' knowledge were complementary in explaining patient outcome heterogeneity. They can provide an efficient approach to patient flow management, for example, in an emergency medicine setting. We further demonstrated the feasibility of creating extensive and precise keyword dictionaries with NLP by medical experts without requiring programming knowledge. Using the dictionary, we could classify short and unstructured clinical texts into diagnostic categories defined by the clinician.

(JMIR Med Inform 2024;12:e49007) doi:10.2196/49007

KEYWORDS

electronic health records; free text; natural language processing; NLP; artificial intelligence; AI

Introduction

Organizational challenges, such as overcrowding in emergency departments (EDs), directly impact patient outcomes. The digitization of health records offers an opportunity to integrate artificial intelligence (AI) into patient management. However, health care workers often prefer to write unstructured text rather than entering structured data [1,2]. This raises the question of how future electronic health records (EHRs) should be designed: what additional value does free text provide?

We propose adding an additional dimension alongside the classic predictive task performed with text—inference to infer characteristics from text entries. Most studies using text analysis with patient records show promising results in predicting patient outcomes, such as in-hospital mortality, unplanned re-admission after 30 days, and prolonged length of hospital stay [3,4]. The benefits of unstructured text in EHRs for the improvement of prediction models have been demonstrated, as underscored by the extensive review by Seinen et al [5]. Indeed, 20% of the trials that were reported were conducted within a hospital ED environment. However, the analysis of the reported studies focused on demonstrating an improvement in predicting clinical outcomes, such as death or rehospitalization. We extend this approach by using the text not primarily to predict outcomes but to explain the correlation of patient subgroups with clinical outcomes. For instance, we show if certain symptoms documented in the ED triage are associated with a higher probability of an inpatient stay. Our results indicate that the information captured by clinical text-based notes is complementary to traditional structured data and can provide clinicians with valuable information about patients.

Overcrowding in the ED is an important case in point where AI supporting the optimization of patient workflows may substantially improve outcomes. It is a recognized challenge facing many EDs worldwide [6,7], adversely impacting patient outcomes [8]. These negative effects are evident during ED resource overload, such as during the COVID-19 pandemic [9]. More recently, senior public health officials in England have attributed up to 500 excess deaths per week during the recent winter months to delays caused by National Health Service capacity constraints [10,11]. Therefore, electronically enabled targeted patient selection could help speed up triage and reduce ED overcrowding. However, the optimal structure of EHRs remains controversial, particularly because clinicians tend to prefer the flexibility of entering unstructured text to structured data entry [12].

By comparing data extracted from 2 fields—1 derived from a structured drop-down menu indicating leading symptoms for ED admission and the other containing unstructured text—we can demonstrate that free text contains additional information beyond structured data and that these 2 types of data complement each other. With our semisupervised topic allocation method, we demonstrate the ability to capture more comprehensive information about a patient's symptom cluster compared with relying solely on a manually attributed single chief complaint. Moreover, we present a transparent approach for extracting topics from short clinical texts based on natural language processing (NLP)–supported annotated clinical libraries, which can be fed into predictive models. In addition to being transparent, our method is language independent and easy to implement for clinical researchers (although the dictionaries we constructed are in German, researchers can easily use our method to construct their own topic dictionaries in any language).

Our approach is based on constructing a dictionary with keywords that define a topic. In contrast to dictionary approaches, unsupervised topic models, such as the latent Dirichlet allocation [13], are often used. However, finding topics in short-text samples using these models is challenging [14]. Moreover, unsupervised models might not capture topics that are of interest to the researcher because these models differentiate between topics based on their statistical difference. For instance, it could be that latent Dirichlet allocation defines topics based on words about the age and gender of the patients because these are the most distinctive features. However, the researcher may be interested in the diagnosis, which is more challenging to classify.

In contrast, supervised machine learning methods require creating a manually classified training data set. The algorithm learns how to classify future data into topics based on the training set. When dealing with a high volume of topics, both human classification and the algorithm's training run the risk of creating noise. Similarly, regression approaches for supervised classifications are not suitable for many topics. Therefore, we chose a dictionary approach based on keywords. To facilitate the selection of the keywords, we developed a preselection of words based on a measure of their semantic similarity. As our presorting of words uses word embeddings, we consider our approach as a hybrid between dictionary- and machine learning–based approaches [15].

Our approach, combined with clinical notes, allow us to address 2 questions:

- What additional information does the free text provide on the patient being admitted compared with the suspected diagnosis from the drop-down menu?
- Could this additional information be useful for clinical or organizational purposes?

Methods

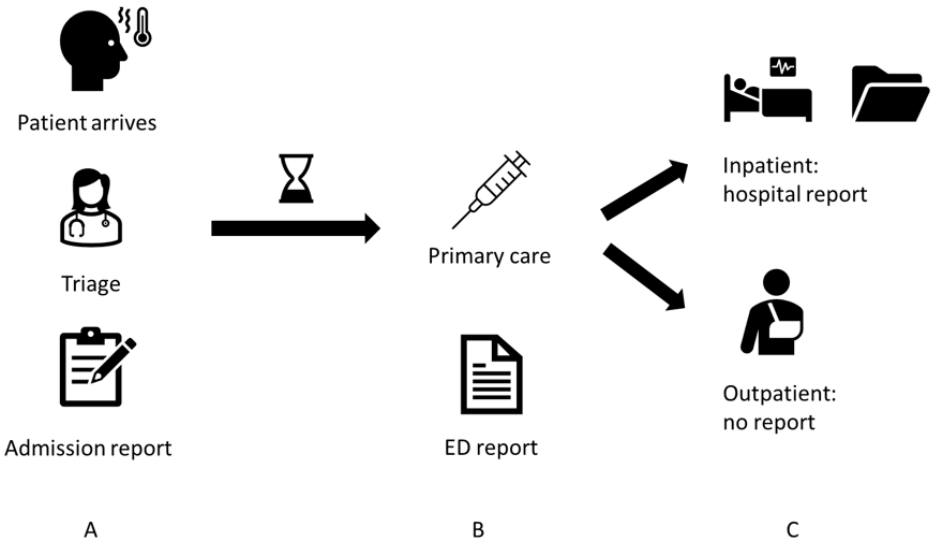
Data

We used data from the ED's admission report. Figure 1 provides a contextual representation of this data type in relation to patient flow and other documents associated with patients. In step 1, patients present themselves at the ED and are admitted in the system. A medical professional conducts the triage by quickly assessing the main symptoms and their severity using the Emergency Severity Index (ESI) score, resulting in an admission report. This report is for the internal patient management within the ED and contains basic patient information (age, gender, and so on) along with the chief complaints and symptoms.

After a waiting time (which depends on the triage score), the patient receives primary care from a medical professional, which

is documented in the ED report. The ED report summarizes the patient's entire stay at the ED and is issued at the end of the patient care from the ED. In the third step, the patient is either discharged into ambulatory care (which does not create any further documents) or is transferred to inpatient care, which results in the classic medical records.

Figure 1. Patient flow in emergency department (ED) and associated reports.



For our analysis, we used the first type of document: the internal ED admission report. Unlike the other types of documents, this report is issued before treatment and provides an opportunity to manage patient flow. Although the ED report from step 2 could also be used for inpatient management, this proves challenging in practice because inpatient care is very heterogeneous and depends on many factors, including different organizational structures in every hospital department. In contrast, the ED admission reports can be used for homogeneous organization within the ED.

Our initial data set contained 293,298 patient visits to the ED of the University Hospital of Zurich, Switzerland, from January 1, 2014, to October 31, 2021 (in German; received in the Excel [Microsoft Corporation] format). For each visit, the data set includes a short text from the triage with the patient's symptoms, along with our 2 outcomes of interest (triage score "ESI," which we further explain below, and type of discharge), basic patient characteristics (patient visit pseudo ID, age, gender, admission type [self, ambulance, or police], and admission reason [accident or illness]), ED organizational variables (average number of patients in ED; average patient waiting time; night, late, or early shift; and treating ED team [internal medicine, surgery, neurology, neurosurgery, or psychiatry]), and the visit's time stamp. The summary statistics of these variables are presented in Table 1.

After excluding cases with no records in the string variable "suspected diagnosis" on admission on which NLP analysis was to be performed, the data set comprised 256,329 (87.4%) of the initial data set of 293,298 patient visits. We only used 2019 to 2021 for comparison as these visits had a recorded chief complaint, reducing the data set to the final sample of 52,222 patient visits. Patients directly admitted to the shock room (ie, ESI score=1) were not considered in our analysis, as no additional triage was performed upon admission. The data structure of our analysis is summarized in Figure 2, and the recorded variables are presented in Textbox 1.

The ESI is an internationally established 5-level triage algorithm widely used in EDs and is based on the acuity as well as the resource intensity of anticipated emergency care, with level 1 denoting acute life-threatening conditions, such as massive trauma warranting immediate, life-saving care, and level 5 denoting non–time-critical conditions of low complexity [13]. Cases triaged as ESI 4 or 5 (approximately 16% of patients) are usually fast-tracked to specialized treatment rooms because the medical resources required to treat these patients are low, and thus, they can be managed in parallel by a dedicated team, which reduces ED congestion. ESI 2 or 3 typically require a more thorough workup. Hence, for the outcome variable "low ESI," we decided to set the cutoff at ESI<4, that is, patients with "low ESI" had been triaged with a score of 2 or 3. Furthermore, the data set included free-text fields (strings), namely, the suspected diagnosis at admission and the diagnosis at discharge.

In the admission process, the clinician performing triage records the patient's symptoms in written form in 2 to 3 sentences. The purpose of this free text is to preregister the patient in the ED and enable all team members to become aware of the impending clinical problems. To our knowledge, all the larger EDs in German-speaking countries with full EHR note the reason for admission in the form of a short, unstructured text upon notification of a pending ED admission.

From May 28, 2019, onward, the symptoms were additionally recorded as so-called chief complaints from a drop-down menu (ordinal variable). The difference between the free text and the chief complaint was that the chief complaint was a fixed category selected from a drop-down menu and was primarily intended to serve administrative and statistical purposes, that is, to allow for post hoc analysis of the patient composition of the ED.

During the entire study period, the list of chief complaints (n=99) varied over time or contained doublets, which we grouped into 58 symptom topics. For patient visits with a

selected chief complaint from the drop-down option "Diverse," it was unclear if a leading symptom had been attributed at triage; hence, we did not include them in the list of chief complaints (referred to as lead symptoms [LS]). Furthermore, we grouped 5 chief complaints with very low occurrences, such as "drowning accident" or "flu vaccine," into our class "diverse." However, we did not use this group in further analysis because of the heterogeneity of the symptoms included. The lead symptom topics were then aggregated into 12 clusters by the authors according to clinical judgment. The complete list of LS can be found in Table S1 in Multimedia Appendix 1.

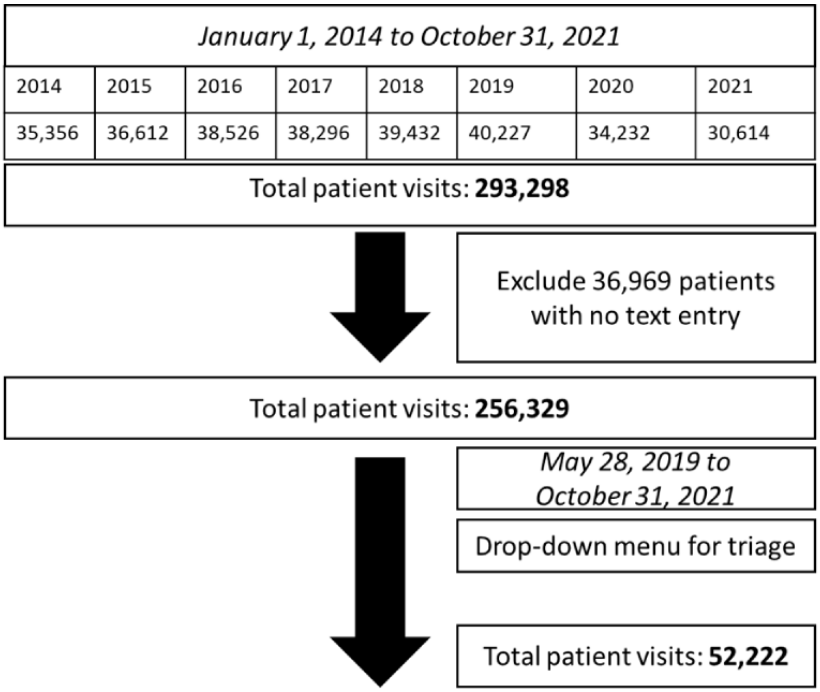
A total of 65 variables from 2014 to 2018 and 69 variables from 2019 to 2021 (including the chief complaint) were recorded in the initial data set. A total of 65 variables from 2014 to 2018 were constant throughout 2014 to 2021 and were retained for preprocessing. The final data table used for the analysis contained the variables listed in Table 1, in addition to the patient ID, year and weekday of the consultation derived from the admission time stamp, the treating ED team (internal medicine, surgery, neurology, or psychiatry), as well as the LS clusters from the drop-down menu and the NLP-extracted topic clusters that were obtained from the field "suspected diagnoses," discussed in detail in *Analysis: Topic Allocation* section. In addition, the table contained the outcomes "inpatient" and "ESI score<4" as binary variables. Two further outcomes were considered, namely, readmission within 30 days and waiting time>30 minutes, but were discarded owing to doubts regarding the quality and consistency of the entered data. We retained the outcomes "inpatient" and "ESI score<4" owing to their direct association with the immediacy of the outcome in the patient pathway within the ED, ensuring robust data quality.

Table 1. Summary statistics of the patient population (n=52,222)[a].

Variable	Values
Age (y), mean (SD)	46.5 (19.7)
Female, n (%)	23,782 (45.54)
Emergency Severity Index score (out of 5), mean (SD)	3.3 (0.6)
Fast track, n (%)	8264 (15.82)
Number of patients in the emergency department, mean (SD)	19.8 (8.3)
Early shift, n (%)	21,644 (41.45)
With emergency medical service, n (%)	9020 (17.27)
With police, n (%)	188 (0.36)
Accident, n (%)	16,845 (32.26)
Inpatient, n (%)	14,112 (27.02)
Night shift, n (%)	7915 (15.16)
Late shift, n (%)	22,663 (43.4)

[a]The total sample contains patient visits for the period from May 28, 2019, to October 31, 2021.

Figure 2. Data structure.



Textbox 1. Variables recorded for our analysis.

Triage

- Suspected diagnosis (free text) and Emergency Severity Index score

Type of discharge

- Hospitalization, ambulatory treatment, or patient has run away

Patient characteristics

- Patient visit pseudo ID, age, gender, admission type (self, ambulance, or police), and admission reason (accident or illness)

Organizational

- Average number of patients in emergency department (ED); average patient waiting time; night, late, or early shift; and treating ED team (internal medicine, surgery, neurology, or psychiatry)

Time

- Time stamp

Analysis: Topic Allocation

We selected the field "suspected diagnosis" to extract the symptoms or complaints that led to ED admission according to the oral report received by the ED physician in charge, as mentioned previously. This field comprises a short-text string entered by the ED physician upon receiving information about the patient's expected arrival at the ED. This information can be transmitted to the ED physician by a referring physician or ambulance well in advance of a patient's arrival. The text is entered before the patient triage is performed by the triage ED nurse. As a clinical note, the physician's text entry is part of the EHR. The information contained in the string "suspected diagnoses" is supposed to be similar to the selected chief complaint from the drop-down menu "lead symptom." Indeed, the latter variable was added later (in 2019) to facilitate the administrative analysis of causes for ED admission, as an analysis using unstructured text was not possible by the hospital administration. Both fields are supposed to contain the medical reason, or chief complaint, leading to ED admission.

We constructed a measure of the semantic distance of all words in the corpus by training a word embedding. Word embeddings are matrices in which each column represents a word and its relative distance to other words (eg, the distance between blood and red is smaller than that between blood and green). Hence, it is possible to find the most similar words for a given keyword using the smallest distance measured with the cosine similarity. To train the word embedding, we used word2vec with the entire text corpus and the continuous bag-of-words algorithm from the Python library Gensim [16], with an embedding size of 300 computed with 100 epochs.

To construct our topic dictionaries, we proceeded in 4 steps, as shown in Figure 3. First, we manually defined topics and selected between 2 and 20 initial seed words (henceforth "keywords") by reading some of the texts and using prior medical knowledge. A smaller number of keywords were used for the design of the topic "infection" (n=1). A larger number of initial keywords were used for the design of the topics "intoxication" (n=40) and "skin" (n=28). In step 2, we then searched for up to 50 of the semantically closest words for each initial list. With the help of the word embedding, it is possible to search for the words that maximize the cosine similarity for the seed keywords. In addition, we only considered keywords that occurred at least 10 times. This list of similar words allowed us to efficiently increase the dictionary for each topic. In step 3, we manually chose words from the preselection of similar words to the seed word, resulting in a separate dictionary per topic (step 4). In some instances, the dictionary used combinations of words. For instance, the topic "chest pain" was allocated to combinations of words such as "pain" or "pressure" with the words "chest" or "thorax."

This table presents the distribution of the diagnosis topics obtained with the NLP-based text annotation before and after the spherical feature annotation. The total number of cases was 52,222, and 20.38% could not be attributed with a diagnosis topic.

The summary of the increase in tags per topic cluster through the NLP-based expansion of our topics library is presented in Table 2. The first column shows the percentage of the sample tagged with a topic using the original keyword approach. The proportion of clinical topics ranged from 0.72% for COVID-19 to 31.6% for trauma-related visits. It should be noted that patient visits can be allocated with multiple topics. The next column shows the share of visits with the spherically increased dictionary, with the percentage increase in topic shares in the last column. Overall, the spherical dictionary enhancement decreased the number of nontagged visits by nearly 25%, from 27.08% of the sample to 20.24%. For the individual topics, the additional keywords increased their share, ranging from 5.29% for trauma to 286.35% for general administrative visits.

In the second procedure, we automatically increased the number of keywords for each topic dictionary. This process is shown in Figure 4, which can be imagined as constructing a multidimensional sphere using the initial keywords. The additional keywords were then located within that sphere.

The "spherical" dictionary enhancement consists of the following steps:

- Compute all distances between the keywords and retain the largest distance (ie, the distance between the 2 least similar words). For each keyword, this distance is the radius of a circle in the embedding space (steps 1 and 2).
- For each of the initial keywords, identify the n-closest words (not in the topic dictionary) using the cosine similarity (step 3).
- Retain these additional words if their distance to all other initial keywords is smaller than the maximum distance computed in the first step, that is, if the new words are in the intersection of all circles (step 4).

Using the abovementioned approach, we could tag 79.76% (41,653/52,222) of the final sample. The remaining texts could not be tagged because they either belonged to small topics that we did not define or because these texts did not contain words that are present in the dictionary.

Once the dictionaries for each topic are constructed, they can be used for additional patient visits and for similar data sets, which makes the approach easily scalable.

Figure 3. Topic dictionaries with semimanual keyword selection. (A) The researcher selects an initial seed word for a topic. (B) Using word embeddings, a list of semantically similar words from the corpus is generated. (C) The researcher manually selects words that are associated with the topic. (D) The topic dictionary is created.

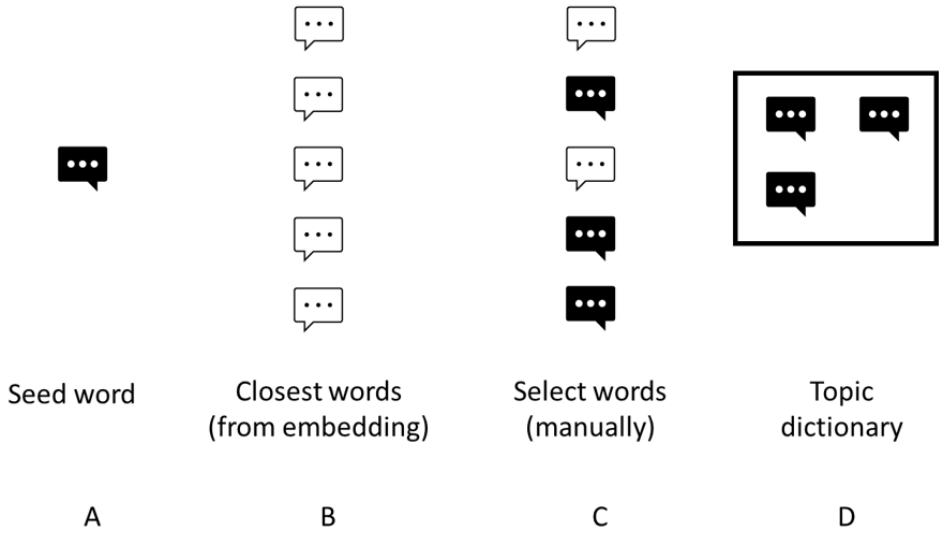


Table 2. Spherical feature annotation and increase in topic share (n=52,222)[a].

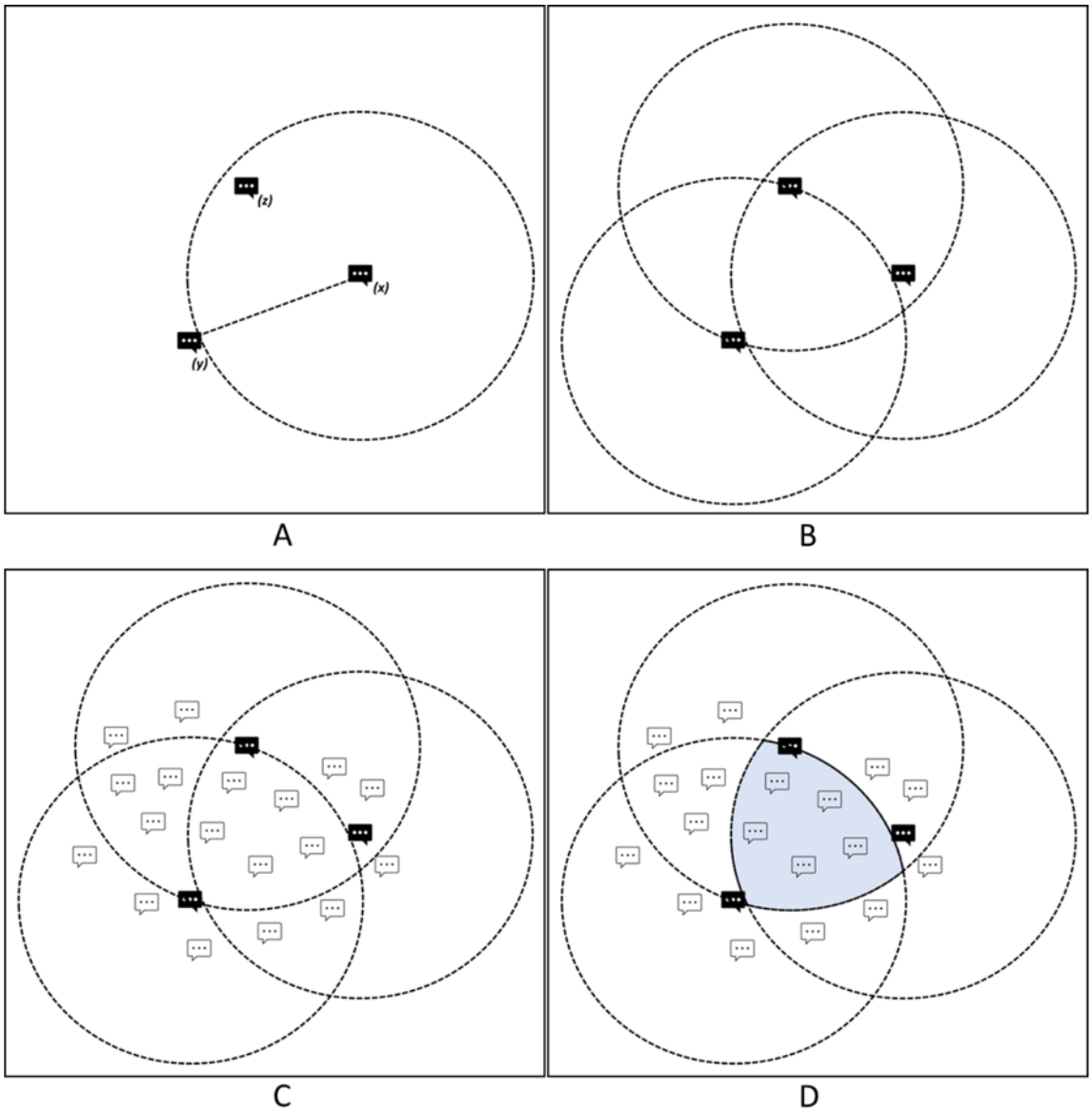
Clinical topic NLP[b]	Records tagged initially, n (%)	Records tagged NLP-augmented, n (%)	Increase in tagged patient records, n (%)[c]
COVID-19	375 (0.72)	405 (0.78)	30 (8)
General symptom	6401 (12.26)	6867 (13.15)	466 (7.28)
General administration	315 (0.6)	1217 (2.33)	902 (286.35)
Systemic clinical	3219 (6.16)	3519 (6.74)	300 (9.32)
Gastrointestinal	3421 (6.55)	4159 (7.96)	738 (21.57)
Respiratory	4040 (6.55)	4159 (7.96)	738 (21.57)
Cardiovascular	2683 (5.14)	5219 (9.99)	2536 (94.52)
Neurological	414 (7.93)	4485 (8.59)	345 (8.33)
Eye; ear, nose, and throat; and derma	1818 (3.48)	2061 (3.95)	243 (13.37)
Gynecology and urology	2712 (5.19)	3004 (5.75)	292 (10.77)
Trauma	16,516 (31.63)	17,389 (33.3)	873 (5.29)
General psychiatric	1989 (3.81)	2627 (5.03)	638 (32.08)
No tag	14,141 (27.08)	10,569 (20.24)	–3572 (–25.26)

[a]This table presents the distribution of the diagnosis topics obtained with the NLP-based text annotation before and after the spherical feature annotation.

[b]NLP: natural language processing.

[c]Percent of initially recorded tags.

Figure 4. Spherical dictionary enhancement. (A) Step A uses the largest distance between 2 words that are already in the topic. The circle around the word (x) shows the region in the embedding space with words closer to (x) than the maximum distance. (B) The same region is circled around the other 2 words (y) and (z). (C) The other words in the embedding space that were initially not included in the topic. (D) The intersection of the 3 circles defines the area in the embedding space where the distance of each word is smaller than the maximum distance.



Ethical Considerations

A waiver from the cantonal ethics committee was obtained before the commencement of this study (BASEC-Nr. Req-2019-00671).

Results

In the first step, we performed a descriptive analysis of the topics. To this end, we first excluded cases without a manually selected LS for further analysis and obtained a data set with 52,222 entries. Of the 52,222 patient visits included in our final analysis, 5994 (11.48%) had a manually recorded chief complaint that was not otherwise specified (eg, “Diverse”) and could not be classified as a symptom Of the 52,222 entries, 10,569 (20.24%) were not tagged with an NLP topic.

The distribution of all NLP topics is shown in Table 3. The distribution ranged from 0.05% of patient visits tagged with the NLP topic “dementia” to 9.89% for “wound.” The largest cluster of aggregated NLP symptom-related topics was “trauma,” with 33.1% of visits, and the smallest was “COVID,” with 0.8% of visits. The distribution of chief complaints can be found in Table S1 in Multimedia Appendix 1. In total, the distribution ranged from 0.01% of patient visits for the recorded chief complaints “melaena,” “hearing problems,” and “contact with chemicals” to 14.6% for “COVID.” The largest cluster of aggregated chief complaints was “trauma” with 23.6% and the smallest was “general organizational” with 1.2% of visits.

For comparability, we grouped all LS and NLP topics into 12 identical symptom clusters, which can be found in Table 4.

Table 3. Clusters for natural language processing–extracted topics (n=52,222)[a].

Cluster and subcluster detail	Values, n (%)
COVID-19	401 (0.77)
General symptoms	6852 (13.12)
Fever	2440 (4.67)
Pain	4505 (8.63)
General weakness	80 (0.15)
Back pain	438 (0.84)
General organizational	1217 (2.33)
Follow-up and prescription	1217 (2.33)
Systemic	3519 (6.74)
Infection not otherwise specified	1239 (2.37)
Sepsis	125 (0.24)
Anaphylaxia and allergy	261 (0.5)
Cancer	1688 (3.23)
Transplantation	227 (0.43)
Glycemia	138 (0.26)
Gastrointestinal	4147 (7.94)
Gastrointestinal bleeding	522 (1)
Abdominal pain	1879 (3.6)
Diarrhea, vomiting, and nausea	2248 (4.3)
Respiratory	4311 (8.26)
Upper airway	1592 (3.05)
Lower airway	1934 (3.7)
Influenza	440 (0.84)
Dyspnea	2197 (4.21)
Cardiovascular	5211 (9.98)
Chest pain	3569 (6.83)
Palpitations and arrythmia	518 (0.99)
Pulmonary embolism	281 (0.54)
Deep venous thrombosis	528 (1.01)
Hypertension	394 (0.75)
Neurological	4466 (8.55)
Headache	1189 (2.28)
Neurological	1737 (3.33)
Vigilance and disorientation	191 (0.37)
Dementia	24 (0.05)
Syncope	453 (0.87)
Vertigo and dizziness	934 (1.79)
Convulsion	226 (0.43)
Eye; ear, nose, and throat; and skin	2061 (3.95)
Epistaxis	58 (0.11)
Eye symptoms	703 (1.35)
Hearing and auricular	18 (0.03)

Cluster and subcluster detail	Values, n (%)
Skin	1311 (2.51)
Urological and gynecological	3004 (5.75)
Urological and kidney	2973 (5.69)
Pregnancy	34 (0.07)
Trauma	17,302 (33.13)
Wound	5163 (9.89)
Fracture and luxation	5375 (10.29)
Trauma and head	2171 (4.16)
Burns	141 (0.27)
Fall	729 (1.4)
Trauma not otherwise specified	9278 (17.77)
Bleeding not otherwise specified	986 (1.89)
Collision	1250 (2.39)
Traffic	314 (0.6)
Psychiatric	2625 (5.03)
Intoxication	1146 (2.19)
Psychiatric	851 (1.63)
Fear	725 (1.39)
Severity	
Nonsevere	113 (0.22)
Severe	235 (0.45)
Chronic	55 (0.11)
Acute	232 (0.44)

[a]This table presents the distribution of the diagnosis topics obtained with the natural language processing–based text annotation. In total, 20.38% of cases could not be attributed with a diagnosis topic.

Table 4. Summary statistics feature annotations (n=52,222)[a].

Cluster	LS[b], (n)	NLP[c] (n)	Correlation (*r*)[d]	Consistency[e]
COVID-19	7623	401	0.18	0.05
General symptom	7993	6852	−0.04	0.10
General administration	642	1217	0.01	0.04
Systemic clinical	1983	3519	0.12	0.22
Gastrointestinal	4063	4147	0.41	0.46
Respiratory	872	4311	0.17	0.44
Cardiovascular	2245	5211	0.28	0.49
Neurological	5123	4466	0.44	0.46
Eye; ear, nose, and throat; and derma	1041	2061	0.26	0.39
Gynecology and urology	1206	3004	0.40	0.67
Trauma	12,337	17,302	0.54	0.79
General psychiatric	1610	2625	0.60	0.78
No tag	5994	10,644	0.07	0.28

[a]This table presents the number of tagged cases for each chief cluster with both the natural language processing–based method and based on the chief complaint tag.

[b]LS: lead symptom.

[c]NLP: natural language processing.

[d]Correlation between LS and NLP.

[e]The number of overlapping LS and NLP tags divided by the total number of LS tags.

In addition to the NLP symptom-related topics, 4 modulating NLP topics, “acute,” “chronic,” “nonsevere,” and “severe,” were recorded, also based on keywords (ie, words in the text indicating severity). The purpose of the modulating topics is to provide more information on severity and control for this dimension in the further analysis.

We found that the correlation between LS clusters and NLP clusters was low (Table 4). Similarly, consistency varies relative to the LS. We also calculated the consistency of the NLP tags relative to the LS groups (the LS groups are the denominator; being more established, we use them as a benchmark). For most clusters, the consistency is approximately 50%, with trauma and psychiatric diagnosis having the highest consistency of 78% and 79%, respectively, and general administration and COVID-19 having the lowest consistency of 4% and 5%, respectively.

Compared with the LS clusters, our NLP topics have the advantage that a patient visit can be tagged to multiple topics. Table S2 in Multimedia Appendix 1 shows the number of NLP topics for each LS cluster. Of the 46,228 patient visits where we could assign a manually recorded chief complaint, 8950 (19.36%) were not tagged with an NLP topic. In contrast, 33.48% (15,477/46,228) of the visits were tagged with at least 2 NLP topics.

We estimated 3 models using logistic regression to show the association of the different symptom groups with the ESI and inpatient indicators:

$$\text{Model 1: } Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i \textbf{ (1)}$$

$$\text{Model 2: } Y_i = \alpha + \beta X_i + \delta W_i + \varepsilon_i \textbf{ (2)}$$

$$\text{Model 3: } Y_i = \alpha + \beta X_i + \gamma Z_i + \delta W_i + \varepsilon_i \textbf{ (3)}$$

where Y_i is either the ESI or inpatient indicator variable for patient visit i, α the intercept, X_i is a vector of demographic and organizational variables for patient visit I (age; gender; admission type; admission reason; average number of patients in ED; average patient waiting time; night, late, or early shift; and treating ED team), Z_i is a vector of the NLP-derived symptom clusters, W_i is a vector of the lead symptom–derived cluster (based on the drop-down menu), and ε_i is the error term.

Tables 5 and 6 present the results. Column 1 shows the NLP-derived groups, with coefficients ranging between 5% and 13% increased or decreased probability of a high ESI score or 5% to 19% increased or decreased probability for hospitalization. The drop-down–based LS in column 2 has similar but slightly larger coefficients. Column 3 shows both variables, as in model 3, in this specification, the coefficients are mostly complementary, meaning that if a patient shows the same symptom in both the NLP and LS measures, the probabilities can be added. Note that this is not owing to multicollinearity because both coefficients remain significant in most cases.

Table 5. Linear probability model on "Inpatient"[a].

Name of cluster[b]	Model 1[c], regression coefficient (SE)	Model 2[c], regression coefficient (SE)	Model 3 including both measures[d], regression coefficient (SE)
NLP[e] cluster: COVID-19	0.048[f] (0.019)	N/A[g]	−0.022 (0.022)
Chief complaint cluster: COVID-19	N/A	0.127[g] (0.007)	0.133[h] (0.008)
NLP cluster: general symptoms	0.011[f] (0.005)	N/A	−0.019[h] (0.005)
Chief complaint cluster: general symptoms	N/A	−0.002 (0.007)	0.000 (0.007)
NLP cluster: general organizational	−0.004 (0.011)	N/A	0.006 (0.011)
Chief complaint cluster: general organizational	N/A	−0.062[g] (0.016)	−0.052[h] (0.016)
NLP cluster: systemic	0.117[h] (0.007)	N/A	0.101[h] (0.007)
Chief complaint cluster: systemic	N/A	0.118[h] (0.010)	0.104[h] (0.010)
NLP cluster: gastrointestinal	0.071[h] (0.006)	N/A	0.040[h] (0.007)
Chief complaint cluster: gastrointestinal	N/A	0.083[h] (0.008)	0.059[h] (0.008)
NLP cluster: respiratory	0.063[h] (0.007)	N/A	−0.017[f] (0.008)
Chief complaint cluster: respiratory	N/A	0.133[h] (0.014)	0.126[f] (0.014)
NLP cluster: cardiovascular	−0.020[h] (0.006)	N/A	−0.009 (0.006)
Chief complaint cluster: cardiovascular	N/A	−0.038[h] (0.010)	−0.031[h] (0.010)
NLP cluster: neurological	−0.046[h] (0.007)	N/A	−0.045[h] (0.007)
Chief complaint cluster: neurological	N/A	−0.058[h] (0.009)	−0.048[h] (0.009)
NLP cluster: eye, ENT[i], or skin	−0.055[h] (0.009)	N/A	−0.044[h] (0.009)
Chief complaint cluster: eye, ENT, or skin	N/A	−0.128[h] (0.013)	−0.112[h] (0.013)
NLP cluster: urological or gynecological	−0.015[f] (0.008)	N/A	−0.004 (0.008)
Chief complaint cluster: urological or gynecological	N/A	−0.033[h] (0.012)	−0.036[h] (0.013)
NLP cluster: trauma	−0.041[h] (0.005)	N/A	−0.038[h] (0.005)
Chief complaint cluster: trauma	N/A	0.011 (0.007)	0.020[h] (0.007)
NLP cluster: psychiatric	−0.079[h] (0.009)	N/A	−0.053[h] (0.010)
Chief complaint cluster: psychiatric	N/A	−0.068[h] (0.013)	−0.039[h] (0.014)

[a]This table presents the results from a linear probability model with inpatients as the dependent variable. All the models include a set of demographic and administrative covariates.

[b]Observation: 52,222; R^2=0.259.

[c]Observation: 52,222; R^2=0.263.

[d]Observation: 52,222; R^2=0.269.

[e]NLP: natural language processing.

[f]P<.05.

[g]N/A: not applicable.

[h]P<.01.

[i]ENT: ear, nose, and throat.

Table 6. Linear probability model on "low Emergency Severity Index (ESI) score"[a].

Name of cluster[b]	Model 1[c], regression coefficient (SE)	Model 2[c], regression coefficient (SE)	Model 3 including both measures[d], regression coefficient (SE)
NLP[e] cluster: COVID-19	0.079[f] (0.019)	N/A[g]	0.023 (0.019)
Chief complaint cluster: COVID-19	N/A	0.214[f] (0.007)	0.172[f] (0.007)
NLP cluster: general symptoms	0.036[f] (0.005)	N/A	−0.023[f] (0.005)
Chief complaint cluster: general symptoms	N/A	−0.142[f] (0.007)	0.127[f] (0.007)
NLP cluster: general organizational	−0.050 (0.011)	N/A	−0.044[f] (0.011)
Chief complaint cluster: general organizational	N/A	0.308[f] (0.016)	0.352[f] (0.016)
NLP cluster: systemic	0.076[f] (0.007)	N/A	0.093[f] (0.007)
Chief complaint cluster: systemic	N/A	0.009 (0.010)	0.009 (0.010)
NLP cluster: gastrointestinal	0.192[f] (0.006)	N/A	0.088[f] (0.007)
Chief complaint cluster: gastrointestinal	N/A	0.305[f] (0.008)	0.262[f] (0.008)
NLP cluster: respiratory	0.114[f] (0.007)	N/A	0.053[f] (0.007)
Chief complaint cluster: respiratory	N/A	0.121[f] (0.014)	0.088[f] (0.014)
NLP cluster: cardiovascular	0.050[f] (0.006)	N/A	0.030[f] (0.006)
Chief complaint cluster: cardiovascular	N/A	0.205[f] (0.009)	0.197[f] (0.010)
NLP cluster: neurological	−0.015[h] (0.007)	N/A	−0.002 (0.007)
Chief complaint cluster: neurological	N/A	−0.038[f] (0.009)	−0.039[f] (0.009)
NLP cluster: eye, ENT[i], or skin	−0.134[f] (0.009)	N/A	−0.061[f] (0.009)
Chief complaint cluster: eye, ENT, or skin	N/A	−0.302[f] (0.013)	−0.279[f] (0.013)
NLP cluster: urological or gynecological	0.055[f] (0.008)	N/A	0.006 (0.008)
Chief complaint cluster: urological or gynecological	N/A	0.193[f] (0.012)	0.187[f] (0.013)
NLP cluster: trauma	−0.129[f] (0.005)	N/A	−0.098[f] (0.005)
Chief complaint cluster: trauma	N/A	−0.011 (0.007)	0.013[j] (0.007)
NLP cluster: psychiatric	0.063[f] (0.009)	N/A	0.080[f] (0.010)
Chief complaint cluster: psychiatric	N/A	0.086[f] (0.012)	0.051[f] (0.013)

[a]This table presents the results from a linear probability model with the low ESI score indicator as the dependent variable (ESI score of 2 or 3). All models included a set of demographic and administrative covariates.

[b]Observation: 52,222; R^2=0.409.

[c]Observation: 52,222; R^2=0.448.

[d]Observation: 52,222; R^2=0.457.

[e]NLP: natural language processing.

[f]P<.01.

[g]N/A: not applicable.

[h]P<.05.

[i]ENT: ear, nose, and throat.

[j]P<.10.

Of the 12 symptom clusters, 11 (92%) in column 1 had a significant regression coefficient for hospitalization (all but "general organizational"). Eight clusters remained significant even when including the cluster of clinician-determined chief complaints in the model. In the model explaining "inpatient," in 10 (83%) out of the 12 symptom cluster pairs, the coefficients of the NLP topic clusters showed the same algebraic sign as the chief complaint clusters. In contrast, for 2 symptom cluster pairs, they did not ("general symptoms" and "trauma"). A change in the algebraic sign of either the chief complaint cluster

or the NLP topics cluster occurred in 4 cluster pairs when both NLP topics and chief complaints were included in the model ("COVID," "general symptoms," "general organizational," and "respiratory"). We obtained similar results when analyzing the low ESI scores. However, a change in the algebraic sign of a coefficient within solely 1 pair of symptom clusters was noted ("trauma"). Interestingly, the clusters "cardiovascular," "neurological," and "trauma" were significantly associated with nonhospitalization, of which "neurological" and "trauma" but not "cardiovascular" were also significantly associated with a lower ESI score.

As a robustness check, we used each of the 3 model specifications to predict the ESI indicator and the inpatient indicator. Using the respective sets of variables of each specification, we used a logistic regression with a 2:1 train-test split to predict both outcomes. Table 7 shows the F_1-score and area under the curve (AUC) score of these predictions. The results show that the 3 specifications have similar predictive power (an AUC of 0.82-0.84 for "inpatient" and an AUC of 0.90-0.92 for ESI indicator).

The inference and prediction results show that the added value of text in this setting is not by increasing the predictive power of the model, where the outcomes are existing process outcomes (eg, discharge type of severity). Instead, unstructured text allows clinicians to access more granular information to optimize patient flows, which cannot be reflected in the inpatient and ESI indicator outcomes.

In a more granular analysis, we estimated models 1 to 3 with the individual NLP topics and the individual LS groups instead of the clusters previously used. The analysis corroborated our clinical presumptions that, for example, age, admission by an ambulance, and "sepsis" as an NLP topic, as well as "chest pain" for a chief complaint, were associated with low ESI scores (2 or 3) or hospital admission. In contrast, the NLP topic or chief complaint cluster "follow-up" was not. The complete results are provided in Tables S3-S6 in Multimedia Appendix 1.

Table 7. Prediction of hospitalization ("Inpatient") and low Emergency Severity Index (ESI) score of 2 or 3 ("Low ESI score").

Variable and model	F_1-score on ones	AUC[a]
Inpatient		
Model 1: NLP[b] clusters	0.57	0.82
Model 2: LS[c] clusters	0.57	0.83
Model 3: NLP+LS clusters	0.59	0.84
Low ESI score		
Model 1: NLP clusters	0.86	0.92
Model 2: LS clusters	0.84	0.90
Model 3: NLP+LS clusters	0.87	0.92

[a]AUC: area under the curve.

[b]NLP: natural language processing.

[c]LS: lead symptom.

Discussion

Principal Findings

Our analysis of patient records showed the additional information extracted from unstructured text and its potential usefulness in the clinical context. We demonstrated that the information extracted from NLP features and the physician's categorization of chief complaints was *complementary*. Indeed, the correlation and consistency between the chief complaint and NLP-derived clusters were low (Table 4). This finding indicates that the free text from the NLP clusters provides additional information than that contained in the symptom clusters from the structured chief complaints.

The complementarity of the information is further emphasized by the results summarized in Tables 5 and 6, and most coefficients remained significant when both types of indicators were included in the model, suggesting that different aspects of patient information appear to be encoded by the 2 approaches. These results support our hypothesis that NLP-derived libraries capture greater depth and breadth of information than a single chief complaint and underscore the relevance of including information captured in unstructured text to address patient populations.

Surprisingly, the "cardiovascular" and "trauma" clusters were not significant features for predicting hospitalization, with "trauma" also significant for predicting a *higher* ESI score. In contrast, the "systemic" cluster, which included sepsis, anaphylaxis, and neoplastic disease, was significant for predicting hospitalization and a lower ESI score, consistent with clinical expectations. Although symptoms suggestive of cardiac dysfunction and trauma may warrant urgent clinical risk assessment, most patients with such complaints would not require hospitalization. Therefore, early allocation of hospital beds for these subgroups is unlikely to reduce overcrowding. Targeting patients with systemic symptoms, in contrast, is likely to do so.

We also proposed a method for analyzing unstructured clinical notes. Our approach has the advantages of speed, simplicity of

implementation, and transparency. The speed at which supervised libraries can be assembled is a strength of the proposed approach. A limitation of implementing supervised NLP algorithms in routine decision support is that they are often resource intensive [17]. In our application, it took an untrained clinician only a few days to assemble the entire library.

Furthermore, using NLP as a tool traditionally requires expertise and the ability to master NLP applications. In fields that require years to decades of training, such as health care, professionals cannot be routinely trained to excel in programming. Thus, a further major barrier to the successful implementation of NLP applications in health care is often the usability of NLP applications [18]. Moreover, the flexibility of the method allows easy adaptation of the created dictionaries to analyze new data sets.

Trust is one of the key benefits of clinician involvement in developing proprietary AI models. Indeed, lack of trust is a recognized major limitation that hinders the potential benefits of using AI in routine clinical practice for organizations and patients [19,20]. Owing to the supervised approach, annotated library compilation is comprehensible and transparent; hence, it is trustworthy for clinicians. This may also become an important advantage if regulation on the implementation of AI use in health care tightens in the future.

The limitation of this study is that our approach still requires manual coding. However, future developments in AI may facilitate this step even further. In addition, human bias was possible because the library was compiled manually. In general, an AI-based text analysis does not achieve perfect precision. However, we primarily advocate using free-text analysis for organizational, not clinical, decision support. Therefore, this limitation is not clinically relevant. A further limitation may lie in the fact that the low correlation between the NLP and chief complaint clusters could stem from errors originating from the manual grouping or NLP clustering. However, we believe these results are plausible. Indeed, the chief complaints “fever” and “pain” were included in the cluster “general symptoms,” as were the NLP-extracted tags “fever” and “pain.” However, as only 1 chief complaint could allocated to a patient, during the COVID-19 pandemic, most patients presenting with fever or influenza-like pain would have most likely been categorized as presenting with the chief complaint “COVID.”

Conclusions

Health care workers on the one side and EHR engineers as well as hospital administration on the other side are caught in a long, ongoing conflict over the extent of structuring the data entered into EHR. Health care workers often argue that entering structured data is a cumbersome task and that the information archived can be of little use in daily clinical practice. In contrast, administrators and EHR engineers often advocate that structuring data is the only reliable solution, enabling a meaningful analysis of the data. Technological advances may help resolve this conflict.

We were able to demonstrate the importance of maintaining free text in EHR. Indeed, using the chief complaints attributed by a physician from a drop-down menu and a corresponding free-text field as a case in point, we were able to show that free text contains a wealth of information that is not routinely captured by structured data.

Moreover, we developed an approach that could enable the information captured in free text to be easily extracted and processed by hospital informatics systems and fed into a workflow, possibly improving the efficiency of patient management.

Therefore, future EHRs should include the possibility of entering free text.

Acknowledgments

The authors would like to thank Professor Michael Krauthammer from the University of Zurich, Switzerland, and Privat-Dozentin Dr Ksenija Slankamenac, PhD, from the University Hospital Zurich for their feedback in helping to prepare this submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary file for web-based publication only.
[PDF File (Adobe PDF File), 240 KB - medinform_v12i1e49007_app1.pdf]

References

1. Hwang JE, Seoung BO, Lee SO, Shin SY. Implementing structured clinical templates at a single tertiary hospital: survey study. JMIR Med Inform 2020 Apr 30;8(4):e13836. [doi: 10.2196/13836] [Medline: 32352392]
2. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
3. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]

4. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020 Feb;145(2):463-469. [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
5. Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assoc 2022 Jun 14;29(7):1292-1302 [FREE Full text] [doi: 10.1093/jamia/ocac058] [Medline: 35475536]
6. Velt KB, Cnossen M, Rood PP, Steyerberg EW, Polinder S, Lingsma HF. Emergency department overcrowding: a survey among European neurotrauma centres. Emerg Med J 2018 Jul;35(7):447-448 [FREE Full text] [doi: 10.1136/emermed-2017-206796] [Medline: 29563151]
7. Di Somma S, Paladino L, Vaughan L, Lalle I, Magrini L, Magnanti M. Overcrowding in emergency department: an international issue. Intern Emerg Med 2015 Mar;10(2):171-175. [doi: 10.1007/s11739-014-1154-8] [Medline: 25446540]
8. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: a systematic review of causes, consequences and solutions. PLoS One 2018 Aug 30;13(8):e0203316 [FREE Full text] [doi: 10.1371/journal.pone.0203316] [Medline: 30161242]
9. Iacobucci G. Overcrowding and long delays in A&E caused over 4000 deaths last year in England, analysis shows. BMJ 2021 Nov 18;375:n2835. [doi: 10.1136/bmj.n2835] [Medline: 34794954]
10. Iacobucci G. Government must "get a grip" on NHS crisis to halt avoidable deaths, say leaders. BMJ 2023 Jan 03;380:12. [doi: 10.1136/bmj.p12] [Medline: 36596573]
11. Boyle A. Unprecedented? The NHS crisis in emergency care was entirely predictable. BMJ 2023 Jan 09;380:46. [doi: 10.1136/bmj.p46] [Medline: 36623878]
12. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010 Aug 06;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
13. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022.
14. Tang J, Meng Z, Nguyen X, Mei Q, Zhang M. Understanding the limiting factors of topic modeling via posterior contraction analysis. In: Proceedings of the 31st International Conference on International Conference on Machine Learning ICML'14. 2014 Presented at: ICML'14; June 21-26, 2014; Beijing, China.
15. Maynard D, Funk A. Automatic detection of political opinions in Tweets. In: Proceedings of the Workshops at the 8th Extended Semantic Web Conference, ESWC 2011. 2011 Presented at: Workshops at the 8th Extended Semantic Web Conference, ESWC 2011; May 29-30, 2011; Heraklion, Greece. [doi: 10.1007/978-3-642-25953-1_8]
16. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010 Presented at: LREC 2010 Workshop on New Challenges for NLP Frameworks; May 22, 2010; Valletta, Malta. [doi: 10.13140/2.1.2393.1847]
17. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med 2019 Dec 17;2(1):130 [FREE Full text] [doi: 10.1038/s41746-019-0208-8] [Medline: 31872069]
18. Zheng K, Vydiswaran VG, Liu Y, Wang Y, Stubbs A, Uzuner Ö, et al. Ease of adoption of clinical natural language processing software: an evaluation of five systems. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S189-S196 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.008] [Medline: 26210361]
19. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
20. Celi LA, Fine B, Stone DJ. An awakening in medicine: the partnership of humanity and intelligent machines. Lancet Digit Health 2019 Oct;1(6):e255-e257 [FREE Full text] [doi: 10.1016/s2589-7500(19)30127-x] [Medline: 32617524]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
ED: emergency department
EHR: electronic health record
ESI: Emergency Severity Index
LS: lead symptoms
NLP: natural language processing

Edited by C Lovis; submitted 15.05.23; peer-reviewed by J Kors, C Gaudet-Blavignac; comments to author 30.06.23; revised version received 30.10.23; accepted 24.11.23; published 17.01.24.

Please cite as:
Mehra T, Wekhof T, Keller DI
Additional Value From Free-Text Diagnoses in Electronic Health Records: Hybrid Dictionary and Machine Learning Classification Study
JMIR Med Inform 2024;12:e49007
URL: https://medinform.jmir.org/2024/1/e49007
doi:10.2196/49007
PMID:38231569

Original Paper

BERT-Based Neural Network for Inpatient Fall Detection From Electronic Medical Records: Retrospective Cohort Study

Cheligeer Cheligeer[1,2], PhD; Guosong Wu[1,3], PhD; Seungwon Lee[1,2], PhD; Jie Pan[1,3], PhD; Danielle A Southern[1], MSc; Elliot A Martin[1,2], PhD; Natalie Sapiro[1], MSc, RN; Cathy A Eastwood[1,3], RN, PhD; Hude Quan[1,3], MD, PhD; Yuan Xu[1,3,4,5], MD, PhD

[1]Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[2]Provincial Research Data Services, Alberta Health Services, Calgary, AB, Canada

[3]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[4]Department of Oncology, University of Calgary, Calgary, AB, Canada

[5]Department of Surgery, University of Calgary, Calgary, AB, Canada

Corresponding Author:
Yuan Xu, MD, PhD
Centre for Health Informatics, Cumming School of Medicine
University of Calgary
3280 Hospital Dr NW
Calgary, AB, T2N 4Z6
Canada
Phone: 1 (403) 210 9554
Email: yuxu@ucalgary.ca

Abstract

Background: Inpatient falls are a substantial concern for health care providers and are associated with negative outcomes for patients. Automated detection of falls using machine learning (ML) algorithms may aid in improving patient safety and reducing the occurrence of falls.

Objective: This study aims to develop and evaluate an ML algorithm for inpatient fall detection using multidisciplinary progress record notes and a pretrained Bidirectional Encoder Representation from Transformers (BERT) language model.

Methods: A cohort of 4323 adult patients admitted to 3 acute care hospitals in Calgary, Alberta, Canada from 2016 to 2021 were randomly sampled. Trained reviewers determined falls from patient charts, which were linked to electronic medical records and administrative data. The BERT-based language model was pretrained on clinical notes, and a fall detection algorithm was developed based on a neural network binary classification architecture.

Results: To address various use scenarios, we developed 3 different Alberta hospital notes-specific BERT models: a high sensitivity model (sensitivity 97.7, IQR 87.7-99.9), a high positive predictive value model (positive predictive value 85.7, IQR 57.2-98.2), and the high F_1-score model (F_1=64.4). Our proposed method outperformed 3 classical ML algorithms and an International Classification of Diseases code–based algorithm for fall detection, showing its potential for improved performance in diverse clinical settings.

Conclusions: The developed algorithm provides an automated and accurate method for inpatient fall detection using multidisciplinary progress record notes and a pretrained BERT language model. This method could be implemented in clinical practice to improve patient safety and reduce the occurrence of falls in hospitals.

(JMIR Med Inform 2024;12:e48995) doi:10.2196/48995

KEYWORDS

accidental falls; electronic medical records; data mining; machine learning; patient safety; natural language processing; adverse event

Introduction

Background

Inpatient falls detrimentally impact patients, leading to extended hospital stays and distress among families and caregivers [1-5]. Studies reflect a varying incidence rate of such falls, with 250,000 annually in England and Wales alone [1], and evidence showing 7.5% of patients experience at least 1 fall during hospitalization [2]. Acute care hospitals also report a range of 1 to 9 falls per 1000 bed days, underscoring the pervasive nature of this problem [4]. Patients who fall may experience injuries that increase the risk of comorbidity or even disability [6,7]. They may also experience psychological effects such as anxiety, depression, or loss of confidence, which can affect their recovery and quality of life [8].

Manual chart review is regarded as one of the most common methods to identify inpatient falls [9]. This process involves the thorough examination of patient medical records to gather relevant information on the details of falls. Existing strategies include the Harvard Medical Practice Study [10] and the Global Trigger Tool [11]. Alternative methodologies, such as Patient Safety Indicators, based on International Classification of Diseases (ICD) codes, are used to identify adverse events (AEs), leveraging systematized health care data for detection [12-14]. However, these methodologies, while widely used, present challenges due to the time-consuming nature of ICD coding and manual chart reviews, potentially causing delays in recording and detecting AEs [15,16].

Free text data in electronic medical records (EMRs) offer rich, up-to-date insights into patients' health status, medications, and various narrative content. Despite its wealth of information, the unstructured nature of this data necessitates chart reviews, a labor-intensive process, to identify inpatient falls [17]. There has been an increasing interest in recent years in applying natural language processing (NLP) techniques to electronic clinical notes to automate disease identification and create clinical support decision systems [18-24].

Previous NLP studies in the detection of patient fall including rule-based algorithms [25,26] and machine learning (ML) methods [27-30] have been explored, but they often struggle with the variety and complexity of clinical language.

The deep learning model Bidirectional Encoder Representation from Transformers (BERT) [31] can effectively address these challenges. It uses transformer architecture to understand text contextually, handling linguistic complexity, abbreviations, and data gaps, thereby augmenting text understanding from EMR [20]. The use of transformer-based methods to understand EMR text data has emerged as a promising new trend in automatic clinical text analysis [32].

Objectives

In this study, we intend to pretrain an existing model, BioClinical BERT [33], with free text data from Alberta hospital EMRs to develop an Alberta hospital notes-specific BERT model (AHN-BERT). The pretrained language model would serve as a feature extraction layer in a neural network to identify inpatient falls. We hypothesize that fine-tuning BERT on local hospital data will enable more accurate fall detection compared with generic models. Additionally, we expect AHN-BERT will outperform conventional rule-based and ML approaches, as well as ICD code methods, in detecting falls from unstructured EMR notes in near real time. By evaluating AHN-BERT against current techniques, we hope to demonstrate the value of transfer learning with BERT for improved efficiency and generalizability in surfacing patient safety events from clinical text. Ultimately, our goal is to advance the detection of inpatient falls, allowing for more detailed and accurate patient safety interventions. An improved fall detection system could potentially enable health care providers to swiftly implement preventive measures, reducing the incidence and severity of falls. Additionally, through the facilitation of access and analysis of fall-related data, our system could become an invaluable resource for researchers investigating fall prevention and associated subjects.

Methods

Overview

In our methodology, we emphasized a detailed and transparent approach, covering all aspects from data collection to model validation. This comprehensive process, reflecting best practices in research reporting [34], ensures clarity and precision in our multivariable prediction model, providing an in-depth understanding of its performance and applicability.

Source of Data

Our study is a retrospective analysis. We used a stratified random sample of adult patients admitted to acute care hospitals in Calgary, Alberta. We linked the extracted EMR data to Sunrise Clinical Manager (SCM) records and ICD-coded discharge abstract database (DAD) using an established mechanism [35]. Both tables are stored and managed by the Oracle database.

The chart reviewer team consists of 6 registered nurses with 1 to 10 years of experience using SCM for clinical care. The nurses followed a training procedure, and 1 trained nurse became the project lead for quality assurance. The training involved learning the condition definitions and practicing reviewing each chart systematically. Reviewers examined the entire record for specified conditions and consulted each other with questions. In the process of training and quality assurance, we tested interrater reliability using Conger, Fleiss, and Light κ methods, with 2 nurses reviewing the same set of 10 charts for consensus on AEs. Where agreement was poor ($\kappa<0.60$), retraining occurred until high agreement ($\kappa>0.80$) was achieved [36]. Reviewers then proceeded independently with REDCap tool (Vanderbilt University).

The chart review data served as the reference standard to develop and evaluate our fall detection model. We focused on multidisciplinary progress records (MPRs) for fall detection, as chart review data showed most falls (115/155, 73.7%) were documented in MPRs by nursing staff. We created supervised data sets for the classification task to identify optimal fall detection timing, including 1-day (fall day MRPs notes), 2-day (fall day + day after), 3-day (fall day + 2 days after), and full hospitalization MRPs. All supervised data sets were labeled to

indicate whether notes were associated with inpatient falls. For the training of our model, we used both cases (falls) and controls (nonfalls) at a ratio of 1:29. This was done to ensure the model was exposed to a balanced representation of both scenarios. Our test set mirrored the real-world data distribution to enable an accurate evaluation of model performance. In addition, we constructed an unsupervised corpus specifically for language model pretraining. This corpus comprises free-text note data and does not rely on any predetermined labels or annotations.

Participants

At the time of the study, a total of 4393 charts were reviewed, among which we identified a total of 155 records as falls and 4238 records as no falls. The study included only the first admission of each patient, even if they had multiple hospitalizations within the study period. We exclusively focused on adults 18 to 100 years of age, thereby excluding minors and centenarians. Furthermore, if a patient had multiple fall incidents, only the most recent record was considered, although no such cases were identified during the study. The temporal framework for the study encompassed a decade, from 2010 to 2020. Exclusion criteria were also clearly defined: patients without unstructured note data or those who could not be linked using our established data linkage mechanism were omitted from the study.

Missing Data and Data Cleaning

Our study implemented rigorous data cleaning to ensure data integrity. After conducting a conflict review and excluding records with inconsistencies in fall status documentation (17 records checked for both falls and no falls), failed data linkage (1 record), temporal conflicts between fall and admission dates (5 records), and missing MPR documentations (47 records), the final clean data set totaled 4323 records (142 falls and 4181 no falls).

Outcome and Variables

The desired outcome of our proposed framework is to predict whether a patient's daily progress note contains hints about inpatient falls. The input to our model is each patient's n-day note. We use a BERT model to represent the textual data in numerical format, also known as contextualized word embeddings.

The input text is represented by 768-dimensional feature vectors, which can be considered as 768 variables. However, due to the distributed representation of neural language models, each variable does not represent a single word. Instead, individual variables preserve contextual information segments for each word, constituting meaningful vector representations of the entire input text.

On a related note, we have also collected and analyzed several demographic and clinical variables for our patient cohort from DAD database. Although not directly used in our predictive modeling, these variables furnish invaluable insights into the characteristics of our study population and contribute to the overall richness of our research data. These include age at the time of admission, sex, the incidence of intensive care unit visits during the hospital stays, the length of the hospital stays, and the hospital's geographical location. The latter was particularly focused on 3 acute care hospitals based in Calgary, Alberta: hospitals "A," "B," and "C." These variables help us understand the context in which the patient notes were written and may influence the interpretation of the model's results.

Sample Size

We included all patient data that has been reviewed by the reviewer team and filtered out from the inclusion-exclusion criteria.

Model Development

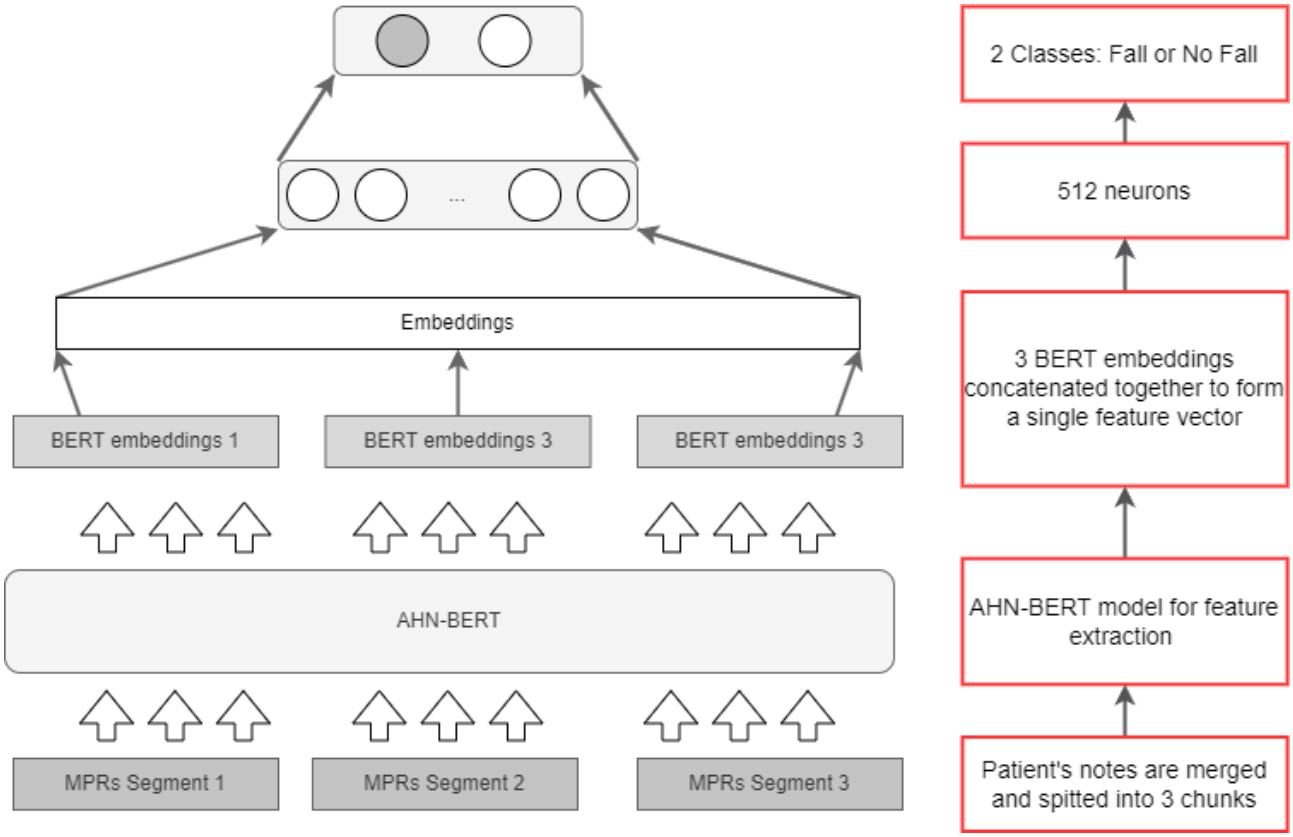
To conform to the BERT input length limit, MPR notes exceeding 400 tokens were programmatically split into segments under 400 words, preserving contextual information. All notes underwent preprocessing including removal of extraneous headers, signatures, whitespaces, and escape characters using regular expressions, and duplicate sentences were eliminated.

Our model architecture comprises 2 key components, an Alberta hospital note-specific BERT model for contextual feature extraction from clinical text, which we term AHN-BERT, and a feedforward neural network classifier to detect falls from the extracted features (as Figure 1).

AHN-BERT was initialized with weights from BioClinical BERT and further pretrained on our corpus of unsupervised hospital notes to adapt to local clinical terminology and language patterns. To prevent bias from overly lengthy documents, notes were filtered to be between 50 and 5000 tokens prior to pretraining. AHN-BERT was pretrained using a masked language modeling technique on 15% of randomly selected input tokens, enabling learning of contextual representations of clinical text without explicit labels. For feature extraction, AHN-BERT processes up to 3 concatenated note segments under 400 tokens. The resulting "[CLS]" vectors summarizing the semantic content of each segment are aggregated via concatenation to represent the full note's contextual information [37].

A feedforward neural network is used as the classifier, taking the concatenated features as input. The network comprises fully connected layers to map the features into class probabilities for fall detection. Dropout regularization is implemented in the classifier to prevent overfitting to the training data. Sigmoid activation in the output layer provides posterior probabilities for the binary fall classification task.

Figure 1. Proposed model architecture. AHN-BERT: Alberta hospital notes-specific BERT; BERT: Bidirectional Encoder Representation from Transformers; MPR: multidisciplinary progress record.



Model Assessment

First, to determine the optimal timeframe for note selection that best represents inpatient falls, we compared model performance using 1-day, 2-day, 3-day, and complete patient note data sets. Since the exact time lapse between an inpatient fall and corresponding documentation is variable, we evaluated these distinct time intervals in a data-driven approach to identify the optimal period for note selection. We used the same model architecture and pretrained AHN-BERT for all data sets, comparing training and validation loss convergence and evaluation metrics to assess performance.

Second, we tuned the classification probability threshold to balance model sensitivity and precision. The threshold denotes the cutoff for determining class membership based on predicted probabilities. By optimizing the threshold, we controlled the tradeoff between correctly identifying true positives and avoiding false positives. We developed three distinct models by threshold tuning for different purposes: (1) a high-sensitivity model that maximizes sensitivity by optimizing the threshold, (2) a high positive predictive value (PPV) model that maximizes PPV through threshold optimization, and (3) a high F_1-score model that balances sensitivity and PPV by optimizing the threshold, serving as a general-purpose model.

Third, we conducted a comparative evaluation between our top-performing neural network model and several other approaches, including 2 alternative BERT-based models, 3 conventional ML models, and an ICD-code–based algorithm. The 2 additional BERT-based models used original pretrained BERT and BioClinical BERT as feature extractors. For the 3 conventional ML models (support vector machine, logistic regression, and decision tree classifiers), we used bag-of-words features and term frequency-inverse document frequency weighting. These models were trained and compared on the 1-day MPRs data set. The ICD-code algorithm was applied to the same patient cohort but relied on administrative diagnosis codes rather than clinical notes. It aimed to demonstrate the efficacy of standard diagnostic codes for identifying falls compared with our neural network model. Falls were identified by the presence of ICD-10 codes W00-W20 when not listed as the primary diagnosis.

Statistical Analysis

The characteristics of the patients included in the study were thoroughly evaluated. These characteristics encompassed age, sex, the incidence of intensive care unit visits, the length of their hospital stay, and their originating hospitals. We summarized categorical variables as frequencies and percentages, while continuous variables were expressed as medians and IQRs. The χ^2 test was used for categorical variables to determine statistical differences, while the Wilcoxon rank-sum test was used for continuous variables. A P value threshold of 5% or lower was set to denote statistical significance.

To evaluate our ML model, we calculated several statistical metrics such as sensitivity, specificity, PPV, negative predictive value, accuracy, and F_1-score.

Computational Environment

Our study harnessed a high-performance computing environment, primarily driven by an NVIDIA GeForce RTX

3080 GPU with 16GB of memory, vital for pretraining and fine-tuning our language model. The statistical analysis and experiment leveraged Python 3.8, and libraries such as NumPy [38], Scikit-learn [39], Pandas [40], and PyTorch [41] for tasks like data processing and modeling.

Ethics Approval

This study was approved by the Conjoint Health Research Ethics Board at the University of Calgary (REB21-0416). Patient consent was waived as part of the ethics board review process.

Results

Participants

Our final study cohort contains 4323 individuals, with 142 (3.28%) patients identified by chart reviewers as having falls recorded in their medical charts during their hospital stay. The remaining 4181 (96.7%) did not fall. All patients were successfully linked to the SCM and DAD by unique identification number and admission date. Table 1 presents the descriptive statistics in general. Multimedia Appendix 1 further stratifies Table 1 into respective hospitals (Multimedia Appendix 1).

Table 1. Descriptive statistics of the study cohort.

	Total (n=4323)	Confirmed fall (n=142)[a]	No fall (n=4181)[b]	*P* value[c]
Age, median (IQRs[d])	62.0 (48.5-75.5)	71.0 (59.6-82.4)	61.0 (47.5-74.5)	<001
Sex (male), n (%)	2169 (50.2)	73 (51.4)	2096 (50.1)	.77
ICU[e] visit, n (%)	163 (3.8)	19 (13.4)	144 (3.4)	<.001
Length of hospital stay (days), median (IQRs)	3.0 (0.5-5.5)	12.0 (1.5-22.5)	3.0 (0.5-5.5)	<.001
Hospitals, n (%)[f]				.04
Hospital “A”	3548 (82.1)	104 (73.2)	3444 (82.4)	
Hospital “B”	651 (15.1)	34 (23.9)	617 (14.8)	
Hospital “C”	124 (2.8)	4 (2.9)	120 (2.8)	

[a]A term used in the study to refer to patients who fell during their hospital stay and were confirmed to have fallen through medical records or other documentation.

[b]A term used in the study to refer to patients who did not fall during their hospital stay.

[c]A measure indicating the statistical significance (P<.05) of the observed difference between groups.

[d]A measure of statistical dispersion representing the difference between the 75th and 25th percentiles of a data set.

[e]ICU: intensive care unit.

[f]Three different hospitals were included in the study: hospitals A, B, and C.

Model and Framework Assessment

First, to determine the optimal timeframe, we compared 1-day, 2-day, 3-day, and complete note data sets using the same model architecture and AHN-BERT pretrained embeddings. We trained each model for 200 epochs, with the primary goal of comparing their overall performance on the test sets. Evaluating performance metrics and training convergence, the 1-day data set was most effective and efficient, achieving 93.0% sensitivity and 83.0% specificity.

Second, we optimized the classification threshold to balance sensitivity and precision. These models maximize sensitivity, PPV, and F_1-score respectively. As results are shown in Table 2, our proposed architecture with AHN-BERT achieved overall the highest metrics among the comparison.

Third, the comparative assessment showed our approach outperformed 2 alternative BERT models, 3 classical ML models (support vector machine, logistic regression, and decision tree), and an ICD-code algorithm. The BERT models used original BERT and BioClinical BERT embeddings, while the ML models used bag-of-words and term frequency-inverse document frequency on the 1-day data set. The ICD method relied on administrative codes rather than text. Our neural network model demonstrated superior inpatient fall detection across different methods and data sources.

Our high sensitivity model exhibited 97.7% sensitivity, enabling near-perfect capture of relevant notes, along with 82.3% accuracy, but a low 26.8% F_1-score. The high PPV model achieved 97.5% accuracy, 85.7 % PPV, and 27.9 % sensitivity. The high F_1-model balanced 66.7% sensitivity and 60.5% PPV to optimize 64.4% F_1-score and 97.7% accuracy. In comparison, the ICD-based method had 27.9% sensitivity, while traditional classifiers achieved 51.2%-76.7% sensitivities and 8.3-15.8 PPVs.

The result of the probability-based threshold adjustment in accordance with PPV, sensitivity, and F_1-score is shown in Figure 2. By adjusting the classification threshold, we can control the trade-off between sensitivity and precision (as Figure 3).

Table 2. Performance of proposed deep learning models, classical machine learning methods, and International Classification of Diseases–based algorithms on fall identification with 1-day data set.

Category and model name	Sensitivity (%), (95% CI)	Specificity (%), (95% CI)	PPV[a] (%), (95% CI)	NPV[b] (%), (95% CI)	Accuracy (%), (95% CI)	F_1-score[c] (%)
BERT[d]-based models						
AHN-BERT[e] (high sensitivity)	97.7 (87.7-99.9)	81.8 (79.6-83.9)	15.6 (14.0-17.3)	99.9 (99.3-100.0)	82.3 (80.1-84.4)	26.8
AHN-BERT (high PPV)	27.9 (15.3-43.7)	99.8 (99.4-100.0)	85.7 (57.2-98.2)	97.6 (96.6-98.4)	97.5(96.5-98.2)	42.1
AHN-BERT (high F_1)	66.7 (49.8-80.9)	99.0 (98.2-99.5)	60.5 (44.4-75.0)	98.7 (97.8-99.2)	97.7 (96.7-98.4)	63.4
BERT-uncased	79.1 (64.0-9 0.0)	61.4 (58.6-64.1)	6.6 (5.6-7.7)	98.8 (97.9-99.4)	62.0 (59.3-64.6)	12.1
BioClinical BERT	74.4 (58.8-86.5)	69.8 (67.2-72.4)	7.8 (6.5-9.3)	98.8 (97.9-99.3)	70.0 (67.4-72.5)	14.1
Classical machine learning classifier						
Support vector machine	76.7 (61.4-88.2)	85.0 (82.9-86.9)	14.9 (12.5-17.8)	99.1 (98.4-99.5)	84.7 (82.6-86.6)	25.0
Logistic regression	74.4 (58.8-86.5)	86.4 (84.3-88.2)	15.8 (13.0-19.0)	99.0 (98.3-99.4)	86.0 (83.9-87.8)	26.0
Decision tree	51.2 (35.5-66.7)	93.9 (92.4-95.1)	22.2 (14.5-31.7)	98.3 (97.3-98.9)	92.4 (90.9-93.8)	31.0
Rule-based classifier						
ICD 10[f]	27.9 (15.3-43.7)	92.3 (90.7-93.8)	11.1 (6.9-17.3)	97.4 (96.9-97.8)	90.2 (88.5-91.8)	15.9

[a]PPV: positive predictive value. The proportion of true positive results among all positive results.

[b]NPV: negative predictive value. The proportion of true negative results among all negative results.

[c]F_1-score: a measure of a model's accuracy that considers both sensitivity and PPV.

[d]BERT: Bidirectional Encoder Representations from Transformers.

[e]AHN-BERT: Alberta hospital notes-specific BERT.

[f]ICD 10: International Classification of Diseases, 10th Revision.

Figure 2. Performance metrics at varying thresholds: PPV, sensitivity, and F1-score. PPV: positive predictive value.

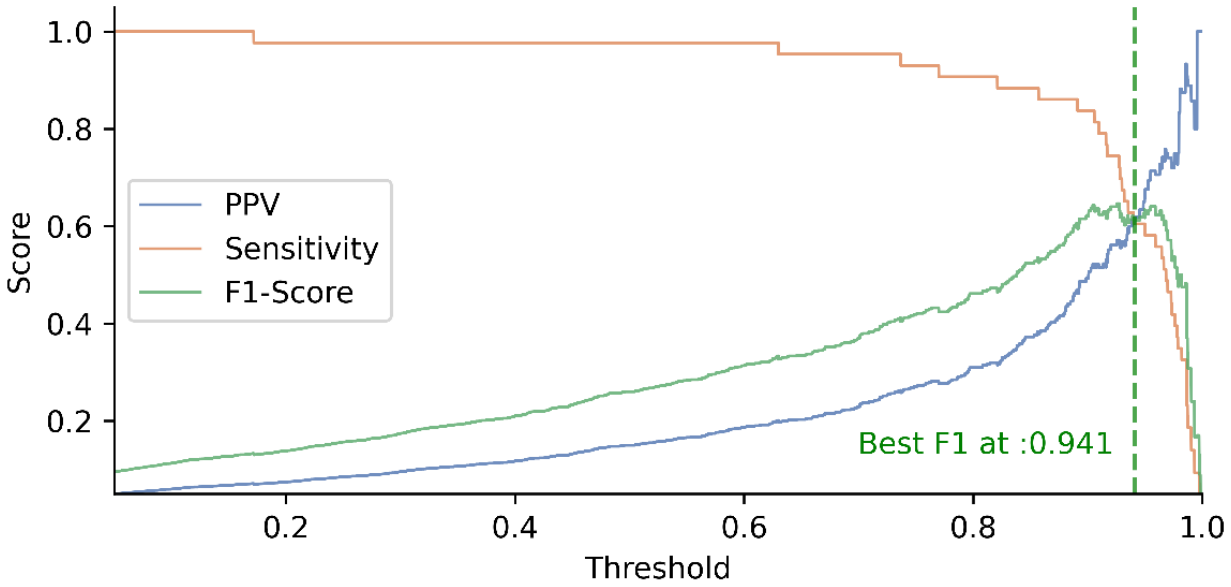
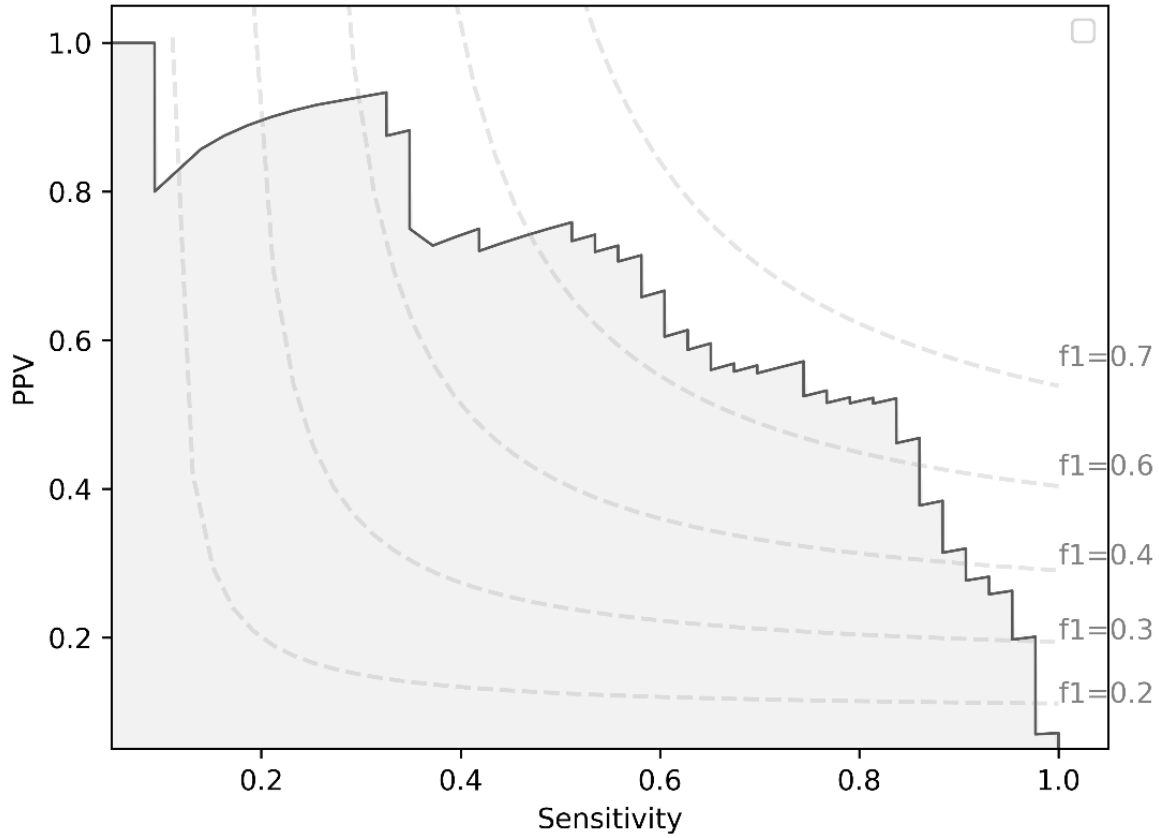


Figure 3. Percision-recall curve. PPV: positive predictive value.



Discussion

Principal Results

In our study, we illustrate how a BERT-based model substantially outperforms the ICD-based algorithm in fall detection within the hospital setting. This superiority stems from the model's ability to process EMR text data in real time, enabling rapid identification of falls. In contrast, ICD codes are assigned retrospectively, leading to delays in fall detection and intervention. Using BERT's advanced NLP facilitates accurate, efficient, and generalizable analysis of clinical notes for surveillance applications.

Specifically, our proposed AHN-BERT model surpasses generic BERT, conventional ML, and ICD-codes. Fine-tuning BERT on local hospital notes better captures local domain-specific language and context, boosting performance. Additionally, combining high-sensitivity and high-PPV models enables optimized 2-stage fall detection by adjusting decision thresholds to balance false positives and negatives. This provides flexibility for different use cases and challenging tasks.

Furthermore, our study provides valuable insights into the optimal time frame for defining falls empirically. This comparison sheds light on the potential benefits of using a finer-grained time interval, which could improve the generalizability and applicability of the model across different populations and settings. Understanding the optimal period for detecting fall incidents can guide the development and implementation of targeted public health interventions. Health surveillance data can be used to evaluate the effectiveness of these interventions and inform future strategies for fall prevention and management. By determining the most suitable time interval for defining fall incidents, health surveillance systems can better allocate resources to areas with a higher risk of falls. This may result in more efficient and effective public health efforts, improving health outcomes for at-risk populations.

Applications

Our models leverage unstructured EMR data to accurately detect inpatient falls, enabling health care systems to enact tailored prevention measures and reduce fall-associated injuries. The automation of extensive clinical documentation review accelerates health care surveillance and quality improvement processes.

Regarding research applications, our algorithms can extract comprehensive fall data from EMR text to support developing evidence-based interventions.

The proposed framework has broad applicability beyond fall detection for tasks like diagnosis prediction, medication adherence monitoring, and adverse drug event identification. This adaptability improves health care outcomes, patient safety, and quality of care.

Strength and Limitations

In our research, the AHN-BERT model has shown remarkable superiority over traditional ICD-based algorithms in fall detection within hospital environments. This enhanced performance is primarily attributed to the model's proficiency in processing and understanding the nuances of EMRs text data. Unlike ICD codes, which can sometimes result in undercoding

or loss of information, the nursing notes processed by our model are more closely aligned with the actual circumstances of inpatient falls. The ability of AHN-BERT to immediately and accurately process this data is a substantial advancement, ensuring that fall detection is not only more precise but also more reflective of the true clinical scenario. Additionally, the combination of high-sensitivity and high-PPV models in our 2-stage fall detection system allows for adjustable decision thresholds, thus balancing false positives and negatives and providing flexibility across different scenarios.

However, the model faces challenges in balancing high sensitivity with a high PPV due to the imbalanced nature of clinical data. The rarity of AEs like falls leads to a higher rate of false positives, as seen in our data set with a significant imbalance ratio. Our test data set, characterized by a significant 29:1 imbalance, aligns more closely with real-world clinical scenarios than balanced data sets used in some prior studies [27], which, while yielding promising results, may not fully represent practical conditions. This intentional choice ensures that our model's performance is tested under conditions typical of rare events like falls, thereby enhancing its relevance and utility in actual clinical settings.

Second, the effectiveness of our models depends on the quality and comprehensiveness of documentation. If fall events or associated risk factors are not well documented, our model, like any data-driven model, may have difficulty detecting them. This underscores the importance of careful, detailed clinical documentation to enhance the effectiveness of monitoring applications. In addition, our study also assumes a certain level of linguistic and terminological consistency within the EMR data. Variations in documentation styles across different health care providers could potentially impact the model's performance, suggesting that future models should incorporate strategies, for example, pretraining the ML, to mitigate such discrepancies. Last, the differentiation between a history of falls and inpatient falls presents a challenge, as it could potentially lead to false positive predictions if falls that occurred prior to hospitalization are documented in the notes. Although the BERT model's contextual understanding can partially alleviate this issue, we acknowledge that more improvements are needed. As part of our future work, we aim to further refine our model to better handle such complexities.

Conclusions

This study developed and evaluated BERT-based NLP models for the automated detection of falls from electronic clinical notes. The developed models provided a more accurate and timely way to detect falls than traditional ML and ICD-codes–based methods. Moreover, we provided a masked language model technique to pretrain a pre-existing BERT model using clinical text data gathered from various health care facilities in Calgary, Alberta, creating a more local institution-specific and effective AHN-BERT model. By using self-supervised language modeling strategies, we can bypass steps that were regarded as vital in standard ML methods, such as the necessity for thorough text preprocessing, complex feature engineering, and a considerable amount of labeled data. In addition, by exploring the optimal period for fall incident detection and selecting 1-day notes for our final architecture, our model contributes to enhanced patient safety and care with less noise.

Acknowledgments

YX and CAE received research support funding from Canadian Institutes of Health Research through grant number DC0190GP. GW was supported by the Canadian Institutes of Health Research postdoctoral fellowship, O'Brien Institute for Public Health Postdoctoral Scholarship, Cumming School of Medicine Postdoctoral Scholarship at the University of Calgary, and the Network of Alberta Health Economists Postdoctoral Fellowship at the University of Alberta.

Authors' Contributions

YX, CAE, HQ, GW, and CC were responsible for the study planning, conceptualization, and coordination. SL, EAM, and GW managed data retrieval, data linkage, and data quality assurance. The design and development of the neural network architecture were carried out by CC and YX. CC conducted clinical note preprocessing, analysis, and model evaluation. The reference standard development and chart review study design was executed by YX, CAE, NS, DAS, and GW. SL and CC drafted the manuscript and GW drafted the Methods (Study Cohort and Data Sources). GW, JP, DAS, EAM, CAE, NS, and YX participated in discussions and provided comments on the manuscript. All authors contributed to the revision and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics for the study cohort from each hospital.
[DOCX File , 19 KB - medinform_v12i1e48995_app1.docx]

References

1. Morris R, O'Riordan S. Prevention of falls in hospital. Clin Med (Lond) 2017;17(4):360-362 [FREE Full text] [doi: 10.7861/clinmedicine.17-4-360] [Medline: 28765417]

2. Schwendimann R, Bühler H, De Geest S, Milisen K. Characteristics of hospital inpatient falls across clinical departments. Gerontology 2008;54(6):342-348. [doi: 10.1159/000129954] [Medline: 18460867]
3. Zeneli A, Montalti S, Masciangelo I, Manieri G, Golinucci M, Nanni O, et al. Fall predictors in hospitalized patients living with cancer: a case-control study. Support Care Cancer 2022;30(10):7835-7843. [doi: 10.1007/s00520-022-07208-x] [Medline: 35705752]
4. Oliver D, Healey F, Haines TP. Preventing falls and fall-related injuries in hospitals. Clin Geriatr Med 2010;26(4):645-692. [doi: 10.1016/j.cger.2010.06.005] [Medline: 20934615]
5. Morello RT, Barker AL, Watts JJ, Haines T, Zavarsek SS, Hill KD, et al. The extra resource burden of in-hospital falls: a cost of falls study. Med J Aust 2015;203(9):367. [doi: 10.5694/mja15.00296] [Medline: 26510807]
6. Miake-Lye IM, Hempel S, Ganz DA, Shekelle PG. Inpatient fall prevention programs as a patient safety strategy: a systematic review. Ann Intern Med 2013;158(5 Pt 2):390-396 [FREE Full text] [doi: 10.7326/0003-4819-158-5-201303051-00005] [Medline: 23460095]
7. King B, Pecanac K, Krupp A, Liebzeit D, Mahoney J. Impact of fall prevention on nurses and care of fall risk patients. Gerontologist 2018;58(2):331-340 [FREE Full text] [doi: 10.1093/geront/gnw156] [Medline: 28011591]
8. Rubenstein LZ. Falls in older people: epidemiology, risk factors and strategies for prevention. Age Ageing 2006;35(Suppl 2):ii37-ii41 [FREE Full text] [doi: 10.1093/ageing/afl084] [Medline: 16926202]
9. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36(1-2):131-143 [FREE Full text] [doi: 10.1016/j.jbi.2003.08.003] [Medline: 14552854]
10. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med 1991 Feb 07;324(6):370-376. [doi: 10.1056/NEJM199102073240604] [Medline: 1987460]
11. Griffin FA, Resar RK. IHI Innovation Series white paper. IHI Global Trigger Tool for Measuring Adverse Events (Second Edition). Cambridge, MA: Institute for Healthcare Improvement; 2009. URL: https://www.ihi.org/ [accessed 2023-03-15]
12. Southern DA, Burnand B, Droesler SE, Flemons W, Forster AJ, Gurevich Y, et al. Deriving ICD-10 codes for patient safety indicators for large-scale surveillance using administrative hospital data. Med Care 2017;55(3):252-260. [doi: 10.1097/MLR.0000000000000649] [Medline: 27635599]
13. Menendez ME, Ring D, Jawa A. Inpatient falls after shoulder arthroplasty. J Shoulder Elbow Surg 2017;26(1):14-19. [doi: 10.1016/j.jse.2016.06.008] [Medline: 27522341]
14. Memtsoudis SG, Danninger T, Rasul R, Poeran J, Gerner P, Stundner O, et al. Inpatient falls after total knee arthroplasty: the role of anesthesia type and peripheral nerve blocks. Anesthesiology 2014;120(3):551-563 [FREE Full text] [doi: 10.1097/ALN.0000000000000120] [Medline: 24534855]
15. Schroll JB, Maund E, Gøtzsche PC. Challenges in coding adverse events in clinical trials: a systematic review. PLoS One 2012;7(7):e41174 [FREE Full text] [doi: 10.1371/journal.pone.0041174] [Medline: 22911755]
16. Golder S, Loke YK, Wright K, Norman G. Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. PLoS Med 2016;13(9):e1002127 [FREE Full text] [doi: 10.1371/journal.pmed.1002127] [Medline: 27649528]
17. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng 2018;2018:4302425 [FREE Full text] [doi: 10.1155/2018/4302425] [Medline: 29849998]
18. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885 [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]
19. Chen T, Dredze M, Weiner JP, Hernandez L, Kimura J, Kharrazi H. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. JMIR Med Inform 2019;7(1):e13039 [FREE Full text] [doi: 10.2196/13039] [Medline: 30862607]
20. Mahajan D, Poddar A, Liang JJ, Lin Y, Prager JM, Suryanarayanan P, et al. Identification of semantically similar sentences in clinical notes: iterative intermediate training using multi-task learning. JMIR Med Inform 2020;8(11):e22508 [FREE Full text] [doi: 10.2196/22508] [Medline: 33245284]
21. Arnaud É, Elbattah M, Gignon M, Dequen G. Learning embeddings from free-text triage notes using pretrained transformer models. 2022 Presented at: Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies—Volume 5 HEALTHINF: Scale-IT-up; February 9-11, 2022; Vienna, Austria p. 835-841. [doi: 10.5220/0011012800003123]
22. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: a review. Trends Anaesth Crit Care 2021;38:4-9. [doi: 10.1016/j.tacc.2021.02.007]
23. Yan H, Rahgozar A, Sethuram C, Karunananthan S, Archibald D, Bradley L, et al. Natural language processing to identify digital learning tools in postgraduate family medicine: protocol for a scoping review. JMIR Res Protoc 2022;11(5):e34575 [FREE Full text] [doi: 10.2196/34575] [Medline: 35499861]

24. Gaviria-Valencia S, Murphy SP, Kaggal VC, McBane Ii RD, Rooke TW, Chaudhry R, et al. Near real-time natural language processing for the extraction of abdominal aortic aneurysm diagnoses from radiology reports: algorithm development and validation study. JMIR Med Inform 2023;11:e40964 [FREE Full text] [doi: 10.2196/40964] [Medline: 36826984]
25. Dolci E, Schärer B, Grossmann N, Musy SN, Zúñiga F, Bachnick S, et al. Automated fall detection algorithm with global trigger tool, incident reports, manual chart review, and patient-reported falls: algorithm development and validation with a retrospective diagnostic accuracy study. J Med Internet Res 2020;22(9):e19516 [FREE Full text] [doi: 10.2196/19516] [Medline: 32955445]
26. Toyabe SI. Detecting inpatient falls by using natural language processing of electronic medical records. BMC Health Serv Res 2012;12:448 [FREE Full text] [doi: 10.1186/1472-6963-12-448] [Medline: 23217016]
27. Fu S, Thorsteinsdottir B, Zhang X, Lopes GS, Pagali SR, LeBrasseur NK, et al. A hybrid model to identify fall occurrence from electronic health records. Int J Med Inform 2022;162:104736 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104736] [Medline: 35316697]
28. Jung H, Park HA, Hwang H. Improving prediction of fall risk using electronic health record data with various types and sources at multiple times. Comput Inform Nurs 2020;38(3):157-164. [doi: 10.1097/CIN.0000000000000561] [Medline: 31498252]
29. Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: case-control study. JMIR Med Inform 2020;8(4):e16970 [FREE Full text] [doi: 10.2196/16970] [Medline: 32319959]
30. Thapa R, Garikipati A, Shokouhi S, Hurtado M, Barnes G, Hoffman J, et al. Predicting falls in long-term care facilities: machine learning study. JMIR Aging 2022;5(2):e35373 [FREE Full text] [doi: 10.2196/35373] [Medline: 35363146]
31. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), Vol 1; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186.
32. Li J, Zhang X, Zhou X. ALBERT-based self-ensemble model with semisupervised learning and data augmentation for clinical semantic textual similarity calculation: algorithm validation study. JMIR Med Inform 2021;9(1):e23086 [FREE Full text] [doi: 10.2196/23086] [Medline: 33480858]
33. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop.: Association for Computational Linguistics; 2019 Presented at: The 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/W19-1909]
34. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015;162(10):735-736 [FREE Full text] [doi: 10.7326/L15-5093-2] [Medline: 25984857]
35. Lee S, Xu Y, Apos Souza AGD, Martin EA, Doktorchik C, Zhang Z, et al. Unlocking the potential of electronic health records for health research. Int J Popul Data Sci 2020;5(1):1123 [FREE Full text] [doi: 10.23889/ijpds.v5i1.1123] [Medline: 32935049]
36. Eastwood CA, Southern DA, Khair S, Doktorchik C, Cullen D, Ghali WA, et al. Field testing a new ICD coding system: methods and early experiences with ICD-11 beta version 2018. BMC Res Notes 2022;15(1):343 [FREE Full text] [doi: 10.1186/s13104-022-06238-2] [Medline: 36348430]
37. Qiao Y, Xiong C, Liu Z, Liu Z. Understanding the behaviors of BERT in ranking. arXiv :1-4 Preprint posted online on April 16, 2019. [FREE Full text] [doi: 10.1090/mbk/121/79]
38. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
40. McKinney W. Data structures for statistical computing in Python. 2010 Presented at: Proceedings of the 9th Python in Science Conference (SciPy 2010); June 28-30, 2010; Austin, TX. [doi: 10.25080/majora-92bf1922-00a]
41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 2019;32:1-12 [FREE Full text]

Abbreviations

AE: adverse event
AHN-BERT: Alberta hospital notes-specific BERT
BERT: Bidirectional Encoder Representation from Transformers
DAD: discharge abstract database
EMR: electronic medical record
ICD: International Classification of Diseases
ML: machine learning

MPR: multidisciplinary progress record
NLP: natural language processing
PPV: positive predictive value
SCM: Sunrise Clinical Manager

Edited by C Lovis; submitted 15.05.23; peer-reviewed by S Musy, M Elbattah; comments to author 05.07.23; revised version received 24.07.23; accepted 23.12.23; published 30.01.24.

Please cite as:
Cheligeer C, Wu G, Lee S, Pan J, Southern DA, Martin EA, Sapiro N, Eastwood CA, Quan H, Xu Y
BERT-Based Neural Network for Inpatient Fall Detection From Electronic Medical Records: Retrospective Cohort Study
JMIR Med Inform 2024;12:e48995
URL: https://medinform.jmir.org/2024/1/e48995
doi:10.2196/48995
PMID:38289643

Original Paper

Mining Clinical Notes for Physical Rehabilitation Exercise Information: Natural Language Processing Algorithm Development and Validation Study

Sonish Sivarajkumar[1], BS; Fengyi Gao[2], MS; Parker Denny[3], DPT; Bayan Aldhahwani[3,4], MS, PT; Shyam Visweswaran[1,5,6], MD, PhD; Allyn Bove[3], DPT, PhD; Yanshan Wang[1,2,5,6], PhD

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, United States

[2]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, United States

[3]Department of Physical Therapy, University of Pittsburgh, Pittsburgh, PA, United States

[4]Department of Physical Therapy, Umm Al-Qura University, Makkah, Saudi Arabia

[5]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[6]Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Yanshan Wang, PhD
Department of Health Information Management
University of Pittsburgh
6026 Forbes Tower
Pittsburgh, PA, 15260
United States
Phone: 1 4123832712
Email: yanshan.wang@pitt.edu

Abstract

Background: The rehabilitation of a patient who had a stroke requires precise, personalized treatment plans. Natural language processing (NLP) offers the potential to extract valuable exercise information from clinical notes, aiding in the development of more effective rehabilitation strategies.

Objective: This study aims to develop and evaluate a variety of NLP algorithms to extract and categorize physical rehabilitation exercise information from the clinical notes of patients who had a stroke treated at the University of Pittsburgh Medical Center.

Methods: A cohort of 13,605 patients diagnosed with stroke was identified, and their clinical notes containing rehabilitation therapy notes were retrieved. A comprehensive clinical ontology was created to represent various aspects of physical rehabilitation exercises. State-of-the-art NLP algorithms were then developed and compared, including rule-based, machine learning–based algorithms (support vector machine, logistic regression, gradient boosting, and AdaBoost) and large language model (LLM)–based algorithms (ChatGPT [OpenAI]). The study focused on key performance metrics, particularly F_1-scores, to evaluate algorithm effectiveness.

Results: The analysis was conducted on a data set comprising 23,724 notes with detailed demographic and clinical characteristics. The rule-based NLP algorithm demonstrated superior performance in most areas, particularly in detecting the "Right Side" location with an F_1-score of 0.975, outperforming gradient boosting by 0.063. Gradient boosting excelled in "Lower Extremity" location detection (F_1-score: 0.978), surpassing rule-based NLP by 0.023. It also showed notable performance in the "Passive Range of Motion" detection with an F_1-score of 0.970, a 0.032 improvement over rule-based NLP. The rule-based algorithm efficiently handled "Duration," "Sets," and "Reps" with F_1-scores up to 0.65. LLM-based NLP, particularly ChatGPT with few-shot prompts, achieved high recall but generally lower precision and F_1-scores. However, it notably excelled in "Backward Plane" motion detection, achieving an F_1-score of 0.846, surpassing the rule-based algorithm's 0.720.

Conclusions: The study successfully developed and evaluated multiple NLP algorithms, revealing the strengths and weaknesses of each in extracting physical rehabilitation exercise information from clinical notes. The detailed ontology and the robust performance of the rule-based and gradient boosting algorithms demonstrate significant potential for enhancing precision

rehabilitation. These findings contribute to the ongoing efforts to integrate advanced NLP techniques into health care, moving toward predictive models that can recommend personalized rehabilitation treatments for optimal patient outcomes.

(JMIR Med Inform 2024;12:e52289) doi:10.2196/52289

KEYWORDS

natural language processing; electronic health records; rehabilitation; physical exercise; ChatGPT; artificial intelligence; stroke; physical rehabilitation; rehabilitation therapy; exercise; machine learning

Introduction

Precision medicine is a promising field of research that aims to provide personalized treatment plans for patients [1]. Recent years have seen a rise in interest in this field, as advances in machine learning and data collection techniques have greatly facilitated this research [2]. However, the principles of precision medicine have primarily been applied to the development of medications, and relatively little research has been conducted on their applications in other areas [3]. For instance, although rehabilitation clinics require individualized treatment procedures for patients, little research has been conducted on methods that use data analysis and machine learning to facilitate the design of such procedures [4]. Although the application of precision medicine to physical therapy has proven effective in improving the health of patients, current methods of creating personalized treatments rarely use automated approaches to facilitate decision support [5]. Thus, there is a need for tools to assist in the development of personalized treatments in physical therapy [6]. In the treatment of patients who had a stroke, the lack of decision support tools is especially pronounced, as the available treatments for this condition have not led to consistent outcomes across patient populations [7].

To develop decision support tools for the design of precision rehabilitation treatments for patients who had a stroke, it would be necessary to use electronic health record data to develop a predictive model of existing treatment options and their impact on patient outcomes [8]. However, physical therapy procedures are typically described in unstructured clinical notes, meaning that simple data extraction methods such as database queries cannot be applied to obtain sufficient information. Additionally, the language used to describe these procedures can differ between clinicians, locations, and periods [9]. More advanced natural language processing (NLP) algorithms are required to extract this information from clinical notes, but such a method has not yet been developed for this application.

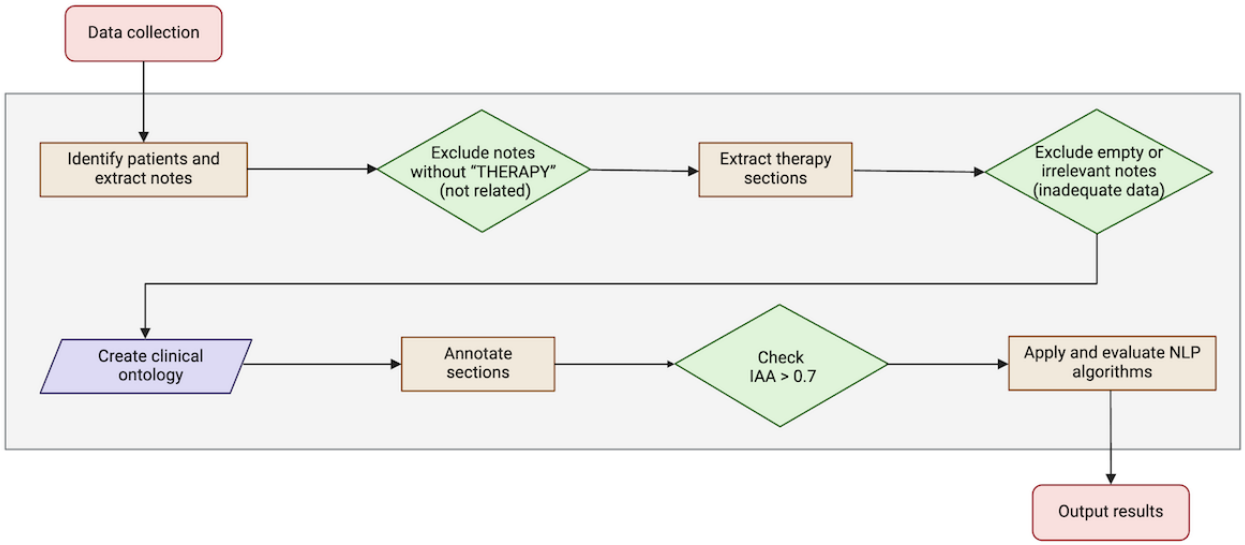
In this paper, we aim to develop and evaluate NLP algorithms to extract physical rehabilitation exercise information from the clinical notes in the electronic health record. Our specific contributions are as follows. First, we created a novel and comprehensive clinical ontology to represent physical rehabilitation exercise information, which includes the type of motion, side of the body, location on the body, the plane of motion, duration, information on sets and reps, exercise purpose, exercise type, and body position. Second, we developed and compared a variety of NLP algorithms leveraging state-of-the-art techniques, including rule-based NLP algorithms, machine learning–based NLP algorithms (ie, support vector machine [SVM], logistic regression [LR], gradient boosting, and AdaBoost), and large language model (LLM)–based NLP algorithms (ie, ChatGPT [OpenAI] [10]) for the extraction of physical rehabilitation exercise from clinical notes. We are among the first to evaluate the capabilities of ChatGPT in extracting useful information from clinical notes.

Methods

Overview

Figure 1 illustrates the data flow and the various stages of the research process. Each of these stages will be described in detail in the following sections.

Figure 1. Flowchart illustrating the data flow throughout the study. IIA: interannotator agreement (IAA); NLP: natural language processing.



Data Collection

The study identified a cohort of patients diagnosed with stroke between January 1, 2016, and December 31, 2016, at University of Pittsburgh Medical Center (UPMC). For these patients, clinical encounter notes created between January 1, 2016, and December 31, 2018, were extracted from the institutional data warehouse. Table 1 provides the demographic characteristics of the patients included in this data set.

Table 1. Demographic information of patients included in the unfiltered data set (N=13,605).

Demographics	Values
Age (years), mean (SD)	75 (16)
Gender, n (%)	
Female	6931 (51)
Male	6673 (49)
Race, n (%)	
Asian	64 (0.5)
Black	1325 (9.7)
White	11,661 (86)
Other	153 (1.1)
Not specified	402 (3)
Ethnicity, n (%)	
Hispanic or Latinx	64 (0.5)
Not Hispanic or Latinx	12,471 (92)
Not specified	984 (7.2)

Ethical Considerations

The study was approved by the University of Pittsburgh's institutional review board (#21040204).

Clinical Ontology for Physical Rehabilitation Exercise

To determine the relevance and hierarchy of extracted information, we developed a clinical ontology consisting of 9 categories of concepts relating to exercise descriptions, informed by consultation with clinical experts (PD, BA, and AB) in the field of physical therapy. In developing our clinical ontology, we also consulted established frameworks such as the International Classification of Functioning, Disability, and Health (ICF) [11] and the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [12]. These comprehensive systems offered valuable insights into the structuring and categorization of health-related concepts, which we adapted for the specific context of physical rehabilitation exercises. Additionally, our ontology incorporates principles from the Unified Medical Language System (UMLS) [13] to ensure compatibility and interoperability with other health care informatics systems.

Each category was given a set of values, as well as examples of how those values might be expressed in clinical notes. The categories are type of motion, side of the body, location on the body, the plane of motion, duration, information on sets and reps, exercise purpose, exercise type, and body position. The ontology also includes examples of indications that the mentioned exercise was not performed during the visit corresponding to the clinical note. This ontology was used to inform both the structure of the annotations and the methods used to extract relevant documents from the data set.

The ontology reflects the complexity and nuance of physical rehabilitation exercises by incorporating terms and categories that are sensitive to the variations and specificities observed in clinical settings. This approach ensures that the ontology not only represents the theoretical model of rehabilitation exercises but also aligns with the practical, real-world application and documentation by health care professionals. Table 2 displays the 9 categories for 3 exercise descriptions (performed in-office, home exercise program, and not performed), with sets and reps split into separate rows and including negations and out-of-office exercises at the bottom.

Table 2. Summary of the clinical ontology used for annotations.

Category	Data type	Concepts
Exercise description	Enumerated	Performed in-office, home exercise program, not performed
Type of motion	Enumerated	ROM[a], active ROM, active-assisted ROM, and passive ROM
Side of body	Enumerated	Right, left, bilateral, unilateral, contralateral, and ipsilateral
Location on body	Enumerated	Upper extremity (arms), lower extremity (legs), hip, thigh, knee, ankle, foot, heel, toe, shoulder, scapula, elbow, forearm, wrist, hand, thumb, head, neck, chest, abdomen, and lower back
Plane of motion	Enumerated	Flexion, extension, abduction, adduction, internal rotation, external rotation, lateral flexion, horizontal abduction, horizontal adduction, protraction, retraction, elevation, depression, inversion, eversion, pronation, supination, plantarflexion, dorsiflexion, radial deviation, ulnar deviation, upward rotation, downward rotation, opposition, forward, backward, lateral, medial, scaption, rotation, closure, clockwise, counterclockwise, distraction, all planes, anterior, posterior, horizontal, vertical, diagonal, and gravity elimination
Duration (seconds)	Integer	N/A[b]
Number of sets	Integer	N/A
Number of reps	Integer	N/A
Exercise purpose	Enumerated	Strength, fine motor, motor control, perception, simulated, power, endurance, joint mobility, joint alignment, muscle flexibility, cardio, pulmonary, agility, and vestibular
Exercise type	Enumerated	Upper extremity strength, lower extremity strength, trunk or core strength, scapular strength, ROM, flexibility or mobility, balance or vestibular, gait training, cardio or aerobic, and functional mobility
Body position	Binary	Weight bearing and non-weight bearing
Negation or hypothetical	Binary	Held or not performed and home exercise program

[a]ROM: range of motion.

[b]N/A: not applicable.

Preprocessing and Section Extraction

Physical therapeutic procedures were usually documented in the section "THERAPY." Therefore, we first filtered out the notes that did not contain a physical therapy visit by excluding files whose names lacked the string "THERAPY." From the resulting set of files, the section on therapeutic procedures was extracted using a regular expression, if such a section existed. This resulted in a total of 23,724 notes, some of which were empty or lacked pertinent information.

The method of section extraction has a few minor limitations. Because the regular expression used to locate these sections assumes a structure in the notes that is not always present, it is possible that a file may contain additional text from other sections of the original note in rare instances. All sections used in the creation of the gold-standard labels were manually examined to ensure the absence of these errors. It is also possible that some therapeutic procedures' sections are completely omitted from the note due to copy-and-paste errors made by their authors.

Because many of the extracted sections were very brief or lacked relevant information, we developed a method to create a more robust set of sections by extracting keywords. Initially, concepts were organized into 9 categories based on the clinical ontology. Each category was then assigned a list of keywords. A section was considered to mention a category if it contained at least 1 of the keywords. Consequently, each section was assigned a score between 0 and 9 based on the number of categories mentioned. All sections with a score of 9 and a random selection of notes with a score of 8 were extracted to generate 300 enriched sections that were anticipated to be relatively dense in information. In addition, 300 random sections were selected,

excluding those with a length of fewer than 200 characters in order to reduce the likelihood of omissions.

Gold-Standard Data Set Creation

Gold standard labels were developed by 2 clinical experts in the field of physical therapy (PD and BA) under the supervision of a senior clinical expert in physical therapy (AB). Each annotator was given a set of guidelines on how to label sections and was told to refer to the clinical ontology for examples of each concept to label. Instructions were given to label explicit mentions of each concept, and inferences were only to be made when specified. For example, the concepts under the categories exercise type and positioning were each given several common keywords that indicate exercises that relate to them. The annotators were given identical batches of 20 randomly selected sections to annotate, and the interannotator agreement was calculated using Fleiss κ. This process was repeated for a total of 3 batches, after which all 3 annotators achieved an interannotator agreement greater than 0.7. Throughout this process, the annotation guidelines were revised, and the structure of the labels was finalized. Once sufficient agreement was reached, 50 sections from the enriched set and 50 more from the random set were given to each annotator, totaling 300 distinct annotated sections. These sections were then split randomly into a training set consisting of 125 sections from each of the original sets and a test set consisting of the remaining 50 sections. The details of this corpus are included in Textbox 1, which outlines the total word count, the number of distinct words, and 2 examples of the data.

Textbox 1. Summary of the annotated corpus.

Total words: 74,104

Total distinct words: 2371

Deidentified note example 1:

- "1: AROM right elbow flx/ext HEP (right arm supported on table) 2: AROM right wrist flx/ext HEP 3: AROM right forearm pronation/supination HEP 4: Thumb opposition HEP 5: Seated AAROM table slide??"

Deidentified note example 2:

- "1: foam balance (heel/toe rocking): x 30 2: step taps with 2 taps from foam 12""" block: x 20 B/L 3: tandem walking: 25' x 2 4: backward walking: 25' x 2 5: foam Lunges: x 20 B/L 6: Dips 4""" stair: 2x10 B/L 7: side stepping green TB 10 ft x5 each direction 9: bridging with LLE leg lift 1""" off mat x10 10: tandem stance on foam x 1' 11: Nustep: L5 x 10' (LEs only)"

Rule-Based NLP

The first NLP method we developed was a named entity recognition (NER) algorithm using MedTagger (OHNLP), which is a software that uses rule-based methods to segment documents and extract named entity information with regular expressions [14]. We used this tool to detect the categories outlined in the ontology by creating explainable rules to extract the physical rehabilitation exercise information and compare it against the gold-standard labels. For each rule defined in the algorithm, MedTagger identified spans of text that matched the expression as well as the corresponding category and concept predicted for that text. We initiated the rules using simple keywords in the clinical ontology as defined in Table 2 and then refined the rules using the training set of the gold-standard notes.

Machine Learning–Based NLP

In addition to attempting to automate the annotation of clinical notes with exercise information, several sequence-level binary classification methods were explored to predict whether a specific concept is mentioned in a given span of text at least once according to the gold-standard labels. Here, a sequence is defined as a string of text within a section that describes an individual exercise. As the therapeutic procedures are documented as numbered lists, it is assumed that each enumerated item that contains text constitutes a single procedure for the purpose of this study. The aim was to extract these procedures from sections and then classify each according to which concepts they mention.

For this task, the sequences provided in the gold-standard data were used as raw input, and targets were defined using the labels that were associated with each sequence. These labels consisted of 101 concepts as given by the clinical ontology in Table 2, excluding duration, sets, and reps since these are numeric types unfit for binary classification tasks. Because the postprocessed output from MedTagger was formatted in a similar manner to the gold-standard data for ease of comparison, a similar method was used to create predictions and directly score MedTagger against the true labels for this task. In this manner, we compared our rule-based NLP algorithm against several other methods by redefining the information extraction task as a sequence classification task. The labels of all predicted spans of text were assigned to the section containing it.

A total of 4 machine learning models were trained to perform binary classification on sections, including SVM [15], LR [16], gradient boosting [17], and AdaBoost [18]. We built different machine learning models for different physical rehabilitation exercise concept extraction tasks. This resulted in 101 distinct SVM, LR, gradient boosting, and AdaBoost models each trained to predict a distinct concept. Each model was created using the *scikit-learn* [19] library in Python (version 3; Python Software Foundation). The input for each model was given in a simple uncased bag-of-words vector space fitted to the training set. The LR was performed with a learning rate of 1×10^{-4} and balanced class weights. The SVM model used a polynomial kernel with a degree of 2 and also used balanced class weights. The AdaBoost and gradient boosting were performed with the default parameters provided by *scikit-learn*, with 100 and 50

estimators, respectively. All unspecified hyperparameters were kept at the default values used by *scikit-learn*.

LLM-Based NLP

Recently, LLMs have gained much interest due to their promising results across many NLP tasks and straightforward development pipelines. To measure a baseline for the performance of LLMs on this data set, this study used OpenAI's ChatGPT, specifically GPT-3.5-turbo [20], which serves as the NLP backend for the first release of ChatGPT. This study accessed ChatGPT through the Microsoft Azure secure cloud service with an approved institutional review board.

For each of the concepts to be identified, a separate prompt was created asking the model to identify whether or not that concept is mentioned in the given sections. We note that the sections used in this exercise were fully deidentified and contained no sensitive information. One of the primary advantages of using LLMs to perform NLP tasks is that they can yield results comparable to other methods with much less time spent on fine-tuning and other manual design tasks. Therefore, in order to evaluate the practicality of using this method in an application with constraints on time or available machine learning expertise, prompts were automatically generated using a template and the name assigned to each label. Two simple prompting techniques were compared: a few-shot prompting technique that included example dialogue for 2 positive examples and 2 negative examples [21] and a zero-shot prompt that omitted any examples [22]. Prompts were created by providing a list of dialogue strings that can be attributed to the user providing the prompt, the model giving the response, and a third party giving instructions to the model, hereafter named System. The following is an example of a few-shot prompt that was automatically generated and sent to ChatGPT.

> *System: You are an assistant assigned to determine if a given text segment from a medical record contains mentions of Active Range of Motion. You must answer yes or no.*
>
> *User: 7: [PERSONALNAME] maze - AROM LUE - 3 rep*
>
> *Model: Yes.*
>
> *User: 23: / Supine on mat / dowel rod no wt above head for triceps*
>
> *Model: No.*
>
> *User: 2: AROM Right wrist ext/flex - 20 x*
>
> *Model: Yes.*
>
> *User: 24: SL HS activation- with tapping and suspended to limit friction x10 DEFERRED*
>
> *Model: No.*
>
> *User: 14: SLR- 2x10 deferred to HEP*

Prompts were made in this manner using 4 dialogue exchanges, switching out the last line for the section to be classified. Any concepts that were labeled fewer than 2 times were excluded from the experiment because not enough examples could be generated for the few-shot prompt. A total of 3 additional concepts were excluded relating to whether a procedure was performed in-office, at home, or not at all, as well as the 3 numeric concepts. This left a total of 82 concepts to be extracted.

Results

Because the clinical ontology covered a broad range of concepts within this domain, many of these concepts were very sparse in our data. More than half of the 101 concepts were present in fewer than 10 exercise descriptions in the train or test sets; these concepts have been omitted from the results. Table 3 contains a breakdown of the F_1-scores for each machine learning method, as well as the performance of the rule-based NLP algorithm on the NER task, for each of the remaining 40 concepts. See Multimedia Appendix 1 for the results on all concepts. The best-performing machine learning model is shown in bold for each concept.

Table 3. Binary F_1-scores of each algorithm on the test set (50 documents).

Category and concept	RBNLP[a] NER[b], n	RBNLP sequence, n	LR[c], n	SVM[d], n	AdaBoost, n	Gradient boosting, n	ChatGPT (few-shot), n	ChatGPT (zero-shot), n	Training set size, n	Test set size, n
Description										
Performed in-office	0.957	0.976	0.970	0.960	0.977	0.983[e]	N/A[f]	N/A	2464	497
Home exercise program	0.986[e]	0.986[e]	0.986[e]	0.938	0.986[e]	0.986[e]	N/A	N/A	93	34
Not performed	0.949	0.949	0.923	0.909	0.936	0.950[e]	N/A	N/A	1295	206
ROM[g]										
Active	0.839	0.830	0.824	0.840	0.863[e]	0.863[e]	0.321	0.109	103	22
Active-assisted	0.769	0.769	0.800	0.791	0.837	0.857[e]	0.543	0.210	160	24
Passive	0.952	0.938	0.970[e]	0.903	0.938	0.970[e]	0.552	0.198	121	16
Side										
Right side	0.912	0.975[e]	0.674	0.851	0.628	0.680	0.912	0.878	548	97
Left side	0.912	0.937[e]	0.763	0.823	0.721	0.752	0.823	0.832	462	134
Bilateral	0.772	0.907[e]	0.559	0.474	0.667	0.659	0.706	0.723	260	51
Location										
Upper extremity	0.847	0.939[e]	0.879	0.847	0.901	0.876	0.291	0.241	285	47
Lower extremity	0.955	0.936	0.936	0.930	0.966	0.978[e]	0.378	0.339	223	44
Hip	0.949	0.947	0.973[e]	0.973[e]	0.943	0.972	0.403	0.806	168	36
Knee	0.950	0.950	0.919	0.882	0.974[e]	0.974[e]	0.469	0.434	108	19
Ankle	1.000[e]	1.000[e]	0.923	0.600	1.000[e]	1.000[e]	0.607	0.262	55	14
Shoulder	0.936	0.977[e]	0.952	0.952	0.953	0.953	0.744	0.548	224	44
Scapula	0.833[e]	0.833[e]	0.783	0.700	0.833[e]	0.833[e]	0.525	0.607	72	10
Elbow	0.967[e]	0.963	0.963	0.943	0.923	0.923	0.848	0.447	147	26
Forearm	0.815	0.833	0.870	0.952[e]	0.870	0.952[e]	0.151	0.204	86	10
Wrist	0.902[e]	0.898	0.826	0.773	0.875	0.875	0.600	0.314	129	23
Hand	0.951[e]	0.944	0.926	0.848	0.925	0.949	0.438	0.574	243	68
Plane										
Abduction	0.976	0.985[e]	0.971	0.937	0.971	0.971	0.576	0.839	170	33
Anterior	0.545	0.545	0.750[e]	0.667	0.750[e]	0.667	0.221	0.195	22	10
Backward	0.727	0.720	0.688	0.800	0.952[e]	0.846	0.720	0.790	92	11
Extension	0.980	0.980	0.979	0.933	0.989[e]	0.989[e]	0.556	0.684	266	48
External rotation	0.897	0.917[e]	0.870	0.818	0.870	0.870	0.655	0.543	74	11
Flexion	0.956	0.947	0.964[e]	0.955	0.964[e]	0.964[e]	0.757	0.615	327	55
Forward	0.977[e]	0.974	0.857	0.865	0.950	0.900	0.667	0.729	148	19

Category and concept	RBNLP[a] NER[b], n	RBNLP sequence, n	LR[c], n	SVM[d], n	AdaBoost, n	Gradient boosting, n	ChatGPT (few-shot), n	ChatGPT (zero-shot), n	Training set size, n	Test set size, n
Lateral	0.577	0.588	0.786	0.837	0.870[e]	0.851	0.546	0.373	132	23
Supination	0.923[e]	0.917	0.880	0.917	0.917	0.917	0.550	0.480	82	11
Exercise type										
Upper extremity strength	0.913[e]	0.913[e]	0.840	0.791	0.913[e]	0.894	0.272	0.166	138	21
Lower extremity strength	0.926	0.969[e]	0.913	0.894	0.924	0.894	0.449	0.332	447	97
Trunk or core strength	0.897	0.889[e]	0.692	0.471	0.471	0.700	0.104	0.090	35	12
Range of motion	0.853	0.876[e]	0.842	0.843	0.725	0.674	0.301	0.153	257	53
Flexibility or mobility	0.962	0.974[e]	0.909	0.857	0.947	0.949	0.279	0.147	178	38
Balance or vestibular	0.787	0.752	0.852	0.809	0.882	0.939[e]	0.597	0.470	351	47
Gait training	0.808	0.837	0.837	0.814	0.851	0.860[e]	0.626	0.529	310	47
Functional mobility	0.775	0.831[e]	0.727	0.750	0.691	0.780	0.220	0.182	204	33
Purpose										
Simulated	0.769	0.769	0.870[e]	0.762	0.857	0.870[e]	0.688	0.667	48	10
Positioning										
Weight bearing	0.788	0.833	0.876[e]	0.867	0.857	0.871	0.197	0.282	255	43
Non-weight bearing	0.931	0.932	0.916	0.918	0.946[e]	0.923	0.283	0.038	539	91
Average	0.878	0.891[e]	0.861	0.835	0.875	0.883	0.502	0.433	283	53

[a]RBNLP: rule-based natural language processing.

[b]NER: named entity recognition.

[c]LR: logistic regression.

[d]SVM: support vector machine.

[e]The best performance for each entity.

[f]N/A: not applicable.

[g]ROM: range of motion.

The rule-based NLP's performance on the sequence classification task was similar to its performance on the NER task. Instances of higher performance in sequence classification compared to NER can be partially explained by mismatches in predicted spans and their labels affecting NER accuracy, yet still allowing for correct overall text section classification. The rule-based algorithm tied with or outperformed the other models on half of the concepts in Table 3. Among the machine learning models, gradient boosting performed nearly as well, achieving the highest F_1-score on 18 concepts.

In addition to these concepts, the rule-based NLP algorithm also predicted the spans of durations, sets, and reps. Since these categories do not have any specific concepts assigned to them, the number presented in each span was used instead as a comparison against the true label, converting minutes to seconds where applicable. This resulted in F_1-scores of 0.65, 0.58, and 0.88, respectively. It is important to note that we limited the experiments for "Duration," "Sets," and "Reps" exclusively to rule-based algorithms because these categories inherently involve numeric data, which align well with the deterministic and pattern-based nature of rule-based approaches.

Gradient boosting demonstrated the best performance for identifying range of motion (ROM) concepts and determining the location of exercise (performed in-office, home exercise program, and not performed) with F_1-scores of 0.863 for active ROM; 0.857 for active-assisted ROM; and 0.977, 0.986, and 0.950, respectively, for the locations. The rule-based natural language processing algorithm outperformed machine learning models in detecting sides of the body with F_1-scores of 0.975 for the right side and 0.937 for the left side, and it also performed the best on most exercise types, except for balance or vestibular and gait training concepts, which were classified best by gradient boosting with F_1-scores of 0.939 and 0.860, respectively. The LR obtained a strictly higher score than other methods in the weight-bearing exercise concept with an F_1-score of 0.876. The AdaBoost got a strictly higher score on 3 concepts, notably on non–weight bearing positioning with an F_1-score of 0.946. The SVM model did not score higher than other models but had 3 ties, indicating competitive performance.

These findings indicate that the rule-based approach is particularly effective for certain types of exercises, with superior performance in most categories. However, gradient boosting demonstrated strength in more complex categorizations such as balance or vestibular and gait training, where understanding nuanced differences is crucial.

For the LLM-based NLP, the results show that both zero-shot prompts and few-shot prompts result in high recall scores that sometimes exceed other methods. However, precision was quite low for most concepts, and F_1-scores did not exceed every other method for any concept. However, ChatGPT did occasionally outperform some of the simpler machine learning models and, on 1 occasion, even outperformed the rule-based algorithm (on the backward plane of motion concept). The average precision over all 82 concepts tested was 0.33 for the zero-shot approach and 0.27 for the few-shot approach. The average recall was 0.8 for the zero-shot approach and 0.82 for the few-shot approach. This resulted in average F_1-scores of 0.37 and 0.35, respectively, indicating that the zero-shot approach was slightly better on average than the few-shot approach. However, the few-shot approach performed the best for all but 10 concepts. The reason the zero-shot method performed better on average is thus due to the fact that it shows significant improvement on a few specific concepts, such as hip, scapula, hand, abduction, and extension.

Discussion

Observations

As indicated by the high performance of the machine learning models on many of the concepts, the task of extracting information from exercise descriptions was not complex. Although some of these concepts could be extracted effectively using straightforward rules or a small machine learning model, there were also many cases where clinical notes appeared ambiguous without context. For instance, the abbreviation "SL" could be interpreted as "single leg" or "side-lying" depending on the exercise being described. In addition, "L" could mean "left" or "lateral," which explains why the rule-based NLP algorithm performed slightly worse when classifying left versus right. The use of single letters as abbreviations, especially "A" as a shorthand for "anterior," could cause issues in machine learning algorithms without careful consideration. It would be possible to increase the performance of the rule-based algorithm by further tuning the rules to search for context clues at other points in the document, but this could potentially cause the rules to overfit the training set. Of particular interest are the numeric data present in duration, sets, and reps. These are particularly tricky to extract since they are expressed in a wide variety of ways by different physicians. It can be difficult to define what sets and reps are depending on the exercise, and sometimes one or both are not well-defined at all. Additionally, the use of apostrophes and quotes can either indicate measurements of time or distance, once again requiring context to disambiguate. Mentions of distance were not annotated in the gold-standard labels, but it is important in measuring the intensity of some exercises, so we plan to include it in the future.

Some of the misclassifications of the rule-based algorithm are due to inaccuracies in the gold-standard data set. For instance, many false positives produced by the rule-based algorithm appeared to be concepts that were missed by the annotators. There were also a few minor errors that could be explained by a mouse slip, including a span of text being assigned the wrong concept or a span excluding the last letter in a word. There were also some spelling mistakes in the notes themselves; common instances were explicitly mentioned in the rules to increase precision. Preprocessing clinical notes to correct spelling mistakes might be useful to improve results, although this creates a risk of incorrect changes being made to uncommon words. All of these errors were not particularly common throughout the labels, but they could have a significant effect on concepts that are already uncommon in the data.

Another obstacle that obscured some of the signals in the data came from the deidentification process. In addition to removing names, addresses, and other protected information from these documents, many other tokens and phrases were mistakenly removed, including equipment names and numbers denoting indices in a list. These were replaced with placeholder tokens such as "[ADDRESS]" or "[PERSONALNAME]." The low precision of the deidentification process caused some relevant information to be obfuscated or entirely erased from notes.

During the data annotation, we found that many of the concepts identified as relevant in this domain were not well documented in the data we extracted for annotation. This could be due in part to the fact that the data were only collected from patients who had a stroke, but this is not expected to be the main reason because patients who had a stroke can have a wide variety of musculoskeletal problems, resulting in a correspondingly wide variety of treatments being mentioned in clinical notes [23]. The other reason the data set lacks many of these concepts could be that they are rarely mentioned in these particular sections of clinical notes, either because they are not common enough to appear in many records at all or because they are mentioned more often in other sections. Thus, future research could focus on improving extraction methods to focus more on these uncommon concepts or include information from outside of the exercise descriptions.

In addition to ChatGPT for the LLM-based NLP approach, we also fine-tuned a Bidirectional Encoder Representations from Transformers (BERT) model with the task of categorizing the physical rehabilitation exercise concept. The BioClinicalBERT model [24] was used, which was pretrained on Medical Information Mart for Intensive Care-III (MIMIC-III) [25]. However, the amount of data collected seemed insufficient to make the model perform comparably to simpler methods. The model with the highest F_1-score on the validation set had an average F_1-score of 0.05 across all concepts on the test set. It accurately predicted in-office exercise performance with an F_1-score of 0.72. However, the performance on the remaining 100 concepts ranged only from 0 to 0.35. Therefore, we did not include this approach in the experimental comparison.

Limitations and Future Work

One limitation in this research was the necessary exclusion of "Duration," "Number of Sets," and "Number of Reps" from our machine learning–based NLP analysis due to their numeric nature, rendering them unsuitable for binary classification tasks. In future work, we plan to incorporate regression models or specialized classification techniques capable of handling numeric data. We also plan to expand our research to include additional variables such as stroke duration and severity, recognizing their potential to significantly enhance the prediction accuracy and effectiveness of rehabilitation strategies.

Furthermore, another limitation of this study is that we did not consider technique names and their association with specific motion types in rehabilitation exercise notes. For instance, we encountered the text "1: Standing AAROM PNF exercise D1/D2 flexion - 20 x" during annotation but did not annotate the technique name PNF (proprioceptive neuromuscular facilitation). To address this, in future work, we intend to develop a supplementary module for our algorithm that can effectively extract and map popular technique names to their corresponding motion types and categories, thereby enhancing the comprehensiveness and applicability of the algorithm.

Moreover, we plan to implement a robust standardized extraction protocol in the next version of our algorithm to mitigate the omission of therapeutic procedure sections due to copy-and-paste errors. This protocol will include multiple checks for consistency and completeness and will be assessed through a pilot study to ensure its reliability and accuracy. To enhance our model's generalizability amid varied note-writing practices across rehabilitation facilities, future research will also focus on diversifying data sources, refining adaptability to diverse writing styles and terminologies, and conducting extensive validation studies in a range of settings to improve performance. Through continuous monitoring and refinement of our extraction process, we are committed to enhancing the reliability and validity of our data, thereby strengthening the overall quality and impact of our research.

Conclusions

In this study, we developed and evaluated several NLP algorithms to extract physical rehabilitation exercise information from clinical notes of patients who had stroke. We first created a novel and comprehensive clinical ontology to represent physical rehabilitation exercise in clinical notes and then developed a variety of NLP algorithms leveraging state-of-the-art techniques, including rule-based NLP algorithms, machine learning–based NLP algorithms, and LLM-based NLP algorithms. The experiments on the clinical notes of a cohort of patients who had a stroke showed that the rule-based NLP algorithm had the best performance for most of the physical rehabilitation exercise concepts. Among all machine learning models, gradient boosting achieved the best performance on a majority of concepts. On the other hand, the rule-based NLP performed well for extracting handled durations, sets, and reps, while gradient boosting excelled in ROM and location detection. The LLM-based NLP achieved high recall with zero-shot and few-shot prompts but low precision and F_1-scores. It occasionally outperformed simpler models and once bet the rule-based algorithm.

Acknowledgments

This work was supported by the School of Health and Rehabilitation Sciences Dean's Research and Development Award.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full binary F_1-scores of each algorithm on the test set (50 documents) and additional results from the ChatGPT experiment.
[DOCX File , 63 KB - medinform_v12i1e52289_app1.docx]

References

1. Ginsburg GS, Phillips KA. Precision medicine: from science to value. Health Aff (Millwood) 2018;37(5):694-701 [FREE Full text] [doi: 10.1377/hlthaff.2017.1624] [Medline: 29733705]
2. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
3. Shin SH, Bode AM, Dong Z. Precision medicine: the foundation of future cancer therapeutics. NPJ Precis Oncol 2017;1(1):12 [FREE Full text] [doi: 10.1038/s41698-017-0016-z] [Medline: 29872700]

4. French MA, Roemmich RT, Daley K, Beier M, Penttinen S, Raghavan P, et al. Precision rehabilitation: optimizing function, adding value to health care. Arch Phys Med Rehabil 2022;103(6):1233-1239. [doi: 10.1016/j.apmr.2022.01.154] [Medline: 35181267]
5. Severin R, Sabbahi A, Arena R, Phillips SA. Precision medicine and physical therapy: a healthy living medicine approach for the next century. Phys Ther 2022;102(1):pzab253 [FREE Full text] [doi: 10.1093/ptj/pzab253] [Medline: 34718788]
6. Lotze M, Moseley GL. Theoretical considerations for chronic pain rehabilitation. Phys Ther 2015;95(9):1316-1320 [FREE Full text] [doi: 10.2522/ptj.20140581] [Medline: 25882484]
7. Blum C, Baur D, Achauer LC, Berens P, Biergans S, Erb M, et al. Personalized neurorehabilitative precision medicine: from data to therapies (MWKNeuroReha)—a multi-centre prospective observational clinical trial to predict long-term outcome of patients with acute motor stroke. BMC Neurol 2022;22(1):238 [FREE Full text] [doi: 10.1186/s12883-022-02759-2] [Medline: 35773640]
8. Zhao Y, Fu S, Bielinski SJ, Decker PA, Chamberlain AM, Roger VL, et al. Natural language processing and machine learning for identifying incident stroke from electronic health records: algorithm development and validation. J Med Internet Res 2021;23(3):e22951 [FREE Full text] [doi: 10.2196/22951] [Medline: 33683212]
9. Newman-Griffis D, Maldonado JC, Ho PS, Sacco M, Silva RJ, Porcino J, et al. Linking free text documentation of functioning and disability to the ICF with natural language processing. Front Rehabil Sci 2021;2:742702 [FREE Full text] [doi: 10.3389/fresc.2021.742702] [Medline: 35694445]
10. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-03-18]
11. International Classification of Functioning, Disability, and Health Children and Youth Version: ICF-CY. Geneva: World Health Organization; 2007.
12. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-290. [Medline: 17095826]
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
14. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:149-153 [FREE Full text] [Medline: 24303255]
15. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273-297 [FREE Full text] [doi: 10.1007/bf00994018]
16. Pregibon D. Logistic regression diagnostics. Ann Statist 1981;9(4):705-724 [FREE Full text] [doi: 10.1214/aos/1176345513]
17. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001;29(5):1189-1232 [FREE Full text] [doi: 10.1214/aos/1013203451]
18. Schapire RE. Explaining adaboost. In: Schölkopf B, Vovk V, Luo Z, editors. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Berlin Heidelberg: Springer; 2013:37-52.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
20. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Red Hook, New York, US: Curran Associates, Inc; 2022:27730-27744.
21. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. ArXiv 2023 [FREE Full text]
22. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981 [FREE Full text] [Medline: 37128372]
23. De Rosario H, Pitarch-Corresa S, Pedrosa I, Vidal-Pedrós M, de Otto-López B, García-Mieres H, et al. Applications of natural language processing for the management of stroke disorders: scoping review. JMIR Med Inform 2023;11:e48693 [FREE Full text] [doi: 10.2196/48693] [Medline: 37672328]
24. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. ArXiv 2019 [FREE Full text] [doi: 10.48550/arXiv.1904.03323]
25. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
ICF: International Classification of Functioning, Disability, and Health
LLM: large language model
LR: logistic regression
MIMIC-III: Medical Information Mart for Intensive Care-III
NER: named entity recognition
NLP: natural language processing

PNF: proprioceptive neuromuscular facilitation
ROM: range of motion
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
SVM: support vector machine
UMLS: Unified Medical Language System
UPMC: University of Pittsburgh Medical Center

Edited by A Benis; submitted 29.08.23; peer-reviewed by Z Alhassan, A Rehan Youssef; comments to author 27.11.23; revised version received 02.01.24; accepted 27.02.24; published 03.04.24.

Please cite as:
Sivarajkumar S, Gao F, Denny P, Aldhahwani B, Visweswaran S, Bove A, Wang Y
Mining Clinical Notes for Physical Rehabilitation Exercise Information: Natural Language Processing Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e52289
URL: https://medinform.jmir.org/2024/1/e52289
doi:10.2196/52289
PMID:38568736

Original Paper

An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study

Sonish Sivarajkumar[1], BS; Mark Kelley[2], MS; Alyssa Samolyk-Mazzanti[2], MS; Shyam Visweswaran[1,3], MD, PhD; Yanshan Wang[1,2,3], PhD

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, United States

[2]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, United States

[3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Yanshan Wang, PhD
Department of Health Information Management
University of Pittsburgh
6026 Forbes Tower
Pittsburgh, PA, 15260
United States
Phone: 1 4123832712
Email: yanshan.wang@pitt.edu

Abstract

Background: Large language models (LLMs) have shown remarkable capabilities in natural language processing (NLP), especially in domains where labeled data are scarce or expensive, such as the clinical domain. However, to unlock the clinical knowledge hidden in these LLMs, we need to design effective prompts that can guide them to perform specific clinical NLP tasks without any task-specific training data. This is known as in-context learning, which is an art and science that requires understanding the strengths and weaknesses of different LLMs and prompt engineering approaches.

Objective: The objective of this study is to assess the effectiveness of various prompt engineering techniques, including 2 newly introduced types—heuristic and ensemble prompts, for zero-shot and few-shot clinical information extraction using pretrained language models.

Methods: This comprehensive experimental study evaluated different prompt types (simple prefix, simple cloze, chain of thought, anticipatory, heuristic, and ensemble) across 5 clinical NLP tasks: clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, and medication attribute extraction. The performance of these prompts was assessed using 3 state-of-the-art language models: GPT-3.5 (OpenAI), Gemini (Google), and LLaMA-2 (Meta). The study contrasted zero-shot with few-shot prompting and explored the effectiveness of ensemble approaches.

Results: The study revealed that task-specific prompt tailoring is vital for the high performance of LLMs for zero-shot clinical NLP. In clinical sense disambiguation, GPT-3.5 achieved an accuracy of 0.96 with heuristic prompts and 0.94 in biomedical evidence extraction. Heuristic prompts, alongside chain of thought prompts, were highly effective across tasks. Few-shot prompting improved performance in complex scenarios, and ensemble approaches capitalized on multiple prompt strengths. GPT-3.5 consistently outperformed Gemini and LLaMA-2 across tasks and prompt types.

Conclusions: This study provides a rigorous evaluation of prompt engineering methodologies and introduces innovative techniques for clinical information extraction, demonstrating the potential of in-context learning in the clinical domain. These findings offer clear guidelines for future prompt-based clinical NLP research, facilitating engagement by non-NLP experts in clinical NLP advancements. To the best of our knowledge, this is one of the first works on the empirical evaluation of different prompt engineering approaches for clinical NLP in this era of generative artificial intelligence, and we hope that it will inspire and inform future research in this area.

(JMIR Med Inform 2024;12:e55318) doi:10.2196/55318

KEYWORDS

large language model; LLM; LLMs; natural language processing; NLP; in-context learning; prompt engineering; evaluation; zero-shot; few shot; prompting; GPT; language model; language; models; machine learning; clinical data; clinical information; extraction; BARD; Gemini; LLaMA-2; heuristic; prompt; prompts; ensemble

Introduction

Clinical information extraction (IE) is the task of identifying and extracting relevant information from clinical narratives, such as clinical notes, radiology reports, or pathology reports. Clinical IE has many applications in health care, such as improving diagnosis, treatment, and decision-making; facilitating clinical research; and enhancing patient care [1,2]. However, clinical IE faces several challenges, such as the scarcity and heterogeneity of annotated data, the complexity and variability of clinical language, and the need for domain knowledge and expertise.

Zero-shot IE is a promising paradigm that aims to overcome these challenges by leveraging large pretrained language models (LMs) that can perform IE tasks without any task-specific training data [3]. In-context learning is a framework for zero-shot and few-shot learning, where a large pretrained LM takes a context and directly decodes the output without any retraining or fine-tuning [4]. In-context learning relies on prompt engineering, which is the process of crafting informative and contextually relevant instructions or queries as inputs to LMs to guide their output for specific tasks [5]. The use of prompt engineering lies in its ability to leverage the powerful capabilities of large LMs (LLMs), such as GPT-3.5 (OpenAI) [6], Gemini (Google) [7], LLaMA-2 (Meta) [8], even in scenarios where limited or no task-specific training data are available. In clinical natural language processing (NLP), where labeled data sets tend to be scarce, expensive, and time-consuming to create, splintered across institutions, and constrained by data use agreements, prompt engineering becomes even more crucial to unlock the potential of state-of-the-art LLMs for clinical NLP tasks.

While prompt engineering has been widely explored for general NLP tasks, its application and impact in clinical NLP remain relatively unexplored. Most of the existing literature on prompt engineering in the health care domain focuses on biomedical NLP tasks rather than clinical NLP tasks that involve processing real-world clinical notes. For instance, Chen et al [9] used a fixed template as the prompt to measure the performance of LLMs on biomedical NLP tasks but did not investigate different kinds of prompting methods. Wang et al [10] gave a comprehensive survey of prompt engineering for health care NLP applications such as question-answering systems, text summarization, and machine translation. However, they did not compare and evaluate different types of prompts for specific clinical NLP tasks and how the performance varies across different LLMs. There is a lack of systematic and comprehensive studies on how to engineer prompts for clinical NLP tasks, and the existing literature predominantly focuses on general NLP problems. This creates a notable gap in the research, warranting a dedicated investigation into the design and development of effective prompts specifically for clinical NLP. Currently, researchers in the field lack a comprehensive understanding of the types of prompts that exist, their relative effectiveness, and the challenges associated with their implementation in clinical settings.

The main research question and objectives of this study are to investigate how to engineer prompts for clinical NLP tasks, identify best practices, and address the challenges in this emerging field. By doing so, we aim to propose a guideline for future prompt-based clinical NLP studies. In this work, we present a comprehensive empirical evaluation study on prompt engineering for 5 diverse clinical NLP tasks, namely, clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, and medication attribute extraction [11,12]. By systematically evaluating different types of prompts proposed in recent literature, including prefix [13], cloze [14], chain of thought [15], and anticipatory prompts [16], we gain insights into their performance and suitability for each task. Two new types of prompting approaches were also introduced: (1) heuristic prompts and (2) ensemble prompts. The rationale behind these novel prompts is to leverage the existing knowledge and expertise in rule-based NLP, which has been prominent and has shown significant results in the clinical domain [17]. We hypothesize that heuristic prompts, which are based on rules derived from domain knowledge and linguistic patterns, can capture the salient features and constraints of the clinical IE tasks. We also conjecture that ensemble prompts, which are composed of multiple types of prompts, can benefit from the complementary strengths and mitigate the weaknesses of each individual prompt.

One of the key aspects of prompt engineering is the number of examples or shots that are provided to the model along with the prompt. Few-shot prompting is a technique that provides the model with a few examples of input-output pairs, while zero-shot prompting does not provide any examples [3,18]. By contrasting these strategies, we aim to shed light on the most efficient and effective ways to leverage prompt engineering in clinical NLP. Finally, we propose a prompt engineering framework to build and deploy zero-shot NLP models for the clinical domain. This study covers 3 state-of-the-art LMs, including GPT-3.5, Gemini, and LLaMA-2, to assess the generalizability of the findings across various models. This work yields novel insights and guidelines for prompt engineering specifically for clinical NLP tasks.

Methods

Tasks

We selected 5 distinct clinical NLP tasks representing diverse categories of natural language understanding: clinical sense disambiguation (text classification) [19], biomedical evidence extraction (named entity recognition) [20], coreference resolution [21], medication status extraction (named entity recognition+classification) [22], and medication attribute

extraction (named entity recognition+relation extraction) [23]. Table 1 provides a succinct overview of each task, an example scenario, and the corresponding prompt type used for each task.

Table 1. Task descriptions.

Task	NLP[a] task category	Description	Example prompt
Clinical sense disambiguation	Text classification	This task involves identifying the correct meaning of clinical abbreviations within a given context.	What is the meaning of the abbreviation CR[b] in the context of cardiology?
Biomedical evidence extraction	Text extraction	In this task, interventions are extracted from biomedical abstracts.	Identify the psychological interventions in the given text?
Coreference resolution	Coreference resolution	The goal here is to identify all mentions in clinical text that refer to the same entity.	Identify the antecedent for the patient in the clinical note.
Medication status extraction	NER[c]+classification	This task involves identifying whether a medication is currently being taken, not taken, or unknown.	What is the current status of [24] in the treatment of [25]?
Medication attribute extraction	NER+RE[d]	The objective here is to identify specific attributes of a medication, such as dosage and frequency.	What is the recommended dosage of [26] for [27] and how often?

[a]NLP: natural language processing.

[b]CR: cardiac resuscitation.

[c]NER: named entity recognition.

[d]RE: relation extraction.

Data Sets and Evaluation

The prompts were evaluated on 3 LLMs, GPT-3.5, Gemini, and LLaMA-2, under both zero-shot and few-shot prompting conditions, using precise experimental settings and parameters. To simplify the evaluation process and facilitate clear comparisons, we adopted accuracy as the sole evaluation metric for all tasks. Accuracy is defined as the proportion of correct outputs generated by the LLM for each task, using a resolver that maps the output to the label space. Table 2 shows the data sets and sample size for each clinical NLP task. The data sets are as follows:

- Clinical abbreviation sense inventories: This is a data set of clinical abbreviations, senses, and instances [28]. It contains 41 acronyms from 18,164 notes, along with their expanded forms and contexts. We used a randomly sampled subset from this data set for clinical sense disambiguation, coreference resolution, medication status extraction, and medication attribute extraction tasks (Table 2).
- Evidence-based medicine-NLP: This is a data set of evidence-based medicine annotations for NLP [29]. It contains 187 abstracts and 20 annotated abstracts, with interventions extracted from the text. We used this data set for the biomedical evidence extraction task.

Table 2. Evaluation data sets and samples for different tasks.

Task	Data set	Data set example	Samples
Clinical sense disambiguation	CASI[a]	The abbreviation "CR[b]" can refer to "cardiac resuscitation" or "computed radiography."	11 acronyms from 55 notes
Biomedical evidence extraction	EBM[c]-NLP[d]	Identifying panic, avoidance, and agoraphobia (psychological interventions)	187 abstracts and 20 annotated abstracts
Coreference resolution	CASI	Resolving references to "the patient" or "the study" within a clinical trial report.	105 annotated examples
Medication status extraction	CASI	Identifying that a patient is currently taking insulin for diabetes.	105 annotated examples with 340 medication status pairs
Medication attribute extraction	CASI	Identifying dosage, frequency, and route of a medication for a patient.	105 annotated examples with 313 medications and 533 attributes

[a]CASI: clinical abbreviation sense inventories.

[b]CR: cardiac resuscitation.

[c]EBM: evidence-based medicine.

[d]NLP: natural language processing.

All experiments were carried out in different system settings. All GPT-3.5 experiments were conducted using the GPT-3.5 Turbo application programming interface as of the September 2023 update. The LLaMA-2 model was directly accessed for

our experiments. Gemini was accessed using the Gemini application (previously BARD)—Google's generative artificial intelligence conversational system. These varied system settings and access methods were taken into account to ensure the reliability and validity of our experimental results, given the differing architectures and capabilities of each LLM.

In evaluating the prompt-based approaches on GPT-3.5, Gemini, and LLaMA-2, we have also incorporated traditional NLP baselines to provide a comprehensive understanding of the LLMs' performance in a broader context. These baselines include well-established models such as Bidirectional Encoder Representations From Transformers (BERT) [30], Embeddings From Language Models (ELMO) [31], and PubMedBERT-Conditional Random Field (PubMedBERT-CRF) [32], which have previously set the standard in clinical NLP tasks. By comparing the outputs of LLMs against these baselines, we aim to offer a clear perspective on the advancements LLMs represent in the field. This comparative analysis is crucial for appreciating the extent to which prompt engineering techniques can leverage the inherent capabilities of LLMs, marking a significant evolution from traditional approaches to more dynamic and contextually aware methodologies in clinical NLP.

Prompt Creation Process

A rigorous process was followed to create suitable prompts for each task. These prompts were carefully crafted to match the specific context and objectives of each task. There is no established method for prompt design and selection as of now. Therefore, we adopted an iterative approach where prompts, which are created by health care experts, go through a verification and improvement process in an iterative cycle, which involved design, experimentation, and evaluation, as depicted in Figure 1.

Figure 1. Iterative prompt design process: a schematic diagram of the iterative prompt creation process for clinical NLP tasks. The process consists of 3 steps: sampling, prompt designing, and deployment. The sampling step involves defining the task and collecting data and annotations. The prompt designing step involves creating and refining prompts using different types and language models. The deployment step involves selecting the best model and deploying the model for clinical use. LLM: large language model; NER: named entity recognition; NLP: natural language processing; RE: relation extraction.

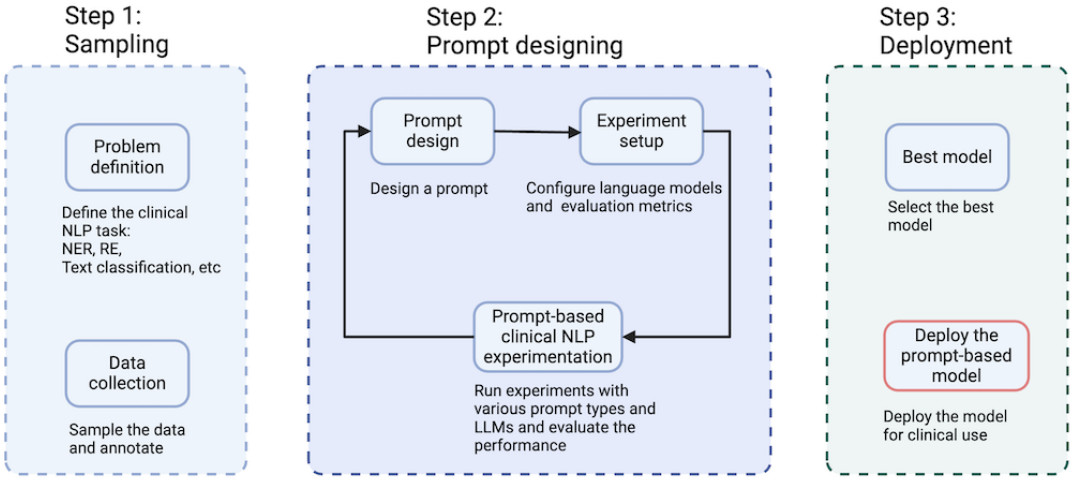


Figure 1 illustrates the 3 main steps of our prompt creation process: sampling, prompt designing, and deployment. In the sampling step (step 1), we defined the clinical NLP task (eg, named entity recognition, relation extraction, and text classification) and collected a sample of data and annotations as an evaluation for the task. In the prompt designing step (step 2), a prompt was designed for the task using one of the prompt types (eg, simple prefix prompt, simple cloze prompt, heuristic prompt, chain of thought prompt, question prompt, and anticipatory prompt). We also optionally performed few-shot prompting by providing some examples along with the prompt. The LLMs and the evaluation metrics for the experiment setup were then configured. We ran experiments with various prompt types and LLMs and evaluated their performance on the task. Based on the results, we refined or modified the prompt design until we achieved satisfactory performance or reached a limit. In the deployment step (step 3), the best prompt-based models were selected based on their performance metrics, and the model was deployed for the corresponding task.

Prompt Engineering Techniques

Overview

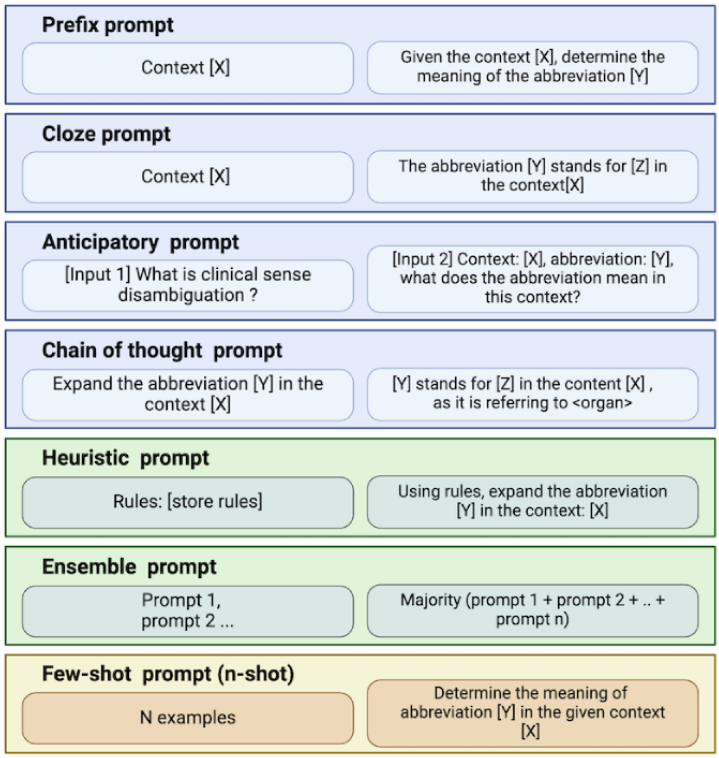
Prompt engineering is the process of designing and creating prompts that elicit desired responses from LLMs. Prompts can be categorized into different types based on their structure, function, and complexity.

Each prompt consists of a natural language query that is designed to elicit a specific response from the pretrained LLM. The prompts are categorized into 7 types, as illustrated in Figure 2 (all prompts have been included in Multimedia Appendix 1). Prefix prompts are the simplest type of prompts, which prepend a word or phrase indicating the type or format or tone of response for control and relevance. Cloze prompts are based on the idea of fill in the blank exercises, which create a masked token in the input text and ask the LLM to predict the missing word or phrase [3]. Anticipatory prompts are the prompts anticipating the next question or command based on experience or knowledge, guiding the conversation. Chain of thought

prompting involves a series of intermediate natural language reasoning steps that lead to the final output [15].

In addition to the existing types of prompts, 2 new novel prompts were also designed: heuristic prompts and ensemble prompts, which will be discussed in the following sections.

Figure 2. Types of prompts: examples of 7 types of prompts that we used to query the pretrained language model for different clinical information extraction tasks. [X]: context; [Y]: abbreviation; [Z]: expanded form.



Heuristic Prompts

Heuristic prompts are rule-based prompts that decompose complex queries into smaller, more manageable components for comprehensive answers. Adopting the principles of traditional rule-based NLP, which relies on manually crafted, rule-based algorithms for specific clinical NLP applications, we have integrated these concepts into our heuristic prompts approach. These prompts use a set of predefined rules to guide the LLM in expanding abbreviations within a given context. For instance, a heuristic prompt might use the rule that an abbreviation is typically capitalized, followed by a period, and preceded by an article or a noun. This approach contrasts with chain of thought prompts, which focus on elucidating the reasoning or logic behind an output. Instead, heuristic prompts leverage a series of predefined rules to direct the LLM in executing a specific task.

Mathematically, we can express a heuristic prompt as $H(x)$, a function applied to an input sequence x. This function is defined as a series of rule-based transformations T_i, where i indicates the specific rule applied. The output of this function, denoted as y_H, is then:

$$y_H=H(x)=T_n(T_{\{n-1\}}(\ldots T_1(x)))$$

Here, each transformation T_i applies a specific heuristic rule to modify the input sequence, making it more suitable for processing by LLMs.

From an algorithmic standpoint, heuristic prompts are implemented by defining a set of rules $R=\{R_1, R_2, \ldots, R_m\}$. Each rule R_j is a function that applies a specific heuristic criterion to an input token or sequence of tokens. Algorithmically, the heuristic prompting process can be summarized as follows:

By merging the precision and specificity of traditional rule-based NLP methods with the advanced capabilities of LLMs, the heuristic prompts offer a robust and accurate system for clinical information processing and analysis.

Ensemble Prompts

Ensemble prompts are prompts that combine multiple prompts using majority voting for aggregated outputs. They use various types of prompts to generate multiple responses to the same input, subsequently selecting the most commonly occurring output as the final answer. For instance, an ensemble prompt might use 3 different prefix prompts, or a combination of other prompt types, to produce 3 potential expansions for an abbreviation. The most frequently appearing expansion is then chosen. For the sake of simplicity, we amalgamated the outputs from all 5 different prompt types using a majority voting approach.

Mathematically, consider a set of m different prompting methods $P_1, P_2, \ldots, P_m$ applied to the same input x. Each method generates

an output y_i for i=1,2, ..., m. The ensemble prompt's output y_E is then the mode of these outputs:

y_E=mode (y_1, y_2, ..., y_m)

Algorithmically, the ensemble prompting process is as follows:

The rationale behind an ensemble prompt is that by integrating multiple types of prompts, we can use the strengths and counterbalance the weaknesses of each individual prompt, offering a robust and potentially more accurate response. Some prompts may be more effective for specific tasks or models, while others might be more resilient to noise or ambiguity. Majority voting allows us to choose the most likely correct or coherent output from the variety generated by different prompt types.

Results

Overview

In this section, we present the results of our experiments on prompt engineering for zero-shot clinical IE. Various prompt types were evaluated across 5 clinical NLP tasks, aiming to understand how different prompts influence the accuracy of different LLMs. Zero-shot and few-shot prompting strategies were also compared, exploring how the addition of context affects the model performance. Furthermore, we tested an ensemble approach that combines the outputs of different prompt types using majority voting. Finally, the impact of different LLMs on task performance was analyzed, and some interesting patterns were observed. Table 3 illustrates that different prompt types have different levels of effectiveness for different tasks and LLMs. We can also observe some general trends across the tasks and models.

Table 3. Performance comparison of different prompt types and language models.

Task and language model	Simple prefix	Simple cloze	Anticipatory	Heuristic	Chain of thought	Ensemble	Few shot
Clinical sense disambiguation							
GPT-3.5	0.88	0.86	0.88	0.96[a]	0.9	0.9	0.82
Gemini	0.76[b]	0.68	0.71	0.75	0.72	0.71	0.67
LLaMA-2	0.88[b]	0.76	0.82	0.82	0.78	0.82	0.78
BERT[c] (from [33])	0.42	0.42	0.42	0.42	0.42	0.42	0.42
ELMO[d] (from [33])	0.55	0.55	0.55	0.55	0.55	0.55	0.55
Biomedical evidence extraction							
GPT-3.5	0.92	0.82	0.88	0.94	0.94	0.88	0.96[a]
Gemini	0.89	0.89	0.91[b]	0.9	0.91[b]	0.9	0.88
LLaMA-2	0.85	0.88[b]	0.87	0.88[b]	0.87	0.88	0.86
PubMedBERT-CRF[e] (from [29])	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Coreference resolution							
GPT-3.5	0.78	0.6	0.74	0.94[a]	0.94[a]	0.74	0.74
Gemini	0.69	0.81[b]	0.73	0.67	0.71	0.69	0.7
LLaMA-2	0.8[b]	0.64	0.74	0.76	0.8[b]	0.78	0.68
Toshniwal et al [34]	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Medication status extraction							
GPT-3.5	0.76[a]	0.72	0.75	0.74	0.73	0.75	0.72
Gemini	0.67[b]	0.51	0.65	0.55	0.59	0.58	0.55
LLaMA-2	0.58	0.48	0.52	0.64[b]	0.52	0.58	0.42
ScispaCy [35]	0.52	0.52	0.52	0.52	0.52	0.52	0.52
Medication attribute extraction							
GPT-3.5	0.88	0.84	0.9	0.96[a]	0.96[a]	0.9	0.96[a]
Gemini	0.68	0.72	0.88[c]	0.7	0.74	0.76	0.88[b]
LLaMA-2	0.6	0.66	0.58	0.66	0.72[b]	0.64	0.6
ScispaCy	0.70	0.70	0.70	0.70	0.70	0.70	0.70

[a]Best performance on a task regardless of the model (ie, for each GPT-3.5 or Gemini or LLaMA-2 triple).

[b]Best performance for each model on a task.

[c]BERT: Bidirectional Encoder Representations From Transformers.

[d]ELMO: Embeddings From Language Models.

[e]PubMedBERT-CRF: PubMedBERT-Conditional Random Field.

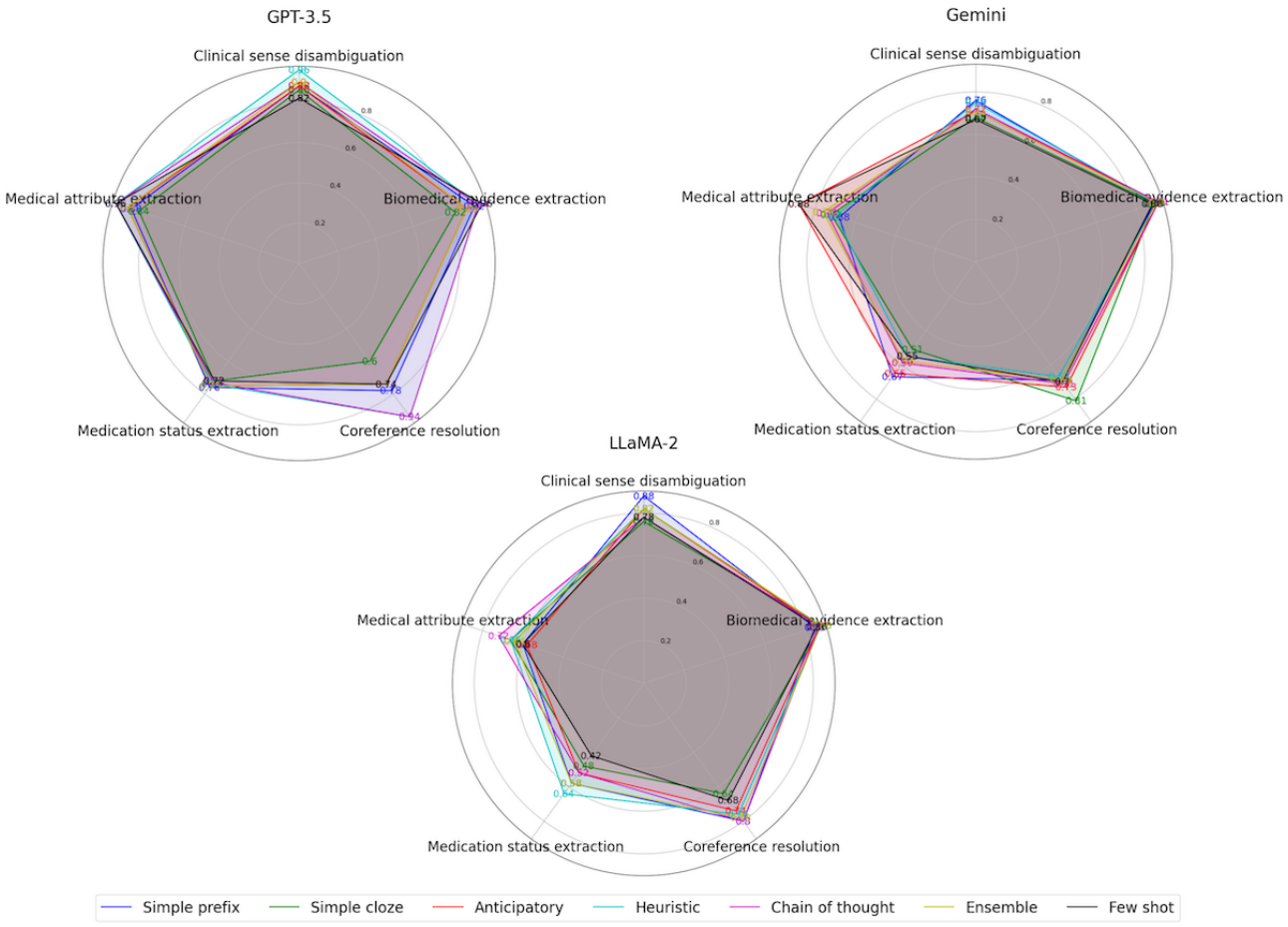
Prompt Optimization and Evaluation

For clinical sense disambiguation, the heuristic and prefix prompts consistently achieved the highest performance across all LLMs, significantly outperforming baselines such as BERT [30] and ELMO, with GPT-3.5 achieving an accuracy of 0.96, showcasing its advanced understanding of clinical context using appropriate prompting strategies. For biomedical evidence extraction, the heuristic and chain of thought prompts excelled across all LLMs in zero-shot setting. This indicates that these prompt types were able to provide enough information and constraints for the model to extract the evidence from the clinical note. GPT-3.5 achieved an accuracy of 0.94 with these prompt types, which was higher than any other model or prompt type combination. For coreference resolution, the chain of thought prompt type performed best among all prompt types with 2 LLMs—GPT-3.5 and LLaMA-2. This indicates that this prompt type was able to provide enough structure and logic for the model to resolve the coreference in the clinical note. GPT-3.5 displayed high accuracy with this prompt type, achieving an

accuracy of 0.94. For medication status extraction, simple prefix and heuristic prompts yielded good results across all LLMs. These prompt types were able to provide enough introduction or rules for the model to extract the status of the medication in relation to the patient or condition. GPT-3.5 excelled with these prompt types, achieving an accuracy of 0.76 and 0.74, respectively. For medication attribute extraction, we found that the chain of thought and heuristic prompts were effective across all LLMs. These prompt types were able to provide enough reasoning or rules for the model to extract and label the attributes of medications from clinical notes. Anticipatory prompts, however, had the best accuracy for Gemini among all the prompts. GPT-3.5 achieved an accuracy of 0.96 with these prompt types, which was higher than any other model or prompt type combination.

Thus, we can see that task-specific prompt tailoring is crucial for achieving high accuracy. Different tasks require different levels of information and constraints to guide the LLM to produce the desired output. The experiments show that heuristic, prefix, and chain of thought prompts are generally very effective for guiding the LLM to produce clear and unambiguous outputs. As shown in Figure 3, it is clear that GPT-3.5 is a superior and versatile LLM that can handle various clinical NLP tasks in zero-shot settings, outperforming other models in most cases.

Figure 3. Graphical comparison of prompt types in the 5 clinical natural language processing tasks used in this study.



Overall, the prompt-based approach has demonstrated remarkable superiority over traditional baseline models across all the 5 tasks. For clinical sense disambiguation, GPT-3.5's heuristic prompts achieved a remarkable accuracy of 0.96, showcasing a notable improvement over baselines such as BERT (0.42) and ELMO (0.55). In biomedical evidence extraction, GPT-3.5 again set a high standard with an accuracy of 0.94 using heuristic prompts, far surpassing the baseline performance of PubMedBERT-CRF at 0.35. Coreference resolution saw GPT-3.5 reaching an accuracy of 0.94 with chain of thought prompts, eclipsing the performance of existing methods such as Toshniwal et al [34] (0.69). In medication status extraction, GPT-3.5 outperformed the baseline ScispaCy (0.52) with an accuracy of 0.76 using simple prefix prompts. Finally, for medication attribute extraction, GPT-3.5's heuristic prompts achieved an impressive accuracy of 0.96, significantly higher than the ScispaCy baseline (0.70). These figures not only showcase the potential of LLMs in clinical settings but also set a foundation for future research to build upon, exploring even more sophisticated prompt engineering strategies and their implications for health care informatics.

Zero-Shot Versus Few-Shot Prompting

The performance of zero-shot prompting and few-shot prompting strategies was compared for each clinical NLP task. The same prompt types and LLMs were used as in the previous experiments, but some context was added to the input in the form of examples or explanations. Two examples or

explanations were used for each task (2-shot) depending on the complexity and variability of the task. Table 3 shows that few-shot prompting consistently improved the accuracy of all combinations for all tasks except for clinical sense disambiguation and medication attribute extraction, where some zero-shot prompt types performed better. We also observed some general trends across the tasks and models.

We found that few-shot prompting enhanced accuracy by providing limited context that aided complex scenario understanding. The improvement was more pronounced compared to simple cloze prompts, which had lower accuracy in most of the tasks. We also found that some zero-shot prompt types were very effective for certain tasks, even outperforming few-shot prompting. These prompt types used a rule-based or reasoning approach to generate sentences that contained definitions or examples of the target words or concepts, which helped the LLM to understand and match the context. For example, heuristic prompts achieved higher accuracy than few-shot prompting for clinical sense disambiguation and medication attribute extraction, while chain of thought prompts achieved higher accuracy than few-shot prompting for coreference resolution and medication attribute extraction. Alternatively, the clinical evidence extraction task likely benefits from additional context provided by few-shot examples, which can guide the model more effectively than the broader inferences made in zero-shot scenarios. This suggests that tasks requiring deeper contextual understanding might be better suited to few-shot learning approaches.

From these results, we can infer that LLMs can be effectively used for clinical NLP in a no-data scenario, where we do not have many publicly available data sets, by using appropriate zero-shot prompt types that guide the LLM to produce clear and unambiguous outputs. However, few-shot prompting can also improve the performance of LLMs by providing some context that helps the LLM to handle complex scenarios.

Other Observations

Ensemble Approaches

We experimented with an ensemble approach by combining outputs from multiple prompts using majority voting. The ensemble approach was not the best-performing strategy for any of the tasks, but it was better than the low-performing prompts. The ensemble approach was able to benefit from the diversity and complementarity of different prompt types and avoid some of the pitfalls of individual prompts. For example, for clinical sense disambiguation, the ensemble approach achieved an accuracy of 0.9 with GPT-3.5, which was the second best–performing prompt type. Similarly, for medication attribute extraction, the ensemble approach achieved an accuracy of 0.9 with GPT-3.5 and 0.76 with Gemini, which were close to the best single prompt type (anticipatory). However, the ensemble approach also had some drawbacks, such as inconsistency and noise. For tasks that required more specific or consistent outputs, such as coreference resolution, the ensemble approach did not improve the accuracy over the best single prompt type and sometimes even decreased it. This suggests that the ensemble approach may introduce ambiguity for tasks that require more precise or coherent outputs.

While the ensemble approach aims to reduce the variance introduced by individual prompt idiosyncrasies, our specific implementation observed instances where the combination of diverse prompt types introduced additional complexity. This complexity occasionally manifested as inconsistency and noise in the outputs contrary to our objective of achieving higher performance. Future iterations of this approach may include refinement of the prompt selection process to enhance consistency and further reduce noise in the aggregated outputs.

Impact of LLMs

Variations in performance were observed among different LLMs (Table 3). We found that GPT-3.5 generally outperformed Gemini and LLaMA-2 on most tasks. This suggests that GPT-3.5 has a better generalization ability and can handle a variety of clinical NLP tasks with different prompt types. However, Gemini and LLaMA-2 also showed some advantages over GPT-3.5 on certain tasks and prompt types. For example, Gemini achieved the highest accuracy of 0.81 with simple cloze prompts and LLaMA-2 achieved the highest accuracy of 0.8 with simple prefix prompts for coreference resolution. This indicates that Gemini and LLaMA-2 may have some domain-specific knowledge that can benefit certain clinical NLP tasks for specific prompt types.

Persona Patterns

Persona patterns are a way of asking the LLM to act like a persona or a system that is relevant to the task or domain. For example, one can ask the LLM to "act as a clinical NLP expert." This can help the LLM to generate outputs that are more appropriate and consistent with the persona or system. For example, one can use the following prompt for clinical sense disambiguation:

> *Act as a clinical NLP expert. Disambiguate the word "cold" in the following sentence: "She had a cold for three days."*

We experimented with persona patterns for different tasks and LLMs and found that they can improve the accuracy and quality of the outputs. Persona patterns can help the LLM to focus on the relevant information and constraints for the task and avoid generating outputs that are irrelevant or contradictory to the persona or system.

Randomness in Output

Most LLMs do not produce the output in the same format every time. There is inherent randomness in the outputs the LLMs produce. Hence, the prompts need to be specific in the way they are done for the task. Prompts are powerful when they are specific and if we use them in the right way.

Randomness in output can be beneficial or detrimental for different tasks and scenarios. In the clinical domain, randomness can introduce noise and errors in the outputs, which can make them less accurate and reliable for the users. For example, for tasks that involve extracting factual information, such as biomedical evidence extraction and medication status extraction, randomness can cause the LM to produce outputs that are inconsistent or contradictory with the input or context.

Guidelines and Suggestions for Optimal Prompt Selection

In recognizing the evolving nature of clinical NLP, we expand our discussion to contemplate the adaptability of our recommended prompt types and LM combinations across a wider spectrum of clinical tasks and narratives. This speculative analysis aims to bridge the gap between our current findings and their applicability to unexplored clinical NLP challenges, setting a foundation for future research to validate and refine these recommendations. In this section, we synthesize the main findings from our experiments and offer some practical advice for prompt engineering for zero-shot and few-shot clinical IE. We propose the following steps for selecting optimal prompts for different tasks and scenarios:

The first step is to identify the type of clinical NLP task, which can be broadly categorized into three types: (1) classification, (2) extraction, and (3) resolution. Classification tasks involve assigning a label or category to a word, phrase, or sentence in a clinical note, such as clinical sense disambiguation or medication status extraction. Extraction tasks involve identifying and extracting relevant information from a clinical note, such as biomedical evidence extraction or medication attribute extraction. Resolution tasks involve linking or matching entities or concepts in a clinical note, such as coreference resolution.

The second step is to choose the prompt type that is most suitable for the task type. We found that different prompt types have different strengths and weaknesses for different task types, depending on the level of information and constraints they provide to the LLM. Table 4 summarizes our findings and recommendations for optimal prompt selection for each task type.

The third step is to choose the LLM that is most compatible with the chosen prompt type. We found that different LLMs have different capabilities and limitations for different prompt types, depending on their generalization ability and domain-specific knowledge. Table 5 summarizes our findings and recommendations for optimal LLM selection for each prompt type.

The fourth step is to evaluate the performance of the chosen prompt type and LLM combination on the clinical NLP task using appropriate metrics such as accuracy, precision, recall, or F_1-score. If the performance is satisfactory, then the prompt engineering process is complete. If not, then the process can be repeated by choosing a different prompt type or LLM or by modifying the existing prompt to improve its effectiveness.

Table 4. Optimal prompt types for different clinical natural language processing task types.

Task type	Prompt type
Classification	Heuristic or prefix
Extraction	Heuristic or chain of thought
Resolution	Chain of thought

Table 5. Optimal language models for different prompt types.

Prompt type	Language model
Heuristic	GPT-3.5
Prefix	GPT-3.5 or LLaMA-2
Cloze	Gemini or LLaMA-2
Chain of thought	GPT-3.5
Anticipatory	Gemini

Discussion

Principal Findings

In this paper, we have presented a novel approach to zero-shot and few-shot clinical IE using prompt engineering. Various prompt types were evaluated across 5 clinical NLP tasks: clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, and medication attribute extraction. The performance of different LLMs, GPT-3.5, Gemini, and LLaMA-2, was also compared. Our main findings are as follows:

1. Task-specific prompt tailoring is crucial for achieving high accuracy. Different tasks require different levels of information and constraints to guide the LLM to produce the desired output. Therefore, it is important to design prompts that are relevant and specific to the task at hand and avoid using generic or vague prompts that may confuse the model or lead to erroneous outputs.
2. Heuristic prompts are generally very effective for guiding the LLM to produce clear and unambiguous outputs. These prompts use a rule-based approach to generate sentences that contain definitions or examples of the target words or concepts, which help the model to understand and match the context. Heuristic prompts are especially useful for tasks that involve disambiguation, extraction, or classification of entities or relations.
3. Chain of thought prompts are also effective for guiding the LLM to produce logical and coherent outputs. These prompts use a multistep approach to generate sentences that contain a series of questions and answers that resolve the

task in the context. Chain of thought prompts are especially useful for tasks that involve reasoning, inference, or coreference resolution.

4. Few-shot prompting can improve the performance of LLMs by providing some context that helps the model to handle complex scenarios. Few-shot prompting can be done by adding some examples or explanations to the input depending on the complexity and variability of the task. Few-shot prompting can enhance accuracy by providing limited context that aids complex scenario understanding. The improvement is more pronounced compared to simple prefix and cloze prompts, which had lower accuracy in most of the tasks.
5. Ensemble approaches can also improve the performance of LLMs by combining outputs from multiple prompts using majority voting. Ensemble approaches can leverage the strengths of each prompt type and reduce the errors of individual prompts. Ensemble approaches are especially effective for tasks that require multiple types of information or reasoning, such as biomedical evidence extraction and medication attribute extraction.

It is noteworthy that context size has a significant impact on the performance of LLMs in zero-shot IE [36]. In the scope of this study, we have avoided the context size dependence on performance, as it is a complex issue that requires careful consideration.

This study serves as an initial exploration into the efficacy of prompt engineering in clinical NLP, providing foundational insights rather than exhaustive guidelines. Given the rapid advancements in generative artificial intelligence and the complexity of clinical narratives, we advocate for continuous empirical testing of these prompt strategies across diverse clinical tasks and data sets. This approach will not only validate the generalizability of our findings but also uncover new avenues for enhancing the accuracy and applicability of LLMs in clinical settings.

Limitations

In this study, we primarily focused on exploring the capabilities and versatility of generative LLMs in the context of zero-shot and few-shot learning for clinical NLP tasks. Our approach also has some limitations that we acknowledge in this work. First, it relies on the quality and availability of pretrained LLMs, which may vary depending on the domain and task. As LLMs are rapidly evolving, some parts of the prompt engineering discipline may be timeless, while some parts may evolve and adapt over time as different capabilities of models evolve. Second, it requires a lot of experimentation and iteration to optimize prompts for different applications, which may be iterative and time-consuming. However, once optimal prompts are identified, the approach offers time savings in subsequent applications by reusing these prompts or making minor adjustments for similar tasks. We may not have explored all the possible combinations and variations of prompts that could potentially improve the performance of the clinical NLP tasks. Third, the LLMs do not release the details of the data set that they were trained on. Hence, the high accuracy could be because the models would have already seen the data during training and not because of the effectiveness of the prompts.

Future Work

We plan to address these challenges and limitations in our future work. We aim to develop more systematic and automated methods for prompt design and evaluation, such as using prompt-tuning or meta-learning techniques. We also aim to incorporate more domain knowledge or external resources into the prompts or the LLMs, such as using ontologies, knowledge graphs, or databases. We also aim to incorporate more quality control or error correction mechanisms into the prompts or the LLMs, such as using adversarial examples, confidence scores, or human feedback.

Conclusions

In this paper, we have benchmarked different prompt engineering techniques for both zero-shot and few-shot clinical NLP tasks. Two new types of prompts, heuristic and ensemble prompts, were also conceptualized and proposed. We have demonstrated that prompt engineering can enable the use of pretrained LMs for various clinical NLP tasks without requiring any fine-tuning or additional data. We have shown that task-specific prompt tailoring, heuristic prompts, chain of thought prompts, few-shot prompting, and ensemble approaches can improve the accuracy and quality of the outputs. We have also shown that GPT-3.5 is very adaptable and precise across all tasks and prompt types, while Gemini and LLaMA-2 may have some domain-specific advantages for certain tasks and prompt types.

We believe that a prompt-based approach has several benefits over existing methods for clinical IE. It reduces the cost and time in the initial phases of clinical NLP application development, where prompt-based methods offer a streamlined alternative to the conventional data preparation and model training processes. It is flexible and adaptable, as it can be applied to various clinical NLP tasks with different prompt types and LLMs. It is interpretable and explainable, as it uses natural language prompts that can be easily understood and modified by humans.

Acknowledgments

This work was supported by the National Institutes of Health (awards U24 TR004111 and R01 LM014306). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

SS conceptualized, designed, and organized this study; analyzed the results; and wrote, reviewed, and revised the paper. MK and AS-M analyzed the results, and wrote, reviewed, and revised the paper. SV wrote, reviewed, and revised the paper. YW conceptualized, designed, and directed this study and wrote, reviewed, and revised the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts for clinical natural language processing tasks.
[DOCX File , 31 KB - medinform_v12i1e55318_app1.docx]

References

1. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
2. Landolsi MY, Hlaoua L, Romdhane LB. Information extraction from electronic medical documents: state of the art and future research directions. Knowl Inf Syst 2023;65(2):463-516 [FREE Full text] [doi: 10.1007/s10115-022-01779-1] [Medline: 36405956]
3. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981 [FREE Full text] [Medline: 37128372]
4. Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, et al. Rethinking the role of demonstrations: what makes in-context learning work? 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; December 7-11, 2022; Abu Dhabi, United Arab Emirates p. 11048-11064 URL: https://aclanthology.org/2022.emnlp-main.759/ [doi: 10.18653/v1/2022.emnlp-main.759]
5. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with chatGPT. ArXiv Preprint posted online on February 21, 2023 [FREE Full text] [doi: 10.48550/arXiv.2302.11382]
6. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Red Hook, NY: Curran Associates, Inc; 2022:27730-27744.
7. Gemini Team Google. Gemini: a family of highly capable multimodal models. ArXiv Preprint posted online on December 19, 2023 [FREE Full text] [doi: 10.48550/arXiv.2312.11805]
8. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. ArXiv Preprint posted online on July 28, 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.09288]
9. Chen Q, Du J, Hu Y, Keloth VK, Peng X, Raja K, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. ArXiv Preprint posted online on May 10, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.16326]
10. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: methodologies and applications. ArXiv Preprint posted online on April 28, 2023 [FREE Full text] [doi: 10.48550/arXiv.2304.14670]
11. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. ArXiv Preprint posted online on March 29, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.16416]
12. Yuan C, Xie Q, Ananiadou S. Zero-shot temporal relation extraction with chatGPT. 2023 Presented at: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; July 13, 2023; Toronto, Canada p. 92-102 URL: https://aclanthology.org/2023.bionlp-1.7/ [doi: 10.18653/v1/2023.bionlp-1.7]
13. Li X, Liang L. Prefix-tuning: optimizing continuous prompts for generation. 2021 Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); August 1-6, 2021; Virtual Event p. 4582-4597 URL: https://aclanthology.org/2021.acl-long.353/ [doi: 10.18653/v1/2021.acl-long.353]
14. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023;55(9):1-35 [FREE Full text] [doi: 10.1145/3560815]
15. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Red Hook, NY: Curran Associates Inc; 2022:24824-24837.
16. Hancock B, Bordes A, Mazare PE, Weston J. Learning from dialogue after deployment: feed yourself, chatbot!. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 3667-3684 URL: https://aclanthology.org/P19-1358/ [doi: 10.18653/v1/p19-1358]
17. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform 2009;42(5):923-936 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.007] [Medline: 19646551]

18. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. : Association for Computational Linguistics; 2022 Presented at: The 2022 Conference on Empirical Methods in Natural Language Processing; December 7-11, 2022; Abu Dhabi, United Arab Emirates p. 1998-2022 URL: https://aclanthology.org/2022.emnlp-main.130.pdf [doi: 10.18653/v1/2022.emnlp-main.130]
19. Wu Y, Xu J, Zhang Y, Xu H. Clinical abbreviation disambiguation using neural word embeddings. 2015 Presented at: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015); July 30, 2015; Beijing, China p. 171-176 URL: https://aclanthology.org/W15-3822.pdf [doi: 10.18653/v1/w15-3822]
20. Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Iorio A, et al. Machine learning approaches to retrieve high-quality, clinically relevant evidence from the biomedical literature: systematic review. JMIR Med Inform 2021;9(9):e30401 [FREE Full text] [doi: 10.2196/30401] [Medline: 34499041]
21. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786-791 [FREE Full text] [doi: 10.1136/amiajnl-2011-000784] [Medline: 22366294]
22. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010;17(4):383-388 [FREE Full text] [doi: 10.1136/jamia.2010.004804] [Medline: 20595304]
23. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019;42(1):99-111 [FREE Full text] [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
24. Chen Y, Wu X, Chen M, Song Q, Wei J, Li X, et al. Dynamic text categorization of search results for medical class recognition in real world evidence studies in the Chinese language. : Association for Computing Machinery, Presented at: Proceedings of the International Conference on Bioinformatics and Computational Intelligence (ICBCI 2017); 2017; Beijing, China p. 40-48. [doi: 10.1145/3135954.3135962]
25. Mallick PK, Balas VE, Bhoi AK, Zobaa AF. Cognitive Informatics and Soft Computing Proceeding of CISC 2017, Advances in Intelligent Systems and Computing (AISC, Volume 768). New York: Springer Verlag; 2019.
26. Ananiadou S, Lee D, Xu H, Song M. DTMBIO'12—The Proceedings of the Sixth ACM International Workshop on Data and Text Mining in Biomedical Informatics. 2012 Presented at: 6th ACM International Workshop on Data and Text Mining in Biomedical Informatics, DTMBIO 2012, in Conjunction with the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012; 2012; New York URL: https://dl.acm.org/action/showFmPdf?doi=10.1145%2F2390068 [doi: 10.1145/2396761.2398758]
27. Elghandour I, State R, Brorsson M, Le L, Antonopoulos N, Xie Y, et al. IEEE/ACM International Symposium on Big Data Computing (BDC). 2016 Presented at: 2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT); December 6-9, 2016; Shanghai, China URL: https://ieeexplore.ieee.org/xpl/conhome/7876287/proceeding [doi: 10.1109/bdcat.2018.00008]
28. Moon S, Pakhomov S, Liu N, Ryan JO, Melton GB. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. J Am Med Inform Assoc 2014;21(2):299-307 [FREE Full text] [doi: 10.1136/amiajnl-2012-001506] [Medline: 23813539]
29. Nye B, Li JJ, Patel R, Yang Y, Marshall I, Nenkova A, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. 2018 Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 15-20, 2018; Melbourne, Australia p. 197-207 URL: https://aclanthology.org/P18-1019/ [doi: 10.18653/v1/p18-1019]
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint posted online on May 24, 2019 [FREE Full text]
31. Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, et al. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Res 2021;304:114135. [doi: 10.1016/j.psychres.2021.114135] [Medline: 34343877]
32. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2021;3(1):1-23. [doi: 10.1145/3458754]
33. Adams G, Ketenci M, Bhave S, Perotte A, Elhadad N. Zero-shot clinical acronym expansion via latent meaning cells. Proc Mach Learn Res 2020;136:12-40 [FREE Full text] [Medline: 34790898]
34. Toshniwal S, Xia P, Wiseman S, Livescu K, Gimpel K. On generalization in coreference resolution. ArXiv Preprint posted online on September 20, 2021 [FREE Full text] [doi: 10.18653/v1/2021.crac-1.12]
35. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. ArXiv Preprint posted online on October 9, 2019 [FREE Full text] [doi: 10.18653/v1/w19-5034]
36. Sivarajkumar S, Wang Y. Evaluation of healthprompt for zero-shot clinical text classification. 2023 Presented at: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI); June 26-29, 2023; Houston, TX, USA. [doi: 10.1109/ichi57859.2023.00081]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
ELMO: Embeddings From Language Models
IE: information extraction
LLM: large language model
LM: language model
NLP: natural language processing
PubMedBERT-CRF: PubMedBERT-Conditional Random Field

Edited by C Lovis; submitted 08.12.23; peer-reviewed by J Zaghir, M Torii, J Zheng; comments to author 04.02.24; revised version received 20.02.24; accepted 24.02.24; published 08.04.24.

Please cite as:
Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y
An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e55318
URL: https://medinform.jmir.org/2024/1/e55318
doi:10.2196/55318
PMID:38587879

Original Paper

Evaluating ChatGPT-4's Diagnostic Accuracy: Impact of Visual Data Integration

Takanobu Hirosawa[1*], MD, PhD; Yukinori Harada[1*], MD, PhD; Kazuki Tokumasu[2*], MD, PhD; Takahiro Ito[3*], MD; Tomoharu Suzuki[4*], MD; Taro Shimizu[1*], MD, MSc, MPH, MBA, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Shimotsuga, Japan

[2]Department of General Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

[3]Satsuki Home Clinic, Tochigi, Japan

[4]Department of Hospital Medicine, Urasoe General Hospital, Okinawa, Japan

[*]all authors contributed equally

Corresponding Author:
Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi, Mibu-cho
Shimotsuga, 321-0293
Japan
Phone: 81 282 87 2498
Email: hirosawa@dokkyomed.ac.jp

Abstract

Background: In the evolving field of health care, multimodal generative artificial intelligence (AI) systems, such as ChatGPT-4 with vision (ChatGPT-4V), represent a significant advancement, as they integrate visual data with text data. This integration has the potential to revolutionize clinical diagnostics by offering more comprehensive analysis capabilities. However, the impact on diagnostic accuracy of using image data to augment ChatGPT-4 remains unclear.

Objective: This study aims to assess the impact of adding image data on ChatGPT-4's diagnostic accuracy and provide insights into how image data integration can enhance the accuracy of multimodal AI in medical diagnostics. Specifically, this study endeavored to compare the diagnostic accuracy between ChatGPT-4V, which processed both text and image data, and its counterpart, ChatGPT-4, which only uses text data.

Methods: We identified a total of 557 case reports published in the *American Journal of Case Reports* from January 2022 to March 2023. After excluding cases that were nondiagnostic, pediatric, and lacking image data, we included 363 case descriptions with their final diagnoses and associated images. We compared the diagnostic accuracy of ChatGPT-4V and ChatGPT-4 without vision based on their ability to include the final diagnoses within differential diagnosis lists. Two independent physicians evaluated their accuracy, with a third resolving any discrepancies, ensuring a rigorous and objective analysis.

Results: The integration of image data into ChatGPT-4V did not significantly enhance diagnostic accuracy, showing that final diagnoses were included in the top 10 differential diagnosis lists at a rate of 85.1% (n=309), comparable to the rate of 87.9% (n=319) for the text-only version (P=.33). Notably, ChatGPT-4V's performance in correctly identifying the top diagnosis was inferior, at 44.4% (n=161), compared with 55.9% (n=203) for the text-only version (P=.002, χ^2 test). Additionally, ChatGPT-4's self-reports showed that image data accounted for 30% of the weight in developing the differential diagnosis lists in more than half of cases.

Conclusions: Our findings reveal that currently, ChatGPT-4V predominantly relies on textual data, limiting its ability to fully use the diagnostic potential of visual information. This study underscores the need for further development of multimodal generative AI systems to effectively integrate and use clinical image data. Enhancing the diagnostic performance of such AI systems through improved multimodal data integration could significantly benefit patient care by providing more accurate and comprehensive diagnostic insights. Future research should focus on overcoming these limitations, paving the way for the practical application of advanced AI in medicine.

(JMIR Med Inform 2024;12:e55627) doi:10.2196/55627

KEYWORDS

artificial intelligence; large language model; LLM; LLMs; language model; language models; ChatGPT; GPT; ChatGPT-4V; ChatGPT-4 Vision; clinical decision support; natural language processing; decision support; NLP; diagnostic excellence; diagnosis; diagnoses; diagnose; diagnostic; diagnostics; image; images; imaging

Introduction

Diagnostic Excellence

Diagnostic excellence involves accurately and efficiently diagnosing a wide range of conditions [1]. Achieving this requires a multifaceted approach [2], including effective collaboration among medical professionals, patients, families, and clinical decision support systems (CDSSs). Each plays a pivotal role, as follows: medical professionals bring their expertise and judgment, patients and families provide essential health information and context, and CDSSs offer data-driven insights, enhancing the collective decision-making process.

CDSSs for Diagnostic Excellence

CDSSs are computer-based tools that assist medical professionals in a wide range of clinical decisions, including diagnosis, treatment planning, medication ordering, preventive care, and patient education [3]. Research has shown that CDSS interventions significantly improve diagnostic accuracy [4], a key aspect of diagnostic excellence [5]. For instance, interventions involving a CDSS in the diagnosis of common chronic diseases demonstrated significant improvements [6]. Accurate diagnosis entails more than identifying a disease; it involves understanding the patient's unique health context, ensuring timely and appropriate treatment, reducing misdiagnosis risk, and ultimately improving patient outcomes [7]. In the rapidly evolving health care environment, maintaining high standards of diagnostic precision becomes increasingly crucial.

Artificial Intelligence in Medicine

CDSSs are broadly categorized into 2 types [3]: knowledge-based systems, which are grounded in medical guidelines and expert knowledge; and non–knowledge-based systems, using artificial intelligence (AI) or statistical pattern recognition for clinical data analysis.

The integration of AI into clinical settings is advancing rapidly. AI systems in medicine range from assisting in diagnostic imaging and analysis to optimizing patient treatment plans [8,9]. These systems are being increasingly adopted in hospitals and clinics [10], significantly contributing to enhanced diagnostic accuracy and efficiency.

However, the integration of AI into clinical settings brings transformative potential but also faces several hurdles. Challenges include ensuring data privacy [11], addressing the lack of large and diverse training data sets, and maintaining the interpretability of AI-generated recommendations to align with ethical standards [12,13]. Real-world obstacles, such as resistance from health care professionals due to trust issues in AI's diagnostic suggestions, underscore the complexity of AI integration into clinical practice.

Advancements in Large Language Models

A notable advancement in AI is the use of large language models (LLMs). As a subset of non–knowledge-based systems, LLMs are specialized forms of generative AI systems that process and generate human-like text based on extensive textual data training [14]. They are adept at tasks like translation, summarization, and even creative writing. In clinical practice, generative AI systems using LLMs have shown promise in summarizing patient history, integrating medical records, analyzing complex data streams, and enhancing communication between patients and medical professionals [15,16], demonstrating their utility in handling complex medical language and concepts. Such advancements not only improve the efficiency of medical documentation but also offer novel approaches to generating differential diagnoses, showcasing the innovative application of LLMs in clinical settings.

Multimodal Artificial Intelligence in Diagnostics

Integrating multimodal data, including text and images, presents technical challenges. Successful integration in other fields, such as autonomous driving technologies that combine multisensory observation data to navigate [17], offers a potential model for health care. Recent developments in generative AI systems, including Google Gemini (previous Google Bard [18]) and ChatGPT-4 with vision (ChatGPT-4V), have enabled the processing of both text and image data. This integration is essential for providing a comprehensive clinical overview. Although effectively combining data from different data sources remains a challenge, the development of multimodal AI models that incorporate data across modalities enabled broad applications that include personalized medicine and digital health [19]. For example, a multimodal model developed from the combination of images and health records could classify pulmonary embolism [20]. Another multimodal model could differentiate between common respiratory failure [21]. Among publicly available generative AI systems, the ChatGPT series, particularly ChatGPT-4V, developed by OpenAI and released in September 2023, stands out [22,23]. It accepts both text and image data [24,25], demonstrating impressive performance in various applications.

Preliminary studies in various fields, including medicine [26-28] and others [29-31] have shown the effectiveness of ChatGPT-4V. Some of these studies have highlighted its efficacy in interpreting medical images [26,28], though they were limited in scope. However, clinical image data includes a wide range of elements, from physical examinations to various investigation results. The full impact of image data integration on diagnostic accuracy is yet to be thoroughly explored.

Study Objectives

This study directly addressed the gaps identified in the current understanding of multimodal AI's application in clinical diagnostics. By comparing the diagnostic accuracy of

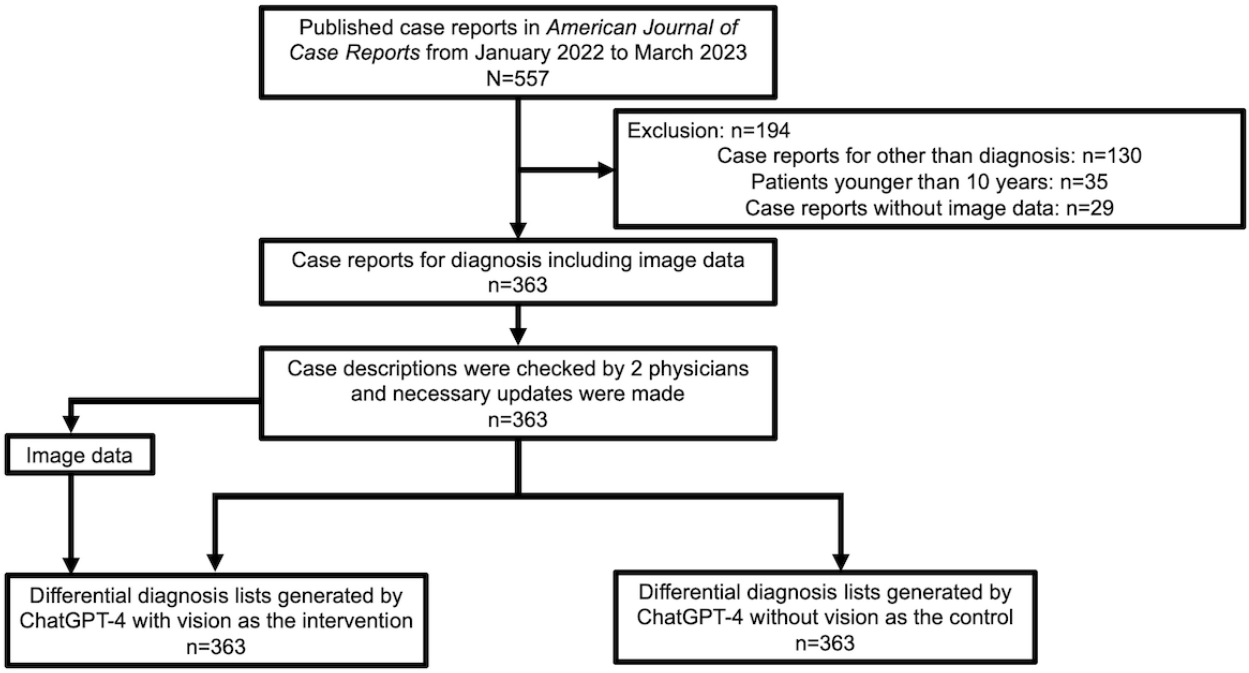
ChatGPT-4V and without vision across detailed case reports, and examining the impact of image data integration, we aimed to provide concrete evidence on the value and challenges of incorporating generative AI into clinical flows. Our objectives were shaped by the need to better understand how multimodal AI can be optimized to support diagnostic excellence, ultimately contributing to the advancement of medical diagnostics through technology.

Methods

Overview

We conducted an experimental study to assess the diagnostic accuracy of multimodal generative AI systems using data from a large number of case reports. This study was conducted in the Department of Diagnostic and Generalist Medicine (General Internal Medicine) at Dokkyo Medical University. This study involved several steps: preparing the data set and control, preparing image data, generating differential diagnosis lists by ChatGPT-4V, and evaluating the diagnostic accuracy of these differential diagnosis lists. A flow chart of the study's methodology is presented in Figure 1.

Figure 1. Study design.



Ethical Considerations

This study used published case reports, and thus ethical committee approval was not applicable.

Preparing Data Set and Control

We used the data set from our previous study (T Hirosawa, Y Harada, K Mizuta, T Sakamoto, K Tokomasu, T Shimizu, unpublished data, November 2023). The data set comprised case descriptions and final diagnoses, sourced from the *American Journal of Case Reports*, spanning January 2022 to March 2023. This peer-reviewed journal covers diagnostically challenging case reports from various medical fields. A total of 557 case reports were identified. The exclusion criteria were carefully chosen based on previous studies for CDSSs [32] and ChatGPT-4V [28] to ensure the focus remained on diagnostically challenging adult cases with relevant image data. Specifically, cases were excluded for the following reasons: nondiagnosis (130 cases), patients younger than 10 years (35 cases), and the absence of image data (29 cases). The included case reports were refined into case descriptions by the primary researcher (TH) and double-checked by another researcher (YH). From the included case reports, we extracted a case description until the final diagnosis was made in the "case report" section. We removed sentences that directly assessed the diagnosis to minimize bias in generating differential diagnoses. This step ensures that the differential diagnoses generated by ChatGPT-4 are based solely on the unbiased clinical presentation of the case. After brush-up, we formatted these case descriptions for input into ChatGPT-4. A typical case description included demographic information, chief complaints, history of present illness, results of physical examinations, and investigative findings leading to diagnoses. The final diagnoses were typically determined by the authors of the case reports. For example, in a case report titled "Levofloxacin-Associated Bullous Pemphigoid in a Hemodialysis Patient After Kidney Transplant Failure" [33] we extracted from "A 27-year-old female with hemodialysis was admitted for evaluation of a worsening bullous rash and shortness of breath over the last 3 days..." to "...Although the swab PCR test for VZV and HSV was negative, there was still concern about disseminated herpes zoster, as the patient was immunosuppressed" as a case description.

Additionally, the final diagnosis was levofloxacin-associated bullous pemphigoid.

In the next step, we used ChatGPT-4 without vision to develop the top 10 differential diagnosis lists based on the data of case descriptions. Two expert physicians independently evaluated whether the final diagnosis was included in the lists, and any discrepancies were resolved through discussion. Therefore, the differential diagnosis lists and data of physicians' evaluation of the lists from a total of 363 case reports were included as the control in this study.

Preparing Image Data

All figures and tables of included case descriptions were standardized to a resolution of 96 dots per inch in JPEG format to balance detail with file size, facilitating efficient processing by ChatGPT-4V without compromising the quality necessary for accurate diagnostic inference. When multiple figures or tables were present in a case description, they were compiled into a single JPEG file, each annotated with a file number in the upper-left corner. If image data exceeded the upload size limit, the images were resized to half their original size while preserving image quality, using the Preview application (version 11.0; Apple Inc) on a Mac computer.

Generating Differential Diagnosis Lists by ChatGPT-4V

We used ChatGPT-4V, a multimodal generative AI system developed by OpenAI, from October 30, 2023, to November 9, 2023. Additional training or reinforcement for diagnosis was not performed. The prompt was constructed as follows: "Identify the top 10 suspected illnesses based on the attached files with file names indicated in the left upper corner of each image, and the provided case description. List these illnesses using only their names, without providing any reasoning AND describe the proportion of the case description and the provided files to develop your suspected illness list (case description + all files = 100%): (copy and paste the case descriptions)." This design was intended to explicitly guide ChatGPT-4V to not only generate a list of possible diagnoses but also reflect on how each type of data influenced its conclusions, providing insights into the AI's diagnostics process. Apart from the prompt and file names, the text data input to ChatGPT-4V remained the same as the control, ChatGPT-4 without vision. The first generated list was used as the differential diagnosis list. The chat history was cleared before entering each new case description. Moreover, the data control settings for chat history were disabled. The details of ChatGPT-4V and ChatGPT-4 without vision are shown in Table 1.

Table 1. The details of ChatGPT-4 with vision and ChatGPT-4 without vision in this study.

Details	ChatGPT-4 with vision (intervention) [24]	ChatGPT-4 without vision (control) [22]
Short name	ChatGPT-4V	ChatGPT-4
Prompt	Identify the top 10 suspected illnesses based on the attached files with file names indicated in the left upper corner of each image, and the provided case description. List these illnesses using only their names, without providing any reasoning AND describe the proportion of the case description and the provided files to develop your suspected illness list (case description + all files =100%): (copy and paste the case descriptions)	Tell me the top 10 suspected illnesses for the following case: (copy and paste the case descriptions)
Text input	Same case descriptions with the above prompt and referred file number	Same case descriptions with the above prompt
Image input	Image data in JPEG format with a resolution of 96 dots per inch	No image data
Output	The top 10 differential diagnosis lists and the proportion of weight between text data and image data contributing to development of the differential diagnosis list	The top 10 differential diagnosis lists
Evaluations	By 2 independent physicians; any discrepancies were resolved by another physician	By 2 independent physicians; any discrepancies were resolved by another physician
Release date	September 2023	March 2023
Access date	From October 30, 2023, to November 9, 2023	From June 22, 2023, to June 29, 2023
Data control for chat history	Off	Off

Evaluation for Differential Diagnosis Lists by Physicians

Two expert physicians, TI and T Suzuki, independently evaluated whether the final diagnoses were included in the differential diagnosis lists. The evaluation was binary, with 1 indicating inclusion and 0 indicating exclusion. A score of 1 indicated that the differential closely matched the final diagnoses. This close match was defined not merely by the presence of the correct diagnosis within the list but by the relevance and clinical appropriateness of the differentials in relation to the final diagnosis. A score of 1 indicated that AI-generated differentials were clinically relevant and could potentially lead to appropriate interventions, thereby aligning with patient safety and standards [34]. Additionally, evaluators ranked the match of differential to the final diagnoses.

Conversely, a score of 0 was given if the differential diagnosis list significantly differed from the final diagnosis, indicating a lack of clinical relevance or potential misdirection in a real-world diagnostic scenario. Any discrepancies were resolved by another expert physician (KT), ensuring objective and consistent evaluation across all included case reports.

Outcome

The study assessed the diagnostic accuracy of ChatGPT-4V, as an intervention and compared it to ChatGPT-4 without vision as a control. The primary outcome was defined as the ratio of cases where the final diagnoses were included within the top 10 differential diagnosis lists. The secondary outcome is defined as the ratio of cases where the final diagnoses were included as top diagnosis. These outcomes were chosen to quantitatively measure diagnostic accuracy and the effectiveness of image data integration in enhancing ChatGPT-4's diagnostics.

Additionally, we assessed the contributing weight between text data (case descriptions) and image data (files) in developing the differential diagnosis lists, as reported by ChatGPT-4V. The total contribution from both elements was set to 100%. Specifically, we analyzed how much the text and image data individually contributed to the formulation of the differential diagnosis list. For example, if the text data (case description) contributed 60% and the image data contributed 40%, the total would sum up to 100%. This method allowed for a comprehensive understanding of the relative impact of textual and image data on AI diagnostics.

Statistical Analysis

For analysis, R (version 4.2.2; R Foundation for Statistical Computing) was used. We present continuous variables as medians and IQRs to accurately reflect the distribution of data. We presented categorical or binary variables as numbers and percentages. Additionally, we used χ^2 tests to compare categorical variables, setting the significance level at a *P* value <.05. The choice of χ^2 tests for comparing categorical variables was based on their ability to handle binary and categorical data effectively, providing a robust measure of association between diagnostic outcomes and ChatGPT-4 with or without vision.

To quantify the impact of each factor on the likelihood of accurate diagnosis inclusion, an univariable logistic regression model was applied. This model allows for the exploration of potential predictors of diagnostic accuracy, offering insights into how different data types contribute to ChatGPT-4's diagnostic processes. For the logistic regression model, the primary and secondary outcomes were treated as binary dependent variables: presence (1) or absence (0) of the correct diagnosis within the top 10 differential diagnosis lists and as the top diagnosis, respectively. Independent variables included the proportion of image data weight, the presence (1) or absence (0) of specific types of image data (eg, computed tomography [CT] images, pathological specimens, laboratory data, magnetic resonance imaging [MRI] scans, and X-ray images), and the number of characters in the text data. Odds ratios (ORs) and associated 95% CIs were used to estimate the relative risks of potential predictors of the final diagnosis included within the top 10 differential diagnosis lists in the univariable logistic regression model.

Results

Case Descriptions and Image Data Profile

A total of 363 case descriptions with additional image data, such as figures or tables, were included. ChatGPT-4V generated the differential diagnosis lists for all case descriptions. Representative final diagnosis, image data, and differential diagnosis lists generated by ChatGPT-4V and ChatGPT-4 without vision are shown in Table 2. The cases included in this study, along with the differential diagnosis lists generated by ChatGPT-4V and without vision, are shown in Multimedia Appendix 1.

Table 2. Representative final diagnoses, image data, and differential diagnosis lists generated by ChatGPT-4 with vision and ChatGPT-4 without vision.

Final diagnoses	ChatGPT-4 with vision (intervention)			ChatGPT-4 without vision (control)
	Image data	Image data weight[a], %	Differential diagnosis lists	Differential diagnosis lists
Olanzapine intoxication	Laboratory data	20%	1. Hospital-acquired pneumonia 2. Sepsis 3. Acute kidney injury 4. Respiratory failure 5. Intensive care unit delirium 6. Hypotension 7. Opioid intoxication 8. Benzodiazepine intoxication 9. Hyperglycemia 10. Cardiovascular complications after surgery	1. Zolpidem overdose or prolonged effect 2. Benzodiazepine overdose 3. Olanzapine overdose or side effect 4. Interaction between zolpidem and olanzapine 5. Delirium 6. Other drug overdose 7. Severe sepsis 8. Neuroleptic malignant syndrome 9. Acute cholinergic crisis 10. Acute stroke
Secondary syphilis	Laboratory data and X-ray images	40%	1. Osteoarthritis 2. Rheumatoid arthritis 3. Gout 4. Pseudogout 5. Systemic lupus erythematosus 6. Psoriatic arthritis 7. Reactive arthritis 8. Osteoporosis 9. Septic arthritis 10. Ankylosing spondylitis	1. Secondary syphilis 2. Reactive arthritis 3. Rheumatoid arthritis 4. Osteoarthritis 5. Gout/pseudogout 6. Psoriatic arthritis 7. Lyme disease 8. Systemic lupus erythematosus 9. Infectious arthritis 10. Vasculitis (eg, granulomatosis with polyangiitis and microscopic polyangiitis)
Sarcomatoid carcinoma	Computed tomography image and pathological specimens	30%	1. Colorectal cancer 2. Adenocarcinoma of the colon 3. Malignant mesothelioma 4. Gastrointestinal stromal tumor 5. Carcinoid tumor of the appendix 6. Lymphoma 7. Desmoid tumor 8. Metastatic carcinoma 9. Leiomyosarcoma 10. Neuroendocrine tumor of the colon	1. Colorectal carcinoma 2. Metastatic malignancy to the colon 3. Peritoneal mesothelioma 4. Sarcomatoid carcinoma 5. Malignant mixed Mullerian tumor (carcinosarcoma) 6. Gastrointestinal stromal tumor 7. Leiomyosarcoma 8. Colonic lymphoma 9. Malignant peripheral nerve sheath tumors 10. Undifferentiated/unclassified malignancies

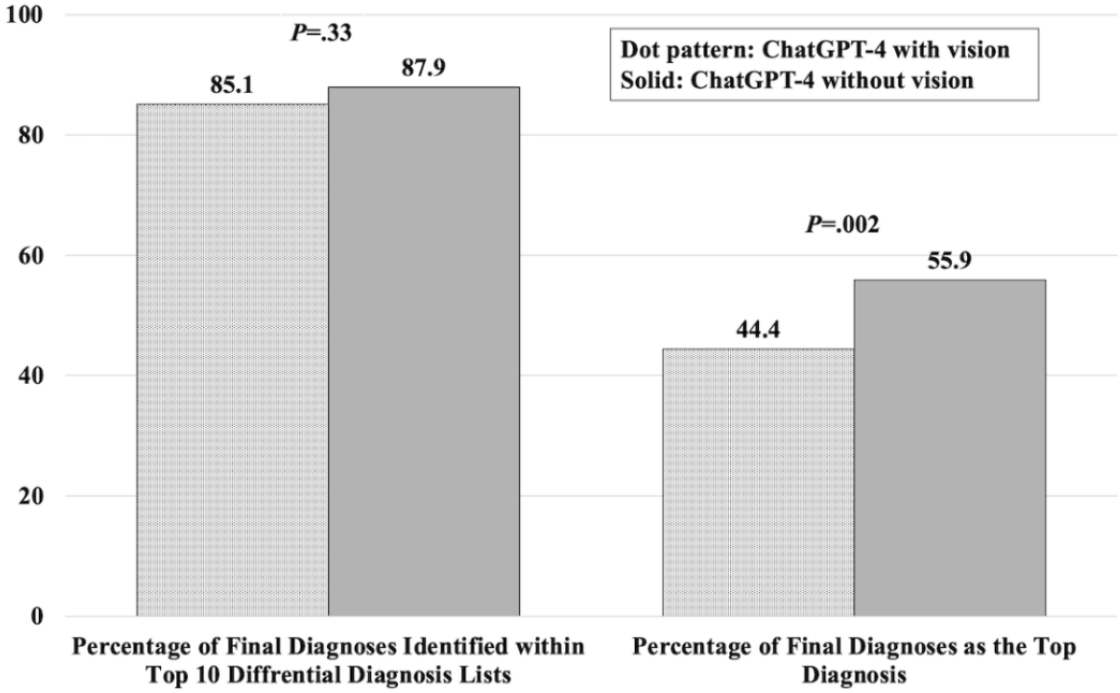
[a]The proportion of image data weight contributing to development of the differential-diagnosis lists.

Among these, the 25th percentile, median, and 75th percentile number of characters in the text data were 1971, 2683, and 3442, respectively. The maximum and minimum number of characters in text data were 7148 and 465, respectively. Regarding image data, CT images, pathological specimens, laboratory data, MRI scans, and X-ray images were included in 163, 124, 98, 77, and 70 case descriptions, respectively. The details of image data are shown in Multimedia Appendix 2.

Diagnostic Performance

For the primary outcome, the rate of final diagnoses within the top 10 differential diagnosis lists generated by ChatGPT-4V was 85.1% (n=363), compared with 87.9% (n=363) by ChatGPT-4 without vision (*P*=.33). For the secondary outcome, the rate of final diagnoses as the top diagnoses generated by ChatGPT-4V was 44.4% (n=363), inferior to 55.9% (n=363) by ChatGPT-4 without vision (*P*=.002). Figure 2 shows the rate of final diagnoses within the top 10 differential diagnosis lists and as the top diagnoses generated by ChatGPT-4V and without vision.

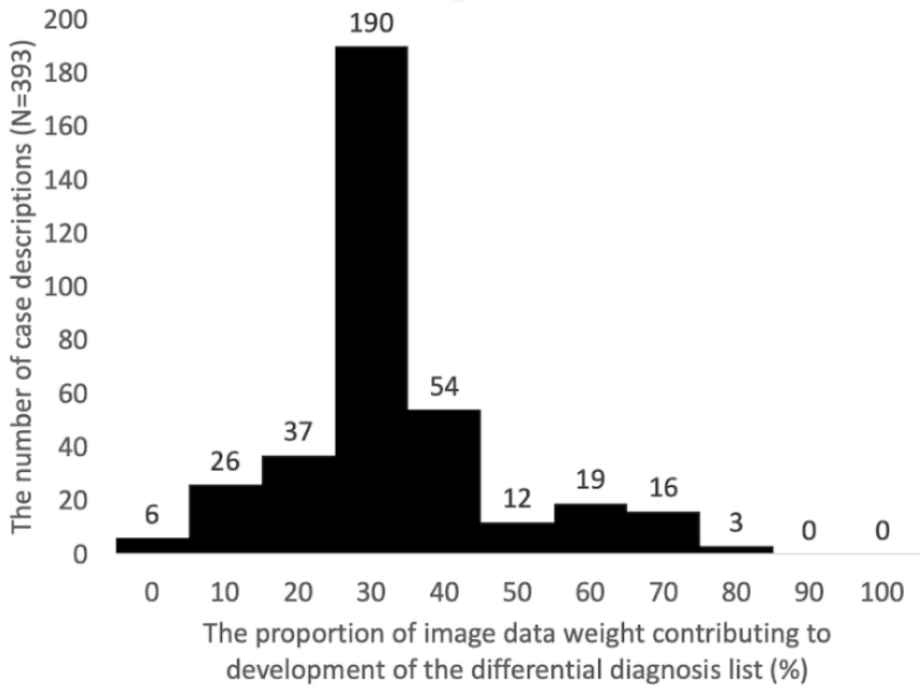
Figure 2. The rate of final diagnoses within the top 10 differential diagnosis lists and as the top diagnoses generated by ChatGPT-4 with vision and without vision.



The Contributing Weight Between Text and Image Data in Developing the Differential Diagnosis Lists

The 25th percentile, median, and 75th percentile proportions of image data weight contributing to the development of the differential diagnosis lists were 30%, 30%, and 40%, respectively, indicating a consistent reliance on image data across a significant portion of cases. The maximum and minimum proportion of image data weight contributing to the development of the differential diagnosis lists were 80% and 0%, respectively, highlighting the wide range of reliance on image data across different case reports. Specifically, in 190 case descriptions of the total 363 included case reports (190/363, 52.3%), the proportion of image data weight contributing to the development of the lists was reported to be 30%. Figure 3 shows the proportion of image data weight contributing to the development of the differential diagnosis lists.

Figure 3. The proportion of image data weight contributing to the development of the differential diagnosis lists by ChatGPT-4 with vision.



The ORs of Variables for Predicting the Outcomes

Laboratory data independently predicted the inclusion of the final diagnoses within the top 10 differential diagnosis lists by ChatGPT-4V: OR 0.52 (95% CI 0.29-0.97; *P*=.03). Additionally, MRI scans were also found to be independent predictive factors: OR 3.87 (95% CI 1.51-13.11; *P*=.01). These results were derived from univariable logistic regression models. Other variables, including the proportion of image data weight contributing to the development of the differential diagnosis lists, CT images, pathological specimens, X-ray images, and the number of characters in text data, were not associated with the final diagnoses included within the top 10 differential diagnosis lists by ChatGPT-4V, as shown in Figure 4.

Additionally, MRI scans (OR 1.93, 95% CI 1.16-3.22; *P*=.01) were independent predictive factors for the final diagnoses as top diagnoses by ChatGPT-4V. Other variables were not associated with the secondary outcome, as shown in Figure 5.

Figure 4. Odds ratios of variables for predicting the final diagnoses included within the top 10 differential diagnosis lists by ChatGPT-4 with vision in univariable regression model. *P* values are derived from the univariable logistic regression model. CT: computed tomography; MRI: magnetic resonance imaging.

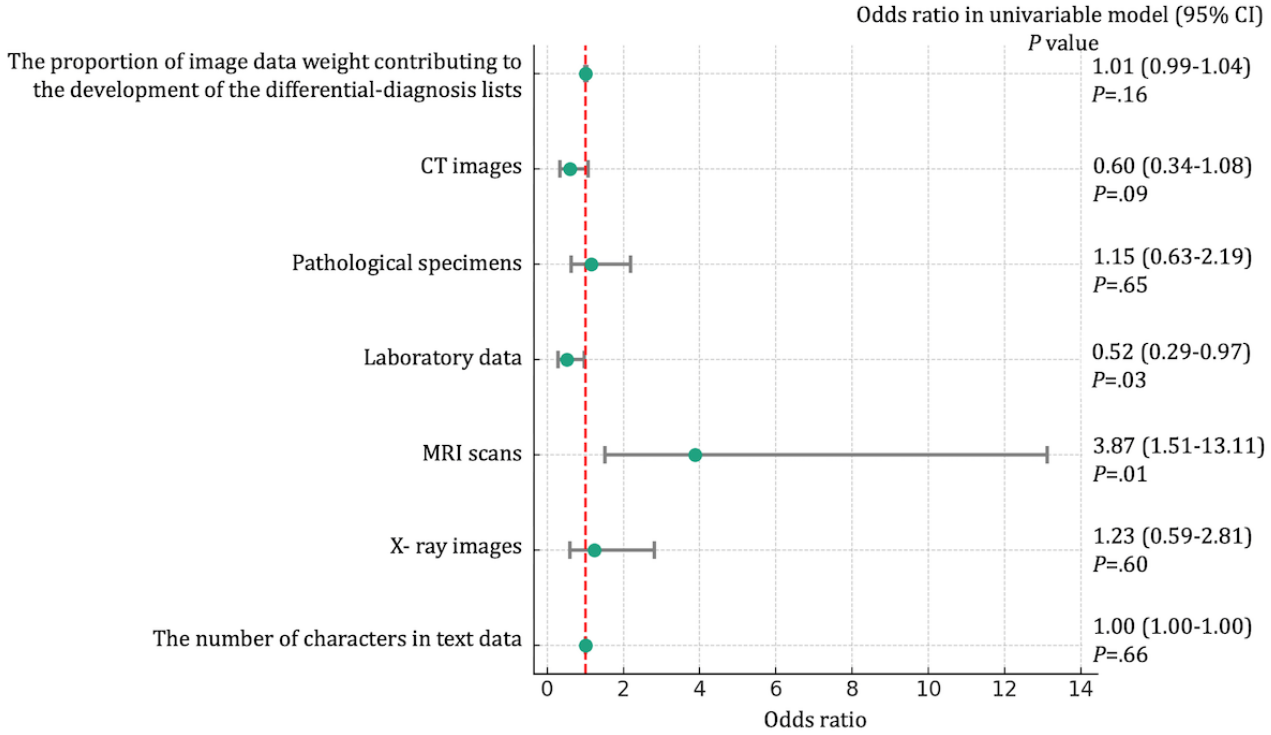
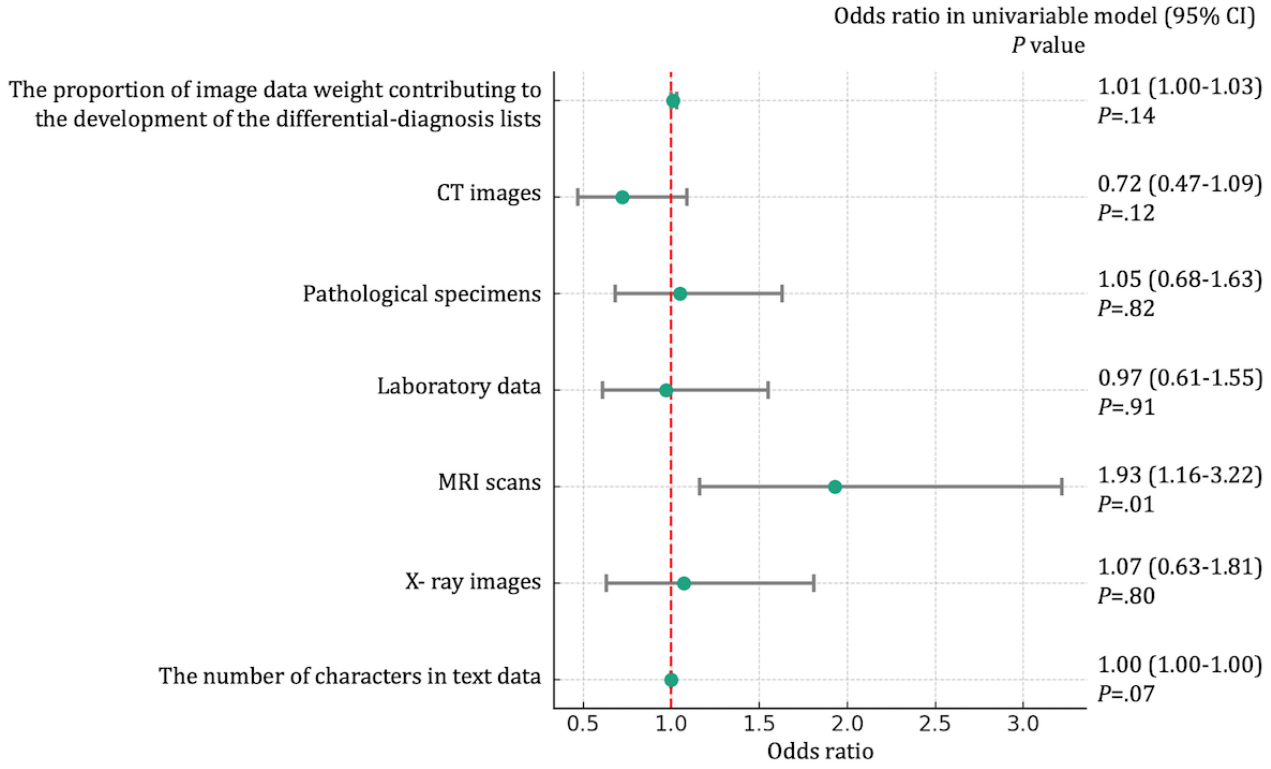


Figure 5. Odds ratios of variables for predicting the final diagnoses as top diagnoses by ChatGPT-4 with vision in univariable regression model. *P* values are derived from the univariable logistic regression model. CT: computed tomography; MRI: magnetic resonance imaging.



Discussion

Principal Results

This study showed several key findings regarding the diagnostic capabilities of ChatGPT-4 with and without vision. The incorporation of image data into ChatGPT-4V did not yield a significant improvement in diagnostic accuracy compared with that without vision. This was evident in the rates of final diagnoses within the top 10 differential diagnosis lists generated by ChatGPT-4V, where ChatGPT-4 without vision actually demonstrated comparable performance. Conversely, the rate of final diagnoses as the top diagnoses generated by ChatGPT-4V was inferior to that without vision. While ChatGPT-4V accepts a wide range of medical images, from physical examinations to various investigation results, its potential to enhance diagnostic accuracy appears underused. This underuse of image processing capabilities could be attributed to the current AI model's limitations in processing and integrating complex image data with textual data. Additionally, the AI system's training regimen, which might have emphasized text data over image data, could have resulted in a bias toward text-based analysis. Future iterations of AI systems should focus on enhancing the model's ability to discern and integrate key diagnostic features from both text and images.

In the univariable logistic regression model, these findings suggest that while the integration of image data by ChatGPT-4V did not uniformly improve diagnostic accuracy across all cases, specific types of image data, particularly MRI scans, play a crucial role in certain diagnostic contexts. MRI scans were associated with significantly higher rates of primary and secondary outcomes. Conversely, laboratory data were associated with significantly lower rates of the primary outcome. These results suggest that MRI scans are typically focused on specific body locations to target particular organs. For example, the inclusion of brain MRI scans led ChatGPT-4V to focus its differential diagnoses on cerebral diseases. The characteristics of MRI scans to focus on anatomical regions could be used to enhance the diagnostic performance of ChatGPT-4V in identifying specific conditions. Moreover, the laboratory data, often presented in tables, typically cover a broader spectrum of information than the case descriptions. For instance, in the case of infectious diseases with elevated blood glucose levels which were included only in the table, ChatGPT-4V considered hyperglycemic condition in addition to the final diagnoses. Incorporating additional laboratory data into the textual analysis could broaden the differential diagnosis lists, potentially reducing the primary outcome. The logistic regression analysis thus provides valuable insights into how different data formats influence the AI's diagnostic capabilities, guiding future improvements in AI design and training to better leverage these inputs.

Focusing on the proportion of image data weight contributing to the development of the differential diagnosis lists, a notable observation emerges regarding ChatGPT-4V's reliance. In more than half of the outputs, image data accounted for 30% of the weight in developing the differential diagnosis lists. This finding leads us to consider the system's internal decision-making process. It is important to consider that the accuracy of the proportion of image data weight in representing the actual process of integrating text and image input in ChatGPT-4V remains uncertain. Despite the consideration, the proportion of image data weight further indicates a dominant dependence on text data. It raises the possibility that ChatGPT-4V may not be

integrating text and image inputs in a balanced way. The implication here is that even with its capability to process image data, the system's diagnostic output might still be mainly influenced by text data.

Given these findings, this unexpected outcome leads us to question why additional image data did not contribute to improvements in diagnostic accuracy. Exploring the reasons behind these results, one plausible explanation emerges related to the potential biases in ChatGPT-4V's use of image data. The biases would be rooted in its training regimen. Rather than aiding in diagnosis, this image data could introduce complexity, leading ChatGPT-4V to rely more on text-based analysis and less on visual clues.

This study highlights the challenges in harnessing the full potential of multimodal AI in medical diagnostics. The findings indicate that despite the advanced capabilities of ChatGPT-4V, its integration of image data is not yet optimizing diagnostic outcomes. This would be partly because of the system's inherent design and training, which could predispose it to prioritize text over image data, despite the latter's potential richness in clinical information. This revelation is crucial for the ongoing development of AI in health care, highlighting a pivotal area for improvement. As AI continues to evolve, focusing on the harmonious integration of text and image data will be essential. This study paves the way for future innovations, guiding efforts to refine multimodal AI systems for more accurate, efficient, and reliable medical diagnostics. Future research should particularly explore the development of more sophisticated methods for image analysis and the optimization of multimodal data integration, aiming to improve the current reliance on text data and enhance the diagnostic power of AI in health care settings.

The findings from our study also raise important considerations for the practical application of AI in health care. While AI systems like ChatGPT-4V hold promise for supporting clinical decision-making, their current limitations necessitate a cautious approach to integration into clinical workflows. For instance, AI could serve as a supporting tool for preliminary analysis, helping triage or providing a second opinion in diagnostic challenges, thereby augmenting the expertise of health care professionals rather than replacing it. Health care professionals should be aware of these systems' strengths and weaknesses, leveraging them as support tools rather than definitive diagnostic solutions.

Limitations

There were several limitations in this study. A major limitation of our study was the reliance on selected image data excerpted from case reports [35], rather than whole slices of image data from clinical settings. This limitation partly arose because the current ChatGPT-4V can only process partial slices of image data [27]. This approach, while necessary for concise reporting in cases, may not accurately reflect the complexity and variability encountered in real-world clinical practice. Moreover, we excluded video files. Although generative AI systems currently do not accept video files, their inclusion could potentially improve diagnostic accuracy. Future research should explore incorporating more comprehensive image data sets and video data, technologies permitting, to enhance the AI system's diagnostic capabilities. Furthermore, the study's reliance on data derived from case reports may not encompass the diversity of real-world clinical scenarios [36]. The specificity of data sources inevitably impacts the generalizability of our findings, highlighting a significant challenge in extending our results to different health care settings and populations. Future studies should consider including complete data from real-patient scenarios with various situations.

Beyond these specific limitations, our study underscores broader concerns regarding the integration of AI in health care, particularly the potential bias inherent in the data sets used to train generative AI systems like ChatGPT-4. These biases may impact the generalizability of the AI's diagnostic and predictive capabilities across diverse populations and clinical settings. The absence of regulatory approval for generative AI systems in clinical practice further complicates their potential adoption, while inconsistencies in ChatGPT-4V interpretations of medical imaging underscore the current limitations of these technologies in performing medical functions [25].

Furthermore, the interpretability and explainability of AI-generated diagnoses remain significant hurdles [16]. The deployment of AI in health care settings also raises practical challenges related to the training of health care professionals in AI use and the integration of AI tools into existing clinical workflows. Ensuring that health care workers are adequately prepared to interpret AI-generated insights and make informed decisions is crucial for the successful adoption of AI technologies.

Last, the rapid evolution of AI technology presents unique challenges, as advancements may quickly outpace the findings of our study. The pace at which AI technologies evolve means that our conclusions may become outdated as new capabilities emerge. This highlights the importance of ongoing research and adaptation in the field of AI and health care, ensuring that studies remain relevant and that AI tools are continually evaluated and updated to reflect the latest technological advancements.

Comparison With Prior Work

Compared with a previous preliminary study for ChatGPT-4V, this study showed higher performance. The previous study assessed the proficiency of ChatGPT-4V for selected medical images from open-source libraries and repositories [27]. The study reported that only 21.7% (n=15) of cases were correctly interpreted with the correct advice. This inconsistency was partly because of the methodological differences between the 2 studies, particularly in terms of data set preparation and evaluation criteria. While the previous study mainly focused on a limited data set with simple prompts and evaluated the system's interpretation and medical advice quality, our study introduced a more comprehensive data set with a rich clinical context. Additionally, we evaluated the diagnostic accuracy, rather than merely assessing interpretation and advice, thereby providing a deeper insight into the AI system's utility in clinical decision-making.

Another study evaluated the performance of ChatGPT-4V for selected clinical cases from the website, including image data [26]. The study showed that ChatGPT-4V heavily relies on the patients' medical history. This result was consistent with this study that additional image data did not improve the diagnostic accuracy. The result was also consistent with this study that approximately half of the outputs reported that the proportion of image data weight contributing to the development of the differential diagnosis lists was 30%.

A critical distinction between our study and previous works is our comparative analysis of ChatGPT-4 with and without vision capabilities. This unique approach allowed us to highlight the impact of image data on diagnostic accuracy, revealing that while ChatGPT-4's vision component does not significantly enhance diagnostic accuracy, it does not detract from it either. This finding is crucial for understanding the role of integrated image data in AI-assisted diagnosis and highlights the potential of AI systems to support health care professionals by providing a comprehensive analysis that includes both text and image data.

Conclusions

The rates of final diagnoses within the differential diagnosis lists generated by ChatGPT-4V did not show improvement over those generated without vision. The rate of final diagnoses as the top diagnosis generated by ChatGPT-4V was inferior to that without vision. Despite its multimodal data processing capabilities, ChatGPT-4V appears to prioritize text data, which may limit its effectiveness in medical diagnostic applications, as highlighted by its system card [25]. The implications of our study for the advancement of multimodal AI systems in health care are profound. It uncovers a pivotal aspect of AI development that requires attention: the nuanced integration and weighted analysis of diverse data types. To emulate the complex reasoning of medical professionals, AI systems must advance beyond simple data incorporation toward a sophisticated synthesis that enhances diagnostic accuracy. For future improvements, we recommend the following: enhanced clinical data fusion techniques; interpretability of AI decisions; and collaborative development efforts with AI developers and medical professionals. In clinical practice, more sophisticated multimodal AI systems have the potential to enhance in providing timely, contextually rich differential diagnoses, serving as educational aids for medical trainees, and enhancing patient care by supporting remote or underserved areas. Through these enhancements, AI tools can ultimately improve patient outcomes.

Acknowledgments

TH, YH, KT, TI, T Suzuki, and T Shimizu contributed to the study concept and design. TH performed the statistical analyses. TH contributed to the drafting of the manuscript. YH, KT, TI, T Suzuki, and T Shimizu contributed to the critical revision of the manuscript for relevant intellectual content. All the authors have read and approved the final version of the manuscript. This study was conducted using resources from the Department of Diagnostics and Generalist Medicine at Dokkyo Medical University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The included cases in this study and the differential diagnosis lists generated by ChatGPT-4 with vision and without vision.
[XLSX File (Microsoft Excel File), 138 KB - medinform_v12i1e55627_app1.xlsx]

Multimedia Appendix 2
The details of image data in this study.
[DOCX File , 20 KB - medinform_v12i1e55627_app2.docx]

References

1. Yang D, Fineberg HV, Cosby K. Diagnostic excellence. JAMA 2021;326(19):1905-1906. [doi: 10.1001/jama.2021.19493] [Medline: 34709367]
2. Singh H, Connor DM, Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. BMJ 2022;376:e068044. [doi: 10.1136/bmj-2021-068044] [Medline: 35172968]
3. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3(1):17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
4. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
5. Watari T, Schiff GD. Diagnostic excellence in primary care. J Gen Fam Med 2023;24(3):143-145 [FREE Full text] [doi: 10.1002/jgf2.617] [Medline: 37261043]

6. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021;18(16):8435 [FREE Full text] [doi: 10.3390/ijerph18168435] [Medline: 34444182]
7. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. In: Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care. Washington, DC: National Academies Press; 2015.
8. Tupasela A, Di Nucci E. Concordance as evidence in the Watson for oncology decision-support system. AI Soc 2020;35(4):811-818 [FREE Full text] [doi: 10.1007/s00146-020-00945-9]
9. Potočnik J, Foley S, Thomas E. Current and potential applications of artificial intelligence in medical imaging practice: a narrative review. J Med Imaging Radiat Sci 2023;54(2):376-385 [FREE Full text] [doi: 10.1016/j.jmir.2023.03.033] [Medline: 37062603]
10. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
11. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021;22(1):122 [FREE Full text] [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
12. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland: World Health Organization; 2021.
13. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
14. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
15. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
16. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
17. Bachute MR, Subhedar JM. Autonomous driving architectures: insights of machine learning and deep learning algorithms. Mach Learn Appl 2021;6:100164 [FREE Full text] [doi: 10.1016/j.mlwa.2021.100164]
18. Hashemi-Pour C, Kerner SM, Patrizio A. Google Gemini (formerly Bard). TechTarget. 2023. URL: https://www.techtarget.com/searchenterpriseai/definition/Google-Bard [accessed 2024-03-26]
19. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28(9):1773-1784 [FREE Full text] [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
20. Huang SC, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep 2020;10(1):22147 [FREE Full text] [doi: 10.1038/s41598-020-78888-w] [Medline: 33335111]
21. Jabbour S, Fouhey D, Kazerooni E, Wiens J, Sjoding MW. Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure. J Am Med Inform Assoc 2022;29(6):1060-1068 [FREE Full text] [doi: 10.1093/jamia/ocac030] [Medline: 35271711]
22. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
23. OpenAI. GPT-4 technical report. ArXiv Preprint posted online on March 15 2024. [doi: 10.48550/arXiv.2303.08774]
24. ChatGPT can now see, hear, and speak. OpenAI. URL: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak [accessed 2024-03-26]
25. GPT-4V(ision) system card. OpenAI. 2023. URL: https://openai.com/research/gpt-4v-system-card [accessed 2024-03-26]
26. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, et al. Can GPT-4V(ision) serve medical applications? case studies on GPT-4V for multimodal medical diagnosis. ArXiv Preprint posted online on December 04, 2023. [doi: 10.48550/arXiv.2310.09909]
27. Senkaiahliyan S, Toma A, Ma J, Chan AW, Ha A, An KR, et al. GPT-4V(ision) unsuitable for clinical care and education: a clinician-evaluated assessment. medRxiv Preprint posted online on November 16, 2023. [doi: 10.1101/2023.11.15.23298575]
28. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in Japanese national medical licensing examination. medRxiv Preprint posted online on November 08, 2023. [doi: 10.1101/2023.11.07.23298133]
29. Driessen T, Dodou D, Bazilinskyy P, de Winter J. Putting ChatGPT Vision (GPT-4V) to the test: risk perception in traffic images. ResearchGate. 2023. URL: https://www.researchgate.net/publication/375238184_Putting_ChatGPT_Vision_GPT-4V_to_the_test_Risk_perception_in_traffic_images [accessed 2024-03-26]
30. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). ArXiv Preprint posted online on October 11, 2023. [doi: 10.48550/arXiv.2309.17421]
31. Yang J, Zhang H, Li F, Zou X, Li C, Gao J. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. ArXiv Preprint posted online on November 06, 2023. [doi: 10.48550/arXiv.2310.11441]

32. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008;23(Suppl 1):37-40 [FREE Full text] [doi: 10.1007/s11606-007-0271-8] [Medline: 18095042]
33. Miao J, Gibson LE, Craici IM. Levofloxacin-associated bullous pemphigoid in a hemodialysis patient after kidney transplant failure. Am J Case Rep 2022;23:e938476 [FREE Full text] [doi: 10.12659/AJCR.938476] [Medline: 36578185]
34. Krupat E, Wormwood J, Schwartzstein RM, Richards JB. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. Med Educ 2017;51(11):1127-1137. [doi: 10.1111/medu.13382] [Medline: 28857266]
35. Riley DS, Barber MS, Kienle GS, Aronson JK, von Schoen-Angerer T, Tugwell P, et al. CARE guidelines for case reports: explanation and elaboration document. J Clin Epidemiol 2017;89:218-235 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.04.026] [Medline: 28529185]
36. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. J Med Internet Res 2022;24(10):e37408 [FREE Full text] [doi: 10.2196/37408] [Medline: 36287594]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
ChatGPT-4V: ChatGPT-4 with vision
CT: computed tomography
LLM: large language model
MRI: magnetic resonance imaging
OR: odds ratio

Edited by A Castonguay; submitted 18.12.23; peer-reviewed by D Hu, D Singh, TAR Sure; comments to author 07.02.24; revised version received 14.02.24; accepted 13.03.24; published 09.04.24.

Please cite as:
Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T
Evaluating ChatGPT-4's Diagnostic Accuracy: Impact of Visual Data Integration
JMIR Med Inform 2024;12:e55627
URL: https://medinform.jmir.org/2024/1/e55627
doi:10.2196/55627
PMID:38592758

Original Paper

Natural Language Processing–Powered Real-Time Monitoring Solution for Vaccine Sentiments and Hesitancy on Social Media: System Development and Validation

Liang-Chin Huang[1], PhD; Amanda L Eiden[2], PhD; Long He[1], MS; Augustine Annan[1], PhD; Siwei Wang[1], MS; Jingqi Wang[1], PhD; Frank J Manion[1], PhD; Xiaoyan Wang[1], PhD; Jingcheng Du[1], PhD; Lixia Yao[2], PhD

[1]Melax Tech, Houston, TX, United States

[2]Merck & Co, Inc, Rahway, NJ, United States

Corresponding Author:
Amanda L Eiden, PhD
Merck & Co, Inc
2025 E Scott Ave
Rahway, NJ, 07065
United States
Phone: 1 7325944000
Email: amanda.eiden@merck.com

Abstract

Background: Vaccines serve as a crucial public health tool, although vaccine hesitancy continues to pose a significant threat to full vaccine uptake and, consequently, community health. Understanding and tracking vaccine hesitancy is essential for effective public health interventions; however, traditional survey methods present various limitations.

Objective: This study aimed to create a real-time, natural language processing (NLP)–based tool to assess vaccine sentiment and hesitancy across 3 prominent social media platforms.

Methods: We mined and curated discussions in English from Twitter (subsequently rebranded as X), Reddit, and YouTube social media platforms posted between January 1, 2011, and October 31, 2021, concerning human papillomavirus; measles, mumps, and rubella; and unspecified vaccines. We tested multiple NLP algorithms to classify vaccine sentiment into positive, neutral, or negative and to classify vaccine hesitancy using the World Health Organization's (WHO) 3Cs (confidence, complacency, and convenience) hesitancy model, conceptualizing an online dashboard to illustrate and contextualize trends.

Results: We compiled over 86 million discussions. Our top-performing NLP models displayed accuracies ranging from 0.51 to 0.78 for sentiment classification and from 0.69 to 0.91 for hesitancy classification. Explorative analysis on our platform highlighted variations in online activity about vaccine sentiment and hesitancy, suggesting unique patterns for different vaccines.

Conclusions: Our innovative system performs real-time analysis of sentiment and hesitancy on 3 vaccine topics across major social networks, providing crucial trend insights to assist campaigns aimed at enhancing vaccine uptake and public health.

(JMIR Med Inform 2024;12:e57164) doi:10.2196/57164

KEYWORDS

vaccine sentiment; vaccine hesitancy; natural language processing; NLP; social media; social media platforms; real-time tracking; vaccine; vaccines; sentiment; sentiments; vaccination; vaccinations; hesitancy; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; machine learning; uptake; willing; willingness; classification

Introduction

Vaccine is an essential public health intervention that has saved millions of lives and achieved a substantial global reduction in cases, hospitalizations, and health care costs associated with vaccine-preventable diseases (VPDs) [1-3]. Yet, despite their value, vaccine hesitancy persists as a barrier to full vaccine uptake. The World Health Organization (WHO) defines vaccine hesitancy as the delay or refusal of vaccination, even when vaccination services are accessible [4]. Additionally, the WHO identifies vaccine hesitancy as one of the top 10 global health threats [5]. Delay or refusal of vaccines due to vaccine hesitancy can have broad-reaching implications; unvaccinated individuals not only put themselves at risk of VPDs, such as COVID-19,

but also pose a threat to the broader community or even global health [6]. This phenomenon has been documented since the advent of vaccines in over 90% of the countries [7]. Considering the case of measles, mumps, and rubella (MMR), it is crucial to uphold community protection or herd immunity, necessitating widespread vaccination to protect those unable to receive the vaccine [8]. A former London study successfully raised MMR vaccination rates from 80% to 94% in under 2 years through incentivized care packages and innovative technology use, approaching the desired herd immunity target [9].

There is a myriad of reasons for vaccine hesitancy, including personal or familial beliefs, concerns about adverse reactions or efficacy, and skepticism toward government and vaccine manufacturers [6,10-17]. This intricate web of motivations makes vaccine hesitancy a complex public health challenge [18].

Understanding vaccine hesitancy is crucial for developing effective interventions, public health education, and vaccination promotion strategies [19-22]. While surveys have traditionally served as a valuable tool for gathering public opinions on vaccination, they possess inherent limitations such as static data collection, resource intensiveness, and potential time lag [23-29]. To address these limitations, real-time tracking of vaccine hesitancy activities and trends offers public health professionals' valuable insights. This approach helps identify critical intervention points before the vaccination uptake wanes, allowing for more targeted and timely communication efforts.

The emergence of social media platforms has enabled billions of users to engage in discussions, information sharing, and opinion expression on various subjects, including health-related topics [30]. While this presents an unprecedented opportunity for public health improvement, it also poses a significant risk linked to the dissemination of vaccine-related misinformation and disinformation [31]. Previous research has used semiautomatic methods such as manual coding and hashtag or keyword analysis to study social media vaccine discussions [32-35]. Nevertheless, these approaches may sometimes encounter potential challenges with scalability and precision. Natural language processing (NLP) is an automated method designed to effectively and accurately decipher the wealth of information in natural language text, addressing challenges such as ambiguities and probabilistic parsing, and enabling applications such as information extraction and discourse analysis [36]. This technique has emerged as a promising solution, holding the potential to mitigate these challenges and improve the precision of vaccine-related public sentiment analysis [37,38].

To address these challenges, this study's principal aim was to create an NLP system for real-time monitoring of vaccine sentiment and hesitancy across English-language social media platforms targeting the US market. Our 3-fold contributions are (1) developing one of the first real-time monitoring systems for social media vaccine discussions that covers 3 major social media platforms and 3 vaccine topic groups [39]; (2) comprehensively evaluating multiple machine learning–based NLP models for social media post classification tasks, thus establishing a benchmark for future research; and (3) analyzing decade-long trends of sentiment and hesitancy and linked real-world events to corresponding points on the trends for multiple vaccine targets.

Methods

Overview

We followed a systematic approach to monitor vaccine sentiment and hesitancy posts on Twitter (subsequently rebranded as X), Reddit, and YouTube. We selected Twitter, Reddit, and YouTube as they are the primary social media platforms offering substantial volumes of posts through application programming interface (API) access [40-43]. We focused exclusively on English language posts given the widespread use of English in the largest market countries for our target vaccines and with regard to the accessibility of English language social media. Other platforms and languages, such as Facebook and Spanish [44], may be of interest for future studies; however, these served as a first approach to research. Figure 1 illustrates our workflow, including data annotation, NLP algorithms, and an online dashboard.

First, we categorized vaccine sentiment into positive, negative, and neutral, which were the labels also used in other sentiment analyses using social media data [45,46]. Then, we aligned vaccine hesitancy with the WHO's 3Cs (confidence, complacency, and convenience) vaccine hesitancy model, described in further detail in the *3Cs Vaccine Hesitancy Annotation* section [4]. The definitions of post sentiment and vaccine hesitancy are comprehensively presented in Table 1. We collected data using vaccine-specific search queries (see Table S1 in Multimedia Appendix 1) for relevant posts from the 3 social media platforms. To ensure the quality and reliability of the data, we collaborated with medical experts to create annotated corpora aligned with the information model. These corpora were then used to train NLP algorithms to automatically extract vaccine sentiment and hesitancy content. Finally, we developed an online dashboard to provide real-time insights into vaccine sentiment and hesitancy trends. Our study focuses on evaluating the vaccine sentiment and hesitancy of human papillomavirus (HPV), MMR, and general or unspecified vaccines. The critical role of the vaccines is exemplified by the HPV vaccine, which has effectively reduced prevalent HPV infections and precancerous lesions, underlining the importance of global implementation [47], and the MMR vaccine is renowned for its safety and efficacy, which has greatly mitigated endemic diseases in the United States [48]. Despite these successes, challenges such as insufficient vaccination coverage, increasing hesitancy, and the resurgence of mumps, attributed to waning immunity and antigenic variation, persist worldwide. Throughout the COVID-19 pandemic up to 2022, HPV and MMR were the vaccines that maintained the greatest negative impact on routine vaccinations in the United States, suggesting a need for proactive efforts to increase vaccination coverage to prevent associated health complications and costs [49].

Figure 1. The overview of study design and classifications used to evaluate vaccine-related posts. 3Cs: confidence, complacency, and convenience; ML: machine learning; WHO: World Health Organization.

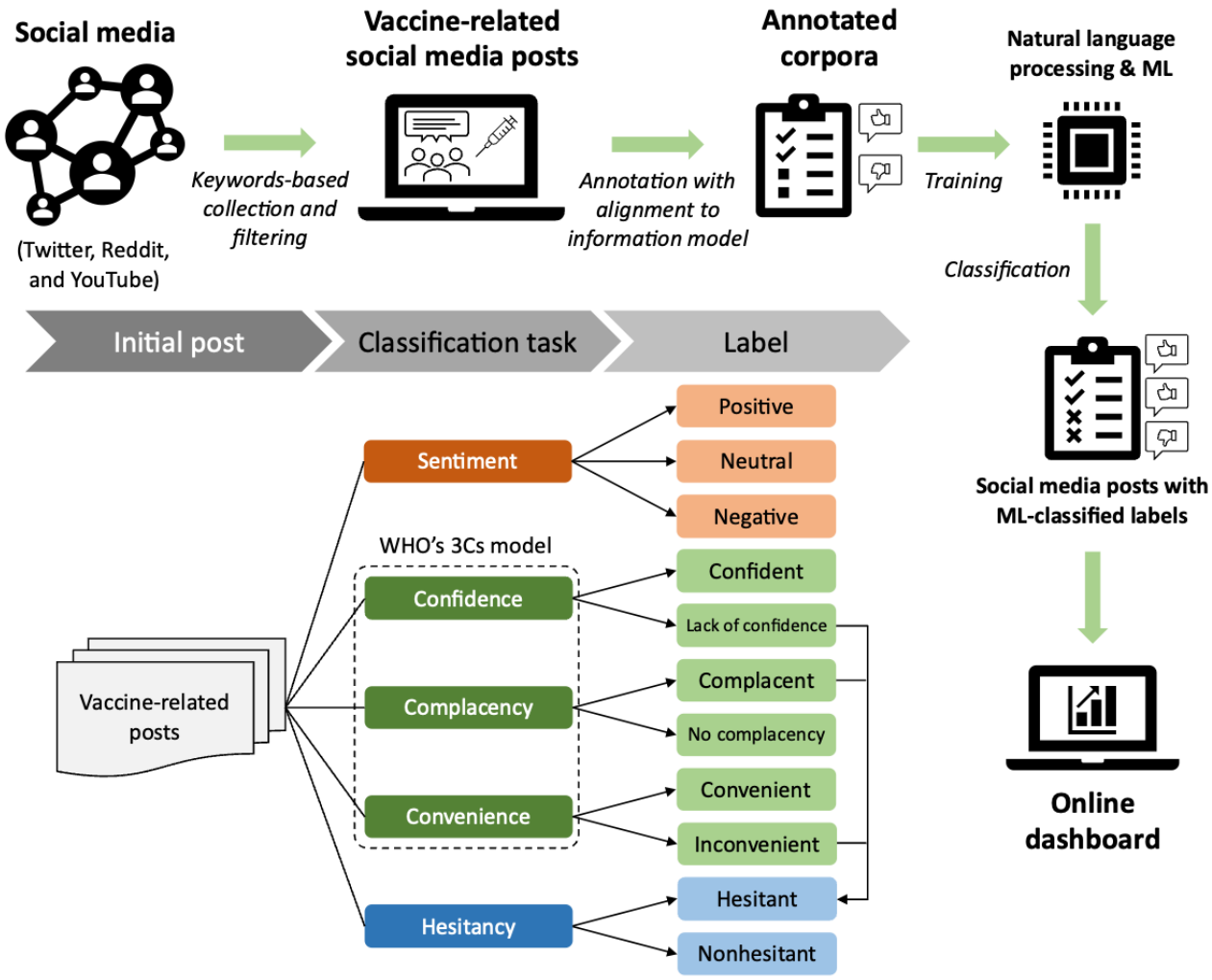


Table 1. Definitions of post sentiment and hesitancy.

Classification task and label	Definition
Sentiment	
Positive	Posts that mention, report, or share positive news, opinions, or stories about vaccines or vaccination.
Neutral	Posts that are related to vaccines or vaccination topics but contain no sentiment, the sentiment is unclear, or they contain both negative and positive sentiments.
Negative	Posts that mention, report, or share negative news, opinions, or stories about vaccines or vaccination, which may discourage vaccination.
Confidence	
Confident	Posts reflecting a trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
Lack of confidence	Posts reflecting a lack of trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
Complacency	
Complacent	Posts where the perceived risks of VPDs[a] are low and vaccination is deemed as an unnecessary preventive action.
No complacency	Posts where the perceived risks of VPDs are high and vaccination is deemed as a necessary preventive action.
Convenience	
Convenient	Posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services do not affect uptake.
Inconvenient	Posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services affect uptake.
Hesitancy	
Hesitant	The post is labeled as lack of confidence, complacent, or inconvenient.
Nonhesitant	The post is not labeled as lack of confidence, complacent, or inconvenient.

[a]VPD: vaccine-preventable disease.

Social Media Data Collection

The systematic collection of social media data spanned from January 1, 2011, to October 31, 2021, across 3 platforms—Twitter, Reddit, and YouTube. During initial exploratory analysis, we recognized variations in text nature and query logic across these platforms, leading us to tailor our search queries for each platform to collect relevant posts while excluding irrelevant ones. Table S1 in Multimedia Appendix 1 lists the customized queries on each platform for each vaccine topic group, which include both inclusion and exclusion keywords. We retrieved the results (relevant posts) using the APIs provided by the 3 platforms. Details about the software versions are described in the Multimedia Appendix 1. To clarify ethical considerations and data privacy issues, when gathering data from Twitter, YouTube, and Reddit, we adhered to their API's data privacy policies and ensured the deidentification of all posts and videos by assigning them a unique random ID.

Ethical Considerations

Ethics board review was not required, as all modelling data came from public sources and there were no ethical issues. The data privacy policies of the application program interfaces (APIs) of Twitter, YouTube, and Reddit were followed when gathering data. We ensured the deidentification of all posts and videos by assigning them a unique random ID.

Data Annotation

From the retrieved results, approximately 90 million posts, we randomly selected 60,000 social media discussions. These posts were manually annotated to build both training and evaluation data sets, which were used for building the text classifiers. We selected 20,000 posts for annotation, including 10,000 tweets, 5000 Reddit posts, and 5000 YouTube comments for each vaccine topic group, including HPV vaccine, MMR vaccine, and general or unspecified vaccines. During annotator training, 4 annotators with a medical training background were recruited for the annotation. An annotation guideline was developed. All annotators first annotated the same 1000 tweets, 1000 Reddit posts, and 1000 YouTube posts independently, and then discussed collectively for any discrepancies. After all discrepancies were resolved through discussions, these annotators began to annotate the rest of the social media posts. A 2-fold annotation strategy was used, where first, we annotated the sentiment of the post as positive, neutral, or negative, assigning only 1 category to each post; and second, we annotated vaccine hesitancy based on the constructs of the WHO 3Cs model, which include confidence, complacency, and convenience (Figure 1). These annotation categories also define each classification task.

Sentiment Annotation

The annotation task involved assigning 1 of 3 sentiment labels to each post, which constituted a multiple-class classification problem. The labels and corresponding illustrative examples are defined in Textbox 1.

Textbox 1. Definitions and examples of sentiment labels.

- Positive: posts that mention, report, or share positive news, opinions, or stories about vaccines or vaccination.
  - Example: "HPV vaccine, prevents against the two HPV types, 16 and 18, which cause 70% of cervical cancers"
  - Example: "Get vaccinated against HPV to protect you in the future for now!"
- Neutral: posts that are related to vaccines or vaccination topics but contain no sentiment, the sentiment is unclear, or they contain both negative and positive sentiments.
  - Example: "The following report is specifically for the MMR vaccine, but you can browse around for others"
  - Example: "I just learned that there are more than 50 strains of HPV...I always thought the vaccine prevented all strains."
- Negative: posts that mention, report, or share negative news, opinions, or stories about vaccines or vaccination, which may discourage vaccination.
  - Example: "According to a report, thousands of kids suffer permanent injury or death by getting vaccines"
  - Example: "Believe it? Vaccines have killed 1000 more kids than any measles!"

3Cs Vaccine Hesitancy Annotation

The annotation task involved assigning multiple labels to each post according to the 3Cs model constructs. Annotators checked each construct to determine whether the post was related to it separately. If any of the constructs were labeled as "lack of confidence," "complacent," or "inconvenient," we considered the post as vaccine hesitant; otherwise, it was considered vaccine nonhesitant. Definitions and examples for each 3Cs model construct are provided in Textbox 2.

Table S2 in Multimedia Appendix 1 provides examples of specific social media posts with annotations for the different categories. The distribution of annotated posts in each sentiment and 3Cs construct for each platform and vaccine topic group is shown in Table S3 in Multimedia Appendix 1.

Textbox 2. Definitions and examples of World Health Organization's 3Cs (confidence, complacency, and convenience) model.

- Lack of confidence: posts reflecting a lack of trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
  - Example: "Fully vaccinated are 30 times more likely to get COVID-19, and 10 times more likely to require hospitalization."
  - Example: "The vaccine label includes all these events. Concerns have been raised about reports of deaths occurring in individuals after receiving that vaccine."
- Complacency: posts where the perceived risks of vaccine-preventable diseases are low, and vaccination is deemed as an unnecessary preventive action.
  - Example: "Why do adults need to know about the measles vaccine? The measles is a benign disease and there is no need for vaccines."
  - Example: "I wasn't vaccinated against a preventable disease. It's not always just a life-or-death dichotomy - I recovered."
- Inconvenience or convenience: posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services affect uptake.
- Example: "I am 30-year-old man and am looking for an HPV vaccine. Unfortunately, my insurance only covers it for women. I am particularly at risk for certain cancers. I really don't understand how insurance companies are allowed to make the gender distinction when the FDA approved it for both."

Text Classification Algorithms

Overview

To classify the sentiment and hesitancy of social media posts, we compared the performance of 5 text classification algorithms—logistic regression (LR) [50], support vector machine (SVM) [51], random forest [52], extreme gradient boosting (XGBoost) [53], and Snorkel [54]. Each of these models has unique characteristics, which are summarized below.

LR Algorithm

LR is a classic statistical methodology that models a binary dependent variable using a logistic function. It is favored in medical research due to its ability to determine the odds ratio, indicating the potential change in outcome probabilities [55].

SVM Algorithm

SVM is one of the most robust classification methods based on statistical learning frameworks. It finds a hyperplane in an N-dimensional space that distinctly classifies data points. In

medical text mining, SVM combined with other algorithms has demonstrated effective performance in extracting and recognizing entities in clinical text, contributing notably to improved patient care [56].

Random Forest Algorithm

Random forest is a classifier that uses ensemble learning to combine decision tree classifiers through bagging or bootstrap aggregating. It has been applied to highly ranked features obtained through suitable ranker algorithms and has shown promising results in medical data classification tasks, enhancing the prediction accuracy for various diseases [57].

XGBoost Algorithm

XGBoost is an ensemble of algorithms that turn weak learners into strong learners by focusing on where the individual models went wrong. In gradient boosting, individual weak models train upon the difference between the classification and the actual results. It has been effective in mining and classifying suggestive sentences from online customer reviews by combining them with a word-embedding approach [58].

Snorkel Algorithm

Snorkel is a system that enables users to train models without hand labeling all training data by writing their labeling functions. Using Snorkel enables the extraction of chemical reaction relationships from biomedical literature abstracts, supporting the understanding of biological processes without requiring a large, labeled training data set [59].

We extracted the term frequency–inverse document frequency vector for each word in all text classification algorithms using *scikit-learn*'s *TfidfTransformer* function with default parameter settings. Term frequency–inverse document frequency evaluates how relevant a word is to a text in a collection of texts [60]. If the model encounters a new post with words or symbols not included in its original bag of words, it will effectively ignore those words during the transformation process. To ensure a balanced training set, the 3 class-balancing methods implemented by Python *imblearn* package applied were (1) random oversampling, (2) synthetic minority over-sampling technique (SMOTE) [61], and (3) SVM-based SMOTE [62] (with the default parameter settings, specifically k_neighbors=5, as they exhibited the optimal performance within the developer's data set [61]). SMOTE randomly selects a minority class instance, finds one of its nearest minority class neighbors, and then synthesizes an instance between these 2 instances in the feature space. SVM-based SMOTE uses support vectors to determine the decision boundaries and then synthesizes a minority class instance along the decision boundary.

NLP Evaluation

The evaluation data sets were created from the annotated corpora and randomly divided into training, validation, and test sets in a 6:2:2 ratio to assess the performance of the 5 text classification algorithms. The models were trained on the training sets, optimized on the validation sets, and then evaluated on the test sets. The following key metrics were calculated to evaluate the models:

A true positive occurs when the model accurately classifies the positive class (positive, negative, or neutral for sentiment; true for 3Cs model constructs). A true negative occurs when the model accurately classifies the negative class (nonpositive, nonnegative, or nonneutral for sentiment; false for 3Cs model constructs). A false positive is an incorrect positive classification, while a false negative is an incorrect negative classification. As the sentiment and hesitancy labels in Tweets, Reddit posts, and YouTube comments are imbalanced, we optimized our models based on F_1-scores, which balance precision and recall, rather than accuracy. The purpose of optimizing a model based on F_1-scores when dealing with imbalanced labels is to achieve a better balance between precision and recall, thereby improving the overall performance of the model. This is especially important in imbalanced data sets where the cost of misclassification can be high.

Dashboard Development

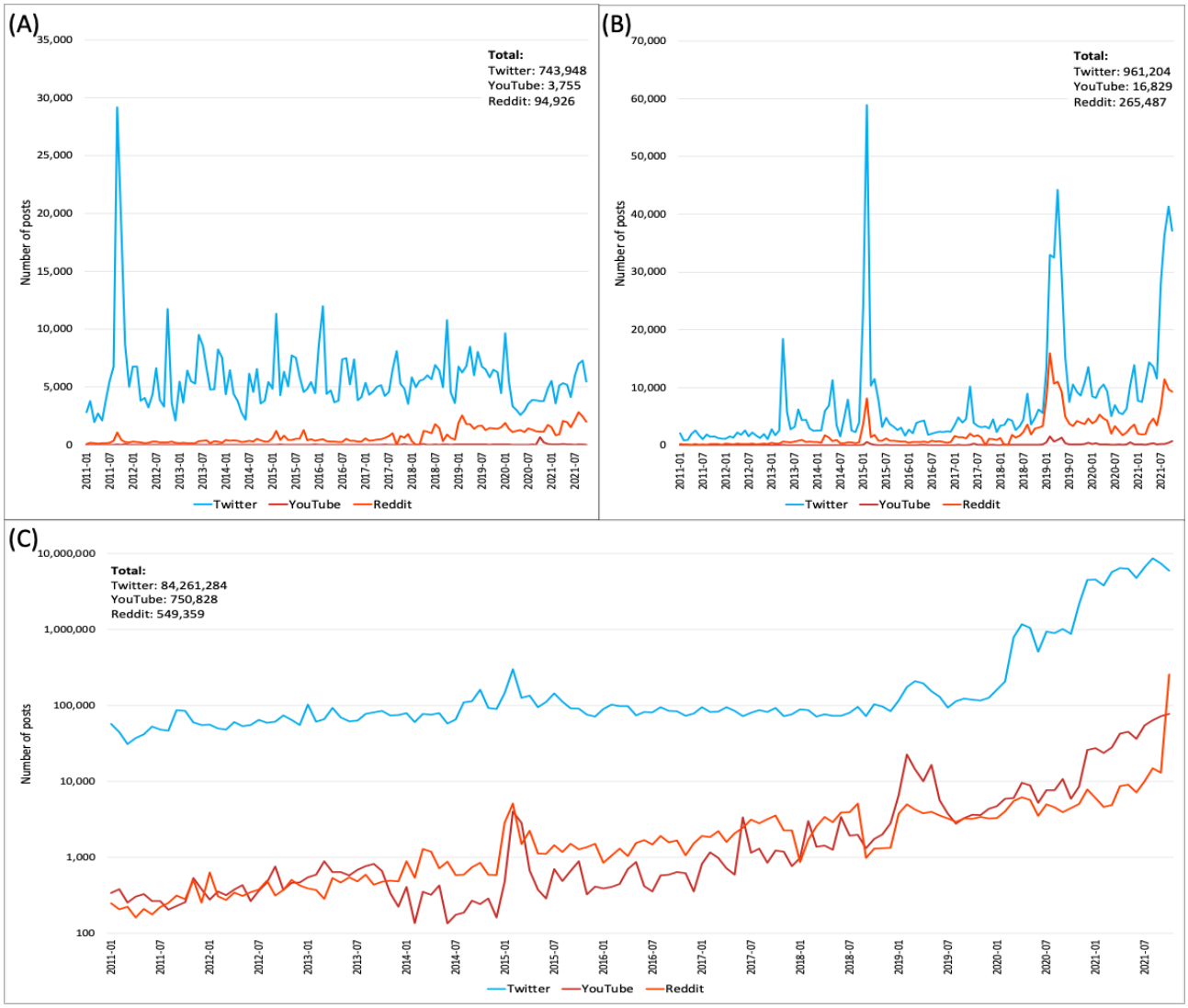
We designed a user-friendly, web-based visualization dashboard [39] for real-time analysis of trends in vaccine sentiment and hesitancy over time and geography (Figure S1A-C in Multimedia Appendix 1). The dashboard also allows for comparisons of sentiment and hesitancy across different social media platforms and vaccine topic groups (Figure S1D in Multimedia Appendix 1). The NLP models were optimized based on their F_1-scores to address the imbalanced labels of sentiment and hesitancy in tweets, Reddit posts, and YouTube comments. The selected models are applied to all unlabeled data collected from 2011 to 2021. Technical details are described and represented in Figure S2 in Multimedia Appendix 1.

Results

Social Media Data Collection Summary

From January 1, 2011, to October 31, 2021, we collected 86 million posts from Twitter, 0.9 million from Reddit, and 76,000 from YouTube, which were related to vaccines. The most widely discussed topic across all 3 platforms was the general or unspecified vaccine, followed by the MMR and then HPV vaccines. We observed a substantial increase in the general vaccine-related discussions on Twitter and Reddit starting in early 2020, coinciding with the onset of the COVID-19 pandemic. The collected social media data and growth trends are plotted in Figure 2.

Figure 2. The long and short-term trends of collected vaccine-related social media post data across 3 different platforms for different vaccine topic groups: (A) human papillomavirus (HPV); (B) measles, mumps, and rubella (MMR); and (C) general or unspecified vaccine.



NLP Performance on Vaccine Sentiment and Hesitancy

We tested all combinations of the 5 NLP algorithms. The performances in sentiment classification, hesitancy classification, and 3Cs classifications are presented in Table 2. The best-performing algorithms (according to F_1-scores) and detailed performance scores for different classification tasks are shown in Tables S4-S8 in Multimedia Appendix 1. In sentiment classification, LR outperformed other algorithms in 7 out of 9 platform–vaccine topic group combinations, with overall accuracies ranging from 0.51 to 0.78 (Table S4 in Multimedia Appendix 1). The macroaveraged F_1-scores of negative, neutral, and positive sentiment classifications across different platforms and vaccine topic groups were 0.43, 0.67, and 0.53, respectively. In hesitancy classification, LR outperformed other algorithms in 6 platform–vaccine topic group combinations, with overall accuracies ranging from 0.69 to 0.91 (Table S5 in Multimedia Appendix 1). The macroaveraged F_1-scores of nonhesitancy and hesitancy classifications were 0.86 and 0.40, respectively. Notably, Reddit users had fewer negative sentiment posts, resulting in lower performance in classifying negative sentiment. In addition, as Reddit had fewer hesitancy posts, classifying hesitancy was more challenging than on Twitter and YouTube.

Our evaluation of various algorithms and class-balancing methods for each platform and vaccine topic group revealed that Snorkel performed best in 3 platform–vaccine topic group combinations in vaccine hesitancy classifications, with overall accuracies ranging from 0.69 to 0.98 (Table S6 in Multimedia Appendix 1). The macroaveraged F_1-scores for lack of confidence and nonlack of confidence classifications were 0.88 and 0.45, respectively. Similarly, for complacency classifications, Snorkel outperformed other algorithms in 4 platform–vaccine topic group combinations, with overall accuracies ranging from 0.64 to 0.99 (Table S7 in Multimedia Appendix 1). The macroaveraged F_1-scores for noncomplacency and complacency classifications were 0.89 and 0.49, respectively. Inconvenience classifications were significantly improved with Snorkel in 8 platform–vaccine topic group combinations, with overall accuracies ranging from 0.89 to 0.99 (Table S8 in Multimedia Appendix 1). However, the results are biased as there were limited posts with convenience information on all 3 social media platforms, which may impact generalizability. The macroaveraged F_1-scores for

noninconvenience and inconvenience classifications were 0.98 and 0.38, respectively. Our findings demonstrate that advanced text classification algorithms such as XGBoost and Snorkel outperformed other algorithms in highly class-imbalanced situations, even when different class-balancing methods were applied.

We have created a web-based dashboard building upon those best-performing NLP algorithms to extract vaccine sentiment and hesitancy from social media posts. The dashboard summarizes posts from the 3 social media platforms and allows users to analyze temporal trends and geographic clustering easily. It offers different views, including 3 social media platform–centric views and a comparison view that enables users to compare selected vaccine topic groups and sentiment or hesitancy (Figure S1 in Multimedia Appendix 1).

When analyzing the sentiment of HPV vaccine posts across 3 social media platforms from January 2011 to October 2021 (Figure 3A), we observed that the ratio of positive sentiment was generally higher than that of neutral and negative sentiment. We also compared vaccine sentiment across 3 social media platforms for MMR vaccines from January 2011 to October 2021 (Figure 3B). Overall, posts expressed positive sentiment toward MMR, with most being neutral. Taking the hesitancy of MMR vaccine as an example, the overall trend shows that the social media posts across 3 social media platforms have a higher ratio of nonhesitancy than hesitancy (Figure 3C).

Table 2. NLP[a] performance (measured by F1-scores and accuracy) on vaccine sentiment and hesitancy.

Performance	Twitter			Reddit			YouTube		
	HPV[b]	MMR[c]	General[d]	HPV	MMR	General	HPV	MMR	General
Sentiment									
Positive F_1-score	0.87	0.57	0.47	0.67	0.50	0.35	0.58	0.53	0.19
Neutral F_1-score	0.71	0.67	0.83	0.67	0.65	0.86	0.51	0.59	0.51
Negative F_1-score	0.41	0.53	0.43	0.32	0.26	0.21	0.60	0.49	0.59
Accuracy	0.78	0.61	0.73	0.63	0.55	0.75	0.56	0.55	0.51
Confidence									
Confident F_1-score	0.35	0.31	0.52	0.35	0.62	0.56	0.44	0.29	0.63
Lack of confidence F_1-score	0.88	0.95	0.79	0.86	0.74	0.84	0.89	0.99	0.98
Accuracy	0.80	0.90	0.71	0.77	0.69	0.77	0.82	0.98	0.95
Complacency									
Complacent F_1-score	0.47	0.36	0.41	0.43	0.68	0.60	0.33	0.50	0.60
No complacency F_1-score	0.94	0.91	0.81	0.93	0.59	0.96	0.91	1.00	0.97
Accuracy	0.89	0.84	0.71	0.88	0.64	0.93	0.84	0.99	0.95
Convenience									
Convenient F_1-score	0.96	0.99	0.99	0.94	0.95	0.98	0.98	1.00	0.99
Inconvenient F_1-score	0.48	0.18	0.55	0.67	0.17	0.50	0.17	0.50	0.20
Accuracy	0.92	0.98	0.98	0.89	0.91	0.97	0.96	0.99	0.98
Hesitancy									
Hesitant F_1-score	0.40	0.44	0.38	0.19	0.23	0.20	0.58	0.53	0.61
Nonhesitant F_1-score	0.94	0.90	0.89	0.87	0.81	0.95	0.81	0.83	0.76
Accuracy	0.90	0.83	0.82	0.78	0.69	0.91	0.73	0.75	0.70

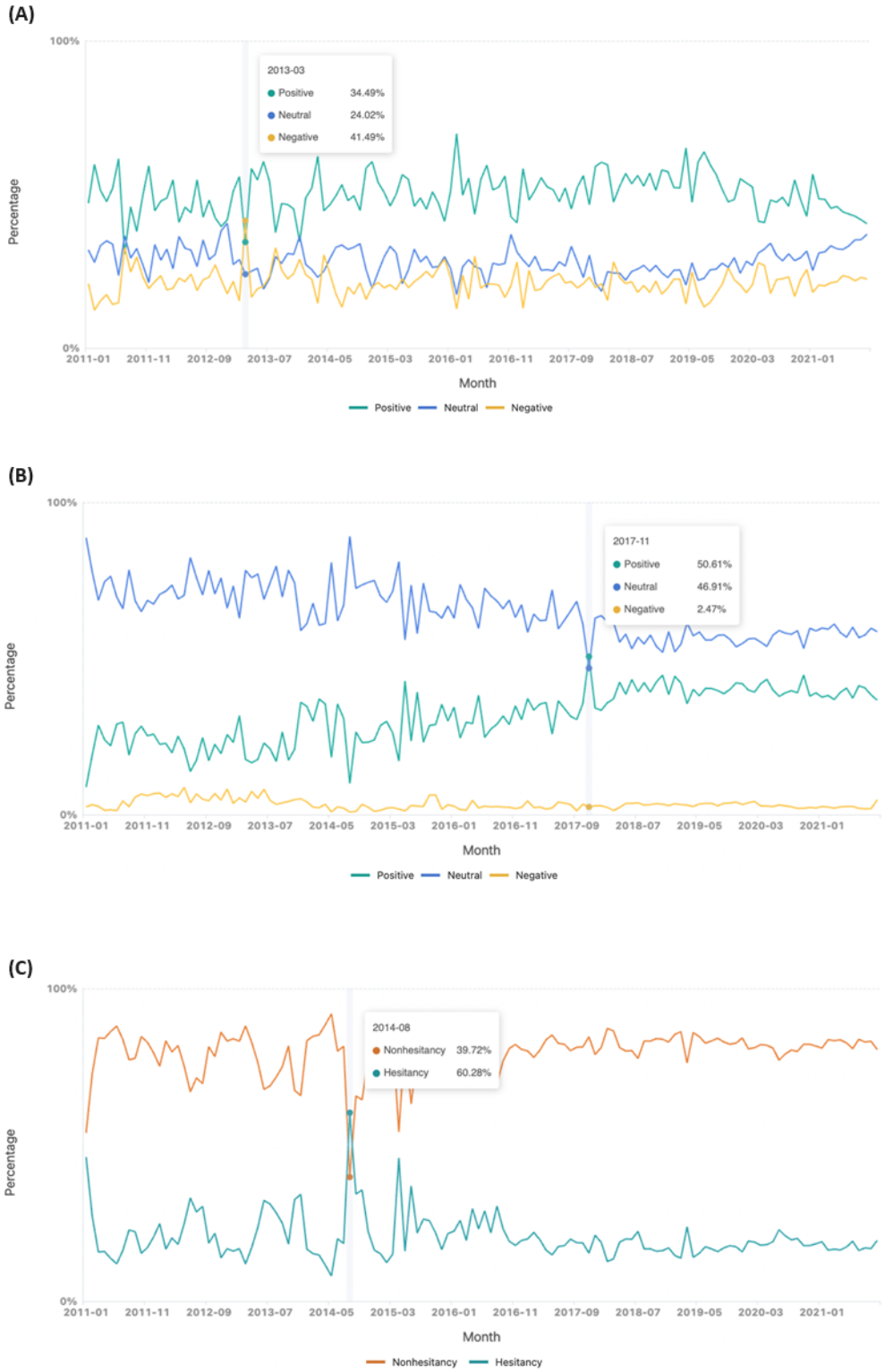
[a]NLP: natural language processing.

[b]HPV: human papillomavirus.

[c]MMR: measles, mumps, and rubella.

[d]General: general or unspecified vaccines.

Figure 3. Temporal trends of vaccine sentiment and hesitancy. (A) Aggregation of 3 social media platform data sources to evaluate vaccine sentiment for HPV vaccine–related posts. (B) Comparison of vaccine sentiment for MMR vaccines. (C) Comparison of vaccine hesitancy for MMR vaccine. HPV: human papillomavirus; MMR: measles, mumps, and rubella.



Discussion

Principal Findings

Our analysis of temporal trends in vaccine-related sentiment on social media platforms yielded valuable insights into the dynamics of public perception. A total of 5 different classification algorithms were subjected to tests for performance in sentiment and hesitancy classifications, revealing that advanced text classification algorithms such as XGBoost and Snorkel outperformed others in classifying hesitancy, complacency, and other factors, while LR had a superior performance for sentiment classification. The superior performance of LR could potentially be attributed to its enhanced ability to effectively handle binary classification

challenges and manage noise variables [63]. As the use of artificial intelligence platforms is increasingly becoming accessible for public use, it is crucial to gain an understanding of their accuracy and limitations. Traditional machine learning algorithms have the ability to predict outcomes but often lack transparency. Hence, enhancing public understanding and advancing toward explainable artificial intelligence is vital for error rectification and improved model efficacy for social media research [64].

When evaluating trends for the HPV vaccine, overall positive sentiment outweighed neutral and negative sentiment (Figure 3A), a notable exception occurred in March 2013. During this period, posts with negative sentiment on all 3 platforms surpassed those with 34% (2270/6582) positive and 24% (1581/6582) neutral sentiment, constituting 41% (2731/6582) of the total. This spike in negative sentiment can be attributed to news articles published in March 2013; for example, "Worried Parents Balk At HPV Vaccine For Daughters" by National Public Radio [65] and "Side Effect Fears Stop Parents from Getting HPV Vaccine for Daughters" by CBS News [66]. These articles highlighted concerns and fears about the HPV vaccine. Afterward, specific studies were conducted and published to further investigate these concerns and fears [67,68]. Notably, the HPV vaccines have been found to be safe in several studies and strongly recommended by the Centers for Disease Control and Prevention (CDC), etc [69,70].

Conversely, overall, posts expressed more neutral sentiment toward MMR than positive sentiment (Figure 3B), with an exception in November 2017. During this month, 51% (2844/5619) of posts expressed positive sentiment and 47% (2636/5619) were neutral. We found that a mumps outbreak was observed right before November 2017, which may have encouraged people to discuss the importance of MMR vaccination. News articles highlighted this outbreak, for example, "Third dose of mumps vaccine could help stop outbreaks, researchers say" by PBS News Hour [71] and "CDC recommends booster shot of MMR vaccine during mumps outbreaks" by CNN [72] mentioned the outbreak and recommended the booster shot of MMR vaccine.

When tracking vaccine hesitancy, we found that the social media posts with a higher ratio of hesitancy were only observed in August 2014 (Figure 3C). During this month, some examples of articles could be associated with vaccine hesitancy: "Journal questions validity of autism and vaccine study" by CNN [73] and "Whistleblower Claims CDC Covered Up Data Showing Vaccine-Autism Link" by TIME [74]. While speculation, particularly among antivaccination subpopulations, continues to surround the discredited study linking MMR vaccines with autism, it is crucial to emphasize that this link has been unequivocally debunked by subsequent research, and organizations such as the CDC and WHO have clarified that no such association exists [75-77]. Nonetheless, these news articles, considered by some as antivaccine propaganda, may partially explain the observed trends in MMR vaccine hesitancy during August 2014.

Strengths and Limitations

In this study, we introduced an NLP-powered online monitoring tool for tracking vaccine-related discussions on multiple social media platforms, covering 3 vaccine topic groups. Our system provides several features that distinguish it from existing tools. It uses NLP algorithms to perform sentiment analysis on social media posts and facilitates the tracking of temporal trends and geographic clustering of vaccine sentiment and hesitancy through visualization. In addition, our system enables users to compare vaccine sentiment and hesitancy across different social media platforms. We have publicly shared our annotated social media vaccine corpora, and we have evaluated several text classification algorithms, providing a benchmark for future research. One of the hypothetical use cases is that our NLP-based tool's application spans from gauging vaccine sentiment during disease outbreaks to when a new vaccine is introduced. During an outbreak, the tool effectively analyzed sentiments toward measles vaccination, facilitating adjustments in public health campaigns.

While our proposed method uses the coarse-grained sentiment model (ie, represents the sentiment as a positive or negative class), fine-grained sentiment models, unlike traditional independent dimensional approaches, beneficially incorporate relations between dimensions, such as valence and arousal, into deep neural networks, thereby providing more nuanced, real-valued sentiment analysis and enhancing prediction accuracy [78-81]. These models prove particularly valuable in language-specific applications and are capable of classifying emotion categories and simultaneously predicting valence, arousal, and dominance scores for specific sentences, providing more nuanced sentiment analysis compared with simple positive or negative classifications.

Beyond the limitations inherent in the sentiment model, our approach also encounters constraints due to the use of traditional machine learning algorithms. Deep learning methods for word or sentiment embedding offer enhanced performance in sentiment analysis tasks by integrating external knowledge such as sentiment polarity and emotional semantics into word vectors [82-87]. They leverage neural networks and multitask learning to create task-specific embeddings, improving the accuracy of tasks such as sentiment and emotion analysis and sarcasm and stress detection [82-84,86]. Furthermore, these methods can adapt to the dynamic nature of language, handling out-of-vocabulary words and context-specific word meanings, proving more accurate and comprehensive than traditional word embeddings [86,87]. In future iterations, we plan to enrich our tool by integrating cutting-edge methods, alongside a more robust evaluation method such as time series cross-validation [88].

While previous studies have used NLP for sentiment analysis on COVID-19 vaccination and information exposure analysis regarding the HPV vaccine using Twitter data sets [40,89], and have investigated the temporal and geographic variations in public perceptions of the HPV vaccine [90], our tool extends its functionality to include a broader spectrum of platforms for tracking different vaccine sentiment and hesitancy on social media. Despite the scientific evidence supporting the safety and

efficacy of vaccines, vaccine hesitancy sentiments on social media can impact public confidence regarding vaccination [91]. Our tool is designed to quickly identify surges in vaccine hesitancy and thereby could be a tool to assist public health professionals in responding promptly with accurate information and effective vaccine promotion strategies.

However, it is essential to acknowledge the inherent limitations of using social media as a public health surveillance tool. These limitations include geography and language restrictions, as well as potential population, age, and gender biases, given that social media users may not represent the general population [92-94]. The user diversity across various social media platforms might partly account for the variation in sentiment and hesitancy label distributions. For example, YouTube has a high volume of users, but Twitter had the most activity in our study because people may view YouTube videos without leaving comments [93]. Moreover, owners of YouTube channels also have the option to disable comments on their uploaded videos. In addition, YouTube comments are highly tied to the content of the videos that the model might not have access to, leading to misinterpretations of sentiment and hesitancy. These biases and variabilities could partly account for the lower prediction accuracy observed for YouTube. Therefore, caution should be exercised when interpreting findings based on social media data, particularly considering the varying distributions of sentiment and hesitancy across different social media platforms in our study. Another limitation pertains to the absence of a weighting system in the dashboard. Currently, the impact of each post, considering variables such as the number of views or reposts, is not considered. In addition, private interactions, specifically on sites such as Facebook, might go unnoticed and this lack of access to private dialogues could limit the comprehensiveness of the responses we capture. Finally, there is the possibility of shifts in user behavior to emerging social media platforms, such as TikTok, introducing additional population bias if such platforms are not included in further analyses.

Conclusions

This study successfully developed an innovative real-time monitoring system for analyzing vaccine sentiment and hesitancy across 3 major social media platforms. This system uses NLP and machine learning to mine and classify social media discussions on vaccines, providing valuable insights into public sentiment and hesitancy trends. The application of this tool presents significant implications for public health strategies, aiding in promptly identifying and mitigating vaccine misinformation, enhancing vaccine uptake, and assisting in the execution of targeted health campaigns. Moreover, it encourages health care professionals to foster an evidence-based discourse around vaccines, thus counteracting misinformation and improving public health outcomes.

Acknowledgments

This work was funded by Merck Sharp & Dohme Corp, a subsidiary of Merck & Co, Inc (Rahway, New Jersey). The content is the sole responsibility of the authors and does not necessarily represent the official views of Merck & Co, Inc or Melax Tech.

Authors' Contributions

JD, ALE, and LY conceptualized and designed the study. LCH and JD performed the experiments. LCH, ALE, JD, and LY drafted the paper. LCH, LH, and JD performed the acquisition, analysis, or interpretation of data. All authors performed critical revision of the paper for important intellectual content. JD, ALE, and LY performed study supervision.

Conflicts of Interest

ALE is a current employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States, who may own stock and stock options in the Company. LY was an employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States during the time of the study. Melax Tech, including JW, JD, and LCH, was compensated for activities related to the execution of the study. FJM was employed by Melax Tech and IMO Health during the research described. IMO Health retains interests in certain software described in this article.

Multimedia Appendix 1
The online dashboard's user interface, architecture, and performances.
[DOCX File , 2065 KB - medinform_v12i1e57164_app1.docx]

References

1. Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. Lancet Infect Dis 2022;22(9):1293-1302 [FREE Full text] [doi: 10.1016/S1473-3099(22)00320-6] [Medline: 35753318]
2. Lindmeier C. Measles vaccination has saved an estimated 17.1 million lives since 2000. World Health Organization. 2015. URL: https://www.who.int/news/item/12-11-2015-measles-vaccination-has-saved-an-estimated-17-1-million-lives-since-2000 [accessed 2024-05-08]
3. Ehreth J. The global value of vaccination. Vaccine 2003;21(7-8):596-600. [doi: 10.1016/s0264-410x(02)00623-0] [Medline: 12531324]

4. MacDonald NE, SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: definition, scope and determinants. Vaccine 2015;33(34):4161-4164 [FREE Full text] [doi: 10.1016/j.vaccine.2015.04.036] [Medline: 25896383]
5. Ten threats to global health in 2019. World Health Organization. URL: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019 [accessed 2024-05-08]
6. Dubé E, Vivion M, MacDonald NE. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. Expert Rev Vaccines 2015;14(1):99-117. [doi: 10.1586/14760584.2015.964212] [Medline: 25373435]
7. Lane S, MacDonald NE, Marti M, Dumolard L. Vaccine hesitancy around the globe: analysis of three years of WHO/UNICEF joint reporting form data-2015-2017. Vaccine 2018;36(26):3861-3867 [FREE Full text] [doi: 10.1016/j.vaccine.2018.03.063] [Medline: 29605516]
8. Black FL. The role of herd immunity in control of measles. Yale J Biol Med 1982;55(3-4):351-360 [FREE Full text] [Medline: 7180027]
9. Cockman P, Dawson L, Mathur R, Hull S. Improving MMR vaccination rates: herd immunity is a realistic goal. BMJ 2011;343:d5703. [doi: 10.1136/bmj.d5703] [Medline: 21971162]
10. Lieu TA, Ray GT, Klein NP, Chung C, Kulldorff M. Geographic clusters in underimmunization and vaccine refusal. Pediatrics 2015;135(2):280-289. [doi: 10.1542/peds.2014-2715] [Medline: 25601971]
11. Omer SB, Pan WKY, Halsey NA, Stokley S, Moulton LH, Navar AM, et al. Nonmedical exemptions to school immunization requirements: secular trends and association of state policies with pertussis incidence. JAMA 2006;296(14):1757-1763 [FREE Full text] [doi: 10.1001/jama.296.14.1757] [Medline: 17032989]
12. Dempsey AF, Schaffer S, Singer D, Butchart A, Davis M, Freed GL. Alternative vaccination schedule preferences among parents of young children. Pediatrics 2011;128(5):848-856. [doi: 10.1542/peds.2011-0400] [Medline: 21969290]
13. Sadaf A, Richards JL, Glanz J, Salmon DA, Omer SB. A systematic review of interventions for reducing parental vaccine refusal and vaccine hesitancy. Vaccine 2013;31(40):4293-4304. [doi: 10.1016/j.vaccine.2013.07.013] [Medline: 23859839]
14. Zhao Z, Luman ET. Progress toward eliminating disparities in vaccination coverage among U.S. children, 2000-2008. Am J Prev Med 2010;38(2):127-137. [doi: 10.1016/j.amepre.2009.10.035] [Medline: 20117568]
15. Zimet GD, Weiss TW, Rosenthal SL, Good MB, Vichnin MD. Reasons for non-vaccination against HPV and future vaccination intentions among 19-26 year-old women. BMC Womens Health 2010;10:27 [FREE Full text] [doi: 10.1186/1472-6874-10-27] [Medline: 20809965]
16. Dredze M, Broniatowski DA, Smith MC, Hilyard KM. Understanding vaccine refusal: why we need social media now. Am J Prev Med 2016;50(4):550-552 [FREE Full text] [doi: 10.1016/j.amepre.2015.10.002] [Medline: 26655067]
17. Peretti-Watel P, Larson HJ, Ward JK, Schulz WS, Verger P. Vaccine hesitancy: clarifying a theoretical framework for an ambiguous notion. PLoS Curr 2015 Feb 25;7:ecurrents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289 [FREE Full text] [doi: 10.1371/currents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289] [Medline: 25789201]
18. Galagali PM, Kinikar AA, Kumar VS. Vaccine hesitancy: obstacles and challenges. Curr Pediatr Rep 2022;10(4):241-248 [FREE Full text] [doi: 10.1007/s40124-022-00278-9] [Medline: 36245801]
19. Larson HJ, Smith DMD, Paterson P, Cumming M, Eckersberger E, Freifeld CC, et al. Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. Lancet Infect Dis 2013;13(7):606-613. [doi: 10.1016/S1473-3099(13)70108-7] [Medline: 23676442]
20. Lawrence HY, Hausman BL, Dannenberg CJ. Reframing medicine's publics: the local as a public of vaccine refusal. J Med Humanit 2014;35(2):111-129. [doi: 10.1007/s10912-014-9278-4] [Medline: 24682632]
21. WHO T. The guide to tailoring immunization programmes. WHO Regional Office for Europe. 2013. URL: https://iris.who.int/handle/10665/351166 [accessed 2024-05-08]
22. Yaqub O, Castle-Clarke S, Sevdalis N, Chataway J. Attitudes to vaccination: a critical review. Soc Sci Med 2014;112:1-11 [FREE Full text] [doi: 10.1016/j.socscimed.2014.04.018] [Medline: 24788111]
23. Cox DS, Cox AD, Sturm L, Zimet G. Behavioral interventions to increase HPV vaccination acceptability among mothers of young girls. Health Psychol 2010;29(1):29-39. [doi: 10.1037/a0016942] [Medline: 20063933]
24. Cates JR, Ortiz R, Shafer A, Romocki LS, Coyne-Beasley T. Designing messages to motivate parents to get their preteenage sons vaccinated against human papillomavirus. Perspect Sex Reprod Health 2012;44(1):39-47 [FREE Full text] [doi: 10.1363/4403912] [Medline: 22405151]
25. Clayton EW, Hickson GB, Miller CS. Parents' responses to vaccine information pamphlets. Pediatrics 1994;93(3):369-372. [Medline: 8115193]
26. Kagashe I, Yan Z, Suheryani I. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data. J Med Internet Res 2017;19(9):e315 [FREE Full text] [doi: 10.2196/jmir.7393] [Medline: 28899847]
27. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci 2015;26(2):159-169 [FREE Full text] [doi: 10.1177/0956797614557867] [Medline: 25605707]
28. Chan B, Lopez A, Sarkar U. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. PLoS One 2015;10(8):e0135072 [FREE Full text] [doi: 10.1371/journal.pone.0135072] [Medline: 26252774]

29. Mitra T, Counts S, Pennebaker J. Understanding anti-vaccination attitudes in social media. 2016 Presented at: Tenth International AAAI Conference on Web and Social Media; May 17-20, 2016; Cologne, Germany p. 269-278 URL: https://ojs.aaai.org/index.php/ICWSM/issue/view/272 [doi: 10.1609/icwsm.v10i1.14729]
30. McDonald L, Malcolm B, Ramagopalan S, Syrad H. Real-world data and the patient perspective: the promise of social media? BMC Med 2019;17(1):11 [FREE Full text] [doi: 10.1186/s12916-018-1247-8] [Medline: 30646913]
31. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
32. Becker BFH, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA, Sturkenboom MCJM. Evaluation of a multinational, multilingual vaccine debate on Twitter. Vaccine 2016;34(50):6166-6171 [FREE Full text] [doi: 10.1016/j.vaccine.2016.11.007] [Medline: 27840012]
33. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The measles vaccination narrative in Twitter: a quantitative analysis. JMIR Public Health Surveill 2016;2(1):e1 [FREE Full text] [doi: 10.2196/publichealth.5059] [Medline: 27227144]
34. Love B, Himelboim I, Holton A, Stewart K. Twitter as a source of vaccination information: content drivers and what they are saying. Am J Infect Control 2013;41(6):568-570. [doi: 10.1016/j.ajic.2012.10.016] [Medline: 23726548]
35. Keelan J, Pavri V, Balakrishnan R, Wilson K. An analysis of the human papilloma virus vaccine debate on MySpace blogs. Vaccine 2010;28(6):1535-1540 [FREE Full text] [doi: 10.1016/j.vaccine.2009.11.060] [Medline: 20003922]
36. Chowdhary KR. Natural language processing. In: Fundamentals of Artificial Intelligence. New York City: Springer; 2020:603-649.
37. Vinet L, Zhedanov A. A 'missing' family of classical orthogonal polynomials. J Phys A Math Theor 2011;44(8):085201. [doi: 10.1088/1751-8113/44/8/085201]
38. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011;7(10):e1002199 [FREE Full text] [doi: 10.1371/journal.pcbi.1002199] [Medline: 22022249]
39. Vaccine Sentiments on Social Media. URL: https://vaccine.social/ [accessed 2024-05-08]
40. Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P. Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Syst Appl 2023;212:118715 [FREE Full text] [doi: 10.1016/j.eswa.2022.118715] [Medline: 36092862]
41. Kumar N, Corpus I, Hans M, Harle N, Yang N, McDonald C, et al. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. BMC Public Health 2022;22(1):446 [FREE Full text] [doi: 10.1186/s12889-022-12824-7] [Medline: 35255881]
42. Li HOY, Pastukhova E, Brandts-Longtin O, Tan MG, Kirchhof MG. YouTube as a source of misinformation on COVID-19 vaccination: a systematic analysis. BMJ Glob Health 2022;7(3):e008334 [FREE Full text] [doi: 10.1136/bmjgh-2021-008334] [Medline: 35264318]
43. Kwon S, Park A. Examining thematic and emotional differences across Twitter, Reddit, and YouTube: the case of COVID-19 vaccine side effects. Comput Human Behav 2023;144:107734 [FREE Full text] [doi: 10.1016/j.chb.2023.107734] [Medline: 36942128]
44. Aleksandric A, Anderson HI, Melcher S, Nilizadeh S, Wilson GM. Spanish Facebook posts as an indicator of COVID-19 vaccine hesitancy in Texas. Vaccines (Basel) 2022;10(10):1713 [FREE Full text] [doi: 10.3390/vaccines10101713] [Medline: 36298580]
45. Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. Knowl Inf Syst 2018;60(2):617-663. [doi: 10.1007/s10115-018-1236-4]
46. Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 2021;11(1):81 [FREE Full text] [doi: 10.1007/s13278-021-00776-6] [Medline: 34484462]
47. Bonanni P, Bechini A, Donato R, Capei R, Sacco C, Levi M, et al. Human papilloma virus vaccination: impact and recommendations across the world. Ther Adv Vaccines 2015;3(1):3-12 [FREE Full text] [doi: 10.1177/2051013614557476] [Medline: 25553242]
48. Bankamp B, Hickman C, Icenogle JP, Rota PA. Successes and challenges for preventing measles, mumps and rubella by vaccination. Curr Opin Virol 2019;34:110-116. [doi: 10.1016/j.coviro.2019.01.002] [Medline: 30852425]
49. Eiden AL, DiFranzo A, Bhatti A, Wang HE, Bencina G, Yao L, et al. Changes in vaccine administration trends across the life-course during the COVID-19 pandemic in the United States: a claims database study. Expert Rev Vaccines 2023;22(1):481-494 [FREE Full text] [doi: 10.1080/14760584.2023.2217257] [Medline: 37218717]
50. Kleinbaum DG, Klein M, Pryor ER. Logistic Regression: A Self-Learning Text. Berlin, Heidelberg, Dordrecht, New York City: Springer; 2002.
51. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24(12):1565-1567. [doi: 10.1038/nbt1206-1565]
52. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens 2007;26(1):217-222. [doi: 10.1080/01431160412331269698]

53. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. New York, NY, United States: Association for Computing Machinery; 2016 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 785-794 URL: https://dl.acm.org/doi/proceedings/10.1145/2939672 [doi: 10.1145/2939672.2939785]
54. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. Proceedings VLDB Endowment 2017;11(3):269-282 [FREE Full text] [doi: 10.14778/3157794.3157797] [Medline: 29770249]
55. Schober P, Vetter TR. Logistic regression in medical research. Anesth Analg 2021;132(2):365-366 [FREE Full text] [doi: 10.1213/ANE.0000000000005247] [Medline: 33449558]
56. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng 2018;2018:4302425 [FREE Full text] [doi: 10.1155/2018/4302425] [Medline: 29849998]
57. Alam MZ, Rahman MS, Rahman MS. A random forest based predictor for medical data classification using feature ranking. Inform Med Unlocked 2019;15:100180 [FREE Full text] [doi: 10.1016/j.imu.2019.100180]
58. Alotaibi Y, Malik MN, Khan HH, Batool A, Alsufyani A, Alghamdi S, et al. Suggestion mining from opinionated text of big social media data. CMC-Comput Mater Con 2021;68(3):3323-3338 [FREE Full text] [doi: 10.32604/cmc.2021.016727]
59. Mallory EK, de Rochemonteix M, Ratner A, Acharya A, Re C, Bright RA, et al. Extracting chemical reactions from text using Snorkel. BMC Bioinformatics 2020;21(1):217 [FREE Full text] [doi: 10.1186/s12859-020-03542-1] [Medline: 32460703]
60. Ramos J. Using TF-IDF to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. 2003. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c [accessed 2024-05-13]
61. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357 [FREE Full text] [doi: 10.1613/jair.953]
62. Tang Y, Zhang YQ, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern B Cybern 2009;39(1):281-288. [doi: 10.1109/TSMCB.2008.2002909] [Medline: 19068445]
63. Kirasich K, Smith T, Sadler B. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review 2018;1(3):9.
64. Mehta H, Passi K. Social media hate speech detection using Explainable Artificial Intelligence (XAI). Algorithms 2022;15(8):291 [FREE Full text] [doi: 10.3390/a15080291]
65. Hensley S. Worried parents balk at HPV vaccine for daughters. NPR. 2013. URL: https://www.npr.org/sections/health-shots/2013/03/18/174617709/worried-parents-balk-at-hpv-vaccine-for-daughters [accessed 2024-05-08]
66. Castillo M. Side effect fears stop parents from getting HPV vaccine for daughters. CBS News. 2013. URL: https://www.cbsnews.com/news/side-effect-fears-stop-parents-from-getting-hpv-vaccine-for-daughters/ [accessed 2024-05-08]
67. Zimet GD, Rosberger Z, Fisher WA, Perez S, Stupiansky NW. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. Prev Med 2013;57(5):414-418 [FREE Full text] [doi: 10.1016/j.ypmed.2013.05.013] [Medline: 23732252]
68. Karafillakis E, Simas C, Jarrett C, Verger P, Peretti-Watel P, Dib F, et al. HPV vaccination in a context of public mistrust and uncertainty: a systematic literature review of determinants of HPV vaccine hesitancy in Europe. Hum Vaccin Immunother 2019;15(7-8):1615-1627 [FREE Full text] [doi: 10.1080/21645515.2018.1564436] [Medline: 30633623]
69. HPV, the vaccine for HPV, and cancers caused by HPV. Centers for Disease Control and Prevention. 2022. URL: https://tinyurl.com/2d45j3jz [accessed 2024-05-08]
70. Meites E, Szilagyi PG, Chesson HW, Unger ER, Romero JR, Markowitz LE. Human papillomavirus vaccination for adults: updated recommendations of the advisory committee on immunization practices. MMWR Morb Mortal Wkly Rep 2019;68(32):698-702 [FREE Full text] [doi: 10.15585/mmwr.mm6832a3] [Medline: 31415491]
71. Branswell H. Third dose of mumps vaccine could help stop outbreaks, researchers say. STAT. 2017. URL: https://www.statnews.com/2017/09/06/mumps-vaccine-study/ [accessed 2024-05-08]
72. Scutti S. CDC recommends booster shot of MMR vaccine during mumps outbreaks. CNN Health. 2017. URL: https://www.cnn.com/2017/10/25/health/cdc-mumps-outbreak-syracuse-university/index.html [accessed 2024-05-08]
73. Goldschmidt D. Journal questions validity of autism and vaccine study. CNN Health. 2014. URL: https://www.cnn.com/2014/08/27/health/irpt-cdc-autism-vaccine-study/index.html [accessed 2024-05-08]
74. Park A. Whistleblower claims CDC covered up data showing vaccine-autism link. TIME. 2014. URL: https://time.com/3208886/whistleblower-claims-cdc-covered-up-data-showing-vaccine-autism-link/ [accessed 2024-05-08]
75. Autism and vaccines. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/vaccinesafety/concerns/autism.html [accessed 2024-05-08]
76. Epidemiological WW. MMR and autism. World Health Organization. 2003. URL: https://www.who.int/groups/global-advisory-committee-on-vaccine-safety/topics/mmr-vaccines-and-autism [accessed 2024-05-08]
77. The Editors of The Lancet, Caplan AL. Retraction—ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 2010;375(9713):445. [doi: 10.1016/s0140-6736(10)60175-4]

78. Xie H, Lin W, Lin S, Wang J, Yu LC. A multi-dimensional relation model for dimensional sentiment analysis. Inf Sci 2021;579:832-844 [FREE Full text] [doi: 10.1016/j.ins.2021.08.052]
79. Park S, Kim J, Ye S, Jeon J, Park HY, Oh A. Dimensional emotion detection from categorical emotion. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; Nov 2021:4367-4380.
80. Lee LH, Li JH, Yu LC. Chinese EmoBank: building valence-arousal resources for dimensional sentiment analysis. ACM Trans Asian Low-Resour Lang Inf Process 2022;21(4):1-18 [FREE Full text] [doi: 10.1145/3489141]
81. Wang J, Yu LC, Lai KR, Zhang X. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 2020;28:581-591 [FREE Full text] [doi: 10.1109/taslp.2019.2959251]
82. Tang D, Wei F, Qin B, Yang N, Liu T, Zhou M. Sentiment embeddings with applications to sentiment analysis. IEEE Trans Knowl Data Eng 2016;28(2):496-509. [doi: 10.1109/tkde.2015.2489653]
83. Xu P, Madotto A, Wu CS, Park JH, Fung P. Emo2Vec: learning generalized emotion representation by multi-task training. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels, Belgium: Association for Computational Linguistics; Oct 2018:292-298.
84. Ye Z, Li F, Baldwin T. Encoding sentiment information into word vectors for sentiment analysis. : Association for Computational Linguistics; 2018 Presented at: Proceedings of the 27th International Conference on Computational Linguistics; August, 2018; Santa Fe, New Mexico, USA URL: https://aclanthology.org/C18-1085/
85. Yu LC, Wang J, Lai KR, Zhang X. Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 2018;26(3):671-681. [doi: 10.1109/taslp.2017.2788182]
86. Wang J, Zhang Y, Yu LC, Zhang X. Contextual sentiment embeddings via bi-directional GRU language model. Knowl-Based Syst 2022;235:107663 [FREE Full text] [doi: 10.1016/j.knosys.2021.107663]
87. Zhu L, Li W, Shi Y, Guo K. SentiVec: learning sentiment-context vector via kernel optimization function for sentiment analysis. IEEE Trans Neural Netw Learn Syst 2021;32(6):2561-2572. [doi: 10.1109/TNNLS.2020.3006531] [Medline: 32673198]
88. Bergmeir C, Hyndman RJ, Koo B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput Stat Data An 2018;120:70-83. [doi: 10.1016/j.csda.2017.11.003]
89. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. Vaccine 2017;35(23):3033-3040 [FREE Full text] [doi: 10.1016/j.vaccine.2017.04.060] [Medline: 28461067]
90. Du J, Luo C, Shegog R, Bian J, Cunningham RM, Boom JA, et al. Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. JAMA Netw Open 2020;3(11):e2022025 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.22025] [Medline: 33185676]
91. Zhang Q, Zhang R, Wu W, Liu Y, Zhou Y. Impact of social media news on COVID-19 vaccine hesitancy and vaccination behavior. Telemat Inform 2023;80:101983 [FREE Full text] [doi: 10.1016/j.tele.2023.101983] [Medline: 37122766]
92. Zhao Y, He X, Feng Z, Bost S, Prosperi M, Wu Y, et al. Biases in using social media data for public health surveillance: a scoping review. Int J Med Inform 2022;164:104804. [doi: 10.1016/j.ijmedinf.2022.104804] [Medline: 35644051]
93. Auxier B, Anderson M. Social media use in 2021. Pew Research Center 2021;1:1-4 [FREE Full text] [doi: 10.4135/9781412963947.n376]
94. Shor E, van de Rijt A, Fotouhi B. A large-scale test of gender bias in the media. SocScience 2019;6:526-550 [FREE Full text] [doi: 10.15195/v6.a20]

Abbreviations

3Cs: confidence, complacency, and convenience
API: application programming interface
CDC: Centers for Disease Control and Prevention
HPV: human papillomavirus
LR: logistic regression
MMR: measles, mumps, and rubella
NLP: natural language processing
SMOTE: synthetic minority over-sampling technique
SVM: support vector machine
VPD: vaccine-preventable disease
WHO: World Health Organization
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 07.02.24; peer-reviewed by M Chatzimina, S Lee, LC Yu, X Vargas Meza; comments to author 25.03.24; revised version received 08.04.24; accepted 11.04.24; published 21.06.24.

Please cite as:
Huang LC, Eiden AL, He L, Annan A, Wang S, Wang J, Manion FJ, Wang X, Du J, Yao L
Natural Language Processing–Powered Real-Time Monitoring Solution for Vaccine Sentiments and Hesitancy on Social Media: System Development and Validation
JMIR Med Inform 2024;12:e57164
URL: https://medinform.jmir.org/2024/1/e57164
doi:10.2196/57164
PMID:38904984

Original Paper

Is Boundary Annotation Necessary? Evaluating Boundary-Free Approaches to Improve Clinical Named Entity Annotation Efficiency: Case Study

Gabriel Herman Bernardim Andrade[1], MSc; Shuntaro Yada[1], PhD; Eiji Aramaki[1], PhD

Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Japan

Corresponding Author:
Eiji Aramaki, PhD
Graduate School of Science and Technology
Nara Institute of Science and Technology
8916-5, Takayama-cho
Ikoma, 630-0192
Japan
Phone: 81 743 72 5250
Email: aramaki@is.naist.jp

Abstract

Background: Named entity recognition (NER) is a fundamental task in natural language processing. However, it is typically preceded by named entity annotation, which poses several challenges, especially in the clinical domain. For instance, determining entity boundaries is one of the most common sources of disagreements between annotators due to questions such as whether modifiers or peripheral words should be annotated. If unresolved, these can induce inconsistency in the produced corpora, yet, on the other hand, strict guidelines or adjudication sessions can further prolong an already slow and convoluted process.

Objective: The aim of this study is to address these challenges by evaluating 2 novel annotation methodologies, *lenient span* and *point annotation*, aiming to mitigate the difficulty of precisely determining entity boundaries.

Methods: We evaluate their effects through an annotation case study on a Japanese medical case report data set. We compare annotation time, annotator agreement, and the quality of the produced labeling and assess the impact on the performance of an NER system trained on the annotated corpus.

Results: We saw significant improvements in the labeling process efficiency, with up to a 25% reduction in overall annotation time and even a 10% improvement in annotator agreement compared to the traditional boundary-strict approach. However, even the best-achieved NER model presented some drop in performance compared to the traditional annotation methodology.

Conclusions: Our findings demonstrate a balance between annotation speed and model performance. Although disregarding boundary information affects model performance to some extent, this is counterbalanced by significant reductions in the annotator's workload and notable improvements in the speed of the annotation process. These benefits may prove valuable in various applications, offering an attractive compromise for developers and researchers.

(JMIR Med Inform 2024;12:e59680) doi:10.2196/59680

KEYWORDS

natural language processing; named entity recognition; information extraction; text annotation; entity boundaries; lenient annotation; case reports; annotation; case study; medical case report; efficiency; model; model performance; dataset; Japan; Japanese; entity; clinical domain; clinical

Introduction

Overview

The electronic health record (EHR) can be an important source of data for health-related research as it contains information on a patient's condition and complaints, performed procedures and administered drugs, the outcome of the treatment, and more [1].

Clinical narratives are a fundamental part of EHRs. Due to their free and unstructured format, natural language processing (NLP) methods are essential for extracting the information from such documents in a way that is comprehensible and useful for

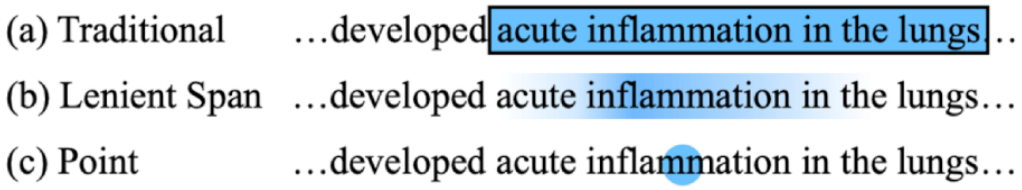
computer systems. Although machine learning–based NLP systems can achieve high performance, these often require large amounts of in-domain annotated data for proper training [2]. Recent few-shot approaches empowered by large language models (LLMs) have also been shown to be performant. Yet, these can also benefit from fine-tuning with in-domain examples, yielding notable improvements [3].

Named entity (NE) annotation, as an inherently manual process, allied to the sheer volume of data that must be meticulously labeled to produce an accurate model, makes it an exhausting and time-consuming task [4]. Particularly when annotating clinical data, workers must possess not only linguistic understanding but specialized medical knowledge is also required. Recruiting such a capable workforce can make the process rather costly [5].

Furthermore, annotation is accompanied by a set of practical issues. For instance, it is natural that contributors disagree on how certain information is annotated or even whether it should be annotated [6]. Determining entity boundaries, meaning where a concept starts and ends, is one of the primary sources of conflict during the process, as so-called *boundary words*, such as articles or adjectives, can induce ambiguity [7].

Especially in medical texts, it is common for annotators to be unsure whether adjectives or modifiers should be included in the annotation. For example, in the sentence presented in Figure 1, some may annotate only the core symptom (“inflammation”).

Figure 1. Example of different annotation paradigms. Traditional annotation (a) requires precisely labeling the beginning and end of the span, while boundary-free (b and c) methods focus on only identifying the core term.



Conversely, others may consider adding all modifiers necessary for a complete encapsulation of the condition.

While entity boundary definition is a problem that affects all languages, scriptio continua languages (which do not have spaces between words), such as Japanese, Chinese, and Korean, are particularly impacted due to the increased difficulty in separating concepts and modifiers.

One can employ strict annotation guidelines to delineate precisely how information should be annotated and even implement adjudication sessions to resolve disagreements. Yet, these can increase the workload and complexity of an already slow and convoluted process.

As an alternative to mitigate such issues, we propose to reformulate the annotation task by eliminating the need to define specific span boundaries when annotating an NE. By demanding less precision from the annotators, we expect to minimize the required decision-making during labeling, thus, improving annotation speed and relieving conflicts.

Although this approach may reduce annotation quality, named entity recognition (NER) performance should not be significantly impacted, as previous research found that models are resilient to a certain amount of boundary imprecision in their training data [8].

In this paper, we leverage this phenomenon by introducing 2 *boundary-free* annotation methodologies: *lenient span*, which relieves the emphasis on entity boundary precision, and *point*, which uses a single position to represent the annotation. Figure 1 presents a visual comparison between the methodologies. We performed a case study to evaluate the efficiency of the proposed methods when annotating a corpus of Japanese medical case reports to create training data for an NER system.

Our contributions are summarized as follows. We present 2 novel boundary-free annotation methodologies, evaluate the efficiency of the annotation process by metrics of annotation time and annotator agreement, and analyze the impact on the performance of an NER system trained with annotated corpora.

Related Work

Annotation Efficiency Improvements

Attempts to improve the annotation workflow are a common theme in NER-related research.

Preannotation depicts the automatic labeling of the text prior to the annotator work [9]. This technique can not only reduce the required annotation time and workload required but also minimize errors [10]. Active learning (AL) [11] can further optimize automatic labeling by iteratively incorporating the data produced during the annotation process to retrain the preannotation model. Kholghi et al [12] ascertained that AL reduced the annotation time by up to 35% (5.6/16 hours) during experiments.

While these are well-established approaches, recent studies also explore alternative ideas. Tokunaga et al [13] analyzed eye-tracking data during NE annotation to identify characteristics that can help design effective features for an annotation tool. Saxena et al [14] introduced a hybrid search-enhanced software that allows users to look for similar terms and annotate related information simultaneously, shortening work time when compared to standard tools.

In recent years, generative LLMs have transformed NLP research and applications, becoming state-of-the-art NLP techniques. While the potential of LLMs to improve the text annotation workflow has also been evaluated in a few different studies [15-17], Tan et al [18] point out that their effectiveness is still strongly affected by model hallucinations and the gap in performance between proprietary and open-source LLMs.

Although crowdsourcing platforms allow the convenient annotation of vast amounts of data [19], they do not improve

task execution or reduce the workload of an individual worker. In addition, as Snow et al [20] noted, inconsistent or low-quality annotations require effective quality control measures. Li [21] found that LLMs can be used to improve the quality of annotation generated by crowdsourcing. Yet LLM annotation quality is still shy of what can be produced manually; thus, combining the automated technique and human effort is still the best approach to creating a high-quality data set [22].

Entity Boundary Imprecision

When addressing boundary imprecision, most studies regard it as a form of noise that should be corrected or circumvented. For instance, Liu et al [23] use confidence scores and normalization techniques based on the labeling structure to estimate the correct span.

Zhu and Li [7] introduced a boundary regularization technique, redistributing a portion of the probability assigned to an annotated span to its neighboring words. This process produces a smooth transition between entity annotations and their nonentity surroundings, mitigating annotation boundary inconsistencies.

Shen et al [24] propose the NER task as a boundary-denoising diffusion process, where a model is trained to derive precise NEs from noisy spans. The authors added controlled noise to gold entity boundaries and used the imprecise data to teach a model to apply a reverse diffusion process to recover the original entity boundaries.

On the other hand, Andrade et al [8] identified that imprecise boundary annotation may not have an extensive impact in some applications. The authors evaluated the effect of various levels of imprecise boundary annotation on NER and entity linking. They identified that models are resilient to a certain amount of noise, showing a small performance drop in that range.

Methods

Data Set

We used the MedTxt-CR-JA corpus [25] in our experiments. This data set comprises 148 open-access case reports in Japanese. Textbox 1 presents an example document from the data set.

A case report is a detailed description of a patient's medical condition, containing, among other information, the temporal progression of the disease and its treatment. Its format is similar to a discharge summary and is frequently used in medical NLP, such as in MIMIC-III [26] or n2c2 shared tasks [27].

This corpus was used in previous studies [28] and contains pre-existing annotations for diseases and symptom names, drugs, anatomical parts, etc. Although we discarded these labels for our experiments, we use them as a gold standard (GS) for evaluation purposes. From now on, this set of annotations is identified as the *gold standard corpus* (GSC).

Textbox 1. Example of a case report from MedTxt-CR-JA and its English translation.

Original:

５８歳，女性．

初診の約２週間前より皮疹が出現，増悪してきたため来院した．

初診時，体幹 四肢に広範囲に浮腫性紅斑が出現し，一部では小水疱を形成していた．

手指背では関節部に一致して角化性紅斑を認め，爪囲には紅斑 紫斑を，眼周囲には軽度の紫紅色斑を認めた．

この時点ではＣＰＫ，ＬＤＨの軽度上昇，抗核抗体２０倍以外，特に異常はなく，確診に至らないため，ステロイド軟膏外用にて経過観察していたところ，３週目頃より体幹 四肢の皮疹が角化性赤色斑へと変化し，１か月目頃より上眼瞼の浮腫性紅斑が著明となり，典型疹となった．

肺癌の合併により発症１年２か月後に死亡した．

臨床経過から，初診時にみられた多形紅斑様あるいは湿疹様の皮疹を皮膚筋炎の早期皮疹と考えた

English translation:

A 58-year-old female.

The patient visited this hospital due to the appearance of a skin rash which worsened about 2 weeks before her first visit.

At the initial examination, the patient had extensive edematous erythema on her torso and extremities, with forming blisters.

Keratinized erythema was uniformly observed around the joints on the back side of the fingers, erythema and purpura were observed around the nails, and mild purplish-red spots were observed around the eyes.

At this point, there were no abnormalities other than mildly elevated CPK and LDH and 20-fold increase in antinuclear antibodies.

Consequently, follow-up with a topical steroid ointment was carried out.

However, by the third week, the skin rash on the torso and extremities changed to keratotic red plaques, and edematous erythema of the upper eyelids became prominent by approximately the first month and became a typical rash.

The patient died 1 year and 2 months after the onset of illness due to complications of lung cancer.

Based on the clinical history, the erythema multiforme or eczema-like skin rash seen at the time of the initial examination is considered to be an early-stage skin rash of dermatomyositis.

We randomly selected a subset of 100 documents from the full corpus, referred to from now on as the *data set*. To minimize the difference in difficulty between texts, we selected documents with similar lengths and quantity of GS entities. Texts are, on average, 554 characters long, roughly equivalent to 250-300 English words, containing around 10 entities per text.

Even though the set of documents for annotation may be considered small, it is worth noting that a scenario with such a small amount of data is not uncommon in the clinical setting, where strong data restrictions usually limit the amount of data available to work with [29].

Annotation Guidelines

It is common to define a set of guidelines before an annotation process to minimize the divergences between annotators and guarantee consistency.

We followed the annotation schema as defined by Yada et al [30]. To simplify the evaluation process, annotators were asked to label only positive (nonnegated) entities of the "Diseases and symptoms" category. We provided the participants with a document describing what should be annotated and some examples, as summarized in Table 1.

Table 1. Annotation guidelines.

Description	Examples[a]
What to annotate	
Reported symptoms, disease names, and clinical findings (pathology, CT[b], and other images)	• Patient visited this hospital due to the appearance of a *skin rash*.
Clinical suspicion, even if there is a slight possibility of disease occurrence	• *Epicarditis* was *suspected* and the patient was hospitalized on July 2.
The locus of a condition, such as an anatomical structure or location, body substance, or physiologic function	• Abdominal CT scan revealed *many enlarged intra-abdominal lymph nodes*.
Adjectives and other modifier words that alter the characteristics or intensity of a condition	• Patient had no subjective symptoms other than a *high fever*. • There was *spotty necrosis* in the lobules.
What should not be annotated	
Absence of symptoms or diseases. Basically, a negation of a clinical concept	• Abdominal findings were unremarkable. • The rash disappeared in about 2 months.
General discussion of a condition merely as a reference and not as a clinical finding	• There is a possibility of primary biliary cholangitis when elevated hepatobiliary enzymes are detected.
Numeric or qualitative findings of an investigation, such as laboratory test values	• The measured blood pressure was abnormal.

[a]In the examples, entities that should be annotated are marked in italics.

[b]CT: computed tomography.

Annotation Methodologies

Our goal is to evaluate whether relieving the emphasis on entity boundary improves annotation speed while maintaining the overall quality of the produced labels. Thus, we compared the *traditional* (boundary-strict) annotation method against 2 proposed boundary-free approaches: *lenient span* and *point annotation*. Figure 1 presents a comparative example of each annotation method.

Traditional Annotation

Traditional annotation requires precise annotation of each NE's exact start and end positions.

Lenient Span Annotation

Lenient span annotation introduces flexibility to the annotation boundaries. While the annotation is still composed of a span, start and end positions are not required to be exactly aligned with the NE boundaries.

Point Annotation

Unlike span-based paradigms, this method requires selecting a single point at any position within the NE span without explicitly specifying the span. It prioritizes speed and simplicity in scenarios where it is not straightforward to determine the NE span precisely. On the other hand, it may introduce ambiguity in the information captured by the annotation.

Note on LLM Annotation

While the use of generative LLMs for text annotation is gaining traction, in this work, we seek ways to aid human annotation and reduce the necessary effort as much as possible where LLMs cannot be used.

The use of LLMs still raises concerns about privacy and security issues; as due to the necessary infrastructure and computational power needed, these models are usually held in the cloud and owned by third-party companies [31]. Given the sensitive nature of clinical data, the usage of LLMs in NLP tasks on real-world data is usually constrained by the policy of medical institutions.

Thus, there is still a need for manual annotations until performant medical LLMs can be accessed through a secure private network or hosted inside hospital facilities at a reasonable cost.

Annotation Task

We asked 4 annotators with medical background and different levels of annotation experience to participate in the experiments. They produced 3 annotated corpora by labeling the documents from the data set using each evaluated methodology. We measured the time taken for each annotation session and computed agreement metrics. We then used each produced corpus to fine-tune a Bidirectional Encoder Representations From Transformers (BERT)–based [32] NER system and evaluated its performance to assess the corpora quality.

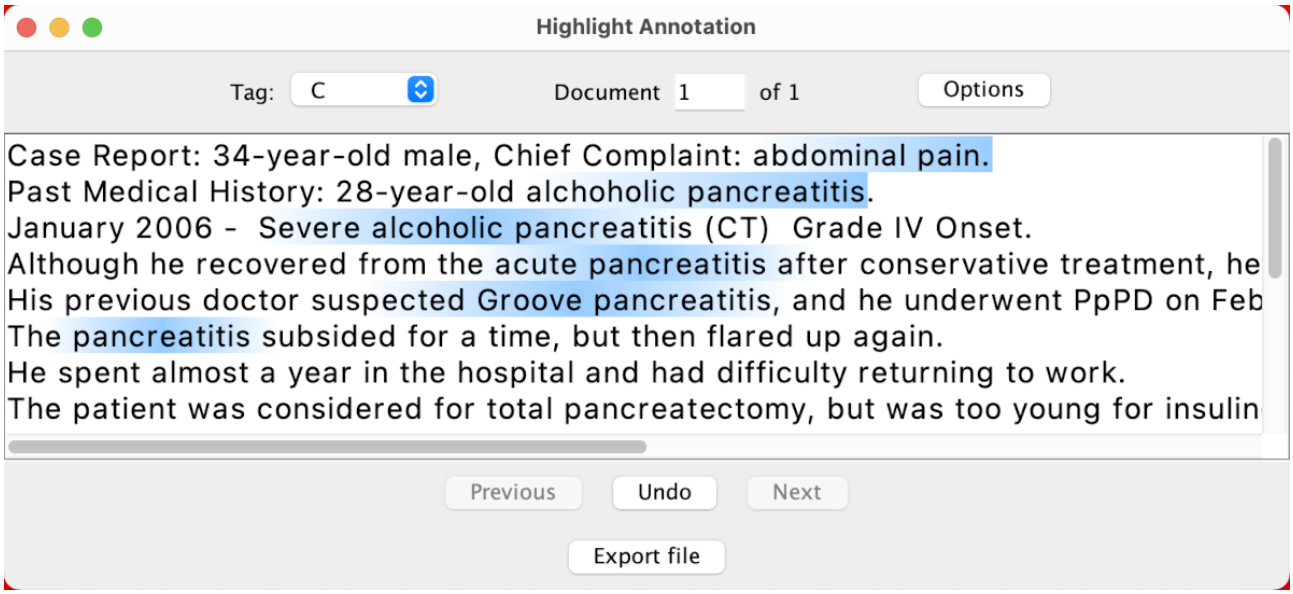
Annotation Tool Development

We developed a Java-based annotation tool to support the proposed boundary-free approaches [33]. Annotations can be presented with smoothed edges using a gradient of color to represent a *soft boundary* and encourage the annotators to be less meticulous when marking the boundaries of the concept. Figure 2 shows a screenshot of the main annotation window.

The text is displayed in its original style, keeping line breaks, spacing, and special characters. Since there is no pretokenization of the texts, annotators can select text spans with character-level precision.

The tool has the following two modes to annotate a concept: (1) *click and drag* and (2) *click-only.*

Figure 2. Screenshot of the annotation tool.



Click and Drag

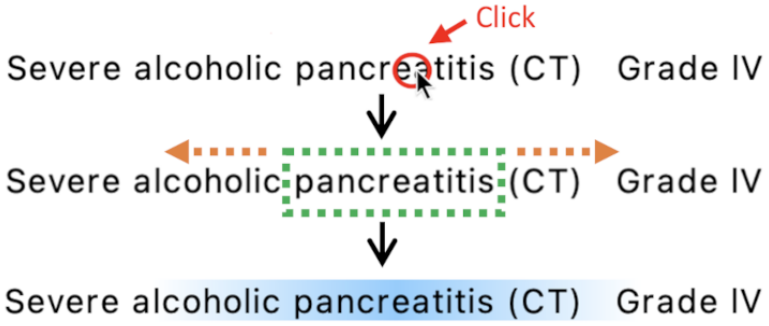
The user clicks on the location where the concept begins and drags the mouse up to where it ends. After releasing the mouse, the area becomes highlighted, representing the labeling.

Click-Only

The user clicks on an entity to label it. While the annotation is stored as a single point, the position will be expanded to a *simulated* span on the interface, representing approximately the labeled concept, as shown in Figure 3.

The annotators received instructions on how to use the tool and a video demonstrating the annotation of a document. They were also supplied with 10 test documents to familiarize themselves with the tool.

Figure 3. Example of a click-only annotation. The selected position, represented by the red circle, is expanded to the word boundaries (in green) plus a random span (orange arrows).



Labeling Workflow

To minimize the number of times each annotator would annotate the same document yet allow us to have at least 2 sets of annotations for a given methodology, we divided our data set of 100 documents into 4 splits.

For each annotation session, each participant received a file containing 2 splits and the annotation methodology that should be used (totaling 50 documents per annotator), as presented in Table 2.

Table 2. Data split for crossover experiment design.

Annotator or annotators	Documents			
	1-25	26-50	51-75	76-100
A	P[a]/T[b]	P	S[c]/T	S
B	S/T	S/T	P	P
C	S	P/T	P/T	S/T
D	P	S	S	P/T

[a]P: point annotation.

[b]T: traditional annotation.

[c]S: lenient span annotation.

We attempted to maximize the mixing between the annotator and the methodology used.

The work was executed in 3 different sessions, the first for point annotation, followed by the lenient span annotation, and lastly, the traditional annotation. During the first 2 sessions, the annotation tool was configured to show smooth edges, and annotators were instructed not to fix slightly incorrect annotations as long as the core concept was highlighted in the tool's interface.

Although the same annotator worked on the same document more than once, the traditional annotation (third) session was conducted 6 months later to avoid memory bias affecting the annotation time measurement. This time annotators were instructed to be as precise as possible when selecting the entity spans and not to refrain from undoing incorrect annotations. The annotation tool was configured beforehand to present the annotations with precise hard boundaries, as any other standard annotation software.

Across all sessions, participants were instructed to annotate the broadest expression whenever in doubt about whether some words should be included in the annotation. Each session produced 2 parallel sets of annotations for each document, unified in a single corpus for each annotation method.

We resolved all disagreements between the 2 sets automatically. We accepted all annotations made by either annotator, even if there is no matching counterpart. Whenever there is boundary disagreement, we choose the broadest span possible when combining the 2 annotations.

For *point* annotations, we grouped annotations that refer to the same NE and averaged their positions. We consider annotations as referring to the same concept when located within 6 characters of distance from each other. The distance limit was chosen based on the average Japanese word length, around 3 characters. We chose a larger value to account for multiword concepts.

Point-to-Span Estimation

Being aware that the single-position label produced by the *point* annotation method may not convey enough information about the adequate range of the NE to be extracted when training the model, we developed a *point-to-span* estimation method [34]. It can complement the annotation with span information without additional manual work.

We used a BERT model (referred to as the *expansion model*) that receives the positional annotation and attempts to predict the original NE span. Effectively, it works as a method to convert Points into Span-based annotations, as illustrated in Figure 4.

The *point-to-span* estimation model is based on the pretrained *tohoku-nlp/bert-base-japanese-char-v2* model [35], and it was fine-tuned using the training parameters presented in Table 3. Training was performed on a server with 2 NVIDIA Quadro RTX 8000 GPUs.

Figure 4. Flow of the point-to-span estimation process. BERT: Bidirectional Encoder Representations From Transformers.

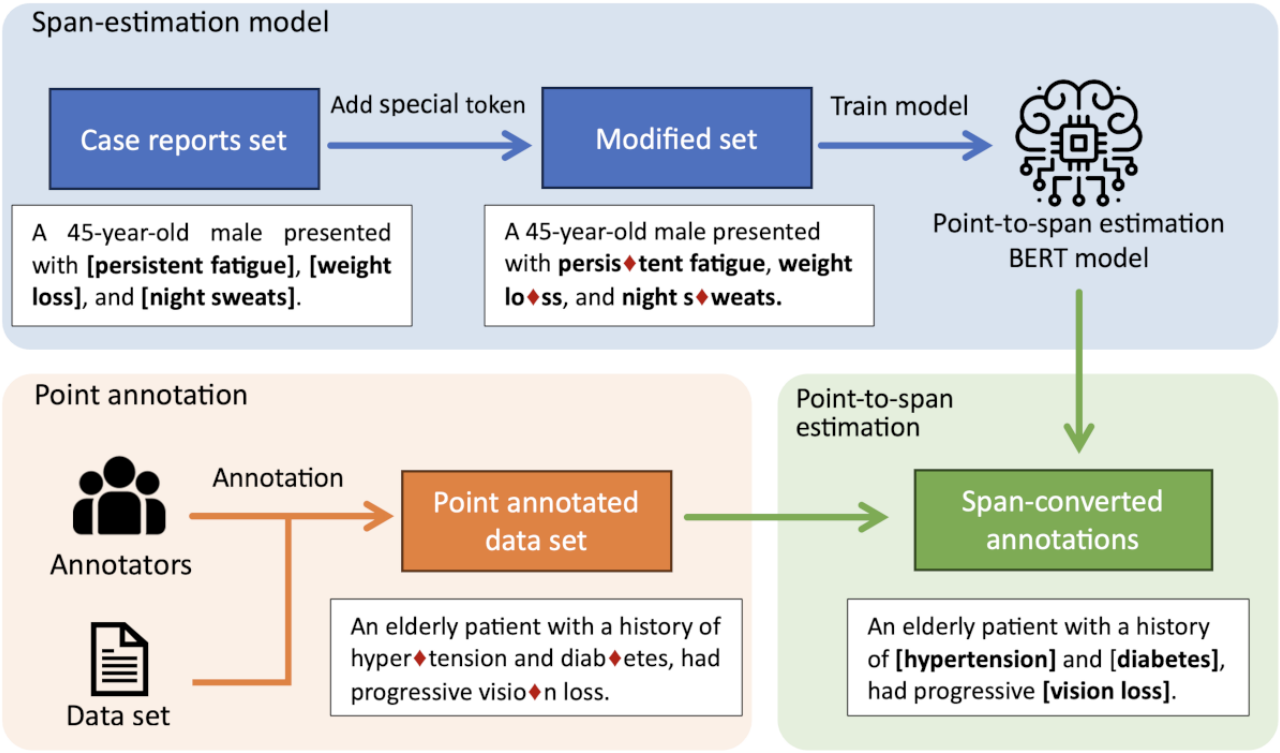


Table 3. Hyperparameters used for model training.

Parameter	Value
Max epochs	10
Training batch size	16
Learning rate	3×10^{-5}
Optimizer	AdamW
Max sentence length	512 characters
Model selection	Early stopping
Training time	Approximately 30 min

As training data, we used a large data set of Japanese medical texts with labeled diseases and symptoms consisting of 1027 synthetic medication history notes generated through crowdsourcing. In total, 10 experienced dispensing pharmacists were hired as writers to craft the corpus. Each writer was assigned 1 of 285 drug names and tasked with creating a "typical" clinical narrative.

Before being fed to the model, each annotation of the training data was replaced by an identifier token in a random location within its span based on a truncated normal distribution. A different distribution was used for each annotation, centered on the middle point, with SD being a sixth of the annotation length. Due to the randomicity of the data, we augmented the data set 10 times by re-executing the annotation replacement module and generating different valid positions for the .

The expansion model was then trained to identify this token and output the start and end positions of the concept based on the word containing the token and its surrounding context.

We evaluated the model by predicting the spans for annotations on the GSC. We preprocessed the GSC annotations using the same method to replace the annotations with . tokens. Our best model was able to achieve an F_1-score of 0.77.

We applied the expansion model to the point-annotated data set to infer spans for each annotation, producing a *point-expanded* corpus. Effectively, the combination of point annotation and expansion allows the generation of a span-annotated data set with less human effort.

Evaluation

Annotation Method Efficiency

We evaluated the annotation methods according to the following:

- Annotation quality: We assessed the percentage of GSC concepts that were correctly annotated. We consider an annotation correct when at least 1 token overlaps with the GS span.

- Annotation time: Annotators manually measured the time they took to work on the data during each session. They were instructed to start the timing after loading the texts in the annotation software.
- Interannotator agreement (IAA): We use Cohen Kappa [36], one of the most common metrics for gauging agreement between annotators. Kappa is a function of the proportion of observed and expected agreement, and it may be interpreted as the proportion of agreement corrected for chance [37].

Given that the *point* annotation methodology allows for multiple correct annotations within the NE span, we computed an additional *adjusted variant* of the metrics specifically for these annotations. In this variant, we considered annotations to agree if they were within a 3-character range of each other, reflecting the average word length in the Japanese language.

Downstream Task Performance

As one of the typical downstream tasks, we developed an NER system to benchmark each annotation approach. We again employed the pretrained *tohoku-nlp/bert-base-japanese-char-v2* model [35] and fine-tuned it using our annotated corpora.

We used the same training parameters for all models, as presented in Table 3. To minimize the variability between results, we used 5-fold cross-validation and averaged the obtained values.

We evaluated model predictions on the MedTxt-CR-JA test set, comprised of 75 documents, by the metrics of *precision*, *recall*, and *F-score*. We employ two variants of the metrics: (1) strict and (2) relaxed.

Strict metrics follow CoNLL criteria [38] and only consider predictions where the span exactly matches the ground truth. These metrics allow us to estimate how closely the model fits the GS.

Relaxed metrics [39] accept partial matches or extra tokens as long as at least 1 token of the predicted span overlaps with the GS span. This variant allows assessing the model's capability of identifying the presence of concepts of interest in the text.

Ethical Considerations

In this study, an annotation process was conducted with the help of human participants. All annotators were provided with detailed information about the purpose, methods, and potential uses of the data they produced, and their informed consent was obtained.

To ensure the privacy of all the patients related to the medical data used in this study, we selected a data set already fully anonymized.

As this research did not use personally identifiable information, it was exempt from institutional review board approval in accordance with the Ethical Guidelines for Medical and Health Research Involving Human Subjects stipulated by the Japanese national government (Chapter 1, Part 3, 1C) [40].

Results

Annotation Method Efficiency

Upon merging the data received from the annotators, we produced the final version of the annotated corpus for each one of the methodologies. Table 4 shows some statistics of the produced corpora.

There is no substantial difference between *traditional* and *lenient span* methods when comparing the average length of the produced annotation. However, both produced annotations slightly larger than the gold annotations due to the disagreement resolution approach adopted in this study.

Table 4. Statistics of the produced corpora.

Method	Total annotations	Average annotation length (character)
Gold standard	1167	6.31
Traditional	1065	7.30
Lenient span	1012	7.30
Point	1066	—[a]

[a]Not applicable.

Annotation Quality

Table 5 shows the average of GSC annotations covered by each corpus.

Although none of the methodologies captured all the ground truth concepts, the percentage of entities captured was similar for every method, with less than a 10% (73 annotations) difference between the best (lenient span) and worst (point).

As the value of missed entities is consistent for all methodologies, we attribute it to some divergence between the guidelines for annotating the GSC and the one used in this study. Differences in the interpretation may have led the annotators to skip some of the entities.

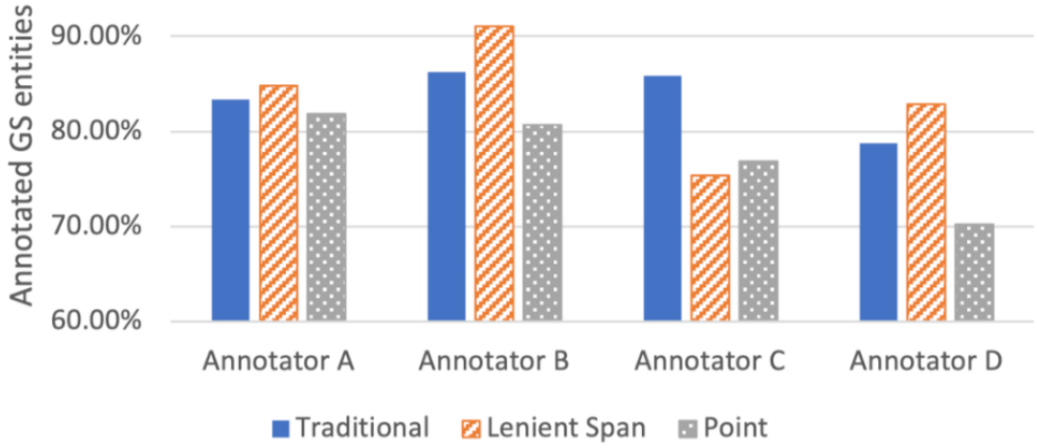
We noticed that the traditional methodology presented a more constant accuracy throughout the annotators, while the boundary-relaxed methods had more variation, especially for annotators C and D.

Figure 5 presents the accuracy of the annotations of each participant in relation to GSC on each methodology.

Table 5. Average number of correctly annotated gold standard (GS) entities per annotation method.

Method	Annotated GS entities, n (%)
Traditional	819 (83.56)
Lenient span	796 (83.65)
Point	746 (77.41)

Figure 5. Annotation accuracy per annotator. GS: gold standard.



Annotation Time

The time measurement results in Table 6 demonstrate that both boundary-free annotation techniques can provide time-saving benefits. On average, reductions of around 25% (around 28 min) and 20% (around 21 min) were observed when using *point* and *lenient span* methods, respectively, compared to the *traditional* annotation process.

Table 6. Comparison of the individual annotation time per annotation method[a].

Annotator	Traditional	Lenient span	Point
A	1:23:44	1:03:23 24%)	0:54:35 (–35%)
B	1:09:14	0:52:07 (–25%)	0:48:45 (–30%)
C	3:16:58	2:10:20 (–34%)	2:15:27 (–31%)
D	1:10:23	1:31:29 (+30%)	1:10:40 (+0%)
Average	1:45:05	1:24:20 (–20%)	1:17:22 (–26%)

[a]Times are presented in the HH:MM:SS format, with the percentage comparison to the traditional method in parenthesis.

Interannotator Agreement

As evidenced by the results presented in Table 7, the IAA measured for both boundary-free annotation methods overcame the *Traditional* methodology.

Point annotations recorded the lowest agreement due to the inherent low probability of annotators precisely pinpointing the exact same position within an NE. Despite that, it achieves the highest measured agreement using the adjusted variant of the metrics.

Table 7. Average interannotator agreement per annotation methodology.

Method	Cohen Kappa
Traditional	0.731
Lenient span	0.774
Point	0.326
Point (adjusted)	0.811

Downstream Task Performance

Table 8 presents the NER model evaluation results.

We trained a GSM using the GS data as a reference for our system's best possible performance.

The data produced in our annotation experiments probably have lower quality due to the lack of proper curation and review

sessions. Thus, when comparing the *Traditional* annotation approach against the GSM, there is a slight decrease in performance: 15% and 11% on strict and relaxed metrics, respectively. Nevertheless, the relation between precision and recall remains the same, as both models were trained on similarly boundary-strict annotations.

Table 8. Evaluation of the trained named entity recognition models.

Method	Strict			Relaxed		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Gold standard model	0.72	0.78	0.75	0.90	0.89	0.89
Traditional	0.60	0.69	0.64	0.77	0.81	0.79
Lenient span	0.56	0.54	0.55	0.67	0.62	0.64
Point	0.00	0.00	0.00	0.60	0.45	0.51
Point (expanded)	0.34	0.35	0.35	0.73	0.71	0.72

Discussion

Principal Findings

Throughout the experiments, it was noticeable that simplifying the annotation process contributed to a more comfortable experience for the participants. We observed increased annotation speed, annotator agreement, and overall positive feedback from the annotators regarding the changes.

Although we showcase our proposal in clinical data, the annotation methodologies are both domain and language-agnostic, so they can be applied to texts of different domains and idioms.

Annotation Speed Improvements

The results in Table 6 show that simplifying the constraints under which annotators work can effectively increase the speed at which they execute the task. By virtually removing the need to decide on entity boundaries, both proposed methodologies allowed the annotation of our data set in less time than the *traditional* method.

However, while an overall decreasing trend in annotation time was observed, different annotators experienced varying degrees of time reduction. Notably, annotator C experienced a significant increase in efficiency when using these methodologies. Conversely, annotator D was quicker with the *traditional* annotation scheme. Still, his precision was lower than other annotators, as shown by the individual accuracy results presented in Figure 5.

Annotator Agreement Improvements

Meanwhile, the IAA evaluation (Table 7) revealed some interesting insights into the annotation consistency of each methodology. Both the *lenient span* and the adjusted *point* agreement overcame the *traditional* methodology by 5.88% and 10.94%, respectively.

While we believe that slightly different interpretations of what information should be annotated may have diminished *traditional* approach agreement, such a finding was still unexpected due to the higher flexibility given to the annotators when removing the need for entity boundaries. However, this improvement can be attributed to the ease with which annotators can consistently agree on the core parts of mentions (or the "main words") compared to determining the precise boundaries of entire entities. Such boundaries may or may not encompass adjectives, modifiers, etc, which often contribute to annotation disagreements.

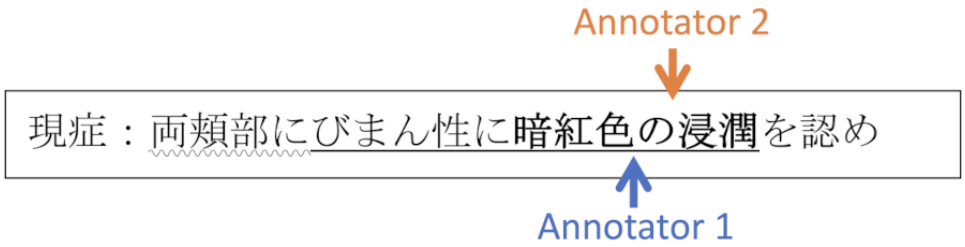
Notably, *point* annotations perceived a large difference in the agreement values measured using the default and *adjusted* variants of the IAA metrics. This is explained by the fact that, even though it is virtually impossible for annotators to select the same character in an NE for all annotations, they generally selected positions close to each other for the same NE. Such finding is evidenced by the distribution of annotation pairs based on the number of characters of difference between them, as depicted in Table 9.

Such a small distance is due to annotators' diligence in positioning the annotation close to the center of the NE's core word. As in the sentence shown in Figure 6 (which translates to "Current symptoms: Diffuse dark red infiltration is observed on both cheeks."), even though the span of the desired annotation is quite large, both annotators placed their labels near the most relevant set of words, "dark red infiltration."

Table 9. Distribution of annotation pairs based on the distance between them.

Number of characters of difference	Annotations, n (%)
0	661 (41.31)
1	640 (40.00)
2	145 (9.06)
3	52 (3.25)
4	25 (1.56)
5	15 (0.94)
6	15 (0.94)
7	14 (0.94)
8	11 (0.69)
9	15 (0.94)
≥10	7 (0.44)

Figure 6. Example of 2 distinct point annotations in a named entity with a large span (underscored). The annotations are located near the center of the core word (in bold).



Annotator's Opinions

Annotator feedback was positive especially regarding the *point* annotation, given its simplicity. The participants highlighted the easiness of the single-click selection mode, particularly due to the reduced mouse manipulation needed.

However, the participants expressed difficulty in understanding the correctness of their annotations and whether the chosen range was indeed accurate. They felt that the soft boundaries displayed by the annotation tool turned the annotations ambiguous, making them unsure whether they matched the range they intended to select.

Impacts on Model Performance

While achieving significant improvements in annotators' work quality, the additional flexibility from boundary-free methods considerably impacted model performance, particularly in strict evaluation, due to the imprecise training data, as seen in Table 8.

The *lenient span*–trained model exhibited a significant subside in its recall, which hindered strict and relaxed evaluations. We did not expect that the ambiguity in NE boundaries could affect the model's capability of locating NEs in the text.

While such performance drop may be acceptable for some applications, we believe additional annotation postprocessing methods could restore the accuracy to levels similar to the *traditional* schema.

Point-to-Span Estimation

In particular, the insights from *point* annotation experiments underscore the potential of automated methods to supplement human annotations. We believe that *point-to-span estimation* can be pivotal for improving annotation speed, but beyond that, it can be proven beneficial to aid in addressing other annotation problems.

Given the lackluster nature of the annotation task, it is not uncommon that annotators make mistakes, such as including punctuation markers or failing to label part of the NE simply for a lack of focus. The span estimation model can be a tool to "normalize" such annotations.

Furthermore, the estimation could be integrated into the actual annotation process by coupling it with our annotation tool, enabling the "click-only" annotation interface to present the predicted span directly and allowing the annotator to correct its mistakes.

However, there is potential for enhancements in the expansion model. Although expanding a point to the expected word seems to be a simple task, as we are evaluating our methods on a scriptio continua language, which makes the definition of the word boundaries not as obvious as in space-delimited languages, such as English.

Through analysis of the model's output, we have observed that the estimation model exhibited a tendency to choose spans larger than the GS entities, particularly when characters that act like

qualitative adjectives (such as “高” for high, “急性” for acute, “巨大” for huge) were connected to the concept of interest.

For instance, the model outputted “高度の肝萎縮“ (Severe liver atrophy) instead of only “肝萎縮” (Liver atrophy). Another example was the expansion of the term “巨大な脾腎シャント“ (Giant splenorenal shunt), where 巨大な (Giant) was included.

Yet, even though the model output in these examples can be regarded as “incorrect” when compared to the GSC, from a clinical point of view, it is not uncommon that some diseases are distinguished by such modifier words. For example, “急性胆嚢炎” (acute cholecystitis) and “慢性胆嚢炎 (chronic cholecystitis), which even have different International Classification of Diseases codes, K81.0 and K81.1, respectively.

Error Analysis

Figures 7 and 8 present example comparisons between all the evaluated models in 2 different sentences.

We could not identify any unusual behavior when inspecting the traditional annotation model output. Yet, we highlight that the lenient span model portrayed a tendency to overly extend the span lengths. In some cases (as shown especially in Figure 8), multiple NEs are “merged” into a single continuous extraction.

As seen in both examples, the model trained with raw point annotations could not extract NE spans, denoting that the single position annotation contains insufficient information to train the model properly.

In contrast, the model trained on expanded point annotations showcases the effectiveness of the *point-to-span* estimation method. Although strict metrics are still substantially lower than other approaches, relaxed results are comparable to the *traditional* annotation approach. The analysis of the model output evidenced that, while it could locate most concepts of interest, it struggled in correctly extracting multiword concepts.

Figure 7. Comparison of model output for the sentence “While waiting for a CT scan, patient went into cardiopulmonary arrest (CA), but could not be resuscitated and died.” Gold standard entities and model extractions are marked in bold and underscored. White space tokenization was added to the Japanese text to enhance readability for non-Japanese readers. The original text does not contain spaces.

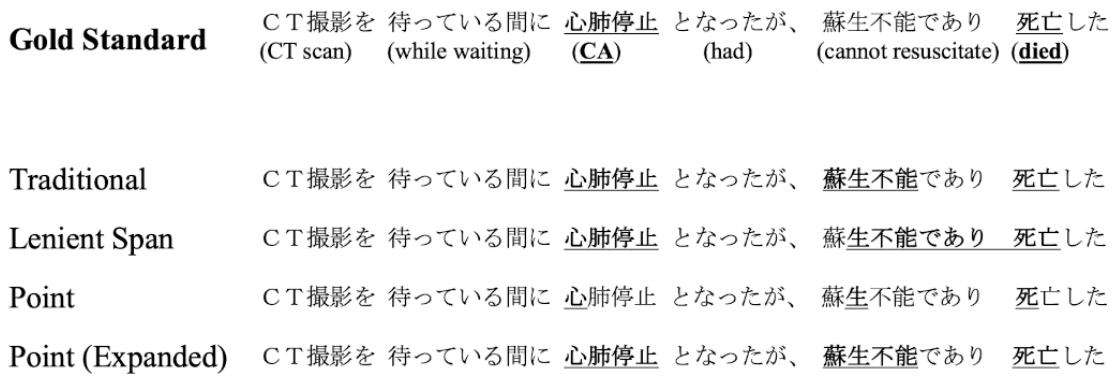
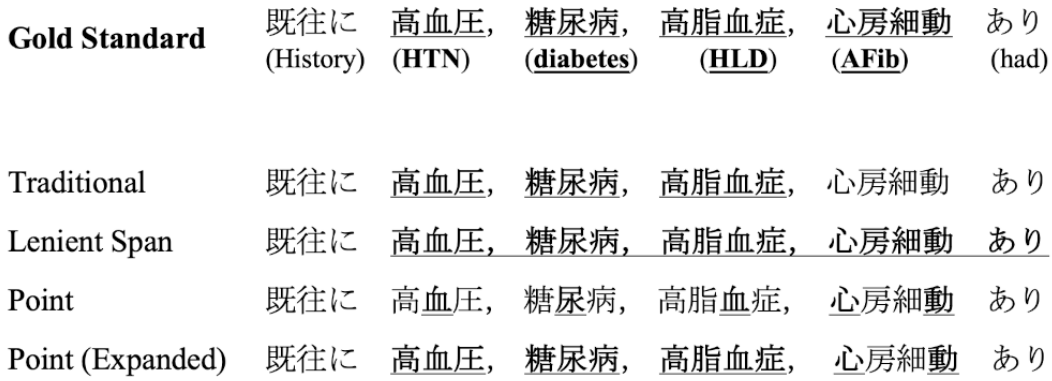


Figure 8. Comparison of model output for the sentence “History of hypertension (HTN), diabetes, hyperlipidemia (HLD), or atrial fibrillation (AFib).” Gold standard entities and model extractions are marked in bold and underscored. White space tokenization was added to the Japanese text to enhance readability for non-Japanese readers. The original text does not contain spaces.



Limitations

While our research focused on exploring novel approaches to text annotation and revealed promising findings, a few concerns and limitations need further investigation. Our investigations were only conducted in the Japanese language. Though our proposal is language independent, applying our techniques in a space-delimited language, such as English, could introduce some bias. Evaluation using different languages is, thus, encouraged. Since our data set in this study has an English variant, we plan to conduct additional experiments.

We concentrated on a singular entity class, disease, and symptom names to streamline the analysis. Even though our texts contain a large number of entities, a single class annotation may not represent a real use case. Exploring our methodologies

in a multiclass scenario would enhance the robustness of our findings and conclusions.

Furthermore, we acknowledge that automated labeling techniques, such as preannotation, can affect the improvements observed in annotation time by adopting boundary-free methodologies. We chose not to incorporate these features in our annotation tool to minimize the number of variables affecting the annotation process.

The observed performance of the trained NER models could have been impacted by our choice of using a simple and automatic approach to solve disagreements. Although it avoids additional annotator work and simplifies the research flow, implementing adjudication or review sessions with the annotations would be preferred, as it could have provided a better annotation quality.

LLMs are prevalent in the current NLP research scenario, and their application has led to the development of systems that push state-of-the-art performance in many different tasks. In the current state of our work, we have not adopted LLMs. Still, we acknowledge that the accuracy of our methods may be improved by employing such methods in our workflow, possibly replacing the Point-to-span BERT model.

Conclusions

In this study, we investigated the effects of reducing the emphasis on entity boundary annotations while labeling NEs in a medical data set. We proposed 2 novel boundary-free annotation methodologies, *lenient span* and *point* annotation. We evaluated the impact of their application in an annotation process regarding annotation efficiency and the quality of the labeling produced.

We also publicly released our developed annotation tool [33] and point-to-span estimation model [34].

Our results demonstrate a trade-off relation between annotation efficiency and model performance. Although not surprising, it unveils the weak points of each methodology and uncovers potential adjustments that can be made to each approach. We underscore that completely disregarding boundary information may ease the annotator's work while it sacrifices performance to some extent.

We plan to evaluate the proposed methodologies in other languages in future work. We also intend to explore the impact of postprocessing techniques, such as normalization or boundary regularization, to enhance model output performance.

Acknowledgments

This work was supported by Japan Science and Technology Agency (JST) Core Research for Evolutionary Science and Technology (CREST) grant JPMJCR22N1, Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) grant JP19H01118, and Cross-Ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" grant JPJ012425, Japan.

Data Availability

The data sets generated during and/or analyzed during this study are available in the MedTxt-CR repository [25].

Authors' Contributions

GHBA designed the study, performed the computational experiments and data analysis, and wrote the manuscript. SY and EA discussed the results and reviewed the manuscript. EA supervised the study. All the authors have approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Computational Stats 2021 Feb 14;13(6):e1549. [doi: 10.1002/wics.1549]
2. Gomes I, Correia R, Ribeiro J, Freitas J. Effort estimation in named entity tagging tasks. In: Proceedings of the 12th Conference on Language Resources and Evaluation. 2020 May Presented at: LREC 2020; May 11-16, 2020; Marseille, France p. 998-306 URL: https://aclanthology.org/2020.lrec-1.37
3. Monajatipoor M, Yang J, Stremmel J, Emami M, Mohaghegh F, Rouhsedaghat M. LLMs in biomedicine: a study on clinical named entity recognition. arXiv Preprint posted online on April 10, 2024. [doi: 10.48550/arXiv.2404.07376]
4. Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís J. Named entity recognition: fallacies, challenges and opportunities. Computer Standards Interfaces 2013 Sep;35(5):482-489. [doi: 10.1016/j.csi.2012.09.004]
5. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011;18(5):540-543 [FREE Full text] [doi: 10.1136/amiajnl-2011-000465] [Medline: 21846785]

6. Baledent A, Mathet Y, Widlöcher A, Couronne C, Manguin JL. Validity, agreement, consensuality and annotated data quality. In: Proceedings of the 13th Conference on Language Resources and Evaluation. 2022 Presented at: LREC 2022; June 20-25, 2022; Marseille, France URL: https://aclanthology.org/2022.lrec-1.315
7. Zhu E, Li J. Boundary smoothing for named entity recognition. 2022 Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); May 22-27, 2022; Dublin, Ireland.
8. Andrade G, Yada S, Aramaki E. Comparative evaluation of boundary-relaxed annotation for entity linking performance. 2023 Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 9-14, 2023; Toronto, ON.
9. Ganchev K, Pereira F, Mandel M, Carroll S, White P. Semi-automated named entity annotation. 2007 Presented at: Proceedings of the Linguistic Annotation Workshop; June 2007; Prague, Czech Republic p. 53-56 URL: https://aclanthology.org/W07-1509 [doi: 10.3115/1642059.1642068]
10. Komiya K, Suzuki M, Iwakura T, Sasaki M, Shinnou H. Comparison of methods to annotate named entity corpora. ACM Trans Asian Low-Resour Lang Inf Process 2018 Jul 21;17(4):1-16. [doi: 10.1145/3218820]
11. Dasgupta SA, Kalai AT. Analysis of perceptron-based active learning. In: Learning Theory. Berlin, Heidelberg: Springer; 2005:249-263.
12. Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning reduces annotation time for clinical concept extraction. Int J Med Inform 2017;106:25-31. [doi: 10.1016/j.ijmedinf.2017.08.001] [Medline: 28870380]
13. Tokunaga T, Nishikawa H, Iwakura T. An eye-tracking study of named entity annotation. 2017 Presented at: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017; September 4-6, 2017; Varna, Bulgaria p. 758-764.
14. Saxena K, Sunkle S, Kulkarni V. Hybrid search based enhanced named entity annotation tool. In: Proceedings of the 15th Innovations in Software Engineering Conference. 2022 Presented at: ISEC '22; February 24-26, 2022; Gandhinagar, India.
15. Kim H, Mitra K, Li CR, Rahman S, Zhang D. MEGAnno+: a Hhuman-LLM Collaborative Annotation System. 2024 Presented at: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations St. Julians; March 17-22, 2024; St. Julians, Malta p. 168-176 URL: https://aclanthology.org/2024.eacl-demo.18
16. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen L, et al. LLMs accelerate annotation for medical information extraction. 2023 Presented at: Proceedings of the 3rd Machine Learning for Health Symposium; December 10, 2022; New Orleans, LA p. 82-100.
17. Kholodna N, Julka S, Khodadadi M, Gumus M, Granitzer M. LLMs in the loop: leveraging large language model annotations for active learning in low-resource languages. arXiv Preprint posted online on April 2, 2024. [doi: 10.48550/arXiv.2404.02261]
18. Tan Z, Beigi A, Wang S, Guo R, Bhattacharjee A, Jiang B, et al. Large language models for data annotation: a survey. arXiv Preprint posted online on February 21, 2024.
19. Sabou M, Bontcheva K, Derczynski L, Scharl A. Corpus annotation through crowdsourcing: towards best practice guidelines. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. 2014 Presented at: LREC'14; May 26-31, 2014; Reykjavik, Iceland p. 859-866 URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf
20. Snow R, O'Connor B, Jurafsky D, Ng A. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. 2008 Presented at: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing Honolulu; October 2008; Honolulu, HI p. 254-263 URL: https://aclanthology.org/D08-1027
21. Li J. A comparative study on annotation quality of crowdsourcing and LLM via label aggregation. arXiv Preprint posted online on January 18, 2024. [doi: 10.48550/arXiv.2401.09760]
22. Pangakis N, Wolken S, Fasching N. Automated annotation with generative AI requires validation. arXiv Preprint posted online on May 31, 2023.
23. Liu K, Fu Y, Tan C, Chen M, Zhang N, Huang S, et al. Noisy-labeled NER with confidence estimation. 2021 Presented at: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6-11, 2021; Online.
24. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. DiffusionNER: boundary diffusion for named entity recognition. 2023 Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 9-14, 2023; Toronto, Canada.
25. Yada S, Nakamura Y, Wakamiya S, Aramaki E. Real-MedNLP: Overview of real document-based medical natural language processing task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies. 2022 Presented at: NTCIR 16 Conference; June 14-17, 2022; Tokyo, Japan p. 285-296 URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/01-NTCIR16-OV-MEDNLP-YadaS.pdf
26. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
27. Mahajan D, Liang JJ, Tsou C. Toward understanding clinical context of medication change events in clinical narratives. AMIA Annu Symp Proc 2021;2021:833-842 [FREE Full text] [Medline: 35308981]

28. Nishiyama T, Nishidani M, Ando A, Yada S, Wakamiya S, Aramaki E. NAISTSOC at the NTCIR-16 real-medNLP task. In: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies. 2022 Presented at: NTCIR 16 Conference; June 14-17, 2022; Tokyo, Japan URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/07-NTCIR16-MEDNLP-NishiyamaT.pdf
29. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
30. Yada S, Joh A, Tanaka R, Cheng F, Aramaki E, Kurohashi S. Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: starting from critical lung diseases. In: Proceedings of the 12th Conference on Language Resources and Evaluation. 2020 Presented at: LREC 2020; May 11-16, 2020; Marseille, France p. 4565-4572 URL: https://aclanthology.org/2020.lrec-1.561
31. Ollion E, Shen R, Macanovic A, Chatelain A. ChatGPT for text annotation? Mind the hype!. SocArXiv Preprint posted online on October 4, 2023. [doi: 10.31235/osf.io/x58kn]
32. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of Deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 1, 2018. [doi: 10.18653/v1/N19-1423]
33. Fuzzy annotation tool. GitHub. URL: https://github.com/gabrielandrade2/FuzzyAnnotationTool [accessed 2024-04-16]
34. Point-to-span estimation BERT model. GitHub. 2023. URL: https://github.com/gabrielandrade2/Point-to-Span-estimation [accessed 2024-04-16]
35. Tohoku NG. BERT base Japanese (character-level tokenization with whole word masking, jawiki-20200831). Hugging Face. URL: https://huggingface.co/tohoku-nlp/bert-base-japanese-char-v2 [accessed 2024-04-16]
36. Cohen J. A coefficient of agreement for nominal scales. Educ Psychological Meas 2016 Jul 02;20(1):37-46. [doi: 10.1177/001316446002000104]
37. Warrens MJ. Five ways to look at Cohen's Kappa. J Psychol Psychother 2015;05(04):1-4. [doi: 10.4172/2161-0487.1000197]
38. Tjong KSE, De MF. Introduction to the conll-2003 shared task: language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. 2003 Presented at: CoNLL-2003; May 31-June 1, 2003; Edmonton, AB p. 142-147 URL: https://aclanthology.org/W03-0419
39. Ghiasvand O, Kate R. Learning for clinical named entity recognition without manual annotations. Inform Med Unlocked 2018;13:122-127. [doi: 10.1016/j.imu.2018.10.011]
40. Ethical Guidelines for Medical and Health Research Involving Human Subjects. Ministry of Health, Labor and Welfare. URL: https://www.mhlw.go.jp/file/06-Seisakujouhou-10600000-Daijinkanboukouseikagakuka/0000080278.pdf [accessed 2024-04-16]

Abbreviations

AL: active learning
BERT: Bidirectional Encoder Representations From Transformers
EHR: electronic health record
GS: gold standard
GSC: gold standard corpus
GSM: gold standard model
IAA: interannotator agreement
NE: named entity
NER: named entity recognition
NLP: natural language processing

Edited by C Lovis, G Eysenbach; submitted 19.04.24; peer-reviewed by C Gaudet-Blavignac, L Raithel; comments to author 09.05.24; revised version received 23.05.24; accepted 25.05.24; published 02.07.24.

Please cite as:
Herman Bernardim Andrade G, Yada S, Aramaki E
Is Boundary Annotation Necessary? Evaluating Boundary-Free Approaches to Improve Clinical Named Entity Annotation Efficiency: Case Study
JMIR Med Inform 2024;12:e59680
URL: https://medinform.jmir.org/2024/1/e59680
doi:10.2196/59680
PMID:

Original Paper

Evaluating Large Language Models for Automated Reporting and Data Systems Categorization: Cross-Sectional Study

Qingxia Wu[1*], MD; Qingxia Wu[2,3*], PhD; Huali Li[4], MM; Yan Wang[1], MM; Yan Bai[1], MD; Yaping Wu[1], PhD; Xuan Yu[1], PhD; Xiaodong Li[1], MM; Pei Dong[2,3], PhD; Jon Xue[5], PhD; Dinggang Shen[5,6], PhD; Meiyun Wang[1,7], MD

[1]Department of Medical Imaging, Henan Provincial People's Hospital & People's Hospital of Zhengzhou University, Zhengzhou, China

[2]Research Intelligence Department, Beijing United Imaging Research Institute of Intelligent Imaging, Beijing, China

[3]Research and Collaboration, United Imaging Intelligence (Beijing) Co, Ltd, Beijing, China

[4]Department of Radiology, Luoyang Central Hospital, Luoyang, China

[5]Research and Collaboration, Shanghai United Imaging Intelligence Co, Ltd, Shanghai, China

[6]School of Biomedical Engineering, Shanghai Tech University, Shanghai, China

[7]Biomedical Research Institute, Henan Academy of Sciences, Zhengzhou, China

[*]these authors contributed equally

Corresponding Author:
Meiyun Wang, MD
Department of Medical Imaging
Henan Provincial People's Hospital & People's Hospital of Zhengzhou University
No 7, Weiwu Road, Jinshui District
Zhengzhou, 450001
China
Phone: 86 037165580267
Email: mywang@zzu.edu.cn

Abstract

Background: Large language models show promise for improving radiology workflows, but their performance on structured radiological tasks such as Reporting and Data Systems (RADS) categorization remains unexplored.

Objective: This study aims to evaluate 3 large language model chatbots—Claude-2, GPT-3.5, and GPT-4—on assigning RADS categories to radiology reports and assess the impact of different prompting strategies.

Methods: This cross-sectional study compared 3 chatbots using 30 radiology reports (10 per RADS criteria), using a 3-level prompting strategy: zero-shot, few-shot, and guideline PDF-informed prompts. The cases were grounded in Liver Imaging Reporting & Data System (LI-RADS) version 2018, Lung CT (computed tomography) Screening Reporting & Data System (Lung-RADS) version 2022, and Ovarian-Adnexal Reporting & Data System (O-RADS) magnetic resonance imaging, meticulously prepared by board-certified radiologists. Each report underwent 6 assessments. Two blinded reviewers assessed the chatbots' response at patient-level RADS categorization and overall ratings. The agreement across repetitions was assessed using Fleiss κ.

Results: Claude-2 achieved the highest accuracy in overall ratings with few-shot prompts and guideline PDFs (prompt-2), attaining 57% (17/30) average accuracy over 6 runs and 50% (15/30) accuracy with k-pass voting. Without prompt engineering, all chatbots performed poorly. The introduction of a structured exemplar prompt (prompt-1) increased the accuracy of overall ratings for all chatbots. Providing prompt-2 further improved Claude-2's performance, an enhancement not replicated by GPT-4. The interrun agreement was substantial for Claude-2 (k=0.66 for overall rating and k=0.69 for RADS categorization), fair for GPT-4 (k=0.39 for both), and fair for GPT-3.5 (k=0.21 for overall rating and k=0.39 for RADS categorization). All chatbots showed significantly higher accuracy with LI-RADS version 2018 than with Lung-RADS version 2022 and O-RADS (P<.05); with prompt-2, Claude-2 achieved the highest overall rating accuracy of 75% (45/60) in LI-RADS version 2018.

Conclusions: When equipped with structured prompts and guideline PDFs, Claude-2 demonstrated potential in assigning RADS categories to radiology cases according to established criteria such as LI-RADS version 2018. However, the current generation of chatbots lags in accurately categorizing cases based on more recent RADS criteria.

(JMIR Med Inform 2024;12:e55799) doi:10.2196/55799

KEYWORDS

Radiology Reporting and Data Systems; LI-RADS; Lung-RADS; O-RADS; large language model; ChatGPT; chatbot; chatbots; categorization; recommendation; recommendations; accuracy

Introduction

Since ChatGPT's public release in November 2022, large language models (LLMs) have attracted great interest in medical imaging applications [1]. Research indicated that ChatGPT showed promising applications in various aspects of the medical imaging process. Even without radiology-specific pretraining, LLMs can pass board examinations [2], provide radiology decision support [3], assist in differential diagnosis [3-6], and generate impressions from findings or structured reports [7-9]. These applications not only accelerate the imaging diagnosis process and alleviate the workload of doctors but also improve the accuracy of diagnosis [10]. However, limitations exist, with 1 study showing ChatGPT-3 producing erroneous answers for a third of daily clinical questions and about 63% of provided references were not found [11]. ChatGPT's dangerous tendency to produce inaccurate responses is less frequent in GPT-4 but still limits usability in medical education and practice at present [12]. Tailoring LLMs to radiology may enhance reliability, as an appropriateness criteria context aware chatbot outperformed generic chatbots and radiologists [12].

The American College of Radiology Reporting and Data Systems (RADS) standardizes communication of imaging findings. As of August 2023, there have been 9 disease-specific systems endorsed by the American College of Radiology, referring to products from the lexicons to report templates [13]. RADS reduces terminology variability, facilitates communication between radiologists and referring physicians, allows consistent evaluations, and conveys clinical significance to improve care. However, complexity and unfamiliarity limit adoption. Consequently, endeavors should be pursued to broaden the implementation of RADS. Therefore, we conducted this study to evaluate LLM's capabilities on a focused RADS assignment task for radiology reports.

A prompt serves as a directive or instruction given to LLMs to generate a particular response. The technique of "prompt tuning" has emerged as a valuable approach to refine the performance of LLMs, particularly for specific domains or tasks [14]. By providing structured queries or exemplary responses, the output of chatbots can be tailored for accurate and relevant answers. Such prompt-tuning strategies leverage LLMs' knowledge while guiding appropriate delivery for particular challenges [14]. Given the complexity and specificity of the RADS categorization, our investigation emphasizes different prompt impacts to assess chatbot capabilities and potential performance enhancement through refined prompting tuning.

In this study, our primary objective was to meticulously evaluate the performance of 3 LLMs (GPT-3.5, GPT-4, and Claude-2) for RADS categorization using different prompt-tuning strategies. We aimed to test their accuracy and consistency in RADS categorization and shed light on the potential benefits and limitations of relying on chatbot-derived information for the categorization of specific RADS.

Methods

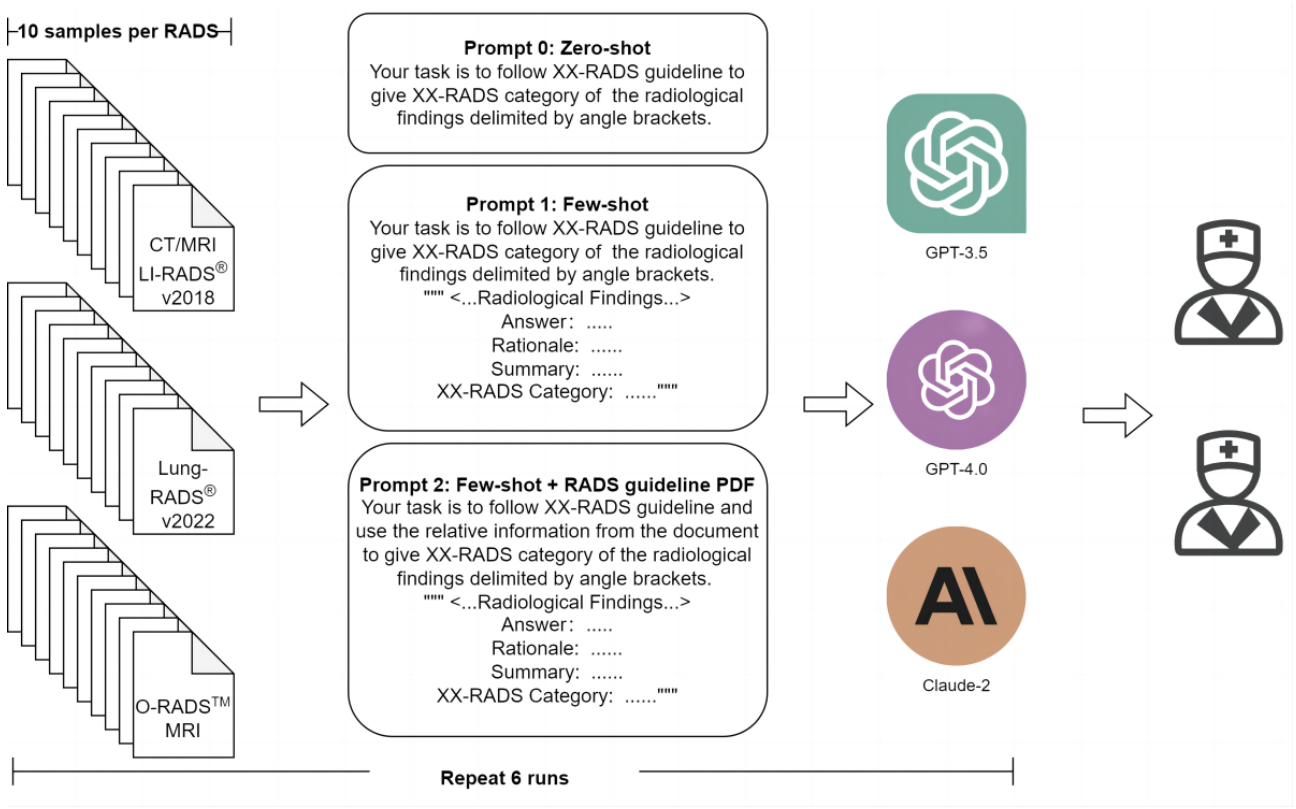
Ethical Considerations

As the study was based on radiological data that were artificially generated by radiologists and did not involve the participation of human subjects, the study was determined to be exempt from ethical review, in accordance with the regulations established by the institutional review board of Henan Provincial People's Hospital.

Study Design

The workflow of the study is shown in Figure 1. We conducted a cross-sectional analysis in September 2023 to evaluate the competency of 3 chatbots—GPT-3.5, GPT-4 (OpenAI, August 30, 2023 version) [15], and Claude-2 (Anthropic) [16]—in the task of assigning 3 RADS categorizations to radiology reports. Given the chatbot's knowledge cessation was as of September 2021, we opted for Liver Imaging Reporting & Data System (LI-RADS) version 2018 [17], Lung CT (computed tomography) Screening Reporting & Data System (Lung-RADS) version 2022 [18], and Ovarian-Adnexal Reporting & Data System (O-RADS) magnetic resonance imaging (MRI) (developed in 2022) [19] as the yardsticks to compare the responses engendered by GPT-3.5, GPT-4, and Claude-2. A total of thirty radiology reports for either CT or MRI examinations were composed for this analysis, with 10 cases representing each of the 3 RADS reporting systems. The radiology reports used for testing were generated by radiologists with more than 10 years' experience to correct the wording styles from real-life cases based on respective RADS systems. For each RADS (ie, LI-, Lung-, and O-RADS), we attempted to reflect the complexity and diversity so that the reports cover typical cases in clinical practice. Therefore, reports with 2-3 simple cases and 7-8 challenging cases were generated for 1 RADS. These include scenarios such as prior examination comparison, the presence of multiple nodules, extensive categorization under different RADS systems, and updates from the most recent LI-RADS and Lung-RADS guidelines. The characteristics of radiology reports for each RADS and the distribution of the number of the reports across the 3 RADS are shown in Multimedia Appendix 1. The objective was to evaluate the performance of chatbots on a highly structured radiology workflow task involving cancer risk categorization based on structured report inputs. The study design focused on a defined use case to illuminate the strengths and limitations of existing natural language-processing technology in this radiology subdomain.

Figure 1. Flowchart of the study design. CT: computed tomography; LI-RADS: Liver Imaging Reporting & Data System; Lung-RADS: Lung CT Screening Reporting & Data System; MRI: magnetic resonance imaging; O-RADS: Ovarian-Adnexal Reporting & Data System; RADS: Reporting and Data Systems.



Prompts

We collected and analyzed responses from GPT-3.5, GPT-4, and Claude-2 for each case. To mitigate bias, the radiological findings were presented individually via separate interactions, with corresponding responses saved for analysis. Three prompt templates were designed to elicit each RADS categorization along with explanatory rationale: Prompt-0 was a zero-shot prompt, merely introducing the RADS assignment task, such as "Your task is to follow Lung-RADS version 2022 guideline to give Lung-RADS category of the radiological findings delimited by angle brackets."

Prompt-1 was a few-shot prompt, furnishing an exemplar of RADS categorization including the reasoning, summarized impression, and final category. The following is an example:

> *Your task is to follow Lung-RADS version 2022 guideline to give Lung-RADS category of the radiological findings delimited by angle brackets. """ < …Radiological Findings… > Answer: Rationale: {…} Overall: {…} Summary: {…} Lung-RADS Category: X """*

Prompt-2 distinctly instructed chatbots to consult the PDF of corresponding RADS guidelines, compensating for these chatbots' lack of radiology-specific pretraining. For Claude-2, the PDF could be directly ingested, while GPT-4 required the use of an "Ask for PDF" plug-in to extract pertinent information [20,21].

Each case was evaluated 6 times with each chatbot across the 3 prompt levels. The representative radiological reports and prompts are shown in Multimedia Appendix 2. The links to all the prompts and guideline PDFs are shown in Multimedia Appendix 3.

Evaluation of Chatbots

Two study authors (QW and HL) independently evaluated the following for each chatbot response in a blinded manner, with any discrepancies resolved by a third senior radiologist (YW). The following were assessed for each response:

1. Patient-level RADS categorization: judged as correct, incorrect, or unsure. "Correct" denotes that the chatbot accurately identified the patient-level RADS category, irrespective of the rationale provided. "Unsure" denotes that the chatbot's response failed to provide a decisive RADS category. For example, a response articulating that "a definitive Lung-RADS category cannot be assigned" would be categorized as "unsure."
2. Overall rating: assessed as either correct or incorrect. A response is judged incorrect if any errors (Es) are identified, including the following:
   - E1: a factual extraction error that denotes the chatbots' inability to paraphrase the radiological findings accurately, consequently misinterpreting the information.
   - E2: hallucination, encompassing the fabrication of nonexistent RADS categories (E2a) and RADS criteria (E2b).
   - E3: a reasoning error, which includes the incapacity to logically interpret the imaging description (E3a) and the RADS category accurately (E3b). The subtype

errors for reasoning imaging description include the inability to reason lesion signal (E3ai), lesion size (E3aii), and enhancement (E3aiii) accurately.
- E4: an explanatory error, encompassing inaccurate elucidation of RADS category meaning (E4a) and erroneous explanation of the recommended management and follow-up corresponding to the RADS category (E4b).

If a chatbot's feedback manifested any of the aforementioned infractions, it was labeled as incorrect, with the specific type of error documented. To assess the consistency of the evaluations, a k-pass voting method was also applied. Specifically, a case was deemed accurately categorized if it met the criteria in a minimum of 4 out of the 6 runs.

Statistical Analyses

The accuracy of the patient-level RADS categorization and overall rating for each chatbot was compared using the chi-square test. The agreement across the 6 repeated runs was assessed using Fleiss κ. Agreement strength was interpreted as follows: <0 signified poor, 0-0.20 indicated slight, 0.21-0.40 represented fair, 0.41-0.60 was interpreted as moderate, 0.61-0.80 denoted substantial, and 0.81-1 was characterized as almost perfect. Statistical significance was defined as 2-sided $P<.05$. All analyses were performed using R statistical software (version 4.1.2; The R Foundation).

Results

Performance of Chatbots

The performance of chatbots is shown in Figure 2 and Tables 1 and 2, with the links to case-level details provided in Multimedia Appendix 4. For the overall rating (Table 1, average row and Figure 2A), Claude-2 with prompt-2 demonstrated significantly higher average accuracy across the 30 cases than Claude-2 with prompt-0 (odds ratio [OR] 8.16; $P<.001$). GPT-4 with prompt-2 also showed improved average accuracy compared with GPT-4 with prompt-0, but the difference was not statistically significant (OR 3.19; $P=.13$). When using the k-pass voting method (Table 1, k-pass voting row), Claude-2 with prompt-2 had significantly higher accuracy than Claude-2 with prompt-0 (OR 8.65; $P=.002$). Similarly, GPT-4 with prompt-2 was significantly more accurate than GPT-4 with prompt-0 (OR 11.98; $P=.01$). For the exact assignment of the patient-level RADS categorization (Table 2, average row and Figure 2B), Claude-2 with Prompt-2 showed significantly more average accuracy than Claude-2 with prompt-0 ($P=.04$).

Figure 2. Bar graphs show the comparison of chatbot performance across 6 runs regarding (A) overall rating and (B) patient-level Reporting and Data Systems categorization.

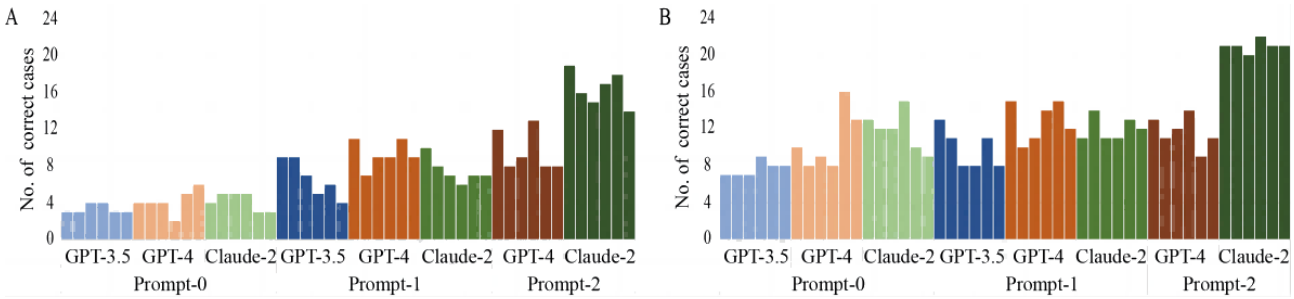


Table 1. Correct overall ratings of different chatbots and prompts.

Chatbots and prompts	Prompt-0, n (%; 95% CI)	Prompt-1, n (%; 95% CI)	Prompt-2, n (%; 95% CI)
GPT-3.5			
Run 1	3 (10; 3-28)	9 (30; 15-50)	N/A[a]
Run 2	3 (10; 3-28)	9 (30; 15-50)	N/A
Run 3	4 (13; 4-32)	7 (23; 11-43)	N/A
Run 4	4 (13; 4-32)	5 (17; 6-35)	N/A
Run 5	3 (10; 3-28)	6 (20; 8-39)	N/A
Run 6	3 (10; 3-28)	4 (13; 4-32)	N/A
Average[b]	3 (10; 3-28)	7 (23; 11-43)	N/A
K-pass voting[c]	1 (3; 0-19)	2 (7; 1-24)	N/A
GPT-4			
Run 1	4 (13; 4-32)	11 (37; 21-56)	12 (40; 23-59)
Run 2	4 (13; 4-32)	7 (23; 11-43)	8 (27; 13-46)
Run 3	4 (13; 4-32)	9 (30; 15-50)	9 (30; 15-50)
Run 4	2 (7; 1-24)	9 (30; 15-50)	13 (43; 26-62)
Run 5	5 (17; 6-35)	11 (37; 21-56)	8 (27; 13-46)
Run 6	6 (20; 8-39)	9 (30; 15-50)	8 (27; 13-46)
Average[b]	4 (13; 4-32)	9 (30; 15-50)	10 (33; 18-53)
K-pass voting[c]	1 (3; 0-19)	6 (20; 8-39)	9 (30; 15-50)[d]
Claude-2			
Run 1	4 (13; 4-32)	10 (33; 18-53)	19 (63; 44-79)
Run 2	5 (17; 6-35)	8 (27; 13-46)	16 (53; 35-71)
Run 3	5 (17; 6-35)	7 (23; 11-43)	15 (50; 33-67)
Run 4	5 (17; 6-35)	6 (20; 8-39)	17 (57; 38-74)
Run 5	3 (10; 3-28)	7 (23; 11-43)	18 (60; 41-77)
Run 6	3 (10; 3-28)	7 (23; 11-43)	14 (47; 29-65)
Average[b]	4 (13; 4-32)	8 (27; 13-46)	17 (57; 38-74)[d]
K-pass voting[c]	3 (10; 3-28)	7 (23; 11-43)	15 (50; 33-67)[d]

[a]N/A: not applicable.

[b]Accuracy by the average method.

[c]Accuracy by k-pass voting (≥4/6 runs correct).

[d]Significant between prompt-0 and prompt-2.

Table 2. The number of correct, incorrect, and unsure responses for patient-level Reporting and Data Systems categorization across different chatbots and prompts.

Chatbots and prompts	Correct/incorrect/unsure patient-level Reporting and Data Systems categories, n/n/n							
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Average[a]	K-pass voting[b]
GPT-3.5								
Prompt-0	7/23/0	7/23/0	7/23/0	9/21/0	8/21/1	8/20/2	8/22/0	7/23/0
Prompt-1	13/15/2	11/19/0	8/21/1	8/21/1	11/19/0	8/22/0	10/20/0	7/23/0
GPT-4								
Prompt-0	10/20/0	8/19/3	9/20/1	8/22/0	16/14/0	13/15/2	11/18/1	8/22/0
Prompt-1	15/14/1	10/18/2	11/18/1	14/15/1	15/14/1	12/18/0	13/16/1	11/19/0
Prompt-2	13/16/1	11/18/1	12/18/0	14/16/0	9/21/0	11/16/3	12/18/0	11/19/0
Claude-2								
Prompt-0	13/17/0	12/18/0	12/18/0	15/15/0	10/20/0	9/21/0	12/18/0	13/17/0
Prompt-1	11/19/0	14/16/0	11/19/0	11/19/0	13/17/0	12/18/0	12/18/1	11/19/0
Prompt-2	21/9/0	21/9/0	20/10/0	22/8/0	21/9/0	2021/8/1	21/9/0[c]	21/9/0

[a]Accuracy by the average method.

[b]Accuracy by k-pass voting (≥4/6 runs correct).

[c]Significant between prompt-0 and prompt-2.

Consistency of Chatbots

As shown in Table 3, among the 30 cases evaluated in 6 runs, Claude-2 with prompt-2 showed substantial agreement (*k*=0.65 for overall rating; *k*=0.74 for RADS categorization). GPT-4, when interfaced with prompt-2, demonstrated moderate agreement (*k*=0.46 for overall rating; *k*=0.41 for RADS categorization). When evaluated with prompt-1, GPT-4 presented moderate agreement (*k*=0.38 for overall rating; *k*=0.42 for RADS categorization). In contrast, Claude-2 showed substantial agreement (*k*=0.63 for overall rating; *k*=0.61 for RADS categorization), while GPT-3.5 exhibited a range from slight to fair agreement. With prompt-0, Claude-2 showed moderate agreement (*k*=0.49) for overall rating and substantial agreement for RADS categorization (*k*=0.65). GPT4 manifested slight agreement (*k*=0.19) for the overall rating and fair agreement for RADS categorization. Meanwhile, GPT-3.5 showed fair agreement (*k*=0.28) for the overall rating and moderate agreement (*k*=0.57) for RADS categorization.

Table 3. The consistency of different chatbots and prompts among 6 runs.

	Prompt-0, Fleiss κ (95% CI)	Prompt-1, Fleiss κ (95% CI)	Prompt-2, Fleiss κ (95% CI)	All, Fleiss κ (95% CI)
Patient-level RADS[a] categorization				
GPT-3.5	0.57 (0.48-0.65)	0.24 (0.15-0.32)	N/A[b]	0.39 (0.33-0.46)
GPT-4	0.33 (0.25-0.42)	0.42 (0.34-0.5)	0.41 (0.33-0.5)	0.39 (0.34-0.44)
Claude-2	0.65 (0.56-0.74)	0.61 (0.52-0.7)	0.74 (0.65-0.83)	0.69 (0.64-0.74)
Overall rating				
GPT-3.5	0.28 (0.19-0.37)	0.14 (0.05-0.23)	N/A	0.21 (0.14-0.27)
GPT-4	0.19 (0.1-0.28)	0.38 (0.29-0.47)	0.46 (0.37-0.55)	0.39 (0.34-0.45)
Claude-2	0.49 (0.4-0.58)	0.63 (0.53-0.72)	0.65 (0.56-0.75)	0.66 (0.61-0.72)

[a]RADS: Reporting and Data Systems.

[b]N/A: not applicable.

Subgroup Analysis

Since the knowledge base for ChatGPT was frozen as of September 2021, accounting for the knowledge limitations of LLMs developed before the latest RADS guideline updates, we compared the responses of different RADS criteria. The total accurate responses across 6 runs were computed for all prompts. Both GPT-4 and Claude-2 demonstrated superior performance in the context of LI-RADS CT/MRI version 2018 as opposed

to Lung-RADS version 2022 and O-RADS MRI (all $P<.05$; Table 4). Figure 3 delineates the performance of various chatbots across different prompts and RADS categories. For the overall rating (Figure 3A), Claude-2 exhibited a progressive trend of enhancement of overall rating accuracy from prompt-0 to prompt-1 to prompt-2, with 20.0% (12/60), 36.7% (22/60), and 75.0% (45/60) for LIRADS; 11.7% (7/60), 18.3% (11/60), and 48.3% (29/60) for Lung-RADS; and 10.0% (6/60), 20.0% (12/60), and 41.7% (25/60) for O-RADS, respectively. Notably, with prompt-2, Claude-2 achieved the highest overall rating accuracy of 75% in older systems such as LI-RADS version 2018. Conversely, GPT-4 improved with prompt-1/2 over prompt-0, but prompt-2 did not exceed prompt-1. For the RADS categorization (Figure 3B), prompt-1 and prompt-2 outperformed prompt-0 for LI-RADS, irrespective of chatbots. However, for Lung-RADS and O-RADS, prompt-0 sometimes superseded prompt-1.

Table 4. The performance of chatbots within different RADS criteria[a].

Chatbots and RADS[b]	Year of development	RADS categorization (correct/incorrect/unsure), n/n/n	*P* value	Overall rating (correct/incorrect), n/n	*P* value
GPT-3.5					
LI-RADS[c] CT[d]/MRI[e]	2018	32/86/2	Reference	22/98	Reference
Lung-RADS[f]	2022	38/78/4	.83	14/106	.15
O-RADS[g] MRI	2022	35/84/1	.46	24/96	.87
GPT-4					
LI-RADS CT/MRI	2018	104/74/2	Reference	78/102	Reference
Lung-RADS	2022	40/128/12	<.001	21/159	<.001
O-RADS MRI	2022	67/110/3	<.001	40/140	<.001
Claude-2					
LI-RADS CT/MRI	2018	93/86/1	Reference	79/101	Reference
Lung-RADS	2022	63/117/0	.001	47/133	<.001
O-RADS MRI	2022	113/67/0	.04	43/137	<.001

[a]Data are aggregate numbers across 6 runs.

[b]RADS: Reporting and Data Systems.

[c]LI-RADS: Liver Imaging Reporting and Data System.

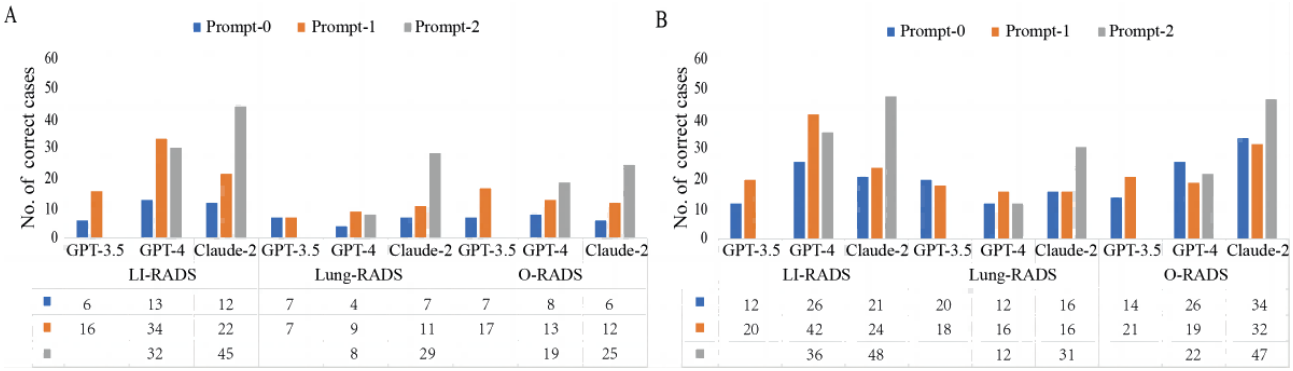
[d]CT: computed tomography.

[e]MRI: magnetic resonance imaging.

[f]Lung-RADS: Lung CT Screening Reporting and Data System.

[g]O-RADS: Ovarian-Adnexal Reporting and Data System.

Figure 3. The performance of chatbots and prompts within different Reporting and Data Systems criteria. (A) Overall rating and (B) patient-level RADS categorization. LI-RADS: Liver Imaging Reporting and Data System; Lung-RADS: Lung CT (computed tomography) Screening Reporting and Data System; O-RADS: Ovarian-Adnexal Reporting and Data System.

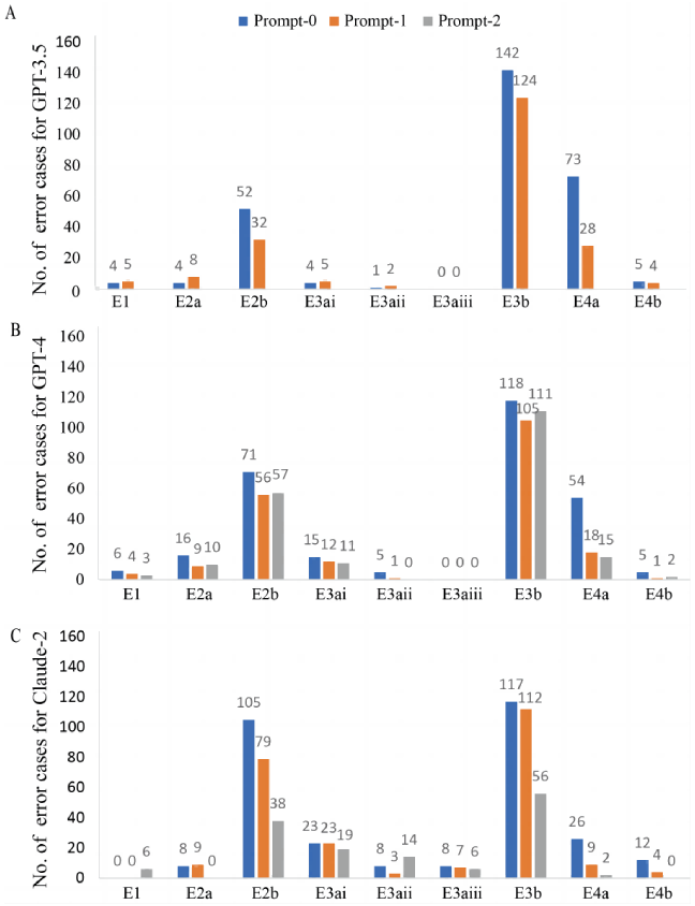


Analysis of Error Types

A total of 1440 cases were analyzed for error types, with details provided in Multimedia Appendix 4. The bar plot illustrating the distribution of errors across the 3 chatbots is shown in Figure 4. A typical example of factual extraction error (E1) occurred in response to the seventh Lung-RADS question. The statement "The 3mm solid nodule in the lateral basal segmental bronchus

is subsegmental" is inaccurate, as the lateral basal segmental bronchus represents one of the 18 defined lung segments and not a subsegment [22].

Figure 4. The number of error types for different chatbots. E1: Factual extraction error denotes the chatbots' inability to paraphrase the radiological findings accurately, consequently misinterpreting the information. E2: Hallucination, encompassing the fabrication of nonexistent Reporting and Data Systems (RADS) categories (E2a) and RADS criteria (E2b). E3: Reasoning error, which includes the incapacity to logically interpret the imaging description (E3a) and the RADS category accurately (E3b). The subtype errors for reasoning imaging description include the inability to reason lesion signal (E3ai), lesion size (E3aii), and enhancement (E3aiii) accurately. E4: Explanatory error, encompassing inaccurate elucidation of RADS category meaning (E4a) and erroneous explanation of the recommended management and follow-up corresponding to the RADS category (E4b).



Hallucination of inappropriate RADS categories (E2a) occurred more frequently with prompt-0 across all 3 chatbots. However, this error rate decreased to zero for Claude-2 when using prompt-2, a trend not seen with GPT-3.5 or GPT-4. A recurrent E2a error in LI-RADS was the obsolete category LR-5V from the 2014 version, now superseded by LR-TIV in subsequent editions [23,24]. Furthermore, hallucination of invalid RADS criteria (E2b) was more prevalent than that of E2a. For instance, the LI-RADS second question response stating "T2 marked hyperintensity is a feature commonly associated with hepatocellular carcinoma (HCC)" is inaccurate, as T2-marked hyperintensity is characteristic of hemangioma and not hepatocellular carcinoma. Despite initial higher E2b rates, Claude-2 demonstrated a substantial reduction with prompt-2 (105 to 38 instances), exceeding the decrement seen with GPT-4 (71 to 57 instances).

Regarding reasoning error, incorrect RADS category reasoning (E3b) was the most frequent error but decreased for all chatbots with prompt-1 and prompt-2 versus prompt-0. Claude-2 reduced errors by almost half with prompt-2, while the GPT-4 decrease was less pronounced. Lesion signal interpretation errors (E3ai) included misinterpreting hypointensity on diffusion-weighted imaging as "restricted diffusion," rather than facilitated diffusion. Lesion size reasoning errors (E3aii) occurred in 34 of 1440 cases, predominantly by Claude-2 (25/34, 73.5%), especially in systems such as Lung-RADS and LI-RADS where size is critical for categorization. Examples were attributing a 12-mm pulmonary nodule to the ≥6-mm but <8-mm range, or assigning a hepatic lesion measuring 2.3 cm × 1.5 cm to the 10- to 19-mm category. Reasoning enhancement errors (E3aiii) were exclusive to Claude-2 in O-RADS, where enhancement significantly impacts categorization. Misclassifying images at 40 seconds postcontrast as early or delayed enhancement exemplifies this error.

Explanatory errors (E4) including incorrect RADS category definitions (E4a) and inappropriate management recommendations (E4b) also substantially declined with prompt-1 and prompt-2. For instance, in the first Lung-RADS question response, the statement "The 4X designation indicates infectious/inflammatory etiology is suspected." is incorrect. Lung-RADS 4X means category 3 or 4 nodules with additional features or imaging findings that increase suspicion of lung cancer [18].

Discussion

Principal Findings

In this study, we evaluated the performance of 3 chatbots—GPT-3.5, GPT-4, and Claude-2—in categorizing radiological findings according to RADS criteria. Using 3 levels of prompts providing increasing structure, examples, and domain knowledge, the chatbots' accuracies and consistencies were quantified across 30 cases. The best performance was achieved by Claude-2 when provided with few-shot prompting and the RADS criteria PDFs. Interestingly, the chatbots tended to categorize better for the relatively older LI-RADS version 2018 criteria in contrast to the more recent Lung-RADS version 2022 and O-RADS guidelines published after the chatbots' training cutoff.

The incorporation of RADS, which standardizes reporting in radiology, has been a significant advancement, although the multiplicity and complexity of these systems impose a steep learning curve for radiologists [13]. Even for subspecialized radiologists at tertiary hospitals, mastering the numerous RADS guidelines poses challenges, requiring familiarity with the lexicons, regular application in daily practice, and ongoing learning to remain current with new versions. While previous studies have shown that LLMs could assist radiologists in various tasks [2-5,7,11], their performance at RADS categorization from imaging findings is untested. We therefore evaluated LLMs for focused RADS categorization of testing cases.

Without prompt engineering (prompt-0), all chatbots performed poorly. However, accuracy improved for all chatbots when provided an exemplar prompt demonstrating the desired response structure (prompt-1). This underscores the use of prompt tuning for aligning LLMs to specific domains such as radiology. Further enriching prompt-1 with the RADS guideline PDFs as a relevant knowledge source (prompt-2) considerably enhanced Claude-2's accuracy, a feat not mirrored by GPT-4. This discrepancy could stem from ChatGPT's reliance on an external plug-in to access documents, while Claude-2's architecture accommodates the direct assimilation of expansive texts, benefiting from its larger-context window and superior long document–processing capabilities.

Notably, we discerned performance disparities across RADS criteria. When queried on older established guidelines such as LI-RADS version 2018 [17], the chatbots demonstrated greater accuracy than more recent schemes such as Lung-RADS version 2022 and O-RADS [18,19,25]. Specifically, GPT-4 and Claude-2 had significantly higher total correct ratings for LI-RADS than for Lung-RADS and O-RADS (all $P<.05$). This could be attributed to their extensive exposure to the voluminous data related to the matured LI-RADS during their pretraining phase. With prompt-2, Claude-2 achieved 75% (45/60) accuracy for overall rating LI-RADS categorization. The poorer performance on newer RADS criteria highlights the need for strategies to continually align LLMs with the most up-to-date knowledge.

A deep dive into the error-type analysis revealed informative trends. Incorrect RADS category reasoning (E3b) constituted the most frequent error across chatbots, decreasing with prompt tuning. Targeted prompting also reduced critical errors such as hallucinations of RADS criteria (E2b) and categories (E2a) likely by constraining output to valid responses. During pretraining, GPT-liked LLMs predict the next word in the unlabeled data set, risking learning fallacious relationships between RADS features. For instance, Lung-RADS version 2022 lacks categories 5 and 6 [18], though some other RADS such as Breast Imaging Reporting and Data System include them [26]. Using prompt-0, chatbots erroneously hallucinated Lung-RADS categories 5 and 6. Explanatory errors (E4) including inaccurate definition of the assigned RADS category (E4a) and inappropriate management recommendations (E4b) also substantially declined with prompt tuning. For instance, when queried on the novel O-RADS criteria with prompt-0, chatbots hallucinated follow-up recommendations from other RADS criteria and responded "O-RADS category 3 refers to an indeterminate adnexal mass and warrants short-interval follow-up." Targeted prompting appears to mitigate these critical errors such as hallucination and incorrect reasoning. Careful prompt engineering is essential to properly shape LLM knowledge for radiology tasks.

Limitations

There are also several limitations in this study. First, only the LI-RADS CT/MRI and O-RADS MRI were included, excluding LI-RADS ultrasound (US) and O-RADS US guidelines, which are often practiced in an independent US department [27,28]. Second, the chatbot's performance was heavily dependent on prompt quality. We test only 3 types of prompts and further prompt strategies studies are warranted to investigate the impact of more exhaustive engineering on chatbots' accuracy. Third, GPT-4-turbo was released on November 6, 2023, representing the latest GPT-4 model with improvements in instruction following, reproducible outputs, and more [29]. Furthermore, its training data extend to April 2023 compared with September 2021 for the base GPT-4 model tested here. We are uncertain about this newest GPT-4-turbo model's performance on the RADS categorization task. Evaluating GPT-4-turbo represents an important direction for future work. Fourth, our study focused on 3 of 9 RADS [13], with a limited 10 cases for each RADS category. Although our choice ensured a blend of old and new guidelines and tried to cover all the RADS scores as much as possible, extending evaluations to all the RADS guidelines and incorporating more radiology reports from real clinical scenarios could offer deeper insights into potential limitations. Nonetheless, this initial study highlights critical considerations of prompt design and knowledge calibration required for safely applying LLMs in radiology. Fifth, evaluating the performance of the LLM in comparison with radiologists of varying expertise levels proves valuable for discerning its strengths and weaknesses in real-world applications. This comparative analysis will be undertaken in our forthcoming studies.

Conclusions

When equipped with structured prompts and guideline PDFs, Claude-2 demonstrates potential in assigning RADS categories

to radiology cases according to established criteria such as LI-RADS version 2018. However, the current generation of chatbots lags in accurately categorizing cases based on more recent RADS criteria. Our study highlights the potential of LLMs in streamlining radiological categorizations while also pinpointing the enhancements necessary for their dependable application in clinical practice for RADS categorization tasks.

Acknowledgments

This study has received funding from the National Natural Science Foundation of China (82371934 and 82001783) and Joint Fund of Henan Province Science and Technology R&D Program (225200810062). The authors thank Chuanjian Lv, MD; Zejun Wen, MM; and Jianghua Lou, MM, for their help in drafting the radiology reports with regard to Lung CT Screening Reporting and Data System, Liver Imaging Reporting and Data System, and Ovarian-Adnexal Reporting and Data System, respectively.

Authors' Contributions

QW (Henan Provincial People’s Hospital & People’s Hospital of Zhengzhou University), QW (Beijing United Imaging Research Institute of Intelligent Imaging), HL, Y Wang, YB, Y Wu, XY, and MW contributed to study design. QW (Henan Provincial People’s Hospital & People’s Hospital of Zhengzhou University) and QW (Beijing United Imaging Research Institute of Intelligent Imaging) contributed to the statistical analysis. All authors contributed to the acquisition, analysis, or interpretation of the data; the drafting of the manuscript; and critical revision of the manuscript.

Conflicts of Interest

QW and PD are senior engineers of Beijing United Imaging Research Institute of Intelligent Imaging and United Imaging Intelligence (Beijing) Co, Ltd. JX and DS are senior specialists of Shanghai United Imaging Intelligence Co, Ltd. The companies have no role in designing and performing the surveillance and analyzing and interpreting the data. All other authors report no conflicts of interest relevant to this article.

Multimedia Appendix 1
The characteristics of radiology reports for each of the Reporting and Data Systems (RADS) and the distribution of the number of the reports across the 3 RADS.
[DOCX File , 107 KB - medinform_v12i1e55799_app1.docx]

Multimedia Appendix 2
Representative radiology reports and prompts.
[DOCX File , 18 KB - medinform_v12i1e55799_app2.docx]

Multimedia Appendix 3
Links to prompts and guideline PDFs.
[DOCX File , 12 KB - medinform_v12i1e55799_app3.docx]

Multimedia Appendix 4
Links to prompt engineering results.
[DOCX File , 11 KB - medinform_v12i1e55799_app4.docx]

References

1. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or Pandora's box? JAMA Intern Med 2023 Jun 01;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]
2. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023 Jun;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
3. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023 Oct;20(10):990-997 [FREE Full text] [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]
4. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. Radiology 2023 Jul;308(1):e231040. [doi: 10.1148/radiol.231040] [Medline: 37462501]
5. Kottlors J, Bratke G, Rauen P, Kabbasch C, Persigehl T, Schlamann M, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. Radiology 2023 Jul;308(1):e231167. [doi: 10.1148/radiol.231167] [Medline: 37404149]

6. Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. Radiology 2023;307(5):e230877. [doi: 10.1148/radiol.230877] [Medline: 37310247]
7. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023 May;307(4):e230725. [doi: 10.1148/radiol.230725] [Medline: 37014240]
8. Wagner MW, Ertl-Wagner BB. Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. Can Assoc Radiol J 2024 Feb;75(1):69-73 [FREE Full text] [doi: 10.1177/08465371231171125] [Medline: 37078489]
9. Ziegelmayer S, Marka AW, Lenhart N, Nehls N, Reischl S, Harder F, et al. Evaluation of GPT-4's chest x-ray impression generation: a reader study on performance and perception. J Med Internet Res 2023 Dec 22;25:e50865 [FREE Full text] [doi: 10.2196/50865] [Medline: 38133918]
10. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology 2023 Jun;307(5):e230987. [doi: 10.1148/radiol.230987] [Medline: 37191491]
11. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
12. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology 2023 Jul;308(1):e230970. [doi: 10.1148/radiol.230970] [Medline: 37489981]
13. American College of Radiology. Reporting and Data Systems (RADS). URL: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems [accessed 2023-08-26]
14. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023 Oct 04;25:e50638 [FREE Full text] [doi: 10.2196/50638] [Medline: 37792434]
15. OpenAI. URL: https://openai.com [accessed 2023-11-08]
16. Anthropic. Claude 2. URL: https://www.anthropic.com/index/claude-2 [accessed 2023-11-08]
17. Chernyak V, Fowler KJ, Kamaya A, Kielar AZ, Elsayes KM, Bashir MR, et al. Liver Imaging Reporting and Data System (LI-RADS) Version 2018: imaging of hepatocellular carcinoma in at-risk patients. Radiology 2018 Dec;289(3):816-830 [FREE Full text] [doi: 10.1148/radiol.2018181494] [Medline: 30251931]
18. Martin MD, Kanne JP, Broderick LS, Kazerooni EA, Meyer CA. Update: Lung-RADS 2022. Radiographics 2023 Nov;43(11):e230037. [doi: 10.1148/rg.230037] [Medline: 37856315]
19. Sadowski EA, Thomassin-Naggara I, Rockall A, Maturen KE, Forstner R, Jha P, et al. O-RADS MRI risk stratification system: guide for assessing adnexal lesions from the ACR O-RADS Committee. Radiology 2022 Apr;303(1):35-47 [FREE Full text] [doi: 10.1148/radiol.204371] [Medline: 35040672]
20. AskYourPDF. URL: https://askyourpdf.com [accessed 2023-11-08]
21. ChatGPT plugins. URL: https://openai.com/blog/chatgpt-plugins [accessed 2023-11-08]
22. Jones J, Rasuli B, Vadera S. Bronchopulmonary segmental anatomy. URL: https://doi.org/10.53347/rID-13644 [accessed 2023-11-08]
23. Mitchell DG, Bruix J, Sherman M, Sirlin CB. LI-RADS (Liver Imaging Reporting and Data System): summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. Hepatology 2015 Mar;61(3):1056-1065. [doi: 10.1002/hep.27304] [Medline: 25041904]
24. Elsayes K, Hooker J, Agrons M, Kielar A, Tang A, Fowler K, et al. 2017 Version of LI-RADS for CT and MR imaging: an update. Radiographics 2017;37(7):1994-2017. [doi: 10.1148/rg.2017170098] [Medline: 29131761]
25. Suarez-Weiss KE, Sadowski EA, Zhang M, Burk KS, Tran VT, Shinagare AB. Practical tips for reporting adnexal lesions using O-RADS MRI. Radiographics 2023;43(7):e220142. [doi: 10.1148/rg.220142] [Medline: 37319025]
26. American College of Radiology. Breast Imaging Reporting & Data System. URL: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads [accessed 2023-11-03]
27. Strachowski LM, Jha P, Phillips CH, Blanchette Porter MM, Froyman W, Glanc P, et al. O-RADS US v2022: an update from the American College of Radiology's Ovarian-Adnexal Reporting and Data System US Committee. Radiology 2023;308(3):e230685. [doi: 10.1148/radiol.230685] [Medline: 37698472]
28. Quaia E. State of the art: LI-RADS for contrast-enhanced US. Radiology 2019;293(1):4-14. [doi: 10.1148/radiol.2019190005] [Medline: 31453768]
29. New models and developer products announced at DevDay. URL: https://openai.com/blog/new-models-and-developer-products-announced-at-devday [accessed 2023-11-09]

Abbreviations

CT: computed tomography
E: error
LI-RADS: Liver Imaging Reporting & Data System

LLM: large language model
Lung-RADS: Lung CT Screening Reporting & Data System
MRI: magnetic resonance imaging
O-RADS: Ovarian-Adnexal Reporting & Data System
OR: odds ratio
RADS: Reporting and Data Systems
US: ultrasound

Edited by C Lovis; submitted 25.12.23; peer-reviewed by Z Liu, D Bu, TAR Sure, S Nuthakki, L Zhu; comments to author 14.01.24; revised version received 02.02.24; accepted 25.05.24; published 17.07.24.

Please cite as:
Wu Q, Wu Q, Li H, Wang Y, Bai Y, Wu Y, Yu X, Li X, Dong P, Xue J, Shen D, Wang M
Evaluating Large Language Models for Automated Reporting and Data Systems Categorization: Cross-Sectional Study
JMIR Med Inform 2024;12:e55799
URL: https://medinform.jmir.org/2024/1/e55799
doi:10.2196/55799
PMID:39018102

Original Paper

Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study

Fadi Aljamaan[1], MD; Mohamad-Hani Temsah[1], MD; Ibraheem Altamimi[1], MD; Ayman Al-Eyadhy[1], MD; Amr Jamal[1], MD; Khalid Alhasan[1], MD; Tamer A Mesallam[2], MD, PhD; Mohamed Farahat[2], MD, PhD; Khalid H Malki[2], MD, PhD

[1]College of Medicine, King Saud University, Riyadh, Saudi Arabia

[2]Department of Otolaryngology, College of Medicine, Research Chair of Voice, Swallowing, and Communication Disorders, King Saud University, Riyadh, Saudi Arabia

Corresponding Author:
Khalid H Malki, MD, PhD
Department of Otolaryngology, College of Medicine
Research Chair of Voice, Swallowing, and Communication Disorders
King Saud University
12629 Abdulaziz Rd
Al Malaz
Riyadh, P.BOX 2925 Zip 11461
Saudi Arabia
Phone: 966 114876100
Email: kalmalki@ksu.edu.sa

Abstract

Background: Artificial intelligence (AI) chatbots have recently gained use in medical practice by health care practitioners. Interestingly, the output of these AI chatbots was found to have varying degrees of hallucination in content and references. Such hallucinations generate doubts about their output and their implementation.

Objective: The aim of our study was to propose a reference hallucination score (RHS) to evaluate the authenticity of AI chatbots' citations.

Methods: Six AI chatbots were challenged with the same 10 medical prompts, requesting 10 references per prompt. The RHS is composed of 6 bibliographic items and the reference's relevance to prompts' keywords. RHS was calculated for each reference, prompt, and type of prompt (basic vs complex). The average RHS was calculated for each AI chatbot and compared across the different types of prompts and AI chatbots.

Results: Bard failed to generate any references. ChatGPT 3.5 and Bing generated the highest RHS (score=11), while Elicit and SciSpace generated the lowest RHS (score=1), and Perplexity generated a middle RHS (score=7). The highest degree of hallucination was observed for reference relevancy to the prompt keywords (308/500, 61.6%), while the lowest was for reference titles (169/500, 33.8%). ChatGPT and Bing had comparable RHS (β coefficient=–0.069; *P*=.32), while Perplexity had significantly lower RHS than ChatGPT (β coefficient=–0.345; *P*<.001). AI chatbots generally had significantly higher RHS when prompted with scenarios or complex format prompts (β coefficient=0.486; *P*<.001).

Conclusions: The variation in RHS underscores the necessity for a robust reference evaluation tool to improve the authenticity of AI chatbots. Further, the variations highlight the importance of verifying their output and citations. Elicit and SciSpace had negligible hallucination, while ChatGPT and Bing had critical hallucination levels. The proposed AI chatbots' RHS could contribute to ongoing efforts to enhance AI's general reliability in medical research.

(JMIR Med Inform 2024;12:e54345) doi:10.2196/54345

KEYWORDS

artificial intelligence (AI) chatbots; reference hallucination; bibliographic verification; ChatGPT; Perplexity; SciSpace; Elicit; Bing

Introduction

Artificial intelligence (AI) evolved from the early twentieth century until Turing [1] conceptualized machine learning in the 1950s and introduced the idea of using machines to process information in order to solve problems and make decisions. Since then, scientists and others have pursued the dream of creating machines that mimic human intelligence through cognitive processes such as thinking, data processing, and planning. AI development requires foundational work in spoken language processing and information storage. Achieving these milestones involves numerous obstacles, including significant technical and financial challenges, which delay the maturation of AI [2].

A major hurdle in AI development was natural language processing, which progressed through multiple phases and culminated with the invention of bidirectional encoder representations from transformers, self-attention, and sequence-to-sequence deep learning technologies. These transformers marked a pivotal advancement in AI development, generating fluent and coherent large language models (LLMs), as they enabled the analysis of each word in an input sequence in context to its neighbors on both sides [3].

In 2018, OpenAI unveiled its first Generative Pre-trained Transformer, ChatGPT, enhancing its utility and potential applications in various human activities, including health care and medical research sectors. Initially, ChatGPT stored data in its database without access to continually updated human literature. It later evolved to employ internet access. Subsequently, other AI chatbots specializing in health care and medical research were introduced. The implementation of AI chatbots in health care is optimistically viewed as a means to improve health care systems, medical education, and patient outcomes [4-6]. They are also viewed as potentially valuable tools in medical research and manuscript writing due to their capabilities in organization, data processing, text generation, and summarization [7,8].

Since the public launch of ChatGPT at the end of 2022, the medical literature surrounding AI has expanded tremendously [9]. Most studies have addressed its potential contributions to medical research and health care practices. However, a critical analysis of the literature reveals that a minority of AI chatbot users realized early that these LLMs do not voluntarily cite credible references to authenticate their information [10-12]. Even when ChatGPT and other AI chatbots were challenged to authenticate their outputs with references [13,14], they generated multiple citations with detailed bibliographic data that seemed perfectly authentic, but most were actually fabricated, contained at least one falsified citation data, or were completely erroneous when verified through medical literature resources [13-16].

These fictitious citations raised major concerns about the AI chatbots' algorithms and methodology of natural language processing, especially regarding references, their bibliographic data credibility, and authenticity. Such erroneous outputs were described in the literature as AI hallucinations or fabrications. The mechanism of these fabrications remains unclear, though some AI developers have identified this faulty output and described it as a misalignment between user expectations and AI chatbot capabilities related to machine learning training issues without proposing definite explanation for that phenomenon or a methodology to assess its references' hallucination degree or impact on AI platform output [17]. Another possible explanation for reference hallucinations is an encoding/decoding transformer glitch or the way AI chatbots handle bibliographic data as text amenable to manipulation. Some investigators assessed the AI-generated medical information, including the ability to verify references and its publication date, but did not pay attention specifically to certain identifiers of references that help to verify their authenticity [18].

Investigating the sources of these hallucinations and taking serious steps to fix the contributing factors to this phenomenon is an urgent need in the current stage to implement AI technology in health care practice and medical research with high credibility. The first step is to identify the degree of these hallucinations, especially in relation to references and their bibliographic data across different AI chatbots.

Utilization of AI-generated medical information is becoming a routine tool for physicians and trainees in medical education and even for diagnosis and management. Therefore, the verification of references for this information is extremely important to avoid adopting or utilizing erroneous unverified information and might be disastrous if used in medical management. Therefore, we proposed to construct a reference evaluation tool that can be universally applied to assess AI chatbot–generated references, which would be a helpful tool for assessing the output, its credibility, and its ability to be implemented into practice. This study aims to introduce this tool as Reference Hallucination Score (RHS) and to demonstrate its application in evaluating and comparing the outputs of 6 AI chatbots, stratify their citation output according to the RHS, and assess the variables that are associated with RHS. This initiative is the first to specifically address the gap in evaluating AI referencing hallucinations, thereby offering a unique approach that fills a critical gap in enhancing the reliability of AI-generated references, improving trust in AI outputs within medical research, and guiding the development of more robust and reliable AI models.

Methods

Medical Prompts

We challenged 6 AI chatbots with 10 medical prompts addressing various medical topics. The selected AI chatbots were ChatGPT 3.5, Bard, Perplexity, Bing, Elicit, and SciSpace [19-24]. We chose them because these are the most widely used AI chatbots in the literature and medical research and because of their ease of access by users. Using the focus group technique, we structured 10 medical prompts with the best possible textual format and clarity to be understood by the AI chatbots. Each of the 2 prompts addressed a similar medical topic: the first in a basic format and the second in a complex or scenario-based format. The prompts are provided in Multimedia Appendix 1. The 5 general topics for the medical prompts were glucose control in gestational diabetes, triggering factors in older adults

with asthma, septic shock in infants with severe combined immunodeficiency, arthritis in patients with sickle cell disease, and substance abuse in patients with personality disorders.

All the prompts were formulated according to the following format: (1) begin by searching PubMed for papers related to the topic; (2) review the search results and select 10 relevant and recent papers; (3) ensure that all information is accurate and up-to-date; (4) format the list of papers in a clear and organized manner, using a consistent style for each entry; (5) include any additional information or notes that may be relevant or helpful for readers; and (6) double-check the accuracy and completeness of the list before publishing or submitting it. The same 10 prompts were applied to each of the 6 AI chatbots. Each prompt requested the AI chatbots to compile a list of 10 references. Each prompt listed 7 items for every reference: reference title, journal name, authors, DOI (digital object identifier), publication date, paper web link, and the reference's relevance to the keyword prompts.

RHS Development

Using the Delphi technique, we proposed the RHS. Score development went through a robust and meticulous methodology as follows:

Literature Review

We commenced with an exhaustive review of existing literature to identify unique identifiers for reference scoring. This step was vital for understanding the current landscape and ensuring that our tool addresses the new gap in the field.

Expert Consultation Using the Delphi Technique

Our approach utilized the Delphi technique—a systematic, multistep process involving rounds of anonymous feedback from experts to reach a consensus. We detail the implementation of this technique as follows.

1. Initial survey of senior librarians: We consulted 2 senior librarians to gather insights on the proposed bibliographic identifiers, seeking their expert suggestions for improvements.
2. Expanded survey of the senior physicians: We further extended our consultation to 12 senior physicians and 11 junior physicians, most of whom are academicians, to assess the relevance and importance of the suggested bibliographic identifiers and to collect additional suggestions.
3. Consensus building among authors: The final step involved synthesizing the feedback received and reaching a consensus among the authors, based on the mean results from the surveyed academics, to finalize the bibliographic identifiers for the hallucination score.

The RHS is an AI chatbot scoring system to evaluate paper references generated by the AI chatbot based on the hallucination severity. The RHS is calculated based on the total score according to the presence of hallucination in any of the 7 parameters (Table 1). The parameters are 4 reference items or identifiers given a score of 2 if they encountered any error in the reference title, journal name, authors' names, or DOI, as the authors judged it as a major degree of hallucination. A score of 1 was given to any error in any of the other 3 identifiers, that is, reference publication date, reference web link, or reference relevance to the keyword prompts, as they were judged as minor degrees of hallucination. The maximum RHS is 11, indicating the maximum degree of reference hallucination, and the minimum RHS is 0, denoting no reference hallucination. To achieve the best outcome from the proposed RHS, the prompts submitted to the AI chatbot should include clear instructions to include all 7 referencing items. The authenticity of the citations' items is evaluated by comparing AI chatbot responses to PubMed and Google Scholar responses. If the AI chatbot could not produce any reference to a specific prompt, the AI chatbot was given a score "N" for that prompt and was scored as failing to generate a result.

Table 1. Reference hallucination score (total score=11).

Reference identifier	Item hallucination score
Erroneous date of publication[a]	1
Erroneous web link of the paper[b]	1
Erroneous citation relevance[c]	1
Erroneous title of the paper[d]	2
Erroneous digital object identifier[e]	2
Erroneous authors' names[f]	2
Erroneous name of the journal[g]	2

[a]The publication date is missing or inaccurate.

[b]The link to the paper is missing or directs to a different paper or an error page.

[c]The keyword prompts are not in the title, abstract, or reference keywords.

[d]The title provided by the artificial intelligence platform is misspelled, incomplete, or nonexistent.

[e]Digital object identifier is missing, inaccurate, nonexistent, or directs to a different paper.

[f]Any author's name is missing, misspelled, or nonexistent.

[g]The journal's name is missing, misspelled, did not publish the paper, or does not exist.

Testing AI Chatbots

Our methodology of judging the hallucinations depended on using a systematic verification process based on PubMed and Google Scholar to ensure legitimacy, accuracy, and transparency. Each reference was initially verified using PubMed by searching for its DOI. If the DOI was inconclusive or unavailable, we leveraged a combination of the paper's title, author names, and other vital details to ascertain its existence and accuracy within the PubMed database. In case of failure to reach the reference through PubMed, Google Scholar was the second step for verification using DOI, the paper's title, authors, and other pertinent details to ensure the reference authenticity.

Each reference's total hallucination score was calculated according to the RHS methodology. Then, the mean RHS of the produced references of each prompt was calculated. This was applied to the 6 AI chatbots. The chatbots' RHS items were compared across the studied chatbots according to hallucination, correct results, and failure to generate results. The median RHS of the complex prompts was compared with basic prompts, and finally, the median RHS of each AI chatbot was compared across all the studied chatbots. Linear regression was used to assess the independent association between mean RHS and other predictors, namely, the studied chatbot, prompt type, and prompt iteration.

Statistical Analysis

The mean (SD) values were used to describe continuous variables. The median (IQR) values were used to describe continuous variables with statistical evidence of skewness. The frequencies and percentages were used to describe categorically measured variables. The Kolmogorov-Smirnov statistical test and histograms were used to assess the statistical normality assumption of the metric variables. The categorical Cronbach α test was used to assess the internal consistency for the reliability of the 7 parameters of the RHS. The nonparametric Kruskal-Wallis analysis of variance test was used to compare the RHS item results across the studied chatbots and to compare the median RHS for statistically significant differences. The Mann-Whitney *U* nonparametric test compared the chatbots' median RHS for statistically significant differences between basic and advanced prompts. The multivariable generalized linear models with gamma regression analysis assessed significant differences in the mean RHS by regressing it against the AI chatbot, prompt complexity, and prompt iterations. The association between the mean RHS and tested predictor independent variables was expressed as an exponentiated β coefficient (risk rate) with its associated 95% CI. The SPSS statistical computing software (IBM Corp) was used for the statistical data analysis, and the α significance level was considered at .05.

Ethical Considerations

This paper did not involve research on living creatures; therefore, no institutional review board approval was required.

Results

A total of 10 prompts were entered into each AI chatbot. Half inquired about the basic medical topics, and the other half about clinical scenarios or complex medical topics. Each prompt requested 10 references related to the prompt topic with their reference data according to our research methodology. Bard was the only AI chatbot that failed to produce any response to all the 10 applied medical prompts. Bard's response to our prompts was, "I'm a language model and don't have the capacity to understand and respond." The AI chatbots failed to generate any reference response for 35 (7%) of the 500 references. The highest hallucination/erroneous output was for the reference relevancy to the prompt content (308/500, 61.6%), followed by publication date (237/500, 47.4%), authors' names (228/500, 45.6%), DOI (227/500, 45.4%), and reference web link (187/500, 37.4%). Regarding the reference title and journal name, the AI chatbots' output had hallucination results of 33.8% (169/500) and 37.6% (188/500), respectively.

Figure S1 in Multimedia Appendix 1 shows each AI chatbot's reference results (correct and hallucinating or erroneous results). SciSpace and Elicit had the highest correct reference identifiers' results, with 629 and 597, respectively. ChatGPT had the highest hallucination results at 592. Bing had the highest rate of failure to generate results at 210. The Kruskal-Wallis nonparametric test showed that the AI chatbots differed significantly with respect to their total hallucination results (χ^2_4=205.9; *P*<.001), correct results (χ^2_4=305.0; *P*<.001), and failure to generate results (χ^2_4=104.3; *P*<.001). The Bonferroni-adjusted post hoc pairwise comparison test was used to compare the chatbots. SciSpace and Elicit did not differ significantly with respect to their hallucination or correct results (*P*>.99). Both had significantly fewer hallucinations and higher correct results compared to the other chatbots (*P*<.001). ChatGPT had the highest rate of hallucination results. Bing and ChatGPT had no significant difference in their correct results (*P*>.99). Bing and Perplexity had no significant difference in their hallucination results (*P*>.99) and were significantly superior to ChatGPT. Perplexity was superior to both regarding its correct results (Bing, *P*=.002; ChatGPT, *P*=.004). Bing had the highest rate of failing to generate results compared to the rest (*P*<.001), followed by Perplexity.

Table 2 displays the detailed results of the studied AI chatbots' reference characteristics. The Kruskal-Wallis nonparametric test showed that the AI chatbots differed significantly in their results. SciSpace and Elicit generated the lowest number of hallucination results regarding title, journal name, authors' names, DOI, and reference web links compared to all the other AI chatbots (*P*<.001). However, Elicit and Perplexity showed no significant difference in hallucinating the publication date results, and both hallucinated more than SciSpace.

Table 2. Artificial intelligence chatbot hallucinating/erroneous reference identifiers' results.

Reference identifier	ChatGPT	Perplexity	SciSpace	Elicit	Bing	Chi-square *(df)*	*P* value
Title	82	41	0	0	46	231.4 (4)	<.001
Digit Object Identifier	89	73	8	4	53	261.26 (4)	<.001
Journal name	83	56	0	0	49	249.4 (4)	<.001
Authors' names	89	74	0	0	65	341.1 (4)	<.001
Publication date	88	48	0	40	61	199.9 (4)	<.001
Reference web link	84	49	3	1	50	231.4 (4)	<.001
Reference relevance to topic prompt	77	63	60	58	50	10.77 (4)	.03

Perplexity, Bing, and ChatGPT had no significant difference regarding their hallucination results for DOI, authors' names, reference web links, and reference titles. Perplexity and Bing hallucinated significantly lesser than ChatGPT regarding the journal name. Perplexity hallucinated slightly lesser than ChatGPT regarding reference title. Bing had significantly fewer hallucination results for references' irrelevancy to the prompt medical topic than ChatGPT (*P*=.045). The remaining AI chatbots, including Bing, did not show a significant difference (*P*>.05).

Table 3 shows AI chatbots' RHS with the descriptive analysis of the median (IQR) for each studied AI chatbot. Kruskal-Wallis analysis showed that chatbots differed significantly with respect to their overall measured hallucination scores (χ^2_4=277.7; *P*<.001). ChatGPT and Bing had the highest median RHS but did not differ significantly (*P*>.99). SciSpace and Elicit had the lowest median RHS (*P*<.001) and did not differ significantly (*P*>.99). Perplexity and Bing did not vary significantly with respect to their median total RHS (*P*=.19). On the other hand, Perplexity had a considerably lower median total RHS than ChatGPT (*P*=.003). For the prompt complexity, the median RHS did not differ significantly for each of the studied AI chatbots (*P*>.05).

Table 3. Reference hallucination scores with bivariate analysis of the artificial intelligence chatbots.

Chatbot	Hallucination score, median (IQR)[a]			*z* test statistic *(df)*	*P* value
	Total RHS[a]	Basic prompts RHS	Advanced prompts RHS		
ChatGPT	11 (1)	11 (1)	11 (0.25)	1.25 (100)	.21
Perplexity	7 (5)	7 (8.25)	8 (5)	0.837 (95)	.40
SciSpace	1 (1)	1 (1)	1 (1)	0.942 (100)	.35
Elicit	1 (2)	1 (2)	1 (2)	0.207 (100)	.84
Bing	11 (6)	11 (6)	11 (5)	0.207 (70)	.84

[a]RHS: reference hallucination score.

Table 4 shows the multivariable generalized linear models with γ regression of the mean total hallucination score based on the chatbot, prompt complexity level, and prompt type. ChatGPT and Bing had the highest mean total hallucination scores and did not differ significantly (*P*=.32). SciSpace had the lowest mean total hallucination score compared to ChatGPT (β coefficient=–1.748; *P*<.001). Elicit had the second lowest total mean hallucination score compared to ChatGPT (β coefficient=–1.63; *P*<.001). Perplexity had the third lowest score compared to ChatGPT (β coefficient=–0.345; *P*<.001).

Table 4. Multivariable generalized linear mixed regression analysis of the artificial intelligence chatbots' total hallucination score[a].

	β coefficient (95% CI)	*P* value
Intercept	2.142 (1.997 to 2.288)	<.001
Chatbot vs Bing	–0.069 (–0.206 to 0.067)	.32
Chatbot vs Elicit	–1.630 (–1.769 to –1.492)	<.001
Chatbot vs SciSpace	–1.748 (–1.880 to –1.617)	<.001
Chatbot vs Perplexity	–0.345 (–0.510 to –0.181)	<.001
Prompt complexity vs advanced level	0.486 (0.326 to 0.645)	<.001
Prompt number	0.018 (–0.006 to 0.041)	.10
Interaction effect: prompt number vs prompt complexity	–0.077 (–0.108 to –0.046)	<.001

[a]Dependent outcome variable: reference hallucination score+1; probability distribution = gamma link function with log shape.

The level of prompt complexity also significantly affected the hallucination score when compared across all the AI chatbots. The advanced prompts' mean total hallucination score was significantly higher than the basic prompts' mean total hallucination score (β coefficient=0.486; *P*<.001). The prompt topic did not correlate significantly with the AI chatbots' mean hallucination score (*P*=.14). However, the interaction term between the prompt medical topic and prompt complexity was found to be statistically significant (β coefficient=–0.077; *P*<.001), indicating that some topics, when presented to the chatbots in a complex scenario, resulted in significantly lower mean total hallucinations compared to the basic presentation of the same topic (*P*<.001). The mean hallucination score based on prompt topic and prompt complexity did not significantly interact with any specific AI chatbots studied.

Discussion

Principal Findings

Our findings showed variations in the RHS across different AI chatbots, ranging from almost null for SciSpace and Elicit to a critically high degree of hallucination for ChatGPT [25]. Among the bibliographic items we studied, the publication date (237/500, 47.4%) showed erroneous or hallucinating results, while the reference title (169/500, 33.8%) showed the least hallucinating or erroneous results. Reference relevancy to the prompt topic was the most common source of hallucination, ranging from 50 erroneous results in case of Bing chatbot up to 77 erroneous results in case of ChatGPT. Bard failed to generate any references for all the studied 10 prompts.

The scientific community uses a transparent, reproducible, and accessible archiving system for its large, evolving research data. For example, FAIR (Findability, Accessibility, Interoperability, and Reuse) guidelines ensure that reference data are archived in a findable, accessible, interoperable, and reusable format to facilitate its citation, maintain researchers' credibility, and ensure data integrity and nonrepudiation [26]. Citation is a vital step for information verification and authentication. AI chatbots that have gained recent widespread use still lack a transparent and robust system for verifying citations and their information sources. Additionally, AI chatbots encounter a hallucination/fabrication phenomenon recognized early in their use in health care. Hallucinations have been encountered in various domains, including the content itself and the cited references, including their bibliographic identifiers [12,27-29]. However, it is worth noting that this obstacle is improving gradually, especially by introducing research, dedicated medical chatbots, and upgrading existing ones [16,30].

A possible explanation for AI chatbots' referencing hallucination is the methodology LLMs use to handle citations. LLMs may deal with citations and their bibliographic identifiers as text, making them vulnerable to paraphrasing and other linguistic manipulations and perhaps as a watermark so that the output that is generated by AI could be identified versus that produced by humans to reduce AI misuse [29,31,32]. Buholayka et al [33] reported that ChatGPT is trained to give uninterrupted flow of conversation even at the cost of giving hallucinating results. Another possible mechanism related to AI chatbots' natural language processing methodology involves encoding and decoding defects during prompt processing, generating errors, and fabricating results [34]. Additional factors might include insufficient training data for the LLM [17,35,36], context misinterpretation, lack of external validation, and overreliance on pattern recognition, all of which could contribute to variable degrees of hallucination, depending on the specific LLM and prompt structure.

Ye et al [37] constructed a systematic approach to AI chatbot hallucination by introducing a unique classification of hallucinations across diverse text generation actions, thereby furnishing theoretical insights, identification techniques, and enhancement strategies. Their methodology consists of 3 domains: (1) comprehensive classification for hallucinations manifested in text-generation tasks, (2) theoretical examinations of hallucinations in LLMs and amelioration, and (3) several research trajectories that hold promise for future exploration. Dhuliawala et al [38] suggested another model to potentially reduce hallucination by the chain-of-verification method. Their 4-step process consists of the chatbot drafting its initial response, formulating verification questions to scrutinize the draft, addressing these questions independently to prevent bias, and finally, generating a thoroughly verified response. Further research needs to be performed to see which model and which LLM receives a better hallucination score with time.

Our findings align with those of Hua et al [25] who evaluated hallucinations in AI-generated ophthalmic scientific abstracts

and references. The uniqueness of our pioneering work is the introduction of a novel RHS to assess AI chatbots. RHS is based on 6 important reference items that make each reference unique and easily trackable for citation, in addition to a seventh item related to the relevance of the reference to the topic prompt [26]. The score was constructed based on any reference's most usable and unique items, with differential weights according to their uniqueness and importance in tracking and identifying any reference.

Variations in hallucination degrees among the different bibliographic items highlight the variations among AI chatbots in handling and identifying references. Having a critically high hallucination rate of the cited reference's relevance to the prompted topic stresses the possibility that AI chatbots identify certain keywords in the prompt and try to search for relevant references but, most of the time, fail to identify the correct and relevant keywords, or as described by a recent OpenAI company statement, they do not align properly with user intentions, leading to the citation of relatively irrelevant sources [17]. Further, hallucinating the paper title and DOI is risky in terms of inability to access the reference. Other identifier hallucinations such as journal name or authors may not greatly block the access to the cited reference.

According to our study, ChatGPT 3.5 had the highest total hallucination score, with the most hallucinations in all aspects of bibliographic items. This echoes other observations. Walters and Wilder [16] described many hallucinations in ChatGPT's bibliographic citations. Their study used ChatGPT 3.5 and ChatGPT 4 to generate literature reviews. They analyzed 636 bibliographic citations across 84 papers, finding a significant number of fabricated citations (55% for GPT 3.5, 18% for GPT 4) and errors in the nonfabricated ones (43% for GPT 3.5 and 24% for GPT 4). Despite GPT 4 showing notable enhancement and insights over GPT 3.5, fabricated references persist [39]. Bing's total hallucination score did not differ significantly from that of ChatGPT as they both had critically high scores. SciSpace and Elicit had comparable individual rates of hallucination regarding the different bibliographic items, although SciSpace had the lowest total number of hallucinations, followed by Elicit. When considering the different studied reference items, ChatGPT, Bing, and Perplexity had comparable hallucinating results apart from the journal name, where Bing and Perplexity hallucinated lesser than ChatGPT. Perplexity stood in the middle, as its overall hallucination score was worse than SciSpace and Elicit on one side and better than ChatGPT and Bing on the other. Our observations align with those of others who investigated AI platforms for writing and research objectives, as they found that Elicit and SciSpace are far more dedicated to searching for scientific papers and summarizing references [40]. This observation stresses scholars' urge to vigilantly examine reference accuracy for any LLM-generated citations, especially if they are not dedicated to research purposes [12].

An interesting observation from our study is the failure of some chatbots to generate the prompted citations, as ChatGPT, Bing, and Perplexity had comparable individual hallucination reference results apart from journal names, where Bing and Perplexity hallucinated lesser than ChatGPT. Bing had a significantly higher rate of failure than ChatGPT, even though it performed comparably in all aspects. Such performance by Bing has been observed in another study that prompted ChatGPT, Bard, and Bing for multiple choice question generation, where Bing had a significant rate of generation failure compared to the other two [41].

Prompt structure and complexity had an interesting association with hallucination score, as complex or clinical scenarios triggered significantly more hallucinations across AI chatbots but not for certain ones. This was also observed when challenging ChatGPT versions 3.5 and 4 with orthopedic questions matched with images, as they both performed far better with simple text multiple choice questions than those with images [42]. On the other hand, when prompted in a complex or scenario format, specific medical topics caused lesser hallucination than when prompted in a basic format. This observation might be explained by the transformer's methodology and their differential performance with different text structures.

Bard performance might point to serious glitches in its text recognition or transformer performance in medical research, at least in certain topics or in the ones used in our study. Previous work has shown serious fabrications encountered in Bard similar to ChatGPT [43]. Overall, our observations extend other findings that AI chatbots provided citations with varying degrees of inaccuracy or hallucination, necessitating users to independently verify the information obtained from these language models [43]. As such, hallucinations put the whole AI chatbot data output under significant question, especially in implementing and applying AI aid into clinical practice [44,45].

Our study incorporates a meticulous design by constructing a novel scoring system to assess AI chatbots' hallucination in relation to references and their reference items with differential weights based on their importance. The multifaceted verification approach that we adopted is a robust method to ascertain the authenticity and relevance of the references generated, providing a replicable model for future studies. That score has proven to differentiate the performance of 5 common chatbots skeptically. The invention of a hallucination score will be a vital step toward systematically evaluating and improving the referencing capabilities of AI chatbots and LLMs. It will also triage AI chatbots' hallucinations, which is a critical step in verifying the authenticity of their content.

Study Limitations

Our study has potential limitations. The methodology used to construct the RHS is novel and is liable for future improvements. The RHS included a limited number of bibliographic parameters, which we believed, based on our consensus and expert colleagues of academics, are the most important and unique reference identifiers. However, other researchers might perceive additional variables or identifiers as crucial. The prompt structure that we used was after extensive trials to reach the prompt structure that produces AI-generated references and their identifiers accurately as much as possible. Still, the prompt design is liable for limitation, as LLM output depends hugely on the prompt structure fed to them. Prompt structure might explain partly the failure of Bard Chatbot to generate any

references, as it might need a special design of prompting. However, we do not believe that the prompt strategy was a major limitation, as it succeeded in generating outputs in almost all the other studied AI chatbots. Additionally, the medical prompts utilized might have a limited scientific scope, which may not cover the full spectrum of medical topics and scenarios that LLMs might encounter in real-world applications. Yet, proposed prompts to assess AI chatbots' referencing hallucination require future refinement, especially with the introduction of new AI chatbots and specifically those that are specialized in the medical field.

Regarding the verification process of the references' identifiers, even though we utilized multiple web-based steps utilizing PubMed and Google Scholar, this might still be suboptimal because although we relied heavily on existing databases and search engines, those engines might have their own set of limitations in indexing or recognizing all published literature, and future researchers might propose more universal agreed-on methodology in that regard. We selected only 6 chatbots for assessment; this might not provide a fair and comprehensive understanding of the hallucination problems encountered across the myriad of AI chatbots that are increasingly becoming available and more specialized. The potential biases in selecting AI chatbots and medical prompts, along with the verification process used, might impact the study results by not fully representing the breadth of AI capabilities and challenges. These limitations should be addressed in future research by expanding the number of chatbots studied, diversifying the medical prompts and their structure addressing medical basic knowledge, diagnosis, and management, in order to strengthen the generalizability of the findings. Finally, the verification processes need to be refined. Our proposed RHS needs future sharpening and application to different and future AI chatbots, especially medical ones, to test its generalizability and sensitivity to assess reference hallucination. Adding more bibliographic parameters might enhance its sensitivity. Furthermore, refining the definition of erroneous citations could potentially improve RHS performance and applicability, particularly in relation to the relevance of the references to the prompted topic.

Conclusion

Our novel RHS tool encompasses a methodology for delineating referencing inaccuracies, crucially in medical domains. It has shown variations across the analyzed AI chatbots. We evaluated the performance of 6 common AI chatbots. Elicit and SciSpace had the least hallucination, with almost none, while ChatGPT and Bing had a critical degree of hallucination. This emphasizes the pressing need for enhanced evaluation mechanisms of AI chatbots' output, particularly the cited references, and highlights the need to verify their output and apply it skeptically, all in order to grade AI chatbots' credibility in terms of their contribution in health care and medical research areas. Additionally, our work establishes a foundation for ensuing research aimed at augmenting the reliability of AI chatbots in academic and clinical landscapes. Improving the LLM mechanism of reference recognition and handling is an important necessity and needs maturation and improvement of the algorithms. Training in user prompt strategy is another trajectory to address to achieve the best performance of these chatbots and improve chatbot-user alignment. Future improvement of RHS or developing new versions will improve AI chatbot assessment and categorization and potentially help AI engineers to evaluate their work. The significance of RHS and its potential impact on improving the reliability of AI-generated references cannot be overstated. The key takeaways highlight the broader implications for the use of AI in medical research, emphasizing the necessity for rigorous evaluations to enhance trust and reliability in AI outputs.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research, King Saud University, for funding through the Vice Deanship of Scientific Research Chairs, Research Chair of Voice, Swallowing, and Communication Disorders. The funders of this study had no role in the design of the study; collection, analyses, or interpretation of data; writing of the manuscript; or decision to publish the results.

Data Availability

The data may be made available upon reasonable request to the corresponding author.

Authors' Contributions

KHM, FA, and M-HT conceptualized this study. FA, IT, and KHM conducted all the investigations. KHM, FA, and M-HT contributed to the methodology of this study and supervised this study. FA and M-HT wrote the original draft. FA, M-HT, IT, AA, AJ, KA, TAM, MF, and KHM performed the reviewing and editing. All authors have read and agreed to the published version of this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data of the medical prompts.

[DOCX File , 131 KB - medinform_v12i1e54345_app1.docx]

References

1. Turing AM. I.—Computing machinery and intelligence. Mind 1950;LIX(236):433-460. [doi: 10.1093/mind/LIX.236.433]
2. Scott IA, Zuccon G. The new paradigm in machine learning - foundation models, large language models and beyond: a primer for physicians. Intern Med J 2024 May;54(5):705-715. [doi: 10.1111/imj.16393] [Medline: 38715436]
3. Kulkarni A, Shivananda A, Kulkarni A, et al. Evolution of neural networks to large language models. In: Applied Generative AI for Beginners. Berkeley, CA: Apress; 2023.
4. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. JMIR Med Educ 2023 Nov 01;9:e51421 [FREE Full text] [doi: 10.2196/51421] [Medline: 37910155]
5. Gödde D, Nöhl S, Wolf C, Rupert Y, Rimkus L, Ehlers J, et al. A SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis of ChatGPT in the medical literature: concise review. J Med Internet Res 2023 Nov 16;25:e49368 [FREE Full text] [doi: 10.2196/49368] [Medline: 37865883]
6. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023 Sep 04;9:e46482 [FREE Full text] [doi: 10.2196/46482] [Medline: 37665620]
7. Vincent J. How artificial intelligence will affect the future of medical publishing. Crit Care 2023 Jul 06;27(1):271 [FREE Full text] [doi: 10.1186/s13054-023-04511-9] [Medline: 37641127]
8. Shan Y, Ji M, Xie W, Qian X, Li R, Zhang X, et al. Language use in conversational agent-based health communication: systematic review. J Med Internet Res 2022 Jul 08;24(7):e37403 [FREE Full text] [doi: 10.2196/37403] [Medline: 35802407]
9. Temsah M, Altamimi I, Jamal A, Alhasan K, Al-Eyadhy A. ChatGPT surpasses 1000 publications on PubMed: envisioning the road ahead. Cureus 2023 Sep;15(9):e44769 [FREE Full text] [doi: 10.7759/cureus.44769] [Medline: 37809155]
10. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus 2023 Aug;15(8):e43271 [FREE Full text] [doi: 10.7759/cureus.43271] [Medline: 37692629]
11. Quinn TP, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc 2021 Mar 18;28(4):890-894 [FREE Full text] [doi: 10.1093/jamia/ocaa268] [Medline: 33340404]
12. Athaluri SA, Manthena SV, Kesapragada VSRKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus 2023 Apr;15(4):e37432 [FREE Full text] [doi: 10.7759/cureus.37432] [Medline: 37182055]
13. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. Mayo Clinic Proceedings: Digital Health 2023 Sep;1(3):226-234 [FREE Full text] [doi: 10.1016/j.mcpdig.2023.05.004]
14. Suppadungsuk S, Thongprayoon C, Krisanapan P, Tangpanithandee S, Garcia Valencia O, Miao J, et al. Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. J Clin Med 2023 Aug 25;12(17):5550 [FREE Full text] [doi: 10.3390/jcm12175550] [Medline: 37685617]
15. Khlaif ZN, Mousa A, Hattab MK, Itmazi J, Hassan AA, Sanmugam M, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR Med Educ 2023 Sep 14;9:e47049 [FREE Full text] [doi: 10.2196/47049] [Medline: 37707884]
16. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023 Sep 07;13(1):14045 [FREE Full text] [doi: 10.1038/s41598-023-41032-5] [Medline: 37679503]
17. Aligning language models to follow instructions. OpenAI. URL: https://openai.com/research/instruction-following [accessed 2023-06-11]
18. Ghanem YK, Rouhi AD, Al-Houssan A, Saleh Z, Moccia MC, Joshi H, et al. Dr. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc 2024 May;38(5):2887-2893 [FREE Full text] [doi: 10.1007/s00464-024-10739-5] [Medline: 38443499]
19. Chat Generative Pre-trained Transformer. ChatGPT 3.5. URL: https://chat.openai.com/chat [accessed 2023-06-11]
20. Gemini. URL: https://bard.google.com/chat [accessed 2023-06-11]
21. Perplexity: Where Knowledge Begins. URL: https://www.perplexity.ai/ [accessed 2023-06-11]
22. Bing. URL: https://www.bing.com/ [accessed 2023-06-11]
23. Elicit. URL: https://elicit.com/ [accessed 2023-06-11]
24. SCISPACE. URL: https://typeset.io/ [accessed 2023-06-11]
25. Hua H, Kaakour A, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol 2023 Sep 01;141(9):819-824. [doi: 10.1001/jamaophthalmol.2023.3119] [Medline: 37498609]
26. Brown R. The importance of data citation. Oxford University Press 2021 Mar 1:211-211. [doi: 10.1093/biosci/biab012]
27. Kumar M, Mani UA, Tripathi P, Saalim M, Roy S. Artificial hallucinations by Google Bard: think before you leap. Cureus 2023 Aug;15(8):e43313 [FREE Full text] [doi: 10.7759/cureus.43313] [Medline: 37700993]

28. Currie GM. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? Semin Nucl Med 2023 Sep;53(5):719-730. [doi: 10.1053/j.semnuclmed.2023.04.008] [Medline: 37225599]
29. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb;15(2):e35179 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
30. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: A comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023 Jun;15(6):e40822 [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
31. Hurrell L. DALL-E 3 watermark launched by OpenAI to reduce AI misuse. TechMonitor. URL: https://techmonitor.ai/technology/ai-and-automation/dall-e-3-watermark [accessed 2023-06-11]
32. Temsah M, Alhuzaimi AN, Almansour M, Aljamaan F, Alhasan K, Batarfi MA, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of AI-generated imagery in DALL·E 3 for illustrating congenital heart diseases. J Med Syst 2024 May 23;48(1):54. [doi: 10.1007/s10916-024-02072-0] [Medline: 38780839]
33. Buholayka M, Zouabi R, Tadinada A. The readiness of ChatGPT to write scientific case reports independently: a comparative evaluation between human and artificial intelligence. Cureus 2023 May;15(5):e39386 [FREE Full text] [doi: 10.7759/cureus.39386] [Medline: 37378091]
34. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput. Surv 2023 Mar 03;55(12):1-38. [doi: 10.1145/3571730]
35. Garvey KV, Thomas Craig KJ, Russell R, Novak LL, Moore D, Miller BM. Considering clinician competencies for the implementation of artificial intelligence-based tools in health care: findings from a scoping review. JMIR Med Inform 2022 Nov 16;10(11):e37478 [FREE Full text] [doi: 10.2196/37478] [Medline: 36318697]
36. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
37. Ye H, Liu T, Zhang A, Hua W, Jia W. Cognitive mirage: A review of hallucinations in large language models. ArXiv. Preprint posted online on September 13, 2023 . [doi: 10.48550/arXiv.2309.06794]
38. Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A. Chain-of-verification reduces hallucination in large language models. ArXiv. Preprint posted online on September 25, 2023 . [doi: 10.48550/arXiv.2309.11495]
39. Al-Tawfiq JA, Jamal A, Rodriguez-Morales AJ, Temsah M. Enhancing infectious disease response: A demonstrative dialogue with ChatGPT and ChatGPT-4 for future outbreak preparedness. New Microbes New Infect 2023 Jun;53:101153 [FREE Full text] [doi: 10.1016/j.nmni.2023.101153] [Medline: 37252334]
40. Giglio AD, Costa MUPD. The use of artificial intelligence to improve the scientific writing of non-native English speakers. Rev Assoc Med Bras (1992) 2023;69(9):e20230560 [FREE Full text] [doi: 10.1590/1806-9282.20230560] [Medline: 37729376]
41. Agarwal M, Sharma P, Goswami A. Analyzing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in Medical Physiology. Cureus 2023 Jun;15(6):e40977 [FREE Full text] [doi: 10.7759/cureus.40977] [Medline: 37519497]
42. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopedic resident performance on orthopedic assessment examinations. J Am Acad Orthop Surg 2023 Dec 01;31(23):1173-1179 [FREE Full text] [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]
43. McGowan A, Gui Y, Dobbs M, Shuster S, Cotter M, Selloni A, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. Psychiatry Res 2023 Aug;326:115334. [doi: 10.1016/j.psychres.2023.115334] [Medline: 37499282]
44. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res 2023 Aug 22;25:e48659 [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]
45. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah M. Artificial intelligence (AI) chatbots in medicine: A supplement, not a substitute. Cureus 2023 Jun;15(6):e40922 [FREE Full text] [doi: 10.7759/cureus.40922] [Medline: 37496532]

Abbreviations

AI: artificial intelligence
DOI: digital object identifier
FAIR: Findability, Accessibility, Interoperability, and Reuse
LLM: large language model
RHS: reference hallucination score

Edited by A Castonguay; submitted 06.11.23; peer-reviewed by M Chatzimina; comments to author 07.12.23; revised version received 05.01.24; accepted 03.07.24; published 31.07.24.

Please cite as:
Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, Mesallam TA, Farahat M, Malki KH
Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study
JMIR Med Inform 2024;12:e54345
URL: https://medinform.jmir.org/2024/1/e54345
doi:10.2196/54345
PMID:

Original Paper

A Case Demonstration of the Open Health Natural Language Processing Toolkit From the National COVID-19 Cohort Collaborative and the Researching COVID to Enhance Recovery Programs for a Natural Language Processing System for COVID-19 or Postacute Sequelae of SARS CoV-2 Infection: Algorithm Development and Validation

Andrew Wen[1,2*], MSc; Liwei Wang[1,2*], PhD; Huan He[1], PhD; Sunyang Fu[1,2], PhD; Sijia Liu[1], PhD; David A Hanauer[3], MSc, MD; Daniel R Harris[4], PhD; Ramakanth Kavuluru[5], PhD; Rui Zhang[6], PhD; Karthik Natarajan[7], PhD; Nishanth P Pavinkurve[7], MSc; Janos Hajagos[8], PhD; Sritha Rajupet[8], MPH, MD; Veena Lingam[8], MD; Mary Saltz[8], MD; Corey Elowsky[8], MSc; Richard A Moffitt[8], PhD; Farrukh M Koraishy[9], MD, PhD; Matvey B Palchuk[10], MSc, MD; Jordan Donovan[10], BS; Lora Lingrey[10], BS; Garo Stone-DerHagopian[10], BS; Robert T Miller[11], MSc; Andrew E Williams[11,12], PhD; Peter J Leese[13], MSPH; Paul I Kovach[13], MPH; Emily R Pfaff[13], PhD; Mikhail Zemmel[14], MSc; Robert D Pates[14], PhD; Nick Guthe[15], BS; Melissa A Haendel[16], PhD; Christopher G Chute[17], DrPH, MD; Hongfang Liu[1,2], PhD; National COVID Cohort Collaborative[18*]; The RECOVER Initiative[18*]

[1]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States
[2]McWilliams School of Biomedical Informatics, University of Texas Health Sciences Center at Houston, Houston, TX, United States
[3]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, United States
[4]Institute for Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Kentucky, Lexington, KY, United States
[5]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, KY, United States
[6]Division of Health Data Science, University of Minnesota Medical School, Minneapolis, MN, United States
[7]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, United States
[8]Department of Biomedical Informatics, Stony Brook Medicine, Stony Brook, NY, United States
[9]Division of Nephrology, Stony Brook Medicine, Stony Brook, NY, United States
[10]TriNetX LLC, Cambridge, MA, United States
[11]Clinical and Translational Science Institute, Tufts Medical Center, Boston, MA, United States
[12]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, United States
[13]North Carolina Translational and Clinical Sciences Institute, University of North Carolina School of Medicine, Chapel Hill, NC, United States
[14]University of Virginia, Charlottesville, VA, United States
[15]Department of Population Health, New York University Grossman School of Medicine, New York, NY, United States
[16]University of Colorado Anschutz Medical Campus, Denver, CO, United States
[17]Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, MD, United States
[18]see Acknowledgments
[*]these authors contributed equally

Corresponding Author:
Hongfang Liu, PhD
McWilliams School of Biomedical Informatics
University of Texas Health Sciences Center at Houston
7000 Fannin Dr Ste 600
Houston, TX, 77030
United States
Phone: 1 713 500 3924
Email: hongfang.liu@uth.tmc.edu

Abstract

Background: A wealth of clinically relevant information is only obtainable within unstructured clinical narratives, leading to great interest in clinical natural language processing (NLP). While a multitude of approaches to NLP exist, current algorithm development approaches have limitations that can slow the development process. These limitations are exacerbated when the task is emergent, as is the case currently for NLP extraction of signs and symptoms of COVID-19 and postacute sequelae of SARS-CoV-2 infection (PASC).

Objective: This study aims to highlight the current limitations of existing NLP algorithm development approaches that are exacerbated by NLP tasks surrounding emergent clinical concepts and to illustrate our approach to addressing these issues through the use case of developing an NLP system for the signs and symptoms of COVID-19 and PASC.

Methods: We used 2 preexisting studies on PASC as a baseline to determine a set of concepts that should be extracted by NLP. This concept list was then used in conjunction with the Unified Medical Language System to autonomously generate an expanded lexicon to weakly annotate a training set, which was then reviewed by a human expert to generate a fine-tuned NLP algorithm. The annotations from a fully human-annotated test set were then compared with NLP results from the fine-tuned algorithm. The NLP algorithm was then deployed to 10 additional sites that were also running our NLP infrastructure. Of these 10 sites, 5 were used to conduct a federated evaluation of the NLP algorithm.

Results: An NLP algorithm consisting of 12,234 unique normalized text strings corresponding to 2366 unique concepts was developed to extract COVID-19 or PASC signs and symptoms. An unweighted mean dictionary coverage of 77.8% was found for the 5 sites.

Conclusions: The evolutionary and time-critical nature of the PASC NLP task significantly complicates existing approaches to NLP algorithm development. In this work, we present a hybrid approach using the Open Health Natural Language Processing Toolkit aimed at addressing these needs with a dictionary-based weak labeling step that minimizes the need for additional expert annotation while still preserving the fine-tuning capabilities of expert involvement.

(JMIR Med Inform 2024;12:e49997) doi:10.2196/49997

KEYWORDS

natural language processing; clinical information extraction; clinical phenotyping; extract; extraction; NLP; phenotype; phenotyping; narratives; unstructured; PASC; COVID; COVID-19; SARS-CoV-2; OHNLP; Open Health Natural Language Processing

Introduction

The advent of the electronic health record (EHR) and the wealth of longitudinal clinical data contained therein have afforded tremendous opportunities for both clinical research and digital health applications. Fundamental to the feasibility of both of these applications is the availability of clinical information, which, despite the plethora of raw data available in the EHR, can be nontrivial to extract in a computationally accessible format. This is particularly the case for information only accessible within unstructured data, such as clinical narratives, due to the intrinsic nature of human language. The same information can be expressed in many different ways, making the task of algorithmic extraction and standardization for computational semantic interpretation very challenging. Concurrently, however, as much as 80% of clinically relevant information has been found to only be accessible in unstructured form [1]. The need for computationally accessible information extracted from unstructured data has been particularly highlighted with recent research efforts surrounding the ongoing COVID-19 pandemic, particularly with respect to its postacute sequelae (PASC) [2-5]. PASC is defined as ongoing, relapsing, or new symptoms or other health effects occurring after the acute phase of the SARS-CoV-2 infection (ie, present 4 or more weeks after the acute infection). A substantial portion of the information of interest relevant to PASC, for instance, signs and symptoms, is often recorded only within narrative text and is not otherwise found within structured EHR data [6].

One proposed solution to computational extraction of the information within unstructured text is natural language processing (NLP). While a multitude of approaches to clinical NLP currently exist, several existing limitations in these approaches that slow down the development process are magnified by the ongoing and evolving nature of the PASC task. In previous work, we introduced the Open Health Natural Language Processing (OHNLP) Toolkit (OHNLPTK), an NLP framework aiming to provide NLP capabilities at scale in a standards-compliant and consensus-driven manner. In this work, we will highlight current limitations in NLP algorithm development approaches and illustrate our approach to addressing these issues by using PASC as an NLP algorithm development use case for the OHNLPTK.

NLP-Based Clinical Information Extraction

Fundamentally, many of the current applications for clinical NLP lie in information extraction [7]: specifically, the identification of the presence of certain clinical concepts within a clinical narrative, determination of whether it applies to the patient to which the clinical document in question pertains (eg, identification of positive or negative, subject, and other clinically relevant contextual information), and normalization such that named entities sharing the same semantic meaning but with differing lexical forms are mapped to a consistent, codified, computationally accessible definition.

In the following subsections, we will briefly discuss existing approaches to each of these tasks as well as several resources

that can be used to augment each of these tasks as relating to PASC.

Clinical Named Entity Recognition

Identification of clinical concept mentions within unstructured text is a named entity recognition (NER) task. Broadly, approaches to this problem can be subdivided into 3 subcategories: symbolic [8-12], statistical (including deep neural) [13-17], and some hybrid of the 2 [18-21], alluding to the approach used to identify the boundaries of the entities within the text itself. Symbolic approaches are typically either expert-driven, where symbolic rule sets are handcrafted by clinical domain experts, or dictionary-based, where various clinical ontologies are mined for lexical variants for matching purposes. Statistical systems bypass such knowledge engineering efforts by training a machine learning system to label concept mentions given a collection of annotated text documents with concept mentions manually annotated by domain experts.

Each of these approaches has its own benefits. Due to their nature of being handcrafted by domain experts, expert-driven approaches can achieve extremely high performance and can be easily fine-tuned to meet application-specific needs and correct any observed errors. Conversely, expert-driven systems tend to be limited in scope to specific concepts due to their need for expert knowledge engineering, which can be expensive both temporally and financially. While this is sufficient for many applications, such an approach is only suitable if sufficient resources are present for the domain experts and the set of concepts that are needed is known. Dictionary-based symbolic systems aim to address this issue. The solution to the domain expertise problem has been through the use of general clinical ontologies and similar vocabulary resources, such as the National Library of Medicine's Unified Medical Language System (UMLS), either as a basis to construct general dictionaries that cover concepts from a much greater breadth of the clinical domain, although without the manual curation that is afforded to expert-driven systems, or to derive a larger set of lexical variants for a specific set of concepts without the need to engage domain expertise. While these systems tend to not perform as well as expertly curated rule-based systems, generally high performance has been shown to be achievable. For statistical systems, the creation of a high-quality annotated corpus is also an expensive and laborious process. The situation becomes more complicated in cases of multisite collaboration as clinical narratives may contain patient identifier information, making data sharing challenging, and at the same time, the local site may not have the necessary resources for creating annotated corpora. Additionally, they are difficult to fine-tune as there is very little control short of additional annotation and training data manipulation to correct any errors. Models used for statistical approaches include conditional random fields [14,21-23], hidden Markov models [24-26], Bidirectional Encoder Representations from Transformers (BERT) (after finetuning specifically to accomplish the NER task) and BERT-like models (which are particularly prevalent in, but not exclusive to, multilingual use cases) [13,15,27-30], and other neural methods such as recurrent neural networks and convolutional neural networks [17,31-34].

Contextual Feature Detection for Clinical Named Entities

Unlike in the general domain, certain contextual features pose a great impact on the relevance and meaning of extracted concepts in the clinical domain. Of particular note is a concept's assertion (asserted vs possible vs hypothetical), negation, temporality, and whether or not it relates to the patient (as opposed to, eg, a family member), as all of these drastically change the relevance of the concept for downstream applications.

Much like for NER, both symbolic and statistical approaches exist for context detection, with many of the same benefits and drawbacks for each. Among the symbolic systems [35-38], the ConText algorithm proposed by Chapman et al [39] is a symbolic system that is widely adopted among clinical NLP implementations. Various statistical approaches have also been proposed, ranging from traditional statistical machine learning methods such as linear kernel support vector machines [40-42] to deep neural methods [43-45]. There is no clear evidence that statistical approaches outperform the widely adopted ConText algorithm [42].

Concept Normalization Approaches and Available Resources

Identification of named entities and whether they apply to the patient is only part of the problem for clinical information extraction tasks: to be computationally accessible, these named entities must first be mapped to some known coding system such that named entities with differing lexical variations but with the same semantic meaning are grouped in some computationally accessible manner; that is, a computationally accessible thesaurus for the extracted named entities must be constructed.

There are several approaches to this concept normalization (also often referred to as entity linking) problem. One of the side benefits of symbolic NER methods is that they will often have this normalization built in, whether as part of the construction process by domain experts for expert-based systems or due to the nature of their lexical variants being derived from structured ontologies that themselves often act as pseudothesauri, as is the case for the UMLS for dictionary-based systems [10,11,21].

The same is not always true for NER approaches based on statistical methods: while some systems, particularly those trained to extract a specific set of clinical concepts, do incorporate normalization by the very nature of their training approach, other systems trained to perform general named entity recognition do not incorporate such an element. A secondary step must then be taken to perform such a normalization, oftentimes again leveraging ontologies by doing similarity matches against ontology entries [21]. Despite this, it is worthwhile to note that normalization performance is typically inferior to that of symbolic approaches.

Irrespective of the approach, a common theme in clinical NLP is that the extracted named entities are typically mapped to some ontology for later ease of computational access, particularly the UMLS due to its breadth of source vocabularies.

PASC and the Emergent Phenotyping Workflow Problem

It is worthwhile to note that NLP-derived data often serves as supplemental information, in that it is used to supply information that, while needed for a particular use case, cannot be found in structured data. This is especially true for phenotyping and similar cohort identification tasks such as clinical trial recruitment, which often have specific inclusion and exclusion criteria that draw from information elements in both structured and unstructured data.

When the inclusion or exclusion criteria and features of interest pertain to emergent entities of interest, however, as was the case with COVID-19, existing approaches to constructing NLP algorithms break down. Statistical methods for NER require data to train, which may not yet exist in sufficiently large volumes due to the fact that the entities of interest themselves are emergent. Additionally, it is worthwhile to note that in certain circumstances surrounding emergent clinical entities, drastic changes to clinical workflow and, by extension, the contents of the documentation itself can occur, as was the case with COVID-19.

Symbolic methods, however, may not necessarily fare better. Ontology-derived dictionary-based approaches can fail in these circumstances due to the fact that ontologies and similar resources used may not yet be updated to contain these emergent entities (or may not yet have new or updated names to refer to existing entities as the terminology used changes over time), and their slow update frequency (biyearly, in the case of the UMLS) results in them being unsuitable for dealing with emergent needs. Expert-driven systems, on the other hand, fare better due to their relative ease of fine-tuning, but the limitations faced by expert-based NER still apply, rendering a purely expert-driven solution infeasible for many use cases.

It is important to note that these problems do not only occur with the introduction of emergent diseases, as was the case with the COVID-19 and PASC studies; rather, attempts to construct NLP systems to address these use cases magnify existing limitations that typically only slow down the development process.

Irrespective of whether these limitations merely slow down or completely hinder NLP development for a particular use case, such limitations are undesirable given the increasing demand for NLP to fulfill a variety of information needs. It thus becomes evident that an approach capable of prototyping and developing NLP systems in a more rapid manner is needed that can combine the fine-grained control and rapid prototyping ability of expert-driven systems with the general applicability afforded by dictionary systems. It is this need that motivates the work presented here.

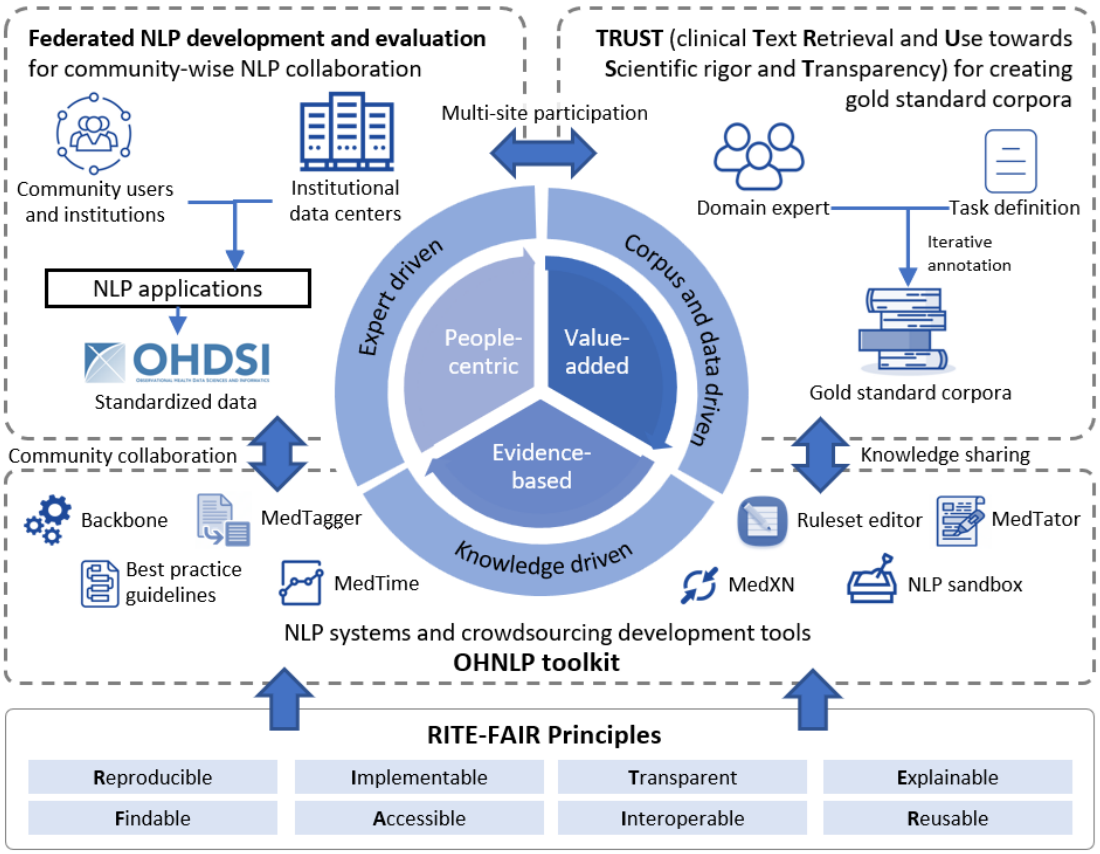
Usage of NLP for PASC-Related Tasks

With regard to NLP usage in the context of PASC, rather than a focus on NLP algorithm development, NLP has primarily been used indirectly for other tasks within the PASC context. For instance, Bhavnani et al [46] extracted 20 signs or symptoms defined by the CDC as being PASC-related from clinical narratives to identify symptom-based phenotypes for patients with PASC, while Zhu et al [47] fine-tuned various BERT-based models to classify documents pertaining to patients with PASC signs or symptoms. More applicable to this work is work primarily focused on identifying what specific NLP-derived signs or symptoms are appropriate for inclusion in a PASC signs or symptoms extraction task. For instance, Wang et al [48] used MTerms [49] to mine existing UMLS concepts from clinical narratives associated with COVID-19 positivity to build a lexicon of 355 long COVID-19 symptoms. We use their developed lexicon as one of the bases for further development in this study.

Clinical NLP to Empower Clinical Research and Translation

We developed the OHNLPTK as part of previous work to enable the rapid development and dissemination of NLP algorithms for empowering clinical research and translation [50]. We follow the RITE-FAIR (reproducible, implementable, transparent, explainable-findable, accessible, interoperable, and reusable) principles [51] to ensure scientific rigor and fairness for resources developed, demonstrated, and disseminated for clinical NLP for health. We have released the OHNLPTK to encourage collaboration across the clinical NLP community to address real-world data problems. The toolkit consists of the following components: (1) a federated NLP deployment framework for privacy-preserving clinical NLP enabled by clinical common data models, (2) a clinical text retrieval and use process toward scientific rigor and transparent (TRUST) process, and (3) an open science collaboration toward real-world clinical NLP (Figure 1).

Figure 1. An overview of the Open Health Natural Language Processing ecosystem. FAIR: findable, accessible, interoperable, and explainable; NLP: natural language processing; OHNLP: Open Health Natural Language Processing; RITE: reproducible, implementable, transparent, and explainable.

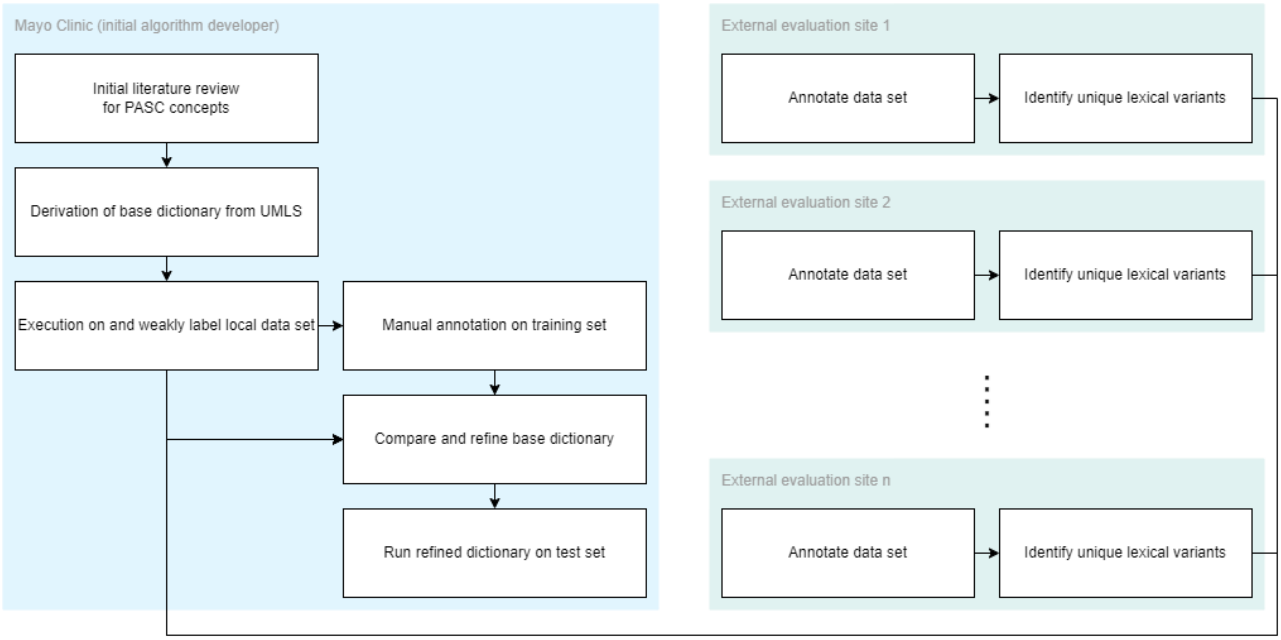


Methods

Overview

In this study, we report on the use of the OHNLPTK to rapidly develop and prototype an NLP system using a case study on PASC signs and symptoms based on PASC resources available in 2 studies. In the ensuing subsections, we will detail each of the steps associated with the creation and evaluation of the final NLP algorithm. For an overview of the entire process, please refer to Figure 2.

Figure 2. An overview of dictionary construction and federated refinement workflows. PASC: postacute sequelae of SARS CoV-2 infection; UMLS: Unified Medical Language System.



PASC Resources

As PASC has been recognized nationally, some useful language resources have been developed. Deer et al [52] identified 303 articles published before April 29, 2021, curated 59 relevant manuscripts that described clinical manifestations in 81 cohorts of individuals 3 weeks or more following acute COVID-19, and mapped 287 unique clinical findings to Human Phenotype Ontology (HPO) terms. Wang et al [48] identified 355 long COVID-19 symptoms corresponding to 1520 UMLS concepts, which in turn resulted in 16,466 synonyms identified from 328,879 clinical notes of 26,117 COVID-19 positive patients from their postacute infection period (days 51-110 from the first positive COVID-19 test).

Dictionary Construction

An initial phenotype definition consisting of textual descriptions, *International Classification of Diseases*, and HPO codes corresponding to clinical entities of interest to PASC was obtained from the aforementioned PASC resources. These entities, where possible, were cross-referenced against the UMLS 2021AB version to obtain an algorithmically derived base dictionary for further refinement. Entities that could not be cross-referenced for enhancement were noted for manual annotation in our second step.

Dictionary Enhancement and Rule Creation

We collected 128 clinical notes from the post–COVID-19 care clinic at Mayo Clinic and split them into a training set (98 notes) and a testing set (30 notes). For the training set, the algorithmically derived base dictionary was run on the document set to generate a weakly labeled data set that was then loaded into the MedTator web annotation tool [53] for manual review and dictionary refinement. Deficiencies (ie, mentions and concepts that were not identified or contained within the algorithmically defined base dictionary) were identified by domain expert review on the training set. Identified missed lexical variants were then manually linked to an appropriate concept ID and added to the base dictionary. Additionally, any concepts missed during base dictionary generation (ie, a clinical concept that was identified as PASC-relevant in the training set but was not part of those concepts identified in the 2 PASC resources used as a baseline) were manually added. To capture complex contextual information, additional rules were created. For example, for "learning difficulty," a rule was created to combine a set of terms meaning difficulty with learning, that is, "(%reDIFFICULTY) to (learn, retain, and gain) new information." The testing set was manually annotated by a single human annotator blindly using the NLP system.

Concept Normalization

Due to the fact that the base dictionary was derived from a structured metathesaurus, the UMLS concept normalizations to UMLS concept unique identifiers (CUIs) were already present. To be compatible with standardization and cross-compatibility, we opted to further map the UMLS CUIs to associated Observational Health Data Sciences and Informatics (OHDSI) Athena concept IDs [54], which are used as part of the OHDSI Observational Medical Outcomes Partnership (OMOP) common data model. This was done algorithmically using text string matching for concept names, followed by manual linking when an exact text match could not be found. For PASC terms that could not be mapped, we manually reviewed and aligned them with the closest Athena concept identifiers, if available, or mapped them to the HPO identifiers. Concept IDs that could not be mapped to either an Athena concept ID or an HPO identifier (due to emergent entities not being yet present in the source vocabularies) were encoded with custom IDs, and these IDs were recorded for later updates once associated ontologies were updated.

Lift Analysis and Local Evaluation

Dictionary coverage was compared before and after refinement to evaluate the effect of our weakly supervised refinement approach, specifically with respect to the number of relevant mentions missed by an ontology-based dictionary-only approach.

Beyond being used in the multisite evaluation phase described later in this study, Mayo Clinic's 30-note testing set was also separately compared against the results from the refined NLP algorithm for in-depth evaluation of NLP and manual annotation mismatches.

Multisite Evaluation

The resulting NLP algorithm was distributed to 5 sites (University of Minnesota, University of Kentucky, University of Michigan, Stony Brook Medicine, and Columbia University) for evaluation to evaluate the federated evaluation component of the OHNLPTK. Each site manually annotated 10 notes for long COVID-19 signs or symptoms using 2 annotators with an independent adjudicator for disagreements. The resulting text annotations (medical concepts that do not contain PHI) were then returned to the Mayo Clinic for dictionary coverage analysis through comparison against the NLP algorithm.

Ethical Considerations

Human subject ethics review and informed consent were handled through the individual institution review boards of participating institutions, which gave final approval for this research on the participating patient cohort. The N3C data transfer to National Center for Advancing Translational Sciences is performed under a Johns Hopkins University Reliance Protocol (IRB00249128) or individual site agreements with the National Institutes of Health. All data transmitted to the coordinating site was either manually deidentified or did not contain PII by definition before transmission. No specific compensation was provided to participants as part of this research.

Results

Dictionary Construction, Lift Analysis, and Local Evaluation

The final dictionary created from previous PASC resources consists of 12,145 unique text strings, mapped to 2343 HPO or OMOP concept identifiers. Within the Mayo Clinic training or development set, the baseline system detected 8090 PASC concept mentions. After manual verification, we identified that 338 PASC concept mentions were missed, rendering the total number of annotated mentions within the training set to be 8428. The final refined dictionary includes 12,234 unique text strings,

mapped to 2366 HPO or OMOP concept identifiers (ie, 23 additional concepts and 89 additional text strings were added). A total of 4 PASC concepts present in the UMLS were captured that could be mapped to neither the OMOP vocabulary nor the HPO (eg, teeth chatter and unrefreshed sleep).

For the local evaluation portion of this study, only the PASC sign and symptom sections of the 30 Mayo Clinic test set notes were compared. In the following report of the results, to facilitate reader understanding, we will use the traditional true positive, false positive, or false negative terminology common to NLP evaluations, despite them not being fully applicable due to varying definitions of what is a true positive depending on what the resulting NLP artifacts are being used for or how PASC-related should be defined for the use case (eg, is it PASC-related if the patient had severe COVID-19 in the past or only if it is explicitly written out as "...consequent to previous COVID-19 infection?"). For more details on this, please refer to the "Discussion" section "On Gaps in the NLP Clinical Information Extraction Subtask." A total of 1560 annotations were produced by the NLP algorithm, while manual annotation produced 1067 annotations. Of these, 1061 (236 unique text strings ignoring capitalization) would be considered true positives in a traditional NLP evaluation, 489 (445 unique text strings ignoring capitalization) false positives, and 6 (4 unique text strings ignoring capitalization) false negatives. It is important, however, to note that due to certain features of this task, a traditional NLP evaluation is not necessarily fully accurate. For instance, among a substantial portion of the "false positives," the NLP algorithm made accurate extraction of a sign or symptom that is PASC-related, but human annotation did not occur as the sign or symptom is preexisting and not resulting due to acute COVID-19 infection. For example, the patient has a previous medical history of some of those signs or symptoms, like a headache or migraine, before COVID-19. Similarly, false negatives may not necessarily be attributed to issues with the NLP algorithm. For instance, of the 4 unique text strings composing the false negatives (sexsomnia, taste or smell changes, burning mouth, and sensitivities to noise and light), one (sexsomnia) was not recorded as a PASC-related concept in either of the source articles from which the NLP concepts to extract were defined, but was suggested to be PASC-related within the textual narrative itself and was therefore annotated as such by the human annotator. The other 3 were either caused by a span mismatch or were lexical variants that were neither in the ontologies used to generate the baseline dictionary nor in the training set for manual expert refinement.

Multisite Evaluation

In Table 1, we present the results from manual annotation after adjudication as well as the dictionary coverage for these annotations.

Table 1. Dictionary coverage statistics.

Results	Site 1	Site 2	Site 3	Site 4	Site 5
Number of annotations, n	126	23	171	118	73
Number in dictionary, n	77	20	138	84	65
Coverage ratio, n/n (%)	77/126 (61.1)	20/23 (87)	138/171 (80.7)	84/118 (71.2)	65/73 (89)

In Table 2, we present an analysis of the annotations not covered by the NLP algorithm. Here, we define a new concept as an annotation that was not included as a concept in the original dictionary or rule set, a new variant as an annotation that is an additional lexical variant of a concept existing in the original dictionary or rule set, and an annotation error as an annotation that falls outside of our task definition: for example, COVID-19 or long COVID-19 is not a sign or symptom of COVID-19 or PASC. For a detailed listing of missed terms, please refer to Multimedia Appendix 1.

Table 2. Statistics of annotations not covered by dictionary.

Results	Site 1	Site 2	Site 3	Site 4	Site 5
Number of annotations not covered by natural language processing, n	49	3	33	34	8
New concept, n/n (%)	23/49 (46.9)	2/3 (66.7)	9/33 (27.3)	4/34 (11.8)	1/8 (12.5)
New variant, n/n (%)	26/49 (53.1)	1/3 (33.3)	17/33 (51.5)	26/34 (76.5)	4/8 (50)
Annotation error, n/n (%)	0 (0)	0 (0)	7/33 (21.2)	4/34 (11.8)	3/8 (37.5)

Discussion

Overview

The COVID-19 pandemic and the associated PASC problem highlighted the importance of having a framework for NLP development that is sufficiently flexible to both provide general concept detection capabilities without requiring extensive domain expertise to craft rule sets or annotate extensive data sets, but also with sufficient flexibility for fine-tuning and addition of concepts that are not present in the base ontologies from which the dictionaries are derived. Additionally, it has highlighted the need for such an NLP development process to be agile in iteration, as the rapidly changing concepts and definitions associated with these emergent concepts are inherently incompatible with the relatively slow process associated with traditional NLP or information extraction (IE) algorithm development (as by the time the algorithm is developed, the definition will have changed). Here, we have presented an approach to NLP algorithm development that

allows us to achieve reasonable results on limited data sets with a rapid turnaround time (the entire process was accomplished over a month) by leveraging federated algorithm development and common infrastructure. The resulting NLP algorithm has been published as part of the open-sourced MedTagger NLP framework [9] and is being executed at 10 academic medical centers as part of their NLP data submissions to the National COVID-19 Cohort Collaborative (N3C) data set [55]. In this section, we will first discuss several gaps in the current NLP development process that were suspected and the degree to which the OHNLPTK was able to address them, and, finally, we will discuss several of this study's limitations and key takeaways.

On Gaps in the NLP Development Process

Our execution of the PASC use case using the OHNLPTK confirmed many of the suspected gaps in the current NLP ecosystem pertaining to emergent diseases.

First, the potential coverage gap of a purely ontology-derived dictionary-based approach to NLP is highlighted: of the 8428 PASC-related concept mentions within our corpus, the baseline ontology-derived dictionary only identified 8090 mentions, missing 338. These 338 mentions span 167 concepts, highlighting the need for secondary expert-based refinement.

Additionally, this case study highlighted the infeasibility of a purely expert-based approach: there were 1018 unique lexical variants covering PASC-related clinical concepts within our training corpus after dictionary-based weak labeling and expert enhancement. Based on this lexicon size, an expert-based approach to implementing an acceptable NLP system would either be prohibitively expensive resource-wise (by horizontally scaling through adding additional experts) or unreasonably time-consuming, to the point of rendering the approach completely infeasible depending on the study's time constraints.

While it is difficult to draw conclusions about the viability of statistical approaches based solely on what was done in this case study, it is worthwhile to note that of the 8428 PASC mentions in our training corpus, the vast majority were weakly labeled through dictionary lookups, while the remainder were manually derived through an expert review. It is very likely that the manual effort required to fully annotate the train data set for a fully supervised approach would have been prohibitively expensive for many studies, given these statistics. Additionally, as these 8428 PASC mentions are split in turn across many concepts, there are very few unique mentions or lexical examples per concept. By extension, there is an insufficient number of examples to support training high-performing models for all concepts, and annotation of further training data would be required. In fact, this data problem is further exacerbated.

It is thus evident that a hybrid approach that integrates each of the three approaches of statistical, symbolic expert-driven, and symbolic dictionary-based would be ideal. Based on the results of the execution of our PASC use case, we believe that we have demonstrated that expert refinement of a rule set on top of a data set already weakly labeled by an ontology-derived dictionary is one such viable hybrid approach.

With our use case demonstration, we also show that the OHNLPTK can be leveraged to execute an NLP development process following this approach. An initial dictionary derived from the UMLS can be directly loaded into deployed instances of the OHNLPTK without any additional software modification. Weakly labeled NLP output can be directly piped to OHNLPTK's MedTator component for expert review, and additional finetuning is autonomously translated into refinement rules that can, in turn, be executed on top of the base dictionary.

On Gaps in the NLP Clinical Information Extraction Subtask

In this case study, we have strictly focused on clinical information extraction, which has thus far been the core focus for a significant portion of clinical NLP applications. It is important, however, to note that even with a clinical information extraction algorithm, the output may not be wholly applicable and/or useful to the use case. There were multiple instances where our NLP system correctly extracted mentions of some PASC-related terms that were marked as false negatives upon comparison against the gold standard.

Upon further investigation, we discovered that the NLP system correctly identified mentions of the entities in question, but they were not annotated by the annotator as they were not specific to PASC. For instance, while headaches are a valid symptom that is associated with PASC, they may also occur independently due to unrelated reasons. While human annotators have the capability to make this distinction, our NLP system cannot currently make such a distinction and simply naively extracts all valid mentions. Strictly speaking, from the perspective of the NLP-based named entity recognition and linkage subtask (ie, "identify mentions of headaches, and annotate them as such"), the algorithm output is correct. Conversely, however, for real-world use cases such as the PASC investigation use case presented here, not making such a distinction has a significant impact on the practical usability of the NLP artifacts produced. While such contextual differentiation would fall under a different subtask and is not strictly information extraction, the ability to perform this differentiation is crucial for many use cases and is an existing gap in current clinical NLP offerings. We aim to further explore approaches to this differentiation task as part of our future work.

On Federated Evaluation and Associated Benefits

Before federated evaluation can be done, a common working definition or annotation guideline must be defined. The addition of other data and sites, however, also introduces its own complexity, specifically with respect to the aforementioned issue of how to define a concept mentioned as being "PASC-related," on which sites had a very broad spectrum of definitions. We concluded as a group that differentiating whether a mention of, for example, a headache is PASC-related ought to be considered a separate task from the clinical information extraction task and that such filtering, if needed, would be done as a separate, post-IE step.

We opted to evaluate the dictionary-based NLP system using dictionary coverage as opposed to a more traditional precision, recall, and F_1-score. This was done because we wished to

evaluate the ability of a generic NLP algorithm to meet the needs of multiple institutions in a cost-effective manner. Due to the recency and evolutionary nature of long COVID-19, it follows that each site's definition (or at least those of the participating annotators) of "PASC or long COVID-19" may differ, thus rendering it difficult to construct a gold standard with a consistent set of concepts. Instead, by allowing individual sites to determine what constitutes a long COVID-19 for their needs, we can evaluate the coverage—that is, to what extent said need is met by the NLP algorithm.

While the initial Mayo-developed NLP system performs reasonably well across the participating sites that returned evaluation results, not all information needs are met in terms of dictionary coverage. We note that site 1 only had 61.1% (77/126) dictionary coverage; of those annotations not covered, 46.9% (23/49) were new concepts. Manual review indicated some of these new concepts were lung-related (pneumothorax, hydropneumothorax, and ground glass attenuation), which suggests site 1's PASC-related symptoms are slightly different than other sites. A total of 9 of these new concepts (9/23, 39.1%) were the result of a very detailed note (32,000 characters in length) of a complex PASC case where other poor-health and pain-related concepts (gallstones and cholelithiasis) were intermixed with chest pain, lung pain, and difficulty breathing. For all sites, more than a third of the missing annotations were categorized as missing lexical variations of included concepts that did not appear within either the autonomously generated ontology-sourced dictionary or the Mayo documentation set used for dictionary refinement.

Such results illustrate the importance of multisite NLP development, particularly in the latter case, which is one of the pitfalls that would be commonly found in traditional NLP tasks. Our results therefore fundamentally demonstrate the advantage of the high degree of portability resulting from a common NLP infrastructure when discussing generalizable NLP solutions. As gaps in data coverage are assessed across a variety of health care institutions, NLP algorithm refinement is vastly simplified, thus granting greater confidence in the wide-range generalizability of the final solution. Similarly, the addition of these additional sites with their own respective annotators also helps mitigate potential bias associated with single-site or single-annotator definitions of the PASC sign or symptom extraction task.

Beyond the issue of what clinical concepts to include is another issue less addressed in this study: a definition of what sorts of data to include. This is a challenge exacerbated by the fact that, much like how emergent diseases will constantly have an evolving associated concept set, the way they are documented (and their extent) is also constantly changing and subject to high levels of disagreement. This phenomenon can be seen in the wildly different annotated concept counts within the evaluation sites despite sharing the same concept set to annotate: certain sites only considered a specific section of a clinical note explicitly dedicated to documented PASC complications to be appropriate as input data, while others used the entire clinical note for any patient that had visited their long COVID-19 clinic. This wide variance in data inclusion definition further supports the need for federated development and evaluation efforts as outlined in this study to further expose the developed algorithm to this wide variety of data types. Furthermore, the fact that the condition is emergent inherently limits data set sizes; at the time this study was conducted, long COVID-19 clinics were still a relatively novel concept in their initial stages of implementation. A key benefit of such a federated evaluation and iterative refinement process would be to help mitigate the limited amount of data available inherently associated with an emergent condition by spreading out the data set to multiple sources and making it available for the development of a wider variety of documentation types.

Limitations

Several limitations exist in this study. First, we compared dictionary coverage as an evaluation, rather than doing a traditional NLP system evaluation. This is primarily due to the lack of a fully annotated gold standard to evaluate against, driven primarily by the ongoing evolutionary nature of the pandemic, causing associated documentation to continuously change. This renders a truly scientifically rigorous gold standard difficult to construct, as annotation guidelines must be constantly updated and there will be very little consensus due to a lack of clear clinical guidelines. Instead, we note that this limitation highlights the need for iterative development, which further emphasizes the need for an agile NLP development, evaluation, and refinement process such as the one we present here.

Additionally, one of the emphases of the OHNLP consortium is the organization of multi-institutional, federated evaluation for NLP algorithm development, enabled by a common NLP system deployment (the OHNLPTK). As such, we are currently in the process of disseminating the algorithm presented here to multiple member sites, who will all conduct a formal evaluation of this algorithm.

Finally, it was previously noted that in order to codify and normalize concepts that do not yet exist in controlled ontologies, we introduced our own coding scheme for several concepts. It is important for standardization's sake, however, to loop back to the original ontologies used and have these concepts incorporated into these source ontologies. This process is ongoing as of the writing of this article.

Conclusion

The PASC NLP problem has highlighted many of the limitations present with current NLP development approaches. The evolutionary and time-critical nature of the PASC NLP task exacerbates many of these limitations, which previously only presented a slowdown of the development process, into limitations that cause many approaches to be outright infeasible. The need for agile and iterative NLP development is thus made evident. Fundamentally, this can be observed as an amalgamation of wanting the benefits of expert-driven systems while minimizing the time and resource expenditure of expert involvement. Here we have presented a hybrid approach that we believe presents such benefits, with a dictionary-based weak labeling step minimizing the need for additional expert annotation while still preserving the fine-tuning capabilities of expert involvement.

Acknowledgments

The research reported in this publication was supported by the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (U01TR002062: AW, LW, HH, SF, SL, HL, KN, and NPP; UL1TR002240: DAH; UL1TR001998: DRH and RK; UL1TR002489: PJL, PIK, and ERP; UL1TR003015: MZ; and UL1TR003098: CGC). This effort was also supported through the National COVID-19 Cohort Collaborative (N3C) and is part of the NIH (National Institutes of Health) Researching COVID to Enhance Recovery (RECOVER) initiative, which seeks to understand, treat, and prevent the postacute sequelae of SARS-CoV-2 infection (PASC). For more information on N3C, refer to [56]. For more information on RECOVER, refer to [57]. N3C efforts were possible because of the patients whose information is included within the data from participating organizations [58] and the organizations and scientists [59] who have contributed to the ongoing development of this community resource [60]. Similarly, we would like to thank the National Community Engagement Group (NCEG), all patients, caregivers, and community representatives, and all the participants enrolled in the RECOVER Initiative. In particular, we gratefully acknowledge Teresa Akintonwa for serving as a patient representative for this manuscript. N3C and RECOVER efforts were funded by the NIH under NCATS (U24TR002306) and the National Heart, Lung, and Blood Institute (OT2HL161847: RZ, KN, NPP, JH, SR, VL, RAM, MS, CE, FMK, MAH, CGC, AEW, and RAM). Portions of the data generated from analyses described in this publication are made accessible through the NCATS N3C data enclave at [61] under the N3C attribution and publication policy (v1.2-2020-08-25b). Provisioning of such data is supported by NCATS (U24TR002306). The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol (IRB00249128) or individual site agreements with NIH. The N3C data enclave is managed under the authority of the NIH; information can be found at [62].

We gratefully acknowledge the following core contributors to N3C: Adam B Wilcox, Adam M Lee, Alexis Graves, Alfred (Jerrod) Anzalone, Amin Manna, Amit Saha, Amy Olex, Andrea Zhou, AEW, Andrew Southerland, Andrew T Girvin, Anita Walden, Anjali A Sharathkumar, Benjamin Amor, Benjamin Bates, Brian Hendricks, Brijesh Patel, Caleb Alexander, Carolyn Bramante, Cavin Ward-Caviness, Charisse Madlock-Brown, Christine Suver, CGC, Christopher Dillon, Chunlei Wu, Clare Schmitt, Cliff Takemoto, Dan Housman, Davera Gabriel, David A Eichmann, Diego Mazzotti, Don Brown, Eilis Boudreau, Elaine Hill, Elizabeth Zampino, Emily Carlson Marti, ERP, Evan French, FMK, Federico Mariona, Fred Prior, George Sokos, Greg Martin, Harold Lehmann, Heidi Spratt, Hemalkumar Mehta, HL, Hythem Sidky, JW Awori Hayanga, Jami Pincavitch, Jaylyn Clark, Jeremy Richard Harper, Jessica Islam, Jin Ge, Joel Gagnier, Joel H Saltz, Joel Saltz, Johanna Loomba, John Buse, Jomol Mathew, Joni L Rutter, Julie A McMurry, Justin Guinney, Justin Starren, Karen Crowley, Katie Rebecca Bradwell, Kellie M Walters, Ken Wilkins, Kenneth R. Gersing, Kenrick Dwain Cato, Kimberly Murray, Kristin Kostka, Lavance Northington, Lee Allan Pyles, Leonie Misquitta, Lesley Cottrell, Lili Portilla, Mariam Deacy, Mark M. Bissell, Marshall Clark, Mary Emmett, MS, MBP, MAH, Meredith Adams, Meredith Temple-O'Connor, Michael G Kurilla, Michele Morris, Nabeel Qureshi, Nasia Safdar, Nicole Garbarini, Noha Sharafeldin, Ofer Sadan, Patricia A Francis, Penny Wung Burgoon, Peter Robinson, Philip RO Payne, Rafael Fuentes, Randeep Jawa, Rebecca Erwin-Cohen, Rena Patel, RAM, Richard L Zhu, Rishi Kamaleswaran, Robert Hurley, Robert T Miller, Saiju Pyarajan, Sam G Michael, Samuel Bozzette, Sandeep Mallipattu, Satyanarayana Vedula, Scott Chapman, Shawn T O'Neil, Soko Setoguchi, Stephanie S Hong, Steve Johnson, Tellen D. Bennett, Tiffany Callahan, Umit Topaloglu, Usman Sheikh, Valery Gordon, Vignesh Subbian, Warren A Kibbe, Wenndy Hernandez, Will Beasley, Will Cooper, William Hillegass, and Xiaohan Tanner Zhang. Details of contributions are available at [63]. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, the N3C, or the Researching COVID-19 to Enhance Recovery Initiatives.

Authors' Contributions

Authorship has been determined according to the International Committee of Medical Journal Editors (ICMJE) recommendations. AW, LW, HH, SF, SL, and HL were responsible for study design, initial pipeline or natural language processing (NLP) algorithm development, initial single-site evaluation, extramural (non-Mayo site) Open Health Natural Language Processing Toolkit or NLP algorithm deployment support, and project coordination among extramural sites. DAH, DRH, RK, RZ, KN, NPP, JH, SR, VL, RAM, MS, CE, FMK, MBP, JD, LL, GSDH, RTM, AEW, PJL, PIK, ERP, MZ, RDP, NG, MAH, and CGC were responsible for NLP algorithm deployment, pipeline execution, feedback for iterative pipeline or implementation development, and experimental procedures. DAH, DRH, RK, RZ, KN, NPP, JH, SR, VL, RAM, MS, CE, and FMK were responsible for federated evaluation efforts. All authors reviewed and contributed expertise to the manuscript.

Conflicts of Interest

MBP, JD, LL, and GS-DH are affiliated with TriNetX LLC. All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
A list of terms not covered by natural language processing (NLP) algorithms.
[DOCX File , 28 KB - medinform_v12i1e49997_app1.docx]

References

1. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearb Med Inform 2014;9(1):14-20 [FREE Full text] [doi: 10.15265/IY-2014-0020] [Medline: 25123716]
2. Thaweethai T, Jolley SE, Karlson EW, Levitan EB, Levy B, McComsey GA, et al. Development of a definition of postacute sequelae of SARS-CoV-2 infection. JAMA 2023;329(22):1934-1946 [FREE Full text] [doi: 10.1001/jama.2023.8823] [Medline: 37278994]
3. Crook H, Raza S, Nowell J, Young M, Edison P. Long covid-mechanisms, risk factors, and management. BMJ 2021;374:n1648. [doi: 10.1136/bmj.n1648] [Medline: 34312178]
4. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. Nature 2021;594(7862):259-264 [FREE Full text] [doi: 10.1038/s41586-021-03553-9] [Medline: 33887749]
5. Merad M, Blish CA, Sallusto F, Iwasaki A. The immunology and immunopathology of COVID-19. Science 2022;375(6585):1122-1127 [FREE Full text] [doi: 10.1126/science.abm8108] [Medline: 35271343]
6. Forbush TB, Gundlapalli AV, Palmer MN, Shen S, South BR, Divita G, et al. "Sitting on pins and needles": characterization of symptom descriptions in clinical notes". AMIA Jt Summits Transl Sci Proc 2013;2013:67-71 [FREE Full text] [Medline: 24303238]
7. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
8. Lacson R, Harris K, Brawarsky P, Tosteson TD, Onega T, Tosteson ANA, et al. Evaluation of an automated information extraction tool for imaging data elements to populate a breast cancer screening registry. J Digit Imaging 2015;28(5):567-575 [FREE Full text] [doi: 10.1007/s10278-014-9762-4] [Medline: 25561069]
9. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:149-153 [FREE Full text] [Medline: 24303255]
10. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
11. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc 2014;21(5):858-865 [FREE Full text] [doi: 10.1136/amiajnl-2013-002190] [Medline: 24637954]
12. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17(1):19-24 [FREE Full text] [doi: 10.1197/jamia.M3378] [Medline: 20064797]
13. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. ArXiv. Preprint posted online on April 6, 2019 2019 [FREE Full text] [doi: 10.1090/mbk/121/79]
14. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26(9):1205-1210 [FREE Full text] [doi: 10.1093/bioinformatics/btq126] [Medline: 20335276]
15. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak 2020;20(Suppl 7):242 [FREE Full text] [doi: 10.1186/s12911-020-01241-8] [Medline: 32998724]
16. Roberts K, Rink B, Harabagiu SM, Scheuermann RH, Toomay S, Browning T, et al. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. AMIA Annu Symp Proc 2012;2012:779-788 [FREE Full text] [Medline: 23304352]
17. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical named entity recognition using deep learning models. AMIA Annu Symp Proc 2017;2017:1812-1819 [FREE Full text] [Medline: 29854252]
18. Gupta K, Thammasudjarit R, Thakkinstian A. A hybrid engine for clinical information extraction from radiology reports. : IEEE; 2019 Presented at: 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE); July 10-12, 2019; Chonburi, Thailand p. 293-297. [doi: 10.1109/jcsse.2019.8864178]
19. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A. Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. JMIR Med Inform 2021;9(3):e17934 [FREE Full text] [doi: 10.2196/17934] [Medline: 33724196]
20. Kim Y, Heider PM, Lally IR, Meystre SM. A hybrid model for family history information identification and relation extraction: development and evaluation of an end-to-end information extraction system. JMIR Med Inform 2021;9(4):e22797 [FREE Full text] [doi: 10.2196/22797] [Medline: 33885370]
21. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018;25(3):331-336 [FREE Full text] [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
22. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak 2015;15:37 [FREE Full text] [doi: 10.1186/s12911-015-0160-8] [Medline: 25943550]

23. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A study of neural word embeddings for named entity recognition in clinical text. AMIA Annu Symp Proc 2015;2015:1326-1333 [FREE Full text] [Medline: 26958273]
24. Morwal S. Named entity recognition using Hidden Markov Model (HMM). IJNLC 2012;1(4):15-23 [FREE Full text] [doi: 10.5121/ijnlc.2012.1402]
25. Skounakis M, Craven M, Ray S. Hierarchical hidden Markov models for information extraction. San Francisco, CA: Morgan Kaufmann Publishers Inc; 2003 Presented at: IJCAI'03: Proceedings of the 18th International Joint Conference on Artificial Intelligence; August 9-15, 2003; Acapulco, Mexico p. 427-433.
26. Todorovic BT, Rancic SR, Markovic IM, Mulalic EH, Ilic VM. Named entity recognition and classification using context Hidden Markov Model. : IEEE; 2008 Presented at: 2008 9th Symposium on Neural Network Applications in Electrical Engineering; September 25-27, 2008; Belgrade, Serbia p. 43-46. [doi: 10.1109/neurel.2008.4685557]
27. Dai Z, Wang X, Ni P, Li Y, Li G, Bai X. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. : IEEE; 2019 Presented at: 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI). IEEE; October 19-21, 2019; Suzhou, China p. 1-5. [doi: 10.1109/cisp-bmei48845.2019.8965823]
28. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
29. Mulyar A, Uzuner O, McInnes B. MT-clinical BERT: scaling clinical information extraction with multitask learning. J Am Med Inform Assoc 2021;28(10):2108-2115 [FREE Full text] [doi: 10.1093/jamia/ocab126] [Medline: 34333635]
30. Nath N, Lee SH, McDonnell MD, Lee I. The quest for better clinical word vectors: ontology based and lexical vector augmentation versus clinical contextual embeddings. Comput Biol Med 2021;134:104433. [doi: 10.1016/j.compbiomed.2021.104433] [Medline: 34004575]
31. Akbik A, Bergmann T, Vollgraf R. Pooled contextualized embeddings for named entity recognition. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2-7, 2019; Minneapolis, Minnesota p. 724-728. [doi: 10.18653/v1/n19-1078]
32. Chowdhury S, Dong X, Qian L, Li X, Guan Y, Yang J, et al. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC Bioinformatics 2018;19(Suppl 17):499 [FREE Full text] [doi: 10.1186/s12859-018-2467-9] [Medline: 30591015]
33. Kocaman V, Talby D. Biomedical named entity recognition at scale. In: Del Bimbo A, Cucchiara R, Sclaroff S, Farinella GM, Mei T, Bertini M, et al, editors. Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part I. Cham: Springer; 2021.
34. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]
35. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc 1999;6(5):393-411 [FREE Full text] [doi: 10.1136/jamia.1999.0060393] [Medline: 10495099]
36. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
37. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, et al. DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform 2015;54:213-219 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.010] [Medline: 25791500]
38. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001;8(6):598-609 [FREE Full text] [doi: 10.1136/jamia.2001.0080598] [Medline: 11687566]
39. Chapman WW, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. Stroudsburg, PA, USA: Association for Computational Linguistics; 2007 Presented at: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; June 29, 2007; Prague, Czech Republic p. 81-88. [doi: 10.3115/1572392.1572408]
40. Cruz NP, Taboada M, Mitkov R. A machine-learning approach to negation and speculation detection for sentiment analysis. J Assoc Inf Sci Technol 2015;67(9):2118-2136. [doi: 10.1002/asi.23533]
41. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557-562 [FREE Full text] [doi: 10.1136/amiajnl-2011-000150] [Medline: 21565856]
42. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS One 2014;9(11):e112774 [FREE Full text] [doi: 10.1371/journal.pone.0112774] [Medline: 25393544]

43. Bhatia P, Celikkaya B, Khalilia M. Joint entity extraction and assertion detection for clinical text. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 954-959.
44. Rumeng L, Abhyuday N, Hong Y. A Hybrid Neural Network Model for joint prediction of presence and period assertions of medical events in clinical notes. AMIA Annu Symp Proc 2017;2017:1149-1158 [FREE Full text] [Medline: 29854183]
45. van Aken B, Trajanovska I, Siu A, Mayrdorfer M, Budde K, Loeser A. Assertion detection in clinical notes: medical language models to the rescue? : Association for Computational Linguistics; 2021 Presented at: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations; July 6, 2021; Online p. 35-40. [doi: 10.18653/v1/2021.nlpmc-1.5]
46. Bhavnani SK, Zhang W, Hatch S, Urban R, Tignanelli C. 364 identification of symptom-based phenotypes in PASC patients through bipartite network analysis: implications for patient triage and precision treatment strategies. J Clin Trans Sci 2022;6(s1):68-68 [FREE Full text] [doi: 10.1017/cts.2022.207]
47. Zhu Y, Mahale A, Peters K, Mathew L, Giuste F, Anderson B, et al. Using natural language processing on free-text clinical notes to identify patients with long-term COVID effects. New York, NY, US: Association for Computing Machinery; 2022 Presented at: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; August 7-10, 2022; Northbrook, Illinois p. 1-9. [doi: 10.1145/3535508.3545555]
48. Wang L, Foer D, MacPhaul E, Lo Y, Bates DW, Zhou L. PASCLex: a comprehensive post-acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. J Biomed Inform 2022;125:103951 [FREE Full text] [doi: 10.1016/j.jbi.2021.103951] [Medline: 34785382]
49. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annu Symp Proc 2011;2011:1639-1648 [FREE Full text] [Medline: 22195230]
50. Liu S, Wen A, Wang L, He H, Fu S, Miller R, et al. An open Natural Language Processing (NLP) framework for EHR-based clinical research: a case demonstration using the National COVID Cohort Collaborative (N3C). J Am Med Inform Assoc 2023;30(12):2036-2040 [FREE Full text] [doi: 10.1093/jamia/ocad134] [Medline: 37555837]
51. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
52. Deer RR, Rock MA, Vasilevsky N, Carmody L, Rando H, Anzalone AJ, et al. Characterizing long COVID: deep phenotype of a complex condition. EBioMedicine 2021;74:103722 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103722] [Medline: 34839263]
53. He H, Fu S, Wang L, Liu S, Wen A, Liu H. MedTator: a serverless annotation tool for corpus development. Bioinformatics 2022;38(6):1776-1778 [FREE Full text] [doi: 10.1093/bioinformatics/btab880] [Medline: 34983060]
54. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
55. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc 2021;28(3):427-443 [FREE Full text] [doi: 10.1093/jamia/ocaa196] [Medline: 32805036]
56. National Center for Advancing Translational Sciences, National Institutes of Health. National COVID Cohort Collaborative (N3C). 2020. URL: https://ncats.nih.gov/n3c [accessed 2024-07-14]
57. Researching COVID to Enhance Recovery. 2024. URL: https://recovercovid.org [accessed 2024-07-14]
58. National Center for Advancing Translational Sciences, National Institutes of Health. Data Transfer Agreement Signatories. 2024. URL: https://covid.cd2h.org/dtas [accessed 2024-07-14]
59. National Center for Advancing Translational Sciences, National Institutes of Health. DUA signatories. 2024. URL: https://covid.cd2h.org/duas [accessed 2024-07-14]
60. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, The N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association 2021;28(3):427-433 [FREE Full text]
61. National Center for Advancing Translational Sciences, National Institutes of Health. National COVID Cohort Collaborative. Enclave essentials. 2024. URL: https://covid.cd2h.org/enclave/ [accessed 2024-07-14]
62. National Center for Advancing Translational Sciences, National Institutes of Health. National COVID Cohort Collaborative Forms and Resources. 2024. URL: https://ncats.nih.gov/research/research-activities/n3c/resources [accessed 2024-07-14]
63. National Center for Advancing Translational Sciences, National Institutes of Health. N3C core contributors. 2024. URL: https://covid.cd2h.org/core-contributors [accessed 2024-07-14]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers

CUI: concept unique identifier
EHR: electronic health record
HPO: Human Phenotype Ontology
IE: information extraction
N3C: National COVID-19 Cohort Collaborative
NER: named entity recognition
NLP: natural language processing
OHDSI: Observational Health Data Sciences and Informatics
OHNLP: Open Health Natural Language Processing
OHNLPTK: Open Health Natural Language Processing Toolkit
OMOP: Observational Medical Outcomes Partnership
PASC: postacute sequelae of SARS-CoV-2 infection
RECOVER: Researching COVID to Enhance Recovery
RITE-FAIR: Reproducible, Implementable, Transparent, Explainable - Findable, Accessible, Interoperable, and Reusable
TRUST: text retrieval and use process toward scientific rigor and transparent
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 15.06.23; peer-reviewed by K Allen, G Zhang; comments to author 19.10.23; revised version received 11.12.23; accepted 01.03.24; published 09.09.24.

Please cite as:
Wen A, Wang L, He H, Fu S, Liu S, Hanauer DA, Harris DR, Kavuluru R, Zhang R, Natarajan K, Pavinkurve NP, Hajagos J, Rajupet S, Lingam V, Saltz M, Elowsky C, Moffitt RA, Koraishy FM, Palchuk MB, Donovan J, Lingrey L, Stone-DerHagopian G, Miller RT, Williams AE, Leese PJ, Kovach PI, Pfaff ER, Zemmel M, Pates RD, Guthe N, Haendel MA, Chute CG, Liu H, National COVID Cohort Collaborative, The RECOVER Initiative
A Case Demonstration of the Open Health Natural Language Processing Toolkit From the National COVID-19 Cohort Collaborative and the Researching COVID to Enhance Recovery Programs for a Natural Language Processing System for COVID-19 or Postacute Sequelae of SARS CoV-2 Infection: Algorithm Development and Validation
JMIR Med Inform 2024;12:e49997
URL: https://medinform.jmir.org/2024/1/e49997
doi:10.2196/49997
PMID:

Original Paper

Automated System to Capture Patient Symptoms From Multitype Japanese Clinical Texts: Retrospective Study

Tomohiro Nishiyama[1*], MEng; Ayane Yamaguchi[2*], MD, PhD; Peitao Han[1], BE; Lis Weiji Kanashiro Pereira[3], PhD; Yuka Otsuki[1], BE; Gabriel Herman Bernardim Andrade[1], PhD; Noriko Kudo[1], PhD; Shuntaro Yada[1], PhD; Shoko Wakamiya[1], PhD; Eiji Aramaki[1], PhD; Masahiro Takada[2,4], MD, PhD; Masakazu Toi[5], MD, PhD

[1]Department of Information Science, Nara Institute of Science and Technology, Ikoma, Japan

[2]Graduate School of Medicine, Kyoto University, Kyoto, Japan

[3]Center for Information and Neural Networks, Advanced ICT Research Institute, Osaka, Japan

[4]Department of Breast Surgery, Kansai Medical University, Hirakata, Japan

[5]Tokyo Metropolitan Cancer and Infectious Disease Center, Komagome Hospital, Tokyo, Japan

[*]these authors contributed equally

Corresponding Author:
Eiji Aramaki, PhD
Department of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho
Ikoma, 630-0192
Japan
Phone: 81 743 72 5250
Email: aramaki@is.naist.jp

Abstract

Background: Natural language processing (NLP) techniques can be used to analyze large amounts of electronic health record texts, which encompasses various types of patient information such as quality of life, effectiveness of treatments, and adverse drug event (ADE) signals. As different aspects of a patient's status are stored in different types of documents, we propose an NLP system capable of processing 6 types of documents: physician progress notes, discharge summaries, radiology reports, radioisotope reports, nursing records, and pharmacist progress notes.

Objective: This study aimed to investigate the system's performance in detecting ADEs by evaluating the results from multitype texts. The main objective is to detect adverse events accurately using an NLP system.

Methods: We used data written in Japanese from 2289 patients with breast cancer, including medication data, physician progress notes, discharge summaries, radiology reports, radioisotope reports, nursing records, and pharmacist progress notes. Our system performs 3 processes: named entity recognition, normalization of symptoms, and aggregation of multiple types of documents from multiple patients. Among all patients with breast cancer, 103 and 112 with peripheral neuropathy (PN) received paclitaxel or docetaxel, respectively. We evaluate the utility of using multiple types of documents by correlation coefficient and regression analysis to compare their performance with each single type of document. All evaluations of detection rates with our system are performed 30 days after drug administration.

Results: Our system underestimates by 13.3 percentage points (74.0%−60.7%), as the incidence of paclitaxel-induced PN was 60.7%, compared with 74.0% in the previous research based on manual extraction. The Pearson correlation coefficient between the manual extraction and system results was 0.87 Although the pharmacist progress notes had the highest detection rate among each type of document, the rate did not match the performance using all documents. The estimated median duration of PN with paclitaxel was 92 days, whereas the previously reported median duration of PN with paclitaxel was 727 days. The number of events detected in each document was highest in the physician's progress notes, followed by the pharmacist's and nursing records.

Conclusions: Considering the inherent cost that requires constant monitoring of the patient's condition, such as the treatment of PN, our system has a significant advantage in that it can immediately estimate the treatment duration without fine-tuning a new NLP model. Leveraging multitype documents is better than using single-type documents to improve detection performance. Although the onset time estimation was relatively accurate, the duration might have been influenced by the length of the data follow-up period. The results suggest that our method using various types of data can detect more ADEs from clinical documents.

(JMIR Med Inform 2024;12:e58977) doi:10.2196/58977

KEYWORDS

natural language processing; named entity recognition; adverse drug reaction; adverse event; peripheral neuropathy; NLP; symptoms; symptom; machine learning; ML; drug; drugs; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications; adverse; neuropathy; cancer; oncology; text; texts; textual; note; notes; report; reports; EHR; EHRs; record; records; detect; detection; detecting

Introduction

Processing large amounts of data using artificial intelligence can help to rapidly obtain a comprehensive understanding of the patient status, which can potentially streamline medical studies focusing on patient stratification, drug safety, and adverse drug event (ADE) detection. Particularly, information on ADEs must be collected prospectively, which is expensive and time-consuming. Even when data are collected retrospectively from electronic health records containing information on various modalities, it is challenging to comprehensively survey the medical details of a large number of patients.

Fortunately, natural language processing (NLP) methods can be used to aid such tasks. Recent advances in NLP have enabled the automatic extraction of contextual information from text. Bidirectional Encoder Representations from Transformers (BERT), a transformer-based model released in 2019, has achieved high accuracy in many NLP tasks [1]. Particularly, using diverse medical records for training machine learning models on multiple aspects of patient information can improve their prediction accuracy in the medical domain, leading to the development of specialized models such as ClinicalBERT and BioBERT [2-5].

The ADE detection systems that use such models have been applied to actual texts in existing research [6-10]. Several studies have also used NLP in retrospective observational studies, similar to the approach used in this study [11,12]. McKenzie et al [11] conducted a retrospective analysis of pneumonia using electronic health records; however, they used rule-based NLP methods for 2 types of documents, clinical notes and radiology reports written by physicians, thus leaving room for further investigation into performance.

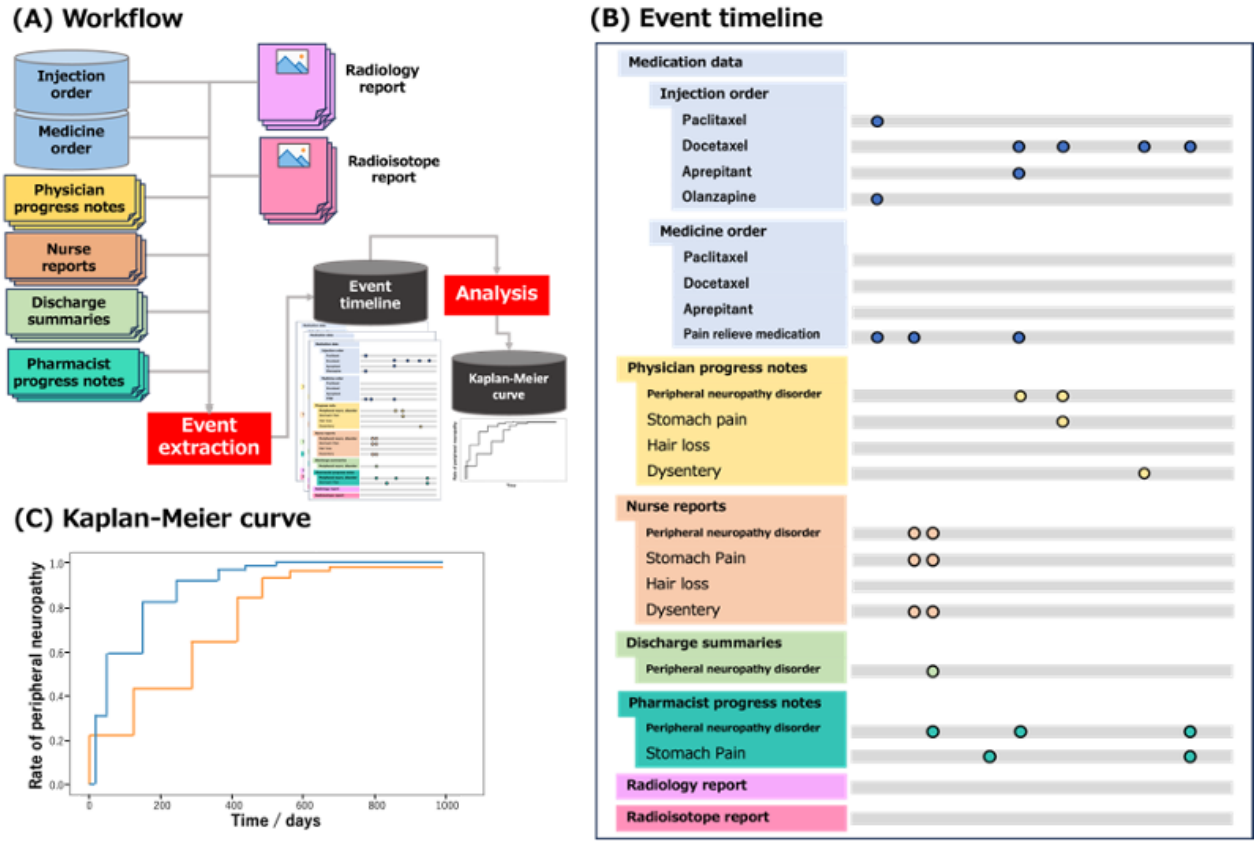
In addition to physicians' records, medical institutions have a wide variety of documents from multiple co-medical personnel, including nursing records, pharmacists' progress notes, and medication orders. Using multiple types of medical documents on retrospective studies requires a comprehensive and robust data analysis because of the expected decrease in missing event detection. While such an analysis is difficult when performed manually due to time requirements and human resource constraints, an NLP-based approach should be more efficient and effective.

A common method for information extraction using NLP is to treat it as a text classification task specific to each document type. However, document-specific text fine-tuning requires that each model be fine-tuned individually for each specific document type, which does not fully demonstrate the strength of automated processing. Fine-tuning a model requires labeled data, and since such data are unlikely to be available beforehand, it requires manual annotation by health care professionals. Even if annotated data are available, privacy concerns and data security restrictions imposed by medical institutions usually make access to them rather difficult. Furthermore, as the transmission of such data over the internet is usually not allowed, the usage of the cloud computational power becomes unfeasible.

For such reasons, fine-tuning models for each individual document type becomes impractical. Therefore, in this study, we used a named entity recognition (NER) model for medical documents, which does not require fine-tuning for each document type. The NER model can be easily used for information extraction without fine-tuning with the target documents since it is already fine-tuned with medical documents to detect symptoms.

In this study, we examined the usefulness of analyzing various medical Japanese documents, including medical records written by physicians and co-medical professionals, to capture the onset and duration of ADEs. Figure 1 shows the basic idea of our approach. Our medical NLP method aims to comprehensively analyze ADE-relevant information contained in medical documents, including nursing records, pharmacist progress notes, and other medical texts, in addition to physicians' records. Our method identified more ADEs from various document types compared to a single type, resulting in a performance similar to that of the typical manual analysis.

Figure 1. Data flow of the proposed system. (A) shows the events from multiple types of documents are extracted. An event timeline (B) is created from each clinical data using the natural language processing method, and then the curve (C) is created based on the aggregated results. The dots in the event timeline indicate the timing at which the description of drug administration or symptom onset is recorded. Based on (B), patients who received the target drug (a taxane drug in this study) are selected, and the Kaplan-Meier curve (C) is generated.



Methods

Overview

Our study uses a retrospective observational approach based on NLP, which enables the handling of a large amount of data. The NLP techniques used were NER and normalization, which extract symptoms from documents and transform them into their standardized forms. We evaluated our method by obtaining the Kaplan-Meier curves based on symptoms that were normalized to peripheral neuropathy (PN). In addition, we also evaluated the duration of PN.

Materials

This study used data from all patients diagnosed with breast cancer (N=2289) treated at the Kyoto University Hospital between 2019 and 2021. The patient data consisted of 2 types of medication orders (structured data) and 6 types of texts written in Japanese (unstructured data). We apply NLP methods to extract information from such unstructured data. Unstructured data require an NLP method to extract information, such as ADEs. Table 1 lists the number of all breast cancer patient orders and text data records included in each document.

Table 1. Amount of order data and text data. The unit of record is per drug for order data and per timing recorded by physicians or co-medicals for text data.

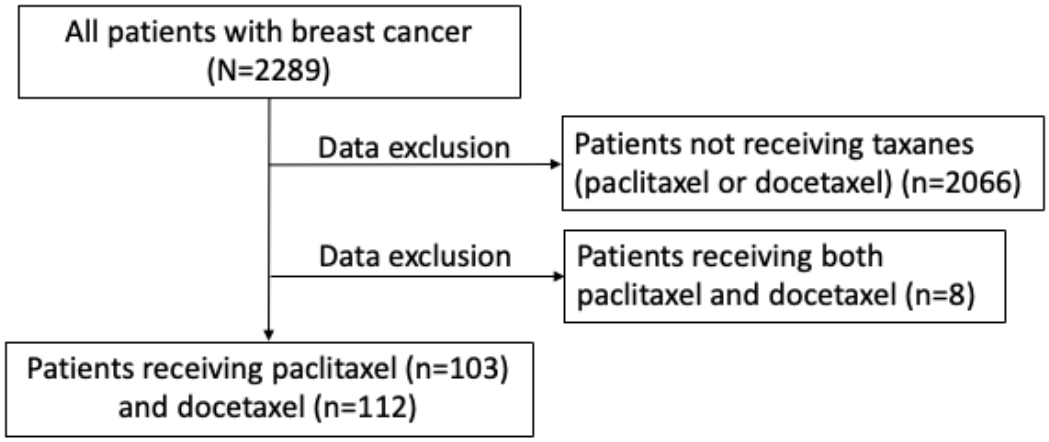
Data type	Records, n
Order data	
Injection order	44,896
Medicine order	63,077
Text data	
Physician progress notes	159,736
Nursing records	40,385
Discharge summaries	23,073
Radiology reports	5663
Radioisotope reports	1147
Pharmacist progress notes	29,148

Inclusion and Exclusion Criteria

The aim of this study was to leverage the strengths of NLP to automatically analyze a large number of documents and evaluate the usefulness of the proposed method. We selected PN as our disease for evaluation, as it satisfies the following conditions: (1) side effects are long-lasting, which means that monitoring many documents is required, and (2) information on the onset of the symptom is not normally included in structural databases.

We selected taxane drugs, such as paclitaxel and docetaxel for the evaluation, as they frequently cause PN. As shown in Figure 2, patients receiving either type of taxane drugs were included in the analysis, whereas those receiving both paclitaxel and docetaxel were excluded. Patients selected according to these criteria were then analyzed for the development of PN as an outcome.

Figure 2. Flowchart describing the procedure for selecting patient data according to criteria.

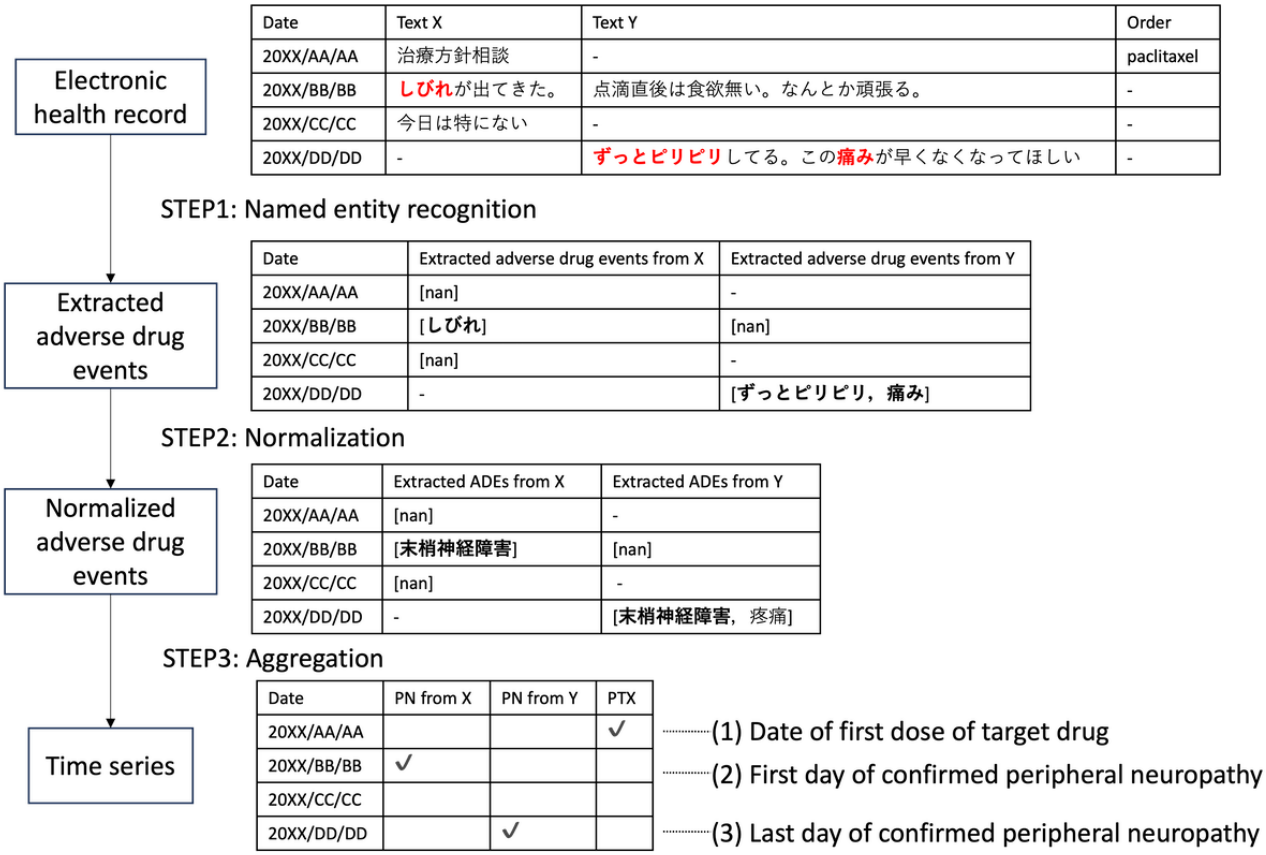


Patients who received both drugs were excluded to prevent them from introducing noise in the analysis of PN onset and duration. Administering a different taxane drug during monitoring sessions, which had not been given previously, could have adversely affected the study results.

Comparison With the Kaplan-Meier Curves

Using information extracted from multitype texts by applying our NLP method, we measured the number of days until the onset of PN after the administration of taxane drugs. As shown in Figure 3, our system is composed of 3 steps: entity recognition, normalization, and aggregation. We compared these results with those reported manually in a previous report [13].

Figure 3. Workflow of our natural language processing system, which is composed of named entity recognition, normalization, and aggregation. Text X and Text Y are examples of 2 types of documents respectively, for example, physician progress notes and pharmacist progress notes.



Named Entity Recognition

The data required for this study, which included the dates on which the symptoms occurred (obtained from text data) and drugs that were administered (obtained from medication orders), were obtained as follows: to obtain symptom data, we applied NER, an NLP method that recognizes and extracts mentions of named entities in text. We used this method to identify symptoms related to PN. We adopted MedNER-CR-JA, which is a BERT-based NER model trained using Japanese case reports [14]. Since BERT can only process a maximum of 512 tokens at a time, sentences were separated by line breaks. Only the symptoms with positive factuality, as extracted by the model, were used in the analysis.

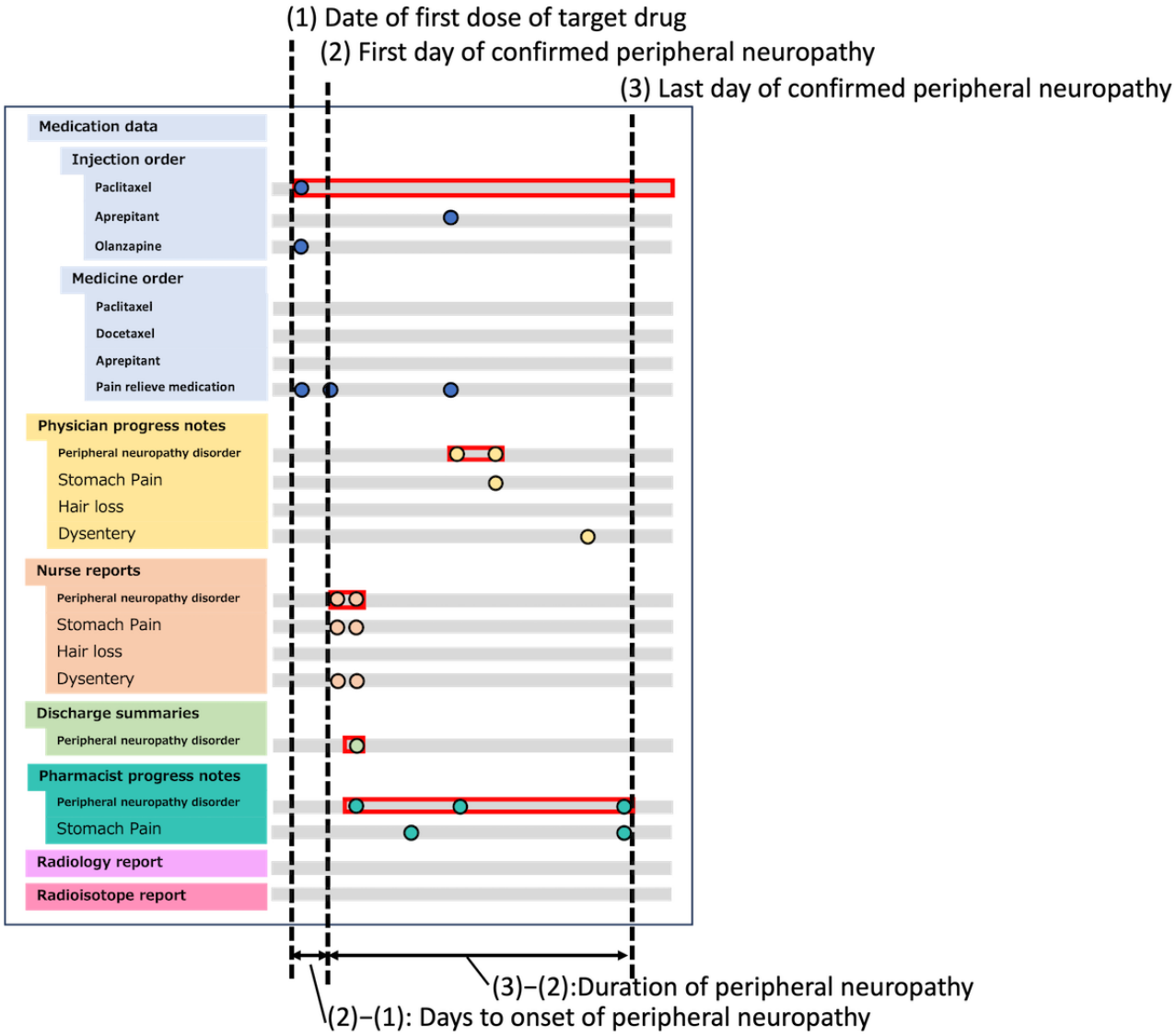
Normalization

The extracted entities are normalized by Levenshtein distance matching using a disease name dictionary (MedDic-CANCER-ADE-JA) [15]. This dictionary contains surface forms and normalized forms with respect to the side effects of anticancer drugs. We select the dictionary surface form that has the lowest Levenshtein distance in relation to the extracted term and then convert it to the related normalized form [16]. The code for this step including NER can be accessed through GitHub [17].

Aggregation

We focused on the expression normalized to PN and conducted the analysis among the converted terms. Specifically, the onset date of PN was defined as the first date on which the expression was normalized to PN in any type of document. The cumulative percentage of patients who developed PN was calculated along the time series. As shown in Figure 4, the onset date was the number of days since the first dose of paclitaxel or docetaxel. We defined the period of residual PN as the period up to the date on which the expression normalized to PN was last identified. The onset date and residual duration for each patient were summed to obtain a Kaplan-Meier plot of onset timing or residual duration, respectively. The onset date and residual duration of each patient were aggregated to obtain a Kaplan-Meier plot of onset timing or residual duration, respectively.

Figure 4. Event timeline from multiple types of data and calculation of the number of days of peripheral neuropathy onset and duration.



We propose that this definition would be more robust if the system analyzed various types of documents reviewed by multiple medical personnel. Increasing the diversity of documents analyzed reduces the risk of overlooking symptoms.

Evaluation

The cumulative percentage of the patients' PN is displayed using the event date on which PN was first identified.

We compared the results produced by our NLP system (Paclitaxel_NLP) with previous results obtained by manual extraction (Paclitaxel_MAN) based on the percentage of PN at 30 days. The detection rate was evaluated by subtracting the percentage of detections achieved by our system from the percentage of detections obtained through manual extraction [13]. We focused on the incidence of PN at 30 days since most patients generally develop the disease after 30 days [13].

In addition, the Pearson correlation coefficient was calculated for the 2 types of paclitaxel results from our system and manual results up to 101 days, the maximum duration in the previous report.

In addition, multiple regression analysis was performed to analyze the results calculated using all records and the results from each record to evaluate which explanatory variables had a greater impact.

Ethical Considerations

This study, which was evaluated and approved by the ethics committee of Kyoto University Graduate School and Faculty of Medicine, Japan (R3723-2), was performed in compliance with the Declaration of Helsinki.

Results

Preliminary Result

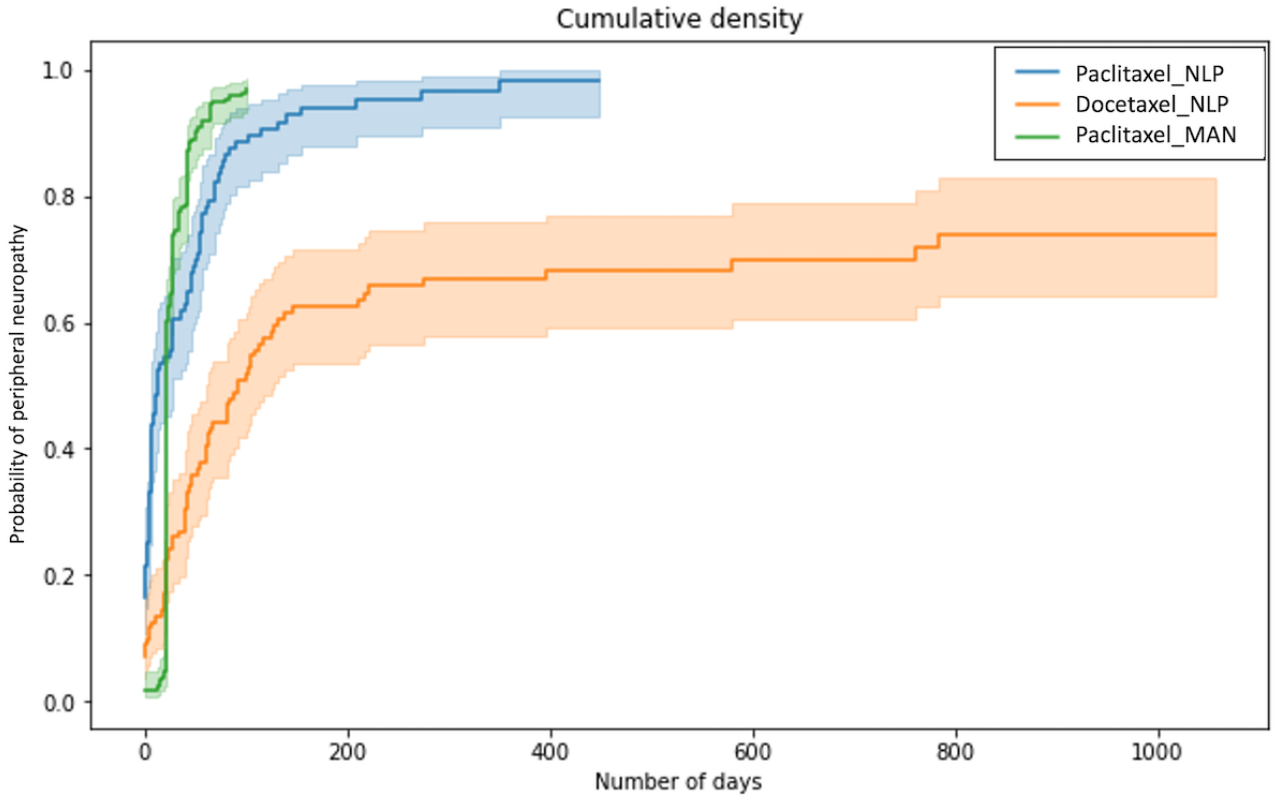
As shown in Figure 2, among the 2289 patients from the data set, 215 were selected (paclitaxel, n=103; docetaxel, n=112). A total of 2066 patients who did not receive paclitaxel nor docetaxel and 8 patients who received both paclitaxel and docetaxel were excluded. The median age of the participants was 59 years (range 33-78) for the paclitaxel-treated patients and 52 (25-73) years for the docetaxel-treated patients, which

is not significantly different from the median age of 53 (range 22-70) years in previous reports. The mean and maximum follow-up periods were, respectively, 380.3 and 1264 days for paclitaxel-treated patients and 545.1 and 1080 days for docetaxel-treated patients.

A total of 7428 symptom expressions were extracted (paclitaxel=3732 and docetaxel=3696), of which 5057 (paclitaxel=2804 and docetaxel=2253) were positive for symptom factuality and 879 (paclitaxel=569 and docetaxel=310) were PN-related.

Figure 5 shows the Kaplan-Meier curves of the results obtained by our system and the previous results obtained using a manual method. Of the 103 patients who received paclitaxel (n=103), 97 had confirmed PN; from the 112 patients who received docetaxel (n=112), 76 had confirmed PN.

Figure 5. Kaplan-Meier curves of the results obtained by our system (Paclitaxel_NLP and Docetaxel_NLP) and the previous results obtained using a manual method (Paclitaxel_MAN). The solid line indicates the proportion of patients who developed peripheral neuropathy among those who received paclitaxel or docetaxel. Filled areas indicate 95% CIs.



Comparison With the Kaplan-Meier Curves

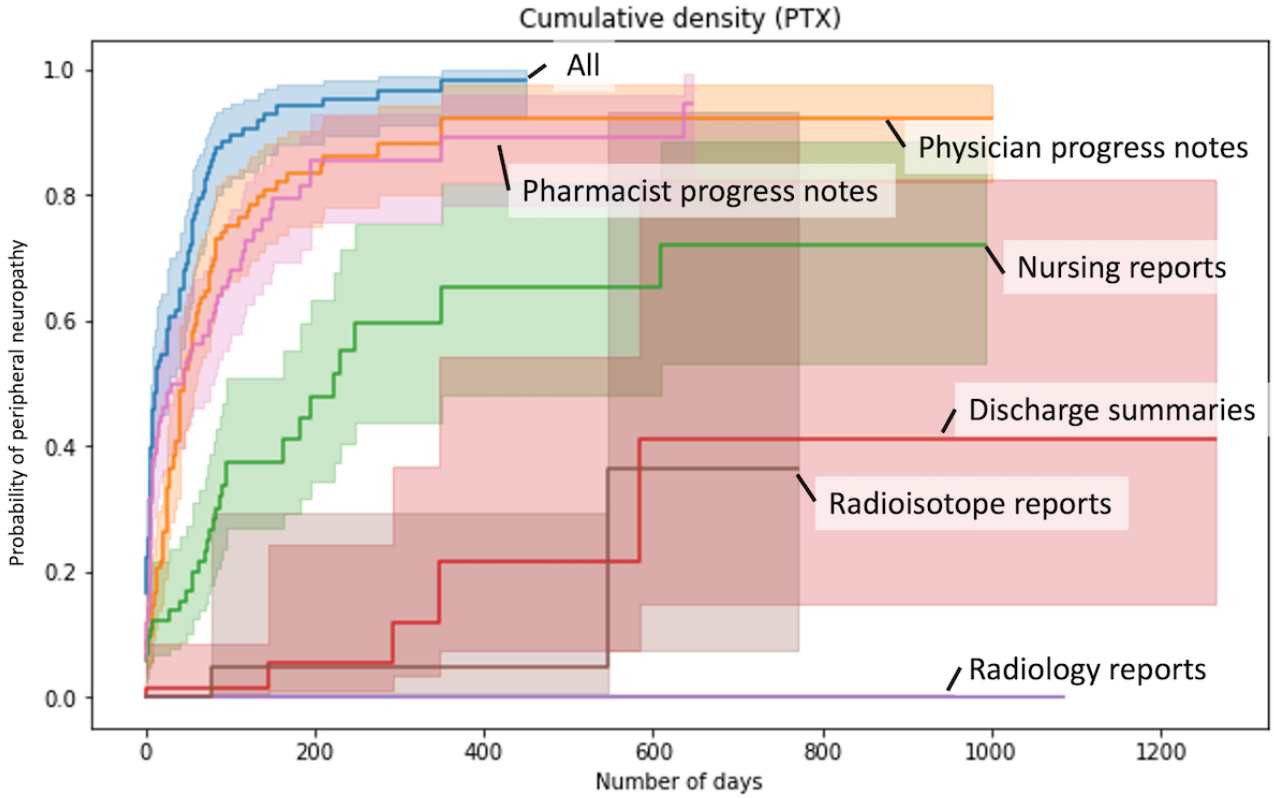
The incidence of PN caused by paclitaxel was 60.7% at 30 days, and as the previous research reported incidence was 74% at 30 days, and the detection gap was 13.3 points (74.0%-60.7%) [13]. The percentages represent the proportions of patients who were determined to have developed PN from documents.

The result does not entirely reflect the actual onset of the disease; however, the system detected PN in almost all patients over 1 year, which seems accurate enough. The correlation coefficient between the results obtained by our system (Paclitaxel_NLP) and those obtained manually (Paclitaxel_MAN) was 0.87, with a P value of 1.72×10^{-32} (<.05), indicating a high correlation.

Figure 6 shows the comparison between the results from per document type and all document types. The percentages of PN identified in each document type, in descending order, were physician progress notes, pharmacist progress notes, nursing records, discharge summaries, radioisotope reports, and radiology reports.

In order to assess which documents influenced the results calculated from all documents, multiple regression analyses were performed. The results from all documents were used as predictor variables, and the results from each document as explanatory variables. The respective regression coefficients and SD (shown in parentheses) were 0.70 (0.04) for pharmacist progress notes, 0.35 (0.03) for physician progress notes, 0.32 (0.09) for nursing records, 1.39 (1.67) for discharge summaries, 1.64×10^{-16} (1.64×10^{-16}) for radiology reports, and –0.53 (0.21) for radioisotope reports. The results suggest the importance of physician progress notes, pharmacist progress notes, and nursing reports among the document types.

Figure 6. Comparison between the results from each document type and all document types.



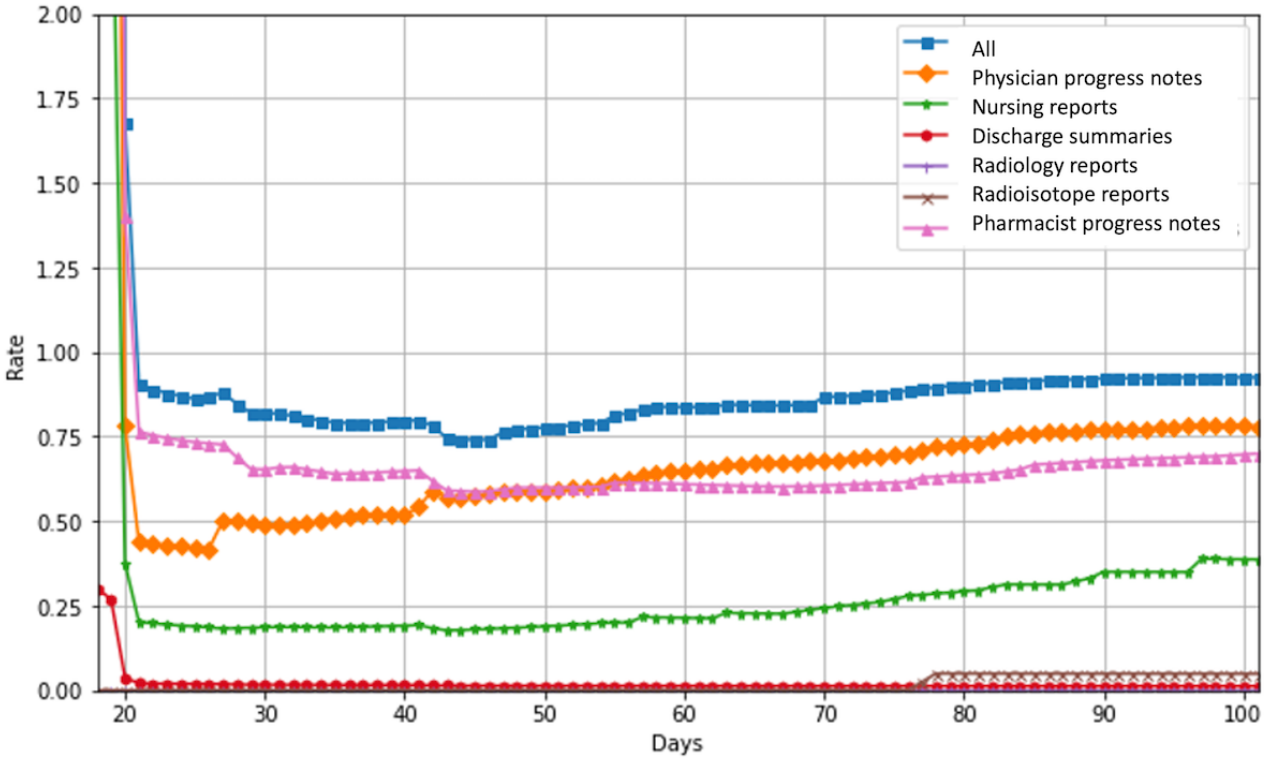
When all records were used, the system was able to detect the onset of all PNs that could be detected at 350 days. On the other hand, the same results could not be obtained, even after 600 days, when each type of documents was used independently.

The results from the pharmacist records were similar to those from all types of documents in the initial period but remained almost constant after 200 days, with few new cases of PN detected. Nursing records, which contain many records of patient care, were expected to be effective in detecting adverse drug reactions such as PN, but the detection rate was less than half that of physician and pharmacist records. The detection rate for discharge summaries, radiology reports, and radioisotope reports was very low (less than half), suggesting that these types of documents are less useful for the target diseases and target drugs in this study.

Figure 7 shows the detection rate of patients with PN in each document compared with the manual results. At 30 days, the detection rate compared with manual was 65.3% for pharmacist progress notes, 49.1% for physician progress notes, 18.6% for nursing records, 1.6% for discharge summaries, 0% for radioisotope reports, and 0% for radiology reports. The detection rates of all records were lower than the combined detection rate of all records (82.0%). This suggests that the use of multiple types of documents is effective.

In the early stage of the observation period, automatic extraction tended to overdetect PN. This is likely due to the incorrect detection of expressions related to side effect descriptions, which will be discussed in the error analysis section. The detection rate decreases in the middle of the period and slowly increases in the latter half. In the first half of the observation period, pharmacist progress notes showed the highest performance in detecting the results from a single type of documents, while physician progress notes showed the highest performance in the second half of the observation period. It is interesting to note that different document types tend to have different detection rates depending on the time of observation.

Figure 7. Rates of patients with peripheral neuropathy detected in each document compared with manual results.



Discussion

Principal Findings

In contrast to the rapid increase in the number of patients developing PN 20 days after beginning treatment in a previous report, our system detected PN at an earlier treatment stage [15].

Our method can extract symptoms from text, determine factuality, and chronologically monitor the patient's symptoms. Therefore, as long as the target symptoms are described in the text, the same method can be applied to any symptom and all drugs other than taxanes, making it a versatile and scalable method. Although there is still room for improvement in accuracy, the analysis can be automated to reduce research costs, particularly in observational studies, where large amounts of text need to be analyzed.

Error Analysis

The detection rate by our system may be affected by false negatives, suggesting that the model overlooks expressions that are difficult to detect, such as onomatopoeic expressions, as we will discuss in this section.

A detailed analysis categorized 3 types of errors, as shown in Table 2, namely, errors in symptom extraction, factuality determination, and normalization. Among these, errors in factuality determination and errors in normalization were found to increase the likelihood of outputting false positives. An example of errors in factuality determination is that explanations such as "this medicine has a risk of PN" can be misinterpreted as PN. Normalization errors included instances of normalizing expressions not limited to PN, such as "numbness" in "Numbness + in upper extremities due to cervical stenosis" to PN. All 3 types of errors were identified as false negatives. As a symptom extraction error, it was confirmed that the onomatopoeic "tingling (ビリビリ, biribiri)" was not extracted. The NER model is not effective at recognizing more informal expressions, such as onomatopoeias, probably because the model is fine-tuned using case reports, which are relatively formal sentences. As for errors in factuality determination, "tingling (ぴりぴり, piripiri)" was not extracted in "Even a rest does not stop tingling sensation in my hands." Although this text implies the positive factuality of the symptoms, the presence of negation in the sentence may have interfered with the model's determination of factuality. Note that this expression is onomatopoeic and translated into the same word in English. However, this is a different expression in Japanese, and the symptoms were properly extracted. Other expressions such as "There is a risk of paralysis (麻痺のリスクあり)" and "Explained side effects of eribulin…PN… (エリブリンの副作用...末梢神経障害...について説明)" were also incorrectly extracted. This is because such expressions are rarely used in case reports.

Table 2. Types of errors in the detection of peripheral neuropathy. Original Japanese texts are in parentheses. Italicized text is the expression extracted by named entity recognition.

Types of errors and examples	Prediction outcome
Errors in extracting symptoms	
Have a tingling sensation in my limbs. （手足がぴりぴりする）	False negative
Errors of factuality determination	
Explained side effects of eribulin: decreased blood counts, risk of infection, *PN*, fatigue, decreased appetite, etc. （エリブリンの副作用：血球減少、感染のリスク、末梢神経障害、倦怠感、食欲低下等について説明）	False positive
Owing to fractured thoracic vertebrae, there is a risk of paralysis during rehabilitation （リハビリは胸椎が骨折で麻痺のリスクあり）	False positive
Even a rest does not stop tingling sensation in my hands.（手のびりびりは休んでもマシにはなりません。）	False negative
Errors of normalization	
Numbness + in upper extremities due to cervical stenosis. （頚椎狭窄で上肢にしびれ＋）	False positive
Cellulitis of the right upper extremity （右上肢蜂窩織炎）	False positive
No abnormal changes of note in blood sampling results（特記すべき異常変化を採血結果に認めない）	False positive
Bilateral supraclavicular lymph nodes, mediastinal lymph nodes, and para-aortic lymph node metastases are considered to be affected, *decreased accumulation* （両側鎖骨上リンパ節、縦隔リンパ節、傍大動脈リンパ節転移は効果ありと考えられる集積低下）	False positive
After wearing a supporter, edema got better, but *pain and numbness* appeared. （サポーターをしたら、浮腫は良くなったが、逆に痛み しびれが出てきた。）	False negative

As a normalization error, the expression such as “pain and numbness” in the sentence “After wearing a supporter, edema got better, but pain and numbness appeared.” was normalized incorrectly because the model extracted not only numbness but also pain and numbness as a coherent expression, and any surface terms in the dictionary did not match sufficiently in this case. False positives in normalization are influenced by the surface form of the dictionary used. “Abnormal change (異常変化)” is matched to “sensory abnormality (感覚異常)” in the dictionary, and “hypoaccumulation (集積低下)” is matched to “hypoalgesia (痛覚低下).”Adjustment of the Levenshtein distance threshold may yield better results.

The false positive result suggests an early overdetection of PN in the automatic detection system, while the false negative result is associated with a decrease in the detection rate in the middle of the graph. As shown in Figure 7, for false negatives, our method of using multiple types of documents compensates for the lower detection rate compared with the use of a single document.

The impact of the error on clinical outcomes is that a false negative in the extracting symptoms and factuality determination represents a significant clinical risk because it means that an adverse drug reaction was missed. However, our method of using multiple types of documents reduces this risk compared with using only 1 type of document because the multiple types of documents complement each other and reduce false negatives. In the case of a false positive, the risk of adverse drug reactions is overestimated, and the patient may not be able to choose an appropriate treatment if the adverse drug reaction is a factor in the drug selection decision. In addition, the same phenomenon may occur in the case of normalization errors. The linking of different symptoms may also lead to incorrect conclusions about adverse drug reactions because the symptoms that occur cannot be accurately captured. For example, an unrelated symptom may be detected as a risk, resulting in unnecessary investigations.

Documents Containing Adverse Drug Event Information

Table 3 shows a breakdown of the number of documents and patients with PN detected in each document. Since large counts of PN detection are seen in nursing records, pharmacist progress notes, and physician progress notes, we assert that analyzing multiple types of documents, such as nursing records and pharmacist progress notes, is as important as physician progress notes. It can be inferred from these results that combining multiple types of medical documents not only enables the detection of more patient events, but also reduces the number of missed events per patient.

Table 3. Counts in each document type.

Document type	Documents, n			Patients, n		
	Total	Paclitaxel	Docetaxel	Total	Paclitaxel	Docetaxel
Physician progress notes	373	246	127	146	85	61
Nursing records	189	117	72	80	49	31
Discharge summaries	24	10	14	22	9	13
Radiology reports	0	0	0	0	0	0
Radioisotope reports	2	2	0	2	2	0
Pharmacist progress notes	291	194	97	137	81	56

Duration of Adverse Drug Event

Duration of PN was calculated as the period from the date of onset to the date of the last PN detection. The median number of days of PN onset by paclitaxel was 12 days, the median number of days of last confirmed onset was 126 days from the start of administration, and the median duration was 92 days. The median number of days of PN onset by docetaxel was 45.5 days, the mean number of last observed days was 135.5 days from the start of administration, and the median duration was 64.0 days. The median duration of PN with paclitaxel reported previously was 727 days, and the results are likely to significantly underestimate the duration because of the nature of the follow-up period of the analyzed data in this study, which was approximately 1000 days at most [13].

Limitations

The results obtained with our method are dependent on the accuracy of the NER model used. Although our model achieved the best performance in a shared task, there is still room for improvement, with an F_1-score of 62.9% for the extraction performance of the relevant tags in this task [18]. This model was fine-tuned based on case reports; however, we expect that fine-tuning using annotated data from the same type of documents as those used in this study, such as nursing records and progress notes, will improve the results. In addition, dictionary matching using the Levenshtein distance is performed for normalization. The normalization may have introduced false positives and false negatives.

The onset of PN was defined as the date when PN was first identified in the text. Therefore, if a PN that occurred in the past is mentioned in the text, it is possible that the onset of PN is assessed late. Similarly, the end of the PN disease period was defined as the date on which PN was last identified. The maximum follow-up period of the studies used in this study was approximately 1000 days, which may be an underestimate of the residual duration of PN.

This method focuses on the presence or absence of PN and does not quantitatively evaluate the common terminology criteria for adverse events grade. Although this model determines only the factuality of the symptoms, a more detailed analysis can be conducted by creating a model that determines the grade.

Conclusions

We proposed a system to detect PN by using NLP methods to allow the analysis of multityped documents automatically and concurrently. Analyses were performed on breast cancer patients receiving paclitaxel and docetaxel. As a result, many PN events were extracted from the nursing records and pharmacists' progress notes as well as physicians' progress notes. This approach is reasonable when considering the multiple types of records used in this study since leveraging multitype documents is better than single-type documents to improve detection performance. Based on the timing of the onset, our system underestimates by 13.3 percentage points.

We also examined persistent PN using a similar approach. Compared with the manual results, it was suggested that the duration of PN was underestimated; however, this may be due to the large difference in the follow-up periods.

Although the accuracy of the system requires further investigation, we believe that our NLP system has great potential to provide an immediate estimate of the persistence of ADEs, which traditionally requires continuous investigation and incurs high costs.

Acknowledgments

This work was supported by the Japan Science and Technology Agency, Advanced Integrated Intelligence Platform (JST AIP), Japanese-German-French AI (artificial intelligence) Research Grant (JPMJCR20G9), Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (JSPS KAKENHI) Grant (JP21H03170) and Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant (JPJ012425).

Conflicts of Interest

MToi has received research grants from Chugai, Takeda, Pfizer, Taiho, Japan Breast Cancer Research Group Association (JBCRG), Kyoto Breast Cancer Research Network (KBCRN), Eisai, Eli-Lilly and companies, Daiichi-Sankyo, AstraZeneca, Astellas, Shimadzu, Yakult, Nippon Kayaku, AFI technology, Luxonus, Shionogi, GL Science, Sanwa Shurui; and lecture fees from Chugai, Takeda, Pfizer, Kyowa-Kirin, Taiho, Eisai, Daiichi-Sankyo, AstraZeneca, Eli Lilly and companies, MSD, Exact Science, Novartis, Shimadzu, Yakult, Nippon Kayaku、Devicore Medical Japan, Sysmex; and advisory fees from Daiichi-Sankyo, Eli Lilly and companies, BMS, Bertis, Terumo, Kansai Medical Net.

References

1. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://aclanthology.org/N19-1423/ [doi: 10.18653/v1/N19-1423]
2. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 2020;3:136 [FREE Full text] [doi: 10.1038/s41746-020-00341-z] [Medline: 33083571]
3. Morin O, Vallières M, Braunstein S, Ginart JB, Upadhaya T, Woodruff HC, et al. An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. Nat Cancer 2021;2(7):709-722. [doi: 10.1038/s43018-021-00236-2] [Medline: 35121948]
4. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv Preprint posted online April 10, 2019. [doi: 10.48550/ARXIV.1904.05342]
5. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
6. Lederman A, Lederman R, Verspoor K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. J Am Med Inform Assoc 2022;29(10):1810-1817 [FREE Full text] [doi: 10.1093/jamia/ocac121] [Medline: 35848784]
7. Laparra E, Mascio A, Velupillai S, Miller T. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearb Med Inform 2021;30(1):239-244 [FREE Full text] [doi: 10.1055/s-0041-1726522] [Medline: 34479396]
8. Magge A, Tutubalina E, Miftahutdinov Z, Alimova I, Dirkson A, Verberne S, et al. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. J Am Med Inform Assoc 2021;28(10):2184-2192 [FREE Full text] [doi: 10.1093/jamia/ocab114] [Medline: 34270701]
9. Wu H, Ji J, Tian H, Chen Y, Ge W, Zhang H, et al. Chinese-Named Entity RecognitionFrom adverse drug event records: radical embedding-combined dynamic embedding-based BERT in a bidirectional long short-term conditional random field (Bi-LSTM-CRF) model. JMIR Med Inform 2021;9(12):e26407 [FREE Full text] [doi: 10.2196/26407] [Medline: 34855616]
10. Murphy RM, Klopotowska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, et al. Adverse drug event detection using natural language processing: a scoping review of supervised learning methods. PLoS One 2023;18(1):e0279842 [FREE Full text] [doi: 10.1371/journal.pone.0279842] [Medline: 36595517]
11. McKenzie J, Rajapakshe R, Shen H, Rajapakshe S, Lin A. A semiautomated chart review for assessing the development of radiation pneumonitis using natural language processing: diagnostic accuracy and feasibility study. JMIR Med Inform 2021;9(11):e29241 [FREE Full text] [doi: 10.2196/29241] [Medline: 34766919]
12. Tsai WC, Tsai YC, Kuo KC, Cheng SY, Tsai JS, Chiu TY, et al. Natural language processing and network analysis in patients withdrawing from life-sustaining treatments: a retrospective cohort study. BMC Palliat Care 2022;21(1):225 [FREE Full text] [doi: 10.1186/s12904-022-01119-8] [Medline: 36550430]
13. Tanabe Y, Hashimoto K, Shimizu C, Hirakawa A, Harano K, Yunokawa M, et al. Paclitaxel-induced peripheral neuropathy in patients receiving adjuvant chemotherapy for breast cancer. Int J Clin Oncol 2013;18(1):132-138. [doi: 10.1007/s10147-011-0352-x] [Medline: 22105895]
14. Nishiyama T, Nishidani M, Ando A, Yada S, Wakamiya S, Aramaki E. NAISTSOC at the NTCIR-16 Real-MedNLP task. Tokyo, Japan; 2022 Presented at: NTCIR 16 Conference: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies; June 14-17, 2022; Tokyo, Japan URL: https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/07-NTCIR16-MEDNLP-NishiyamaT.pdf
15. MedDic-CANCER-ADE-JA_202306.: SOCIOCOM Social Computing Laboratory since 2015; 2023. URL: https://sociocom.naist.jp/download/meddic-cancer-ade-ja_202306/ [accessed 2024-08-23]
16. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady 1966;10(8):707 [FREE Full text]

17. Social Computing Lab. MedNERN-CR-JA.: Hugging Face; 2023. URL: https://github.com/sociocom/MedNERN-CR-JA [accessed 2024-08-23]
18. Shuntaro Y, Yuta N, Shoko W, Eiji A. Real-MedNLP: overview of REAL document-based MEDical natural language processing task. Tokyo; 2022 Presented at: Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies; June 14-17, 2022; Tokyo, Japan URL: https://www.researchgate.net/publication/361410507_Real-MedNLP_Overview_of_REAL_document-based_MEDical_Natural_Language_Processing_Task_SUBTASKS

Abbreviations

ADE: adverse drug event
BERT: Bidirectional encoder representations from transformers
NER: named entity recognition
NLP: natural language processing
PN: peripheral neuropathy

Edited by G Eysenbach, C Lovis; submitted 29.03.24; peer-reviewed by D Chrimes, J Zaghir, JP Goldman, A Wani; comments to author 01.05.24; revised version received 31.05.24; accepted 17.08.24; published 24.09.24.

Please cite as:
Nishiyama T, Yamaguchi A, Han P, Pereira LWK, Otsuki Y, Andrade GHB, Kudo N, Yada S, Wakamiya S, Aramaki E, Takada M, Toi M
Automated System to Capture Patient Symptoms From Multitype Japanese Clinical Texts: Retrospective Study
JMIR Med Inform 2024;12:e58977
URL: https://medinform.jmir.org/2024/1/e58977
doi:10.2196/58977
PMID:39316418

Original Paper

Evaluating Medical Entity Recognition in Health Care: Entity Model Quantitative Study

Shengyu Liu[1], MS; Anran Wang[1], MS; Xiaolei Xiu[1], MS; Ming Zhong[1], MS; Sizhu Wu[1], PhD

Department of Medical Data Sharing, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Corresponding Author:
Sizhu Wu, PhD
Department of Medical Data Sharing
Institute of Medical Information & Library
Chinese Academy of Medical Sciences & Peking Union Medical College
3 Yabao Road
Chaoyang District
Beijing, 100020
China
Phone: 86 10 5232 8760
Email: wu.sizhu@imicams.ac.cn

Abstract

Background: Named entity recognition (NER) models are essential for extracting structured information from unstructured medical texts by identifying entities such as diseases, treatments, and conditions, enhancing clinical decision-making and research. Innovations in machine learning, particularly those involving Bidirectional Encoder Representations From Transformers (BERT)–based deep learning and large language models, have significantly advanced NER capabilities. However, their performance varies across medical datasets due to the complexity and diversity of medical terminology. Previous studies have often focused on overall performance, neglecting specific challenges in medical contexts and the impact of macrofactors like lexical composition on prediction accuracy. These gaps hinder the development of optimized NER models for medical applications.

Objective: This study aims to meticulously evaluate the performance of various NER models in the context of medical text analysis, focusing on how complex medical terminology affects entity recognition accuracy. Additionally, we explored the influence of macrofactors on model performance, seeking to provide insights for refining NER models and enhancing their reliability for medical applications.

Methods: This study comprehensively evaluated 7 NER models—hidden Markov models, conditional random fields, BERT for Biomedical Text Mining, Big Transformer Models for Efficient Long-Sequence Attention, Decoding-enhanced BERT with Disentangled Attention, Robustly Optimized BERT Pretraining Approach, and Gemma—across 3 medical datasets: Revised Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), BioCreative V CDR, and Anatomical Entity Mention (AnatEM). The evaluation focused on prediction accuracy, resource use (eg, central processing unit and graphics processing unit use), and the impact of fine-tuning hyperparameters. The macrofactors affecting model performance were also screened using the multilevel factor elimination algorithm.

Results: The fine-tuned BERT for Biomedical Text Mining, with balanced resource use, generally achieved the highest prediction accuracy across the Revised JNLPBA and AnatEM datasets, with microaverage (AVG_MICRO) scores of 0.932 and 0.8494, respectively, highlighting its superior proficiency in identifying medical entities. Gemma, fine-tuned using the low-rank adaptation technique, achieved the highest accuracy on the BioCreative V CDR dataset with an AVG_MICRO score of 0.9962 but exhibited variability across the other datasets (AVG_MICRO scores of 0.9088 on the Revised JNLPBA and 0.8029 on AnatEM), indicating a need for further optimization. In addition, our analysis revealed that 2 macrofactors, entity phrase length and the number of entity words in each entity phrase, significantly influenced model performance.

Conclusions: This study highlights the essential role of NER models in medical informatics, emphasizing the imperative for model optimization via precise data targeting and fine-tuning. The insights from this study will notably improve clinical decision-making and facilitate the creation of more sophisticated and effective medical NER models.

(JMIR Med Inform 2024;12:e59782) doi:10.2196/59782

KEYWORDS

natural language processing; NLP; model evaluation; macrofactors; medical named entity recognition models

Introduction

Background

The importance of robust named entity recognition (NER) models in medical informatics has become increasingly evident; these models systematically extract structured information from unstructured textual data such as clinical notes and research articles. This capability is crucial for processing large volumes of clinical data, facilitating early disease detection and supporting personalized medicine. NER models transform complex medical data into usable information, significantly enhancing clinical decision-making and medical research [1-3]. The performance of these models directly impacts the quality of information retrieval and data analysis as medical texts often contain complex and diverse terminologies. Precision in identifying and categorizing entities such as diseases, treatments, anatomical structures, and medications is foundational to many health care applications, making NER an indispensable tool in advancing medical informatics [4].

The critical role of NER in electronic health records further illustrates its importance. NER models automatically extract essential patient information, such as symptoms and medical conditions, that is crucial for differential diagnosis and ensuring that clinicians have rapid access to critical patient history. Errors in entity classification or omission can lead to severe consequences, including misdiagnosis or inappropriate treatment, highlighting the need for highly accurate NER models. Moreover, NER facilitates medical research by enabling efficient data mining from extensive medical literature, aiding the organization and retrieval of information on various medical entities. This capability is essential in drug discovery and identifying disease biomarkers, where systematic analysis and synthesis of large amounts of text are required. For instance, in pharmacovigilance, NER models identify adverse drug reactions from clinical notes and reports, contributing to drug safety monitoring and public health initiatives [5,6].

However, the application of NER in medical informatics presents unique challenges. The complexity and specificity of medical language, including synonyms, acronyms, and context-specific meanings, necessitate the continuous refinement of NER models. Recent studies have revealed significant variability in the effectiveness of NER models across different medical datasets, potentially limiting their real-world applicability. Despite these challenges, there have been promising advancements. For example, the BBC-Radical model, which integrates Bidirectional Encoder Representations From Transformers (BERT) with Bidirectional Long Short-term Memory, and conditional random field (CRF), has demonstrated high precision, recall, and F_1-scores in extracting adverse drug reaction–related information from Chinese adverse drug event records [7]. This example illustrates the potential of combining advanced embedding techniques with traditional machine learning methods to enhance NER performance in specific contexts. These findings underscore the critical need for domain-specific adaptations and the integration of a more advanced linguistic and contextual understanding into these models. Such enhancements are essential for improving the prediction accuracy and applicability of NER models across diverse medical contexts.

Given these complexities and the need for continuous improvement, understanding the factors influencing model performance is crucial. Optimizing these models for specific tasks can significantly enhance their efficiency and accuracy. In medical informatics, this optimization is vital as the quality of data analysis directly impacts clinical decision-making. Researchers can make targeted improvements by identifying key elements of model design or aspects of training data that significantly affect performance, such as better handling of rare or complex medical terms. This knowledge assists in developing new models and refining existing ones to meet the specific needs of health care applications, thereby improving the precision of data extraction and analysis. Such improvements are crucial for the reliability of clinical information systems, reducing the risk of misdiagnosis or inappropriate treatment. Ultimately, these advancements support more accurate and informed clinical decision-making [8].

To achieve these improvements, it is essential to thoroughly evaluate the performance of various NER models across different contexts. First, this study identified 3 categories of NER models: statistical machine learning models, deep learning natural language processing (NLP) models based on BERT architecture, and large language models (LLMs). We selected these categories based on their unique strengths and potential to address specific challenges in medical NER.

Statistical machine learning models, such as hidden Markov models (HMMs) and CRF, were chosen for this study due to their established methodologies and effectiveness in sequence prediction tasks. These models leverage probabilistic approaches to capture the sequential nature of language data. However, they often struggle with the complexities of medical terminologies without extensive feature engineering. This limitation necessitates continuous refinement and adaptation to effectively handle medical texts' intricate and specialized language [9].

Deep learning NLP models, especially those based on the BERT architecture, represent a significant advancement in NER capabilities. Variants of the BERT model, such as BioBERT, Robustly Optimized BERT Pretraining Approach (RoBERTa), Big Transformer Models for Efficient Long-Sequence Attention (BigBird), and Decoding-enhanced BERT with Disentangled Attention (DeBERTa), have demonstrated exceptional performance in capturing the intricate context of medical language. We selected these specific variants for this study because they can leverage deep contextual embeddings and undergo large-scale pretraining on medical corpora, enabling them to handle the complexities of medical terminologies effectively. For instance, a model using RoBERTa with whole-word masking and convolutional neural networks achieved high F_1-scores in Chinese clinical NER tasks,

indicating its effectiveness in processing complex medical terminologies within electronic medical records [10]. This success is primarily due to their ability to leverage deep contextual embeddings and undergo large-scale pretraining on medical corpora. These models are highly effective in identifying entities across diverse medical datasets, although they require substantial computational resources and meticulous fine-tuning for optimal performance. The introduction of BERT-based NER model variants marks a significant breakthrough in medical informatics, notably enhancing medical data analysis [11]. By leveraging advanced feature extraction techniques from masked language models such as embeddings from language models and the transformer architecture, these models set new standards for precise analysis of diverse medical datasets [12]. Specifically, RoBERTa incorporates enriched training data and refined masking patterns [13], whereas BigBird addresses previous models' scalability and comprehension challenges by efficiently processing extended sequences [14]. DeBERTa's innovative attention mechanisms further enhance these capabilities [15]. BioBERT, with its specialized pretraining on extensive medical texts, exemplifies the efficacy of domain-specific adaptations, achieving unparalleled precision in recognizing medical entities [16]. This strategic focus on contextual nuances and specialized terminology significantly improves the quality of patient care decisions.

LLMs, such as those based on the generative pretrained transformer architecture, have shown significant promise in NLP, particularly in understanding complex language structures. However, as the studies by Tian et al [17], Zhao et al [18], and Hu et al [19] have highlighted, their application in medical NER presents challenges, including limited prediction efficiency, extended runtimes, and substantial hardware requirements. These challenges are compounded by the complexity and specificity of medical terminology, often resulting in suboptimal accuracy with standard medical datasets such as BioCreative V CDR (BC5CDR), Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA), and NCBI (National Center for Biotechnology Information Disease Corpus). The general optimization strategies of LLMs do not align well with the specialized needs of NER tasks in medical informatics [17-19]. Despite these challenges, we selected Gemma, a fine-tuned version adapted explicitly for medical NER tasks, for this study [20]. Gemma was chosen over other open-source LLMs such as Large Language Model Meta AI 3 [21] or Open Pretrained Transformer [22] due to its advanced capabilities in contextual understanding and efficiency in processing specialized vocabulary. Its fine-tuning process specifically targets medical terminologies, making it more adept at handling the nuances of medical language. In addition, Gemma's tailored adaptations include specific optimizations that align with the unique challenges of medical datasets, allowing it to achieve higher accuracy and reliability in recognizing entities within complex medical texts [23]. The proven effectiveness of fine-tuned models such as Gemma, as demonstrated by systems such as the Medical Named Entity Recognition–Japanese developed for analyzing pharmaceutical care records in Japanese, further underscores its suitability for medical NER tasks [24].

Building on this classification, we integrated various NER models, including HMM, CRF, RoBERTa, BigBird, DeBERTa, BioBERT, and Gemma. Following this integration, evaluating these models comprehensively to enhance their accuracy and durability is essential. Recent studies by Freund et al [25], Ahmad et al [26], and others highlight a shift toward more comprehensive evaluation metrics tailored to the specific needs of medical informatics. This shift facilitates the development of advanced NER applications, significantly improving patient care by enhancing clinical data management. This marks a substantial advancement in the integration of technology and health care. However, the evaluations' primary focus has been assessing the predictive capabilities of NER models using metrics such as precision, recall, and F_1-score. Research by Yoon et al [27], Yu et al [28], and Yadav and Bethard [29] has used these metrics to evaluate these capabilities. The study by Erdmann et al [30] also compared various NER tools for literary text corpora with human annotators using the same metrics, highlighting the importance of such evaluations in the digital humanities. Furthermore, the work by Usha et al [31] and Nagaraj et al [32], which includes advanced techniques such as confusion matrices and receiver operating characteristic and precision-recall curves, offers a more nuanced understanding of classification accuracy and errors in NER models. As the studies by Ozcelik and Toraman [33] and Akhtyamova [34] explored, error analysis is critical for understanding performance nuances, especially in identifying and categorizing short- and long-term entities. This comprehensive evaluation approach underscores the complexities and challenges in developing accurate and efficient NER models for medical informatics [33,34].

The aforementioned studies predominantly used standard metrics such as precision, recall, receiver operating characteristic curve, and F_1-score to evaluate model performance. However, these metrics fall short in capturing performance variations across different medical NER datasets and conducting a detailed analysis of how dataset characteristics affect model performance. Moreover, although these metrics are easy to compute, their interpretation proves challenging. This difficulty mainly arises from the metrics' failure to explain the broader reasons behind model outcomes as they often depend on processing microvector features that are not intuitively understandable. Consequently, researchers focused on enhancing medical NER datasets through NER models encounter significant hurdles in devising effective optimization strategies. In response, some researchers have shifted to customizing macrofactors for evaluating explanatory NER models. For example, Fu et al [35] developed an evaluation framework that outlines 8 distinct factor types to analyze their correlation with the models' F_1-score rankings. Zhou et al [36] proposed an ant colony optimization algorithm based on parameter adaptation. They designed a new dynamic parameter adjustment mechanism to adaptively adjust the pheromone importance factor. This algorithm is also suitable for selection of macrofactors. By adaptively changing the macrofactors, the algorithm can determine which macrofactors affect the prediction accuracy of the NER model [36]. Yao et al [37] also enhanced this domain with their groundbreaking a scale-adaptive mathematical morphology spectrum entropy algorithm, which

adjusts the scale of structural elements to measure macrofactors' impact on model prediction accuracy. These advancements have led to increasingly sophisticated NER model evaluations, resulting in more precise and resilient models.

Objectives

Given this context, the purpose of this study was to systematically evaluate the comprehensive performance of various NER models in the medical field focusing on both general medical texts and specific medical entity types. In addition, this study aimed to explore key macrofactors affecting model prediction performance. To achieve these goals, we used a comprehensive evaluation approach combining traditional and innovative techniques to enhance the accuracy and reliability of NER in medical informatics. This approach included analyzing hardware performance indicators such as training duration, central processing unit (CPU), and graphics processing unit (GPU) use and assessing the precision of models such as HMM, CRF, RoBERTa, BigBird, DeBERTa, BioBERT, and Gemma across different medical entity types. Furthermore, we proposed the multilevel factor elimination (MFE) algorithm, which integrates linear and machine learning strategies to filter multilayer factors and evaluate their impact on prediction accuracy. Through this comprehensive evaluation, we aimed to provide targeted recommendations for researchers, ultimately leading to the development of more accurate and reliable NER models for broader applications in the medical field.

Methods

Overview

This section outlines 2 methods: *training and evaluating medical NER models* and *further assisted evaluation*. The first method evaluates the prediction accuracy of the statistical machine learning models (HMM and CRF), the deep learning NLP models based on BERT architecture (BioBERT, BigBird, DeBERTa, and RoBERTa), and the Gemma LLM across different medical entity types, as well as overall model effectiveness. The second method further assesses the prediction accuracy of merged entity types within these models' postclassification and examines the macrofactors' influence on model performance.

Training and Evaluating Medical NER Models

Overview

This method involved training, validating, and testing NER models using hyperparameter tuning techniques. For models such as RoBERTa, BioBERT, BigBird, and DeBERTa, we used the Adaptive Moment Estimation (ADAM) optimizer for hyperparameter tuning. These strategies were applied using datasets such as the Revised JNLPBA, focusing on optimizing parameters such as learning rates, batch sizes, and other model-specific settings to enhance prediction accuracy. In addition, Gemma was fine-tuned using low-rank adaptation (LoRA) to improve its predictive performance. In contrast, models such as HMM and CRF were not subjected to hyperparameter tuning. This decision was based on the inherent simplicity of these models, which are less sensitive to hyperparameter variations than more complex deep learning models. Consequently, fine-tuning HMM and CRF would likely yield marginal improvements that did not justify the additional computational resources and time investment.

Dataset Selection

In total, 3 medical NER datasets were used to evaluate the prediction accuracy of models across different medical contexts (Table 1). These datasets use the "beginning, inside, outside" sequence annotation method (Figure 1); among them, the Revised JNLPBA dataset, provided by Huang et al [38], was selected because it retains the original semantic annotation type while addressing known vulnerabilities from the original JNLPBA dataset. It features 5 entity types (DNA, RNA, protein, cell line, and cell type), offering a focused scope on biological entities. The BC5CDR dataset, officially released by Li et al [39], was chosen for its 2 distinct entity types—disease and chemical—which present unique challenges due to their complex and overlapping terminology. Finally, the Anatomical Entity Mention (AnatEM) dataset [40], focusing on anatomical entities in medical fields, was used due to its broad range of 12 different entity types, providing a wide spectrum of medical terms. This diversity in entity numbers and types—most of which are distinct across the datasets—strengthens the evaluation by exposing the models to varied linguistic challenges, thereby reducing the randomness and contingency of the experimental results and enhancing the overall credibility of the experiment.

Table 1. Descriptive statistics of the medical named entity recognition datasets.

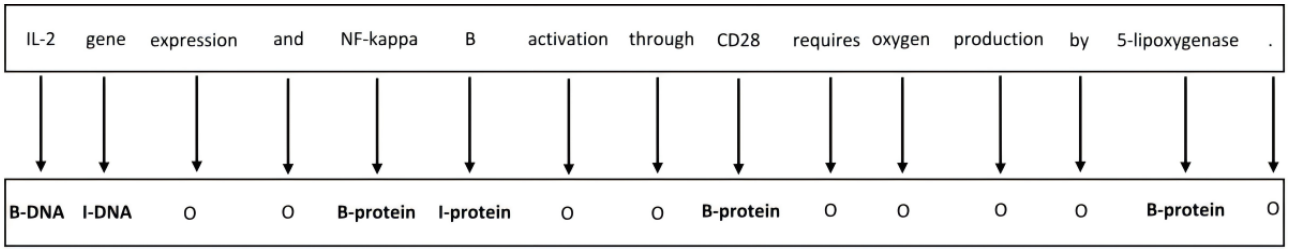
Dataset	Medical entity type	Entity types, n	Annotations, n
Revised JNLPBA[a]	DNA, RNA, protein, cell line, and cell type	5	52,785
BC5CDR[b]	Disease and chemical	2	38,596
AnatEM[c]	Organism subdivision, anatomical system, organ, multi-tissue structure, tissue, cell, developing anatomical structure, cellular component, organism substance, immaterial anatomical entity, pathological formation, and cancer	12	11,562

[a]JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

[b]BC5CDR: BioCreative V CDR.

[c]AnatEM: Anatomical Entity Mention.

Figure 1. An instance of the beginning, inside, outside (BIO) sequence annotation method. CD28: cluster of differentiation 28; IL-2: interleukin-2; NF-kappa: nuclear factor kappa-light-chain-enhancer of activated B cells.



Training and Testing of the Deep Learning NLP Models With BERT Architecture

Overview

This section discusses achieving optimal prediction accuracy through meticulous hyperparameter tuning during the model training phase. We trained models such as RoBERTa, BigBird, DeBERTa, and BioBERT on datasets divided into training, validation, and test sets sourced from the Revised JNLPBA, BC5CDR, and AnatEM collections. The ADAM optimizer was used to fine-tune key hyperparameters, including learning rate, batch size, epochs, and dropout rate. This optimizer adjusts the learning rate for each parameter based on estimates of lower-order moments of the gradients, enabling faster convergence by adapting to the characteristics of the data and gradients. This adaptive adjustment enhances convergence speed and overall model performance, aiming to identify the optimal hyperparameter combination that maximizes performance on the validation set [41,42].

The models' prediction accuracy on the test set was evaluated primarily using F_1-scores for specific entity types and overall performance metrics. The F_1-scores, calculated as the harmonic mean of precision and recall, were chosen as the primary evaluation metric because they provide a balanced assessment of model performance in detecting relevant entities. This balance is crucial in medical NER, where precision measures the proportion of correctly identified entities and recall measures the proportion of actual entities correctly identified. In medical information extraction, where false positives and negatives can have significant implications, the F_1-score's ability to balance these metrics is essential. Missing important medical terms can lead to incomplete patient records or misunderstandings in medical literature, whereas false positives can introduce erroneous information into clinical decision-making processes. Thus, the F_1-score is particularly suitable and reliable for evaluating model performance in medical NER tasks, where accuracy and reliability are paramount [43].

While the ADAM optimizer dynamically adjusted learning rates during model training to improve convergence, cross-validation was used separately to evaluate and optimize hyperparameters. Cross-validation, a robust method for model validation that involves dividing the dataset into multiple subsets for validation, ensures that the chosen hyperparameters generalize well to unseen data. This distinction underscores the complementary roles of the optimizer in training and cross-validation in validating and tuning the model's hyperparameters. The selection and tuning of hyperparameters were guided by cross-validation. In this context, hyperparameter optimization aims to identify a global or satisfactory local optimum that maximizes medical NER model performance by adjusting several key hyperparameters, described in the following sections.

Learning Rate Adaptation Strategy

We used a dynamic learning rate adjustment strategy to enhance the model's efficiency in converging to an optimal local minimum—testing rates of 0.0001 and 0.00001. This approach is similar to the findings of Wu and Liu [44] on the benefits of adaptive learning rates in NLP applications.

Batch Size Considerations

Guided by research on neural network training dynamics and computational constraints [44], we selected batch sizes of 10 and 50. This allowed for more frequent model updates and a finer approach to convergence.

Epoch Configuration

The number of training epochs was set based on the dataset's complexity and initial performance metrics [45] with values of 1, 5, and 10. This adaptive approach minimized the risk of overfitting while ensuring the duration of practical training.

Dropout for Regularization

To prevent overfitting, we applied dropout rates of 0.1, 0.2, and 0.5 as informed by interim validation performance [46]. This regularization technique enhanced generalization across unseen medical texts, ensuring model reliability.

Training and Testing of the Gemma LLM

In this section, the Gemma 7B model was fine-tuned using the LoRA technique to optimize its performance on medical NER tasks. The LoRA approach was selected for its ability to efficiently adapt LLMs to specific tasks while reducing computational and memory overhead. This method involves freezing the pretrained model weights and introducing trainable low-rank decomposition matrices in each layer of the transformer architecture, significantly decreasing the number of trainable parameters and GPU memory requirements [47].

The fine-tuning process was conducted using a carefully structured setup. The Gemma 7B model was trained using datasets such as the Revised JNLPBA, BC5CDR, and AnatEM with a token cutoff length of 512 and a maximum of 20,000 samples. Preprocessing_num_workers was set to 16, and a mixture of depths was converted to facilitate diverse learning dynamics. The output configuration included saving checkpoints every 1000 steps and logging progress every 200 steps, ensuring thorough monitoring of the training process.

Training parameters included a per-device batch size of 8, a gradient accumulation step of 1, and a learning rate of 0.0001, managed by a cosine learning rate scheduler with a warm-up ratio of 0.1. The training spanned 3 epochs, using mixed precision with bfloat16 for efficiency. A weight decay of 0.01 was applied to prevent overfitting. A 10% validation split was used for evaluation, with assessments conducted every 200 steps using a batch size of 4 per device.

LoRA-specific settings included a dropout rate of 0.05 and a rank of 128 for the low-rank matrices applied across all model layers. This configuration enabled significant reductions in trainable parameters and GPU memory use. The LoRA technique allowed the Gemma 7B model to be fine-tuned effectively for complex NER tasks in the medical domain without compromising performance. These adjustments provided a scalable approach for adapting large-scale language models to specialized tasks with efficient computational resources.

Finally, the F_1-score was used to evaluate the prediction accuracy of each entity type under the Gemma 7B model. To further demonstrate the effectiveness of the fine-tuning process, metrics such as the microaverage (AVG_MICRO) and macroaverage (AVG_MACRO) were leveraged to compare the performance of the fine-tuned model against a baseline without fine-tuning.

Training and Testing of the Statistical Machine Learning Models

This section used HMM and CRF for medical NER tasks using the Revised JNLPBA, BC5CDR, and AnatEM datasets. Unlike deep learning methods and LLMs, these statistical models did not require hyperparameter tuning. HMMs were trained by estimating transition and emission probabilities, whereas CRFs used feature functions to capture relationships between labels and features in the data. The prediction accuracy for each entity type was also evaluated using the F_1-score.

Evaluative Metrics and Model Assessment

After the training, cross-validation, and testing phases, we used F_1-score metrics across various hyperparameter combinations to evaluate the model's prediction accuracy for different entity types. These F_1-scores were aggregated into AVG_MICRO and AVG_MACRO as leveraged metrics to assess the model's prediction accuracy under different hyperparameter configurations. The AVG_MICRO metric provides a comprehensive measure of overall model performance, capturing precision and recall across the dataset. The AVG_MACRO metric highlights the model's ability to handle varying entity distributions, including rare entities. This dual-metric approach ensures a balanced evaluation, preventing biases toward more common entities and maintaining consistent performance across diverse entity types [42].

To enhance our evaluation framework, we assessed training time efficiency and computational resource use, including CPU and GPU use. These additional metrics provide insights into each model's operational demands and feasibility, allowing us to evaluate their prediction accuracy and practicality for real-world applications.

Further Assisted Evaluation

Overview

This method further evaluated the relationship between medical data and NER models, focusing on categorical relaxation to reduce entity classification complexity and improve prediction accuracy. We detailed the process of merging similar entity types into broader categories and evaluated the impact on model performance through a comprehensive analysis of macrofactors such as sentence length (sLen) and entity density (eDen). These methods were applied across several datasets, such as Revised JNLPBA, BC5CDR, and AnatEM, to systematically assess NER models.

Academic Revision on Categorical Relaxation

Categorical relaxation in NER classification, particularly within the medical domain, entails merging similar medical entity types to diminish ambiguity and enhance the classification performance of NER models. This technique simplifies the classification landscape and bolsters the models' capacity to generalize from training data to unseen clinical examples. In this study, we implemented categorical relaxation by consolidating specific medical entity types, such as merging different DNA or RNA names. This method was guided by empirical evidence suggesting that merging reduces misclassifications and boosts prediction accuracy, particularly in diagnosing conditions and recommending treatments.

In the Revised JNLPBA dataset, we adopted a merging strategy based on the principles described by Tsai et al [48]. Biologically related categories such as DNA, RNA, and proteins were consolidated into a single "macromolecule" category. Similarly, entities categorized as cell lines and cell types were combined into a single "cell" entity category.

In the AnatEM dataset, we reclassified 12 entity types into 4 broader categories relevant to human health following the classification guidelines from Pyysalo and Ananiadou [40]. This reorganization is predicated on the premise that broader categories more effectively capture essential information and reduce the noise caused by particular and infrequently occurring entities.

Following categorical relaxation, we comprehensively evaluated the RoBERTa, BigBird, DeBERTa, and BioBERT models. This assessment compared the prediction accuracy of these models on the newly consolidated entity types. The objective was to determine the effects of entity type consolidation on model performance, particularly whether simplifying categories enhanced prediction accuracy across different model architectures. This method simplifies entity classification and capitalizes on the inherent similarities among entity types to improve model training and evaluation. By reducing the granularity of entity types, we posited that the models would achieve higher accuracy and more effectively address the complexities of medical texts. The specific outcomes of this categorical relaxation are detailed in Table 2.

Table 2. Merged entity types.

Dataset and medical entity type	Merged entity type
Revised JNLPBA[a]	
DNA, RNA, and protein	Macromolecule
Cell line and cell type	Cell
BC5CDR[b]	
Disease	Disease
Chemical	Chemical
AnatEM[c]	
Organism subdivision, anatomical system, organ, multi-tissue structure, tissue, cell, developing anatomical structure, and cellular component	Anatomical structure
Organism substance	Organism substance
Immaterial anatomical entity	Immaterial anatomical entity
Pathological formation and cancer	Pathological formation

[a]JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

[b]BC5CDR: BioCreative V CDR.

[c]AnatEM: Anatomical Entity Mention.

Constructing and Evaluating NER Macrofactor Datasets

Macrofactor Metrics Definition

Entity macrofactor metrics were defined in an entity (entity phrase fragment) or on the entire dataset, and dataset's different attributes were described. On the basis of the 8 factor types defined by Fu et al [35], we categorized these into 6 macrofactor metrics. This categorization was derived from a detailed examination of the characteristics of 3 medical datasets, including the length of medical terminology and other relevant factors. This allowed us to distill the most pertinent metrics for our study, which included *sLen*, entity phrase length *(eLen)*, *eDen*, number of entity words in each entity phrase (*eNum*), total entity word count in each entity type *(tEWC)*, and entity label consistency *(elCon)* (Textbox 1)

Textbox 1. Definitions of macrofactor metrics.

- *sLen*: this metric refers to the string length of the sentence containing the entity phrase. It quantifies the contextual space in which entities appear. Notably, extreme values are excluded to prevent distortion in average calculations.
- *eLen*: this metric quantifies the string length of each entity phrase, which can comprise one or more entity words. Similar to *sLen*, extreme values are excluded to yield a more precise average *eLen*for each entity type.
- *eDen*: this metric is calculated as the ratio of *eLen* to *sLen*; this metric quantifies how dense entities are populated within the text.
- *eNum*: for datasets such as Revised JNLPBA labeled with the "beginning, inside, outside" sequence, this metric counts the entity words in a phrase, adjusting for labeling specifics such as combining "B-Entity label" and "I-Entity label" into a single count to avoid underrepresentation in the data.
- *tEWC*: this metric quantifies the cumulative number of entity words within each specific entity type across datasets such as the Revised JNLPBA, BC5CDR, and AnatEM.
- *elCon*: this metric evaluates the consistency of entity-type assignments to medical terms across various contexts, which is essential in datasets such as the Revised JNLPBA dataset, where terms can possess multiple semantic interpretations. For instance, "lymphocyte" may be categorized as "B-cell_type" in discussions about receptor counts and as "I-cell_type" in contexts involving blood samples. To calculate*elCon*, each term from the dataset is cataloged along with its associated entity types. For example, "lymphocyte" would be documented with both "B-cell_type" and "I-cell_type." A weight ω is assigned to each term based on the inverse number of its entity types (ω=1/number of entity types), indicating that terms linked to fewer entity types tend to demonstrate higher prediction accuracy. Terms associated with a single entity type are assigned a weight of 1, whereas those with 2 types are given a weight of 0.5. We calculate the average weight using the following formula—average weight value  —excluding extreme values to avoid data skew.

The New Macrofactor Dataset Creation

To construct the macrofactor datasets, we computed metrics such as *sLen*, *eLen*, *eNum*, *eDen*, and *elCon* for each entity word within each entity type across the Revised JNLPBA, BC5CDR, and AnatEM datasets. These metrics quantify specific linguistic and structural attributes of the datasets. In addition, we included *tEWC* as a separate column to represent the total count of entity words for every kind across the datasets, quantifying the overall entity word volume. Figure 2 shows the systematic computation of these metrics except *tEWC* to augment the macrofactor datasets.

Due to the supervised nature of these datasets, labels were assigned based on the F_1-scores achieved by the top-performing models within 3 distinct categories: statistical machine learning models (HMM and CRF), deep learning NLP models based on BERT architecture (RoBERTa, BigBird, DeBERTa, and BioBERT), and LLMs (Gemma). The model with the highest AVG_MICRO and AVG_MACRO scores was identified for each model category. The F_1-scores for each entity type, as achieved by these best-performing models, were then used as labels for the datasets. This method ensures that the labels accurately reflect the prediction accuracy for each entity type. Consequently, 9 unique macrofactor datasets were generated, each representing a combination of the computed metrics and the derived labels. Figure 3A illustrates the use of these F_1-scores as labels, whereas Figure 3B provides an example of how the dataset structure was revised based on these macrofactors.

Figure 2. Sentence length (sLen), entity phrase length (eLen), number of entity words in each entity phrase (eNum), entity density (eDen), and entity label consistency (elCon) values of an entity word.

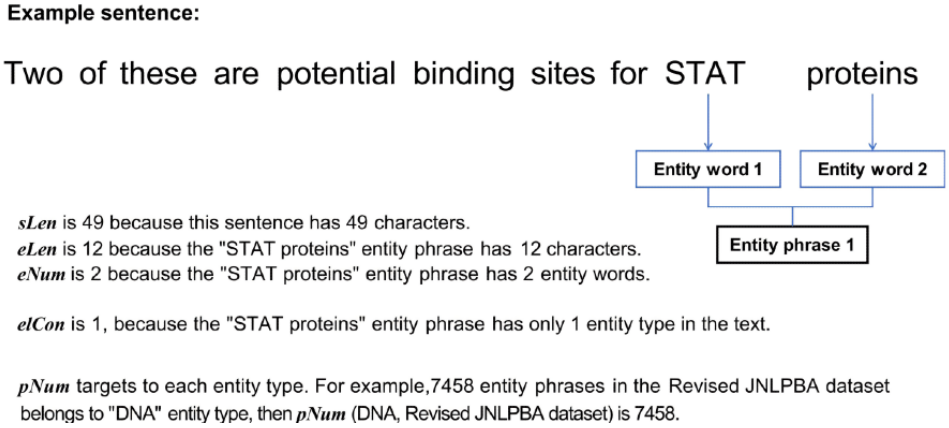
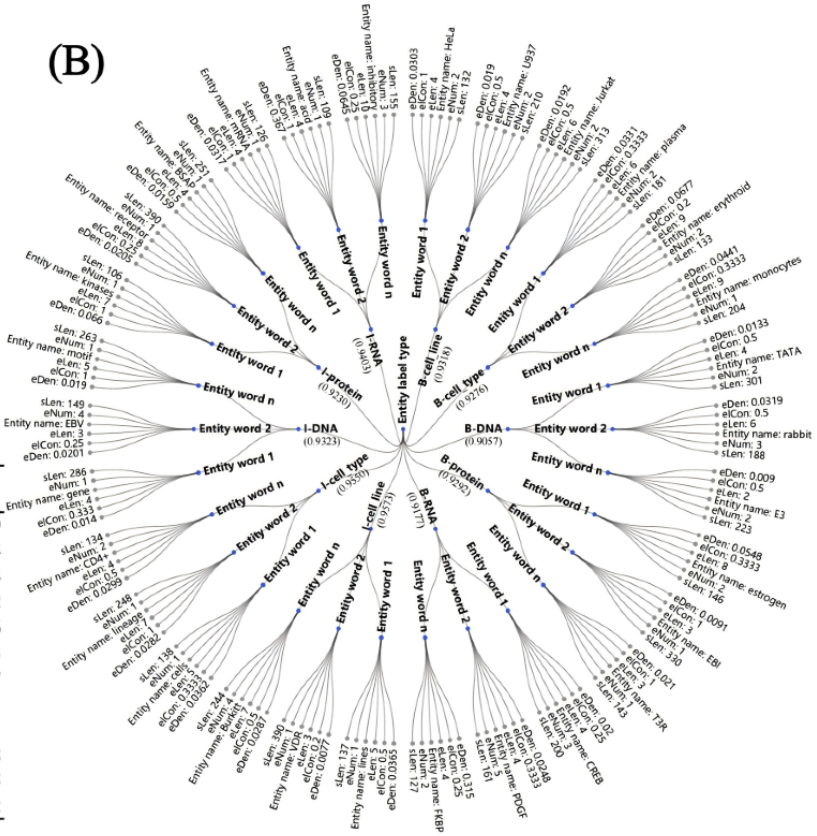


Figure 3. An example of 2 parts illustrating different components of each macrofactor dataset. Panel (A) displays the labels used in the datasets, whereas panel (B) presents a visualization of the macrofactor metrics—sentence length, entity phrase length, number of entity words in each entity phrase, entity density, and entity label consistency—using the radial tree layout algorithm. This visualization offers a structured view of the interrelationships among these metrics. AVG_MACRO: macroaverage; AVG_MICRO: microaverage.

(A)

	Level	F1-Score
0	B-cell_type	0.926899
1	I-cell_type	0.952954
2	B-protein	0.926224
3	I-protein	0.914452
4	B-RNA	0.936306
5	I-DNA	0.930413
6	B-cell_line	0.940837
7	I-cell_line	0.963761
8	B-DNA	0.916821
9	I-RNA	0.944578
	AVG_MACR	0.935325
	AVG_MICR	0.928272



Preliminary Macrofactor Evaluation

Our preliminary analysis adopted a comprehensive method to assess the interactions between prediction accuracy for various entity types (eg, disease, DNA, and RNA) and 6 macrofactors. Initially, we extracted prediction accuracy data for each entity type from these NER models (HMM, CRF, RoBERTa, BigBird, DeBERTa, BioBERT, and Gemma), and we calculated the average values for macrofactor metrics such as *sLen*, *eLen*, *eNum*, *eDen*, *elCon*, and *tEWC*. We then explored the impact of each metric on the prediction accuracy for each entity type. We considered trend analysis to provide deeper insights into the relationships between these metrics and accuracy levels. In addition, the visualization of multimodel predictive trends offered a comprehensive view of model robustness across different entity types.

In-Depth Macrofactor Selection

Overview

Our subsequent research focused on identifying which macrofactors significantly impacted the models' prediction accuracy. We developed an MFE algorithm and conducted macrofactor selection. MFE is an improved algorithm based on recursive feature elimination [49] divided into 3 layers.

First Layer: Factor Ranking and Selection

The MFE algorithm inputs the calculated values for *sLen*, *eLen*, *eNum*, *eDen*,*tEWC*, and *elCon* for each entity word.

Linear correlation analysis is performed for each factor using the following function. The correlation coefficient (r) is a statistical measure used to quantify the strength and direction of the relationship between 2 variables. x_i represents individual macrofactor measurements, is the mean of all macrofactor measurements, y_i represents individual model performance measurements, and is the mean of all model performance measurements. This calculation quantifies the relationship between each macrofactor and the model's prediction accuracy.

Using a method similar to the Pearson correlation method [50], factors are ranked by their correlation scores, and the top 4 are retained for further analysis. This step ensures that only the factors with the highest potential impact are advanced, thereby efficiently streamlining the feature space.

Second Layer: Random Forest Evaluation

A random forest model performs multiple training iterations using the selected macrofactors from the first layer.

After each training session, cross-validation is used to evaluate model performance. The macrofactor with the lowest feature importance score, indicating minimal contribution to prediction accuracy, is systematically excluded from subsequent analyses. This iterative refinement process prioritizes factors that consistently enhance model accuracy [51].

Third Layer: Regression Model Optimization

The refined set of macrofactors is integrated into a regression model.

The least impactful macrofactors are sequentially eliminated based on the influence of their coefficients on the model, which is calculated as follows:

$$\text{Coefficient influence} = \beta \quad \textbf{(2)}$$

β is the coefficient related to the macrofactors; the model is retrained after each removal, continuing until no further improvement in performance is detected. This final step ensures that only the most impactful factors are retained, optimizing the model's efficiency and effectiveness.

Ethical Considerations

Ethics approval was obtained from the Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College (ethics approval code: IMICAMS/01/22/HREC). After obtaining ethics approval, the Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College, wrote an official ethics approval statement.

All medical NER datasets used in this paper are public datasets; no personal or sensitive information was collected, and the datasets complied with local institutional guidelines and legislation. It was unnecessary to obtain written or verbal informed consent from the participants. The experimental protocols and datasets in this study were approved by the Institute of Medical Information and Library, Chinese Academy of Medical Sciences and Peking Union Medical College. All methods were performed according to the relevant guidelines and regulations.

Results

Model Prediction Results

The predicted outcomes of the NER models following extensive training, testing, and hyperparameter optimization are detailed in Table 3. The statistical machine learning models, HMM and CRF, exhibited moderate performance. HMM showed relatively lower scores across all datasets, with AVG_MICRO scores of 0.7265, 0.825, and 0.6112 for the Revised JNLPBA, BC5CDR, and AnatEM datasets, respectively. Similarly, its AVG_MACRO was also lower, indicating limited effectiveness in capturing diverse entity types. On the other hand, the CRF model performed better than HMM, achieving higher AVG_MICRO scores of 0.8476, 0.9019, and 0.743 for the same datasets, respectively. The corresponding AVG_MACRO for CRF was also higher, reflecting its better overall accuracy and balance in entity recognition.

In contrast, the BERT-based deep learning models (BioBERT, RoBERTa, BigBird, and DeBERTa) consistently demonstrated a strong performance across multiple datasets. Notably, BioBERT achieved the highest results on the Revised JNLPBA and AnatEM datasets, with an AVG_MICRO of 0.932 and an AVG_MACRO of 0.9298 on the Revised JNLPBA dataset and an AVG_MICRO of 0.8494 and an AVG_MACRO of 0.6975 on the AnatEM dataset. However, on the BC5CDR dataset, these models performed slightly worse, with AVG_MICRO and AVG_MACRO values generally below 0.8726 and 0.858, respectively. This indicated that, while BERT-based models

exhibited robust performance in specific contexts, their generalization capabilities varied depending on the dataset characteristics.

The Gemma model demonstrated a strong performance on the BC5CDR dataset, achieving the highest AVG_MICRO of 0.9962 and an AVG_MACRO of 0.981. However, its results were less pronounced on the Revised JNLPBA and AnatEM datasets compared to BioBERT and other models.

Table 3. Model prediction accuracy comparison.

Model	Revised JNLPBA[a] dataset		BC5CDR[b] dataset		AnatEM[c] dataset	
	AVG_MICRO[d]	AVG_MACRO[e]	AVG_MICRO	AVG_MACRO	AVG_MICRO	AVG_MACRO
Statistical machine learning models						
HMM[f]	0.7265	0.6815	0.8250	0.7002	0.6112	0.5013
CRF[g]	0.8476	0.8258	0.9019	0.8883	0.7430	0.5114
Deep learning NLP[h] models based on BERT[i] architecture						
BioBERT[j]	0.932	0.9298	0.8726	0.858	0.8494	0.6975
RoBERTa[k]	0.9133	0.9133	0.8313	0.8152	0.8201	0.6501
BigBird[l]	0.9277	0.9218	0.8461	0.8321	0.8147	0.6451
DeBERTa[m]	0.9256	0.921	0.8471	0.8335	0.806	0.6131
Large language model						
Gemma	0.9088	0.8298	0.9962	0.9810	0.8029	0.6496

[a]JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

[b]BC5CDR: BioCreative V CDR.

[c]AnatEM: Anatomical Entity Mention.

[d]AVG_MICRO: microaverage.

[e]AVG_MACRO: macroaverage.

[f]HMM: hidden Markov model.

[g]CRF: conditional random fields.

[h]NLP: natural language processing.

[i]BERT: Bidirectional Encoder Representations From Transformers.

[j]BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining.

[k]RoBERTa: Robustly Optimized BERT Pretraining Approach.

[l]BigBird: Big Transformer Models for Efficient Long-Sequence Attention.

[m]DeBERTa: Decoding-enhanced BERT with Disentangled Attention.

In addition, Figures 4 and 5 highlight the F_1-scores for each entity type achieved by the best-performing models in each category: statistical machine learning models (represented by CRF), deep learning NLP models based on BERT architecture (represented by BioBERT), and the Gemma LLM.

The performance of the CRF, BioBERT, and Gemma models on the Revised JNLPBA dataset showed distinct differences across various entity types. BioBERT consistently outperformed CRF and Gemma across most entity types. For instance, in the "B-DNA" entity type, BioBERT attained the highest score of 0.9057 compared to CRF's 0.7733 and Gemma's 0.8883. Gemma also demonstrated a high performance in several entity types, such as "B-protein" (0.9368) and "I-DNA" (0.9503). Nonetheless, it showed variability in other entity types, such as "B-RNA" (0.7925), where BioBERT achieved a significantly higher score (0.9177). While showing promising results in entity types such as "I-cell_line" (0.8911), the CRF model generally performed lower than BioBERT and Gemma. Notably, CRF's performance in the "I-RNA" entity type (0.7655) was significantly lower than that of both BioBERT (0.9403) and Gemma (0.8652).

In the BC5CDR dataset, Gemma consistently outperformed CRF and BioBERT across most entity types. For instance, in the "B-Disease" entity type, Gemma achieved an F_1-score of 0.9806, significantly higher than that of CRF (0.8993) and BioBERT (0.8615). Similarly, in the "I-Disease" entity type, Gemma attained the highest score of 0.9688 compared to CRF's 0.8497 and BioBERT's 0.7990. Gemma also demonstrated exceptional performance in the "B-Chemical" entity type (0.9926), surpassing CRF (0.9429) and BioBERT (0.9310). The "I-Chemical" entity type showed similar trends, with Gemma achieving an F_1-score of 0.9650, higher than that of CRF (0.8611) and BioBERT (0.8405). The CRF model exhibited moderate performance, achieving lower F_1-scores than Gemma but often outperforming BioBERT, particularly in the "B-Disease" and "B-Chemical" entity types.

In the AnatEM dataset, BioBERT consistently outperformed CRF and Gemma across most entity types. For instance, in the "B-Cell" entity type, BioBERT achieved an F_1-score of 0.9053, higher than that of both CRF (0.7879) and Gemma (0.8972). Similarly, in the "I-Cell" entity type, BioBERT attained the highest score of 0.9139 compared to CRF's 0.81 and Gemma's 0.8907. Gemma, while showing high performance in specific entity types such as "B-Cancer" (0.9259) and "I-Cancer" (0.9132), demonstrated more variability across different entity types. For example, Gemma's performance in the "I-Organ" entity type was lower (0.5714) than BioBERT's (0.6780). Although the CRF model showed promising results in specific entity types such as "B-Cancer" (0.844), it generally performed lower than BioBERT and Gemma. Mainly, CRF's performance in types such as "I-Organ" (0.2917) and "I-Developing_anatomical_structure" (0.00) was significantly lower than that of both BioBERT and Gemma.

Figure 4. Prediction accuracy of various models on two datasets. (A) The figure shows the prediction accuracy on the Revised Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) dataset. (B) The figure shows the prediction accuracy on the BioCreative V CDR (BC5CDR) dataset. AVG_MACRO: macroaverage; AVG_MICRO: microaverage; BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; CRF: conditional random fields.

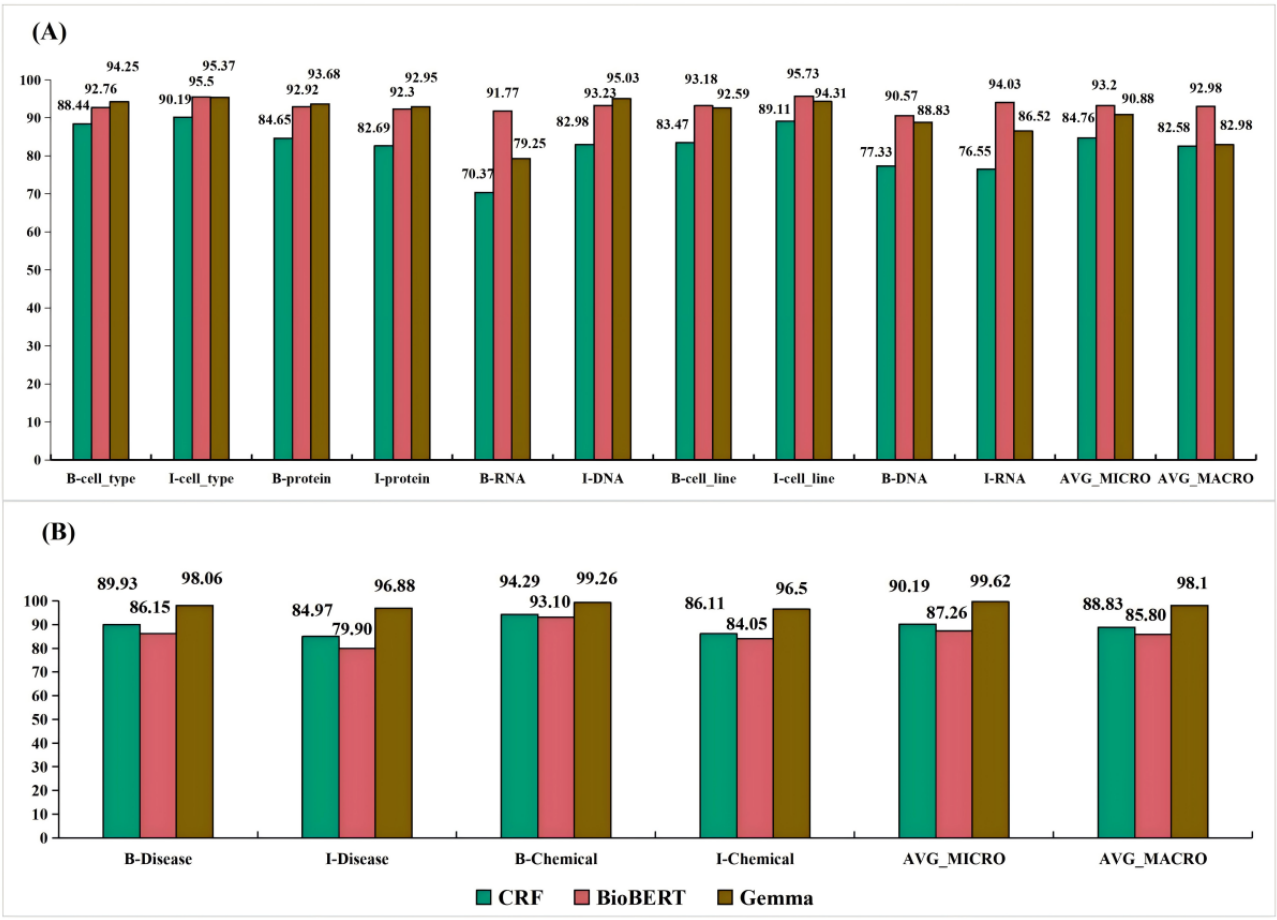
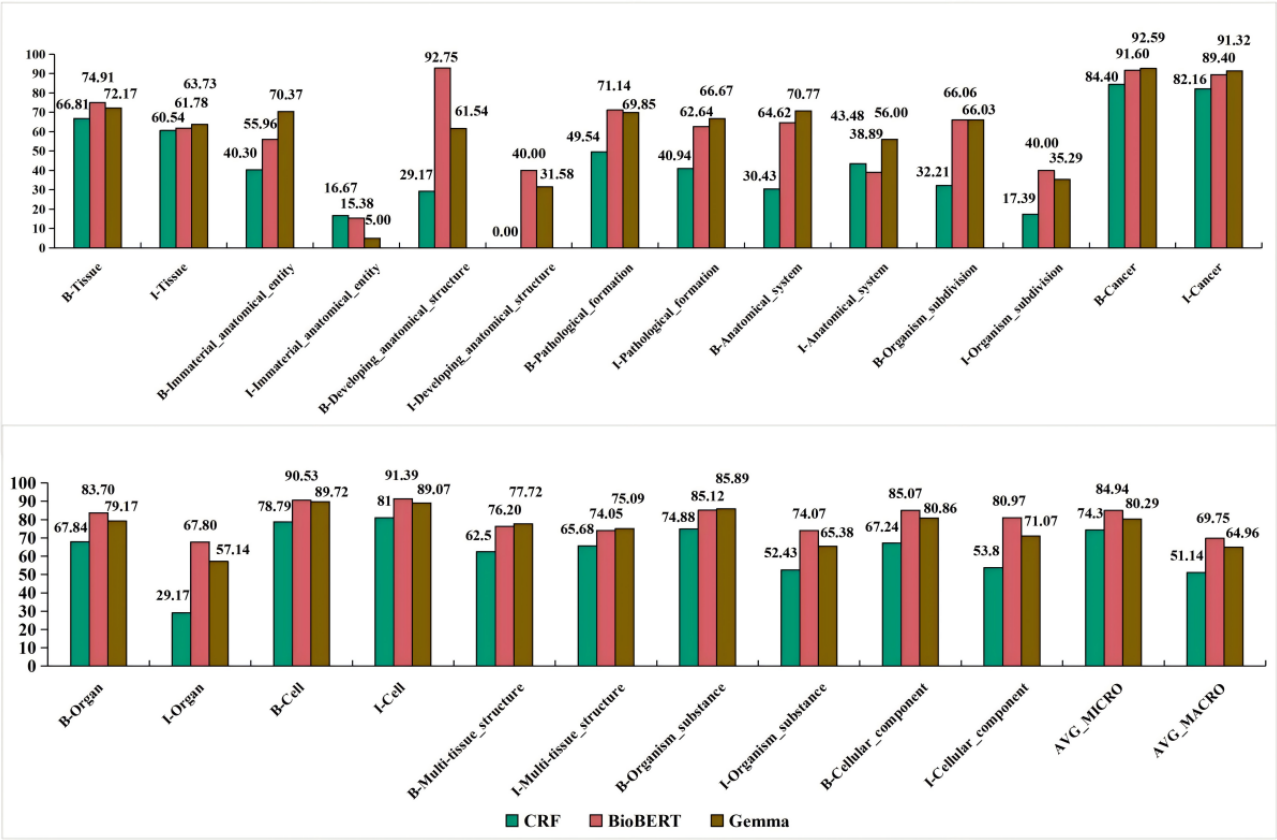


Figure 5. Prediction accuracy of various models on the Anatomical Entity Mention (AnatEM) dataset. AVG_MACRO: macroaverage; AVG_MICRO: microaverage; BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; CRF: conditional random fields.



Fine-Tuning Results of the BERT-Based Models

On the basis of the observations from Figures 6 and 7, the fine-tuning of models based on BERT architecture (BioBERT, BigBird, DeBERTa, and RoBERTa) across the Revised JNLPBA, BC5CDR, and AnatEM datasets demonstrated that a learning rate of 0.0001 consistently yielded the best prediction accuracy. However, the optimal configurations for batch size, epochs, and dropout rate varied significantly among the models and datasets.

BioBERT achieved an AVG_MICRO of 0.932 on the Revised JNLPBA dataset with a batch size of 50, a dropout rate of 0.5, and 5 epochs. On the BC5CDR dataset, it reached an AVG_MICRO of 0.8726 with a lower dropout rate of 0.1, a batch size of 10, and 10 epochs, whereas on the AnatEM dataset, an AVG_MICRO of 0.8494 was obtained with a dropout rate of 0.1, a batch size of 50, and 5 epochs.

BigBird consistently performed best with a batch size of 50 and 10 epochs, achieving an AVG_MICRO of 0.9277 on the Revised JNLPBA dataset. It also showed a strong performance on the BC5CDR and AnatEM datasets, with AVG_MICRO scores of 0.8461 and 0.8147, respectively.

Figure 6. Fine-tuning results of Bidirectional Encoder Representations from Transformers (BioBERT) and Big Transformer Models for Efficient Long-Sequence Attention (BigBird)—microaverage (AVG_MICRO) scores across datasets. (A) The figure shows AVG_MICRO scores for BioBERT across datasets. (B) The figure shows AVG_MICRO scores for BigBird across datasets. AVG_MICRO is used as the sole leveraged metric because both the AVG_MICRO and macroaverage metrics exhibit similar trends, generally increasing or decreasing. This similarity indicates that using AVG_MICRO alone is sufficient to understand the overall performance of the models, making the results more straightforward and focused. AnatEM: Anatomical Entity Mention; BC5CDR: BioCreative V CDR; BERT: Bidirectional Encoder Representations From Transformers; JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

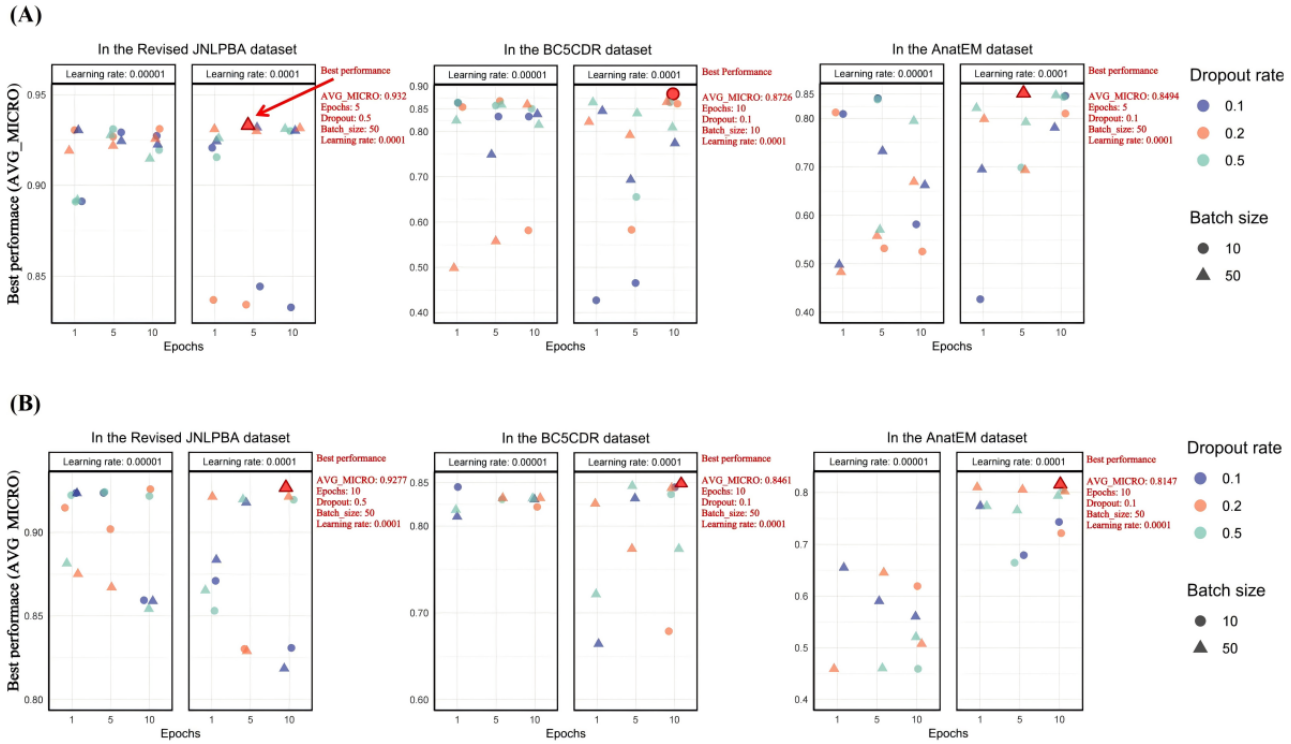
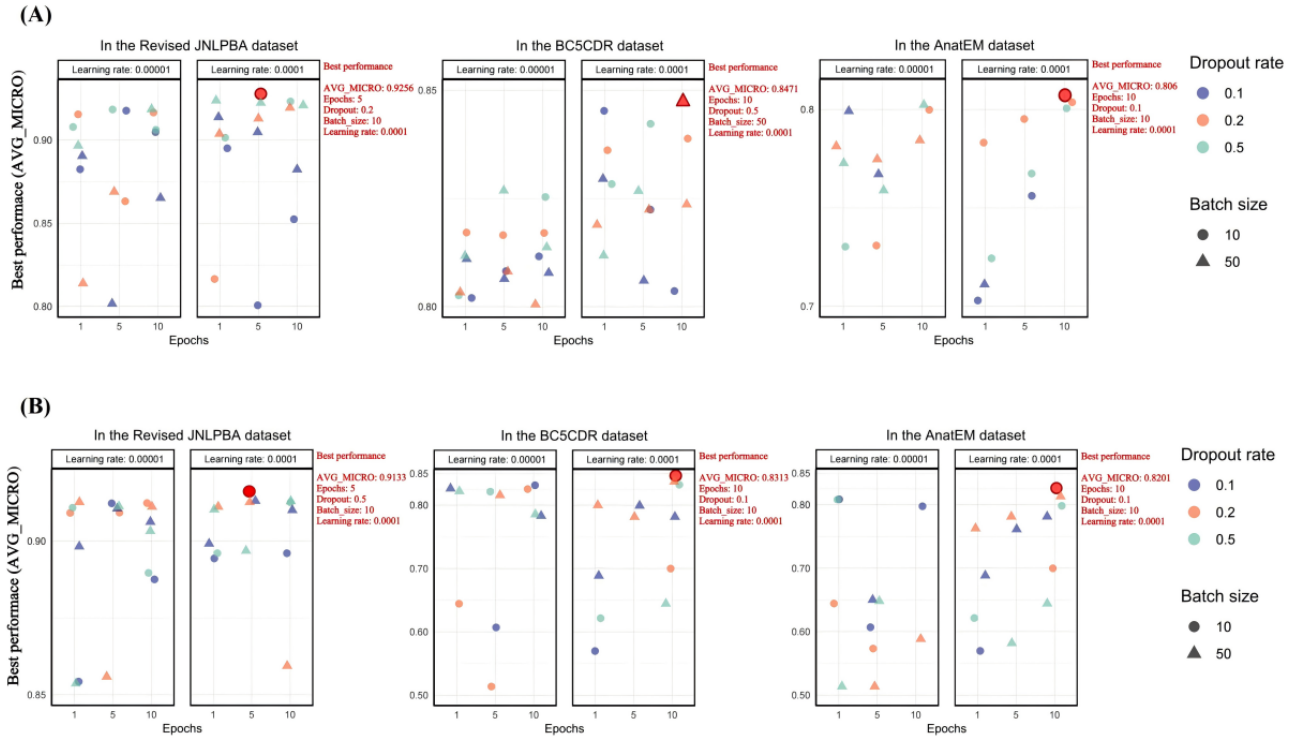


Figure 7. Fine-tuning results of the Decoding-enhanced Bidirectional Encoder Representations from Transformers (BERT) with Disentangled Attention (DeBERTa) and Robustly Optimized BERT Pretraining Approach (RoBERTa) models—microaverage (AVG_MICRO) scores across datasets. (A) The figure shows AVG_MICRO scores for DeBERTa across datasets. (B) The figure shows AVG_MICRO scores for RoBERTa across datasets. AnatEM: Anatomical Entity Mention; BC5CDR: BioCreative V CDR; JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.



DeBERTa recorded an AVG_MICRO of 0.9256 on the Revised JNLPBA dataset with a batch size of 10, a dropout rate of 0.2, and 5 epochs. On the BC5CDR dataset, it achieved an AVG_MICRO of 0.8471 with a higher dropout rate of 0.5, a batch size of 50, and 10 epochs. For the AnatEM dataset, an AVG_MICRO of 0.806 was achieved with a dropout rate of 0.1, a batch size of 10, and 10 epochs.

RoBERTa demonstrated optimal performance across the 3 datasets using a batch size of 10. On the Revised JNLPBA dataset, RoBERTa achieved an AVG_MICRO of 0.9313 with a dropout rate of 0.5 and 5 epochs. On the BC5CDR dataset, it obtained an AVG_MICRO of 0.8313 with a dropout rate of 0.1 and 10 epochs. Similarly, on the AnatEM dataset, RoBERTa achieved an AVG_MICRO of 0.8201 with a dropout rate of 0.1 and 10 epochs.

These results underscore the critical role of hyperparameter tuning as each model and dataset required specific configurations to achieve optimal performance. While a consistent learning rate of 0.0001 was effective across all models, batch size, epoch, and dropout rate variations were necessary to adapt to the specific characteristics of different datasets and model architectures.

Fine-Tuning Results of the Gemma Model

Table 4 shows significant improvements in the Gemma model's performance metrics after fine-tuning using the LoRA technique across various datasets. The AVG_MICRO metric increased by 0.0245 for the Revised JNLPBA dataset, by 0.0111 for the BC5CDR dataset, and by 0.0083 for the AnatEM dataset. Similarly, the AVG_MACRO metric increased by 0.0173 for the Revised JNLPBA dataset, decreased by 0.0027 for the BC5CDR dataset, and increased by 0.0192 for the AnatEM dataset.

Table 4. Comparison of prediction accuracy (as microaverage [AVG_MICRO] and macroaverage [AVG_MACRO] metrics) for the Gemma model before and after fine-tuning on different datasets.

Model, dataset, and leveraged metrics	Before fine-tuning	After fine-tuning
Revised JNLPBA[a]		
AVG_MICRO	0.8843	0.9088
AVG_MACRO	0.8125	0.8298
BC5CDR[b]		
AVG_MICRO	0.9851	0.9962
AVG_MACRO	0.9837	0.9810
AnatEM[c]		
AVG_MICRO	0.7946	0.8029
AVG_MACRO	0.6304	0.6496

[a]JNLPBA:Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

[b]BC5CDR: BioCreative V CDR.

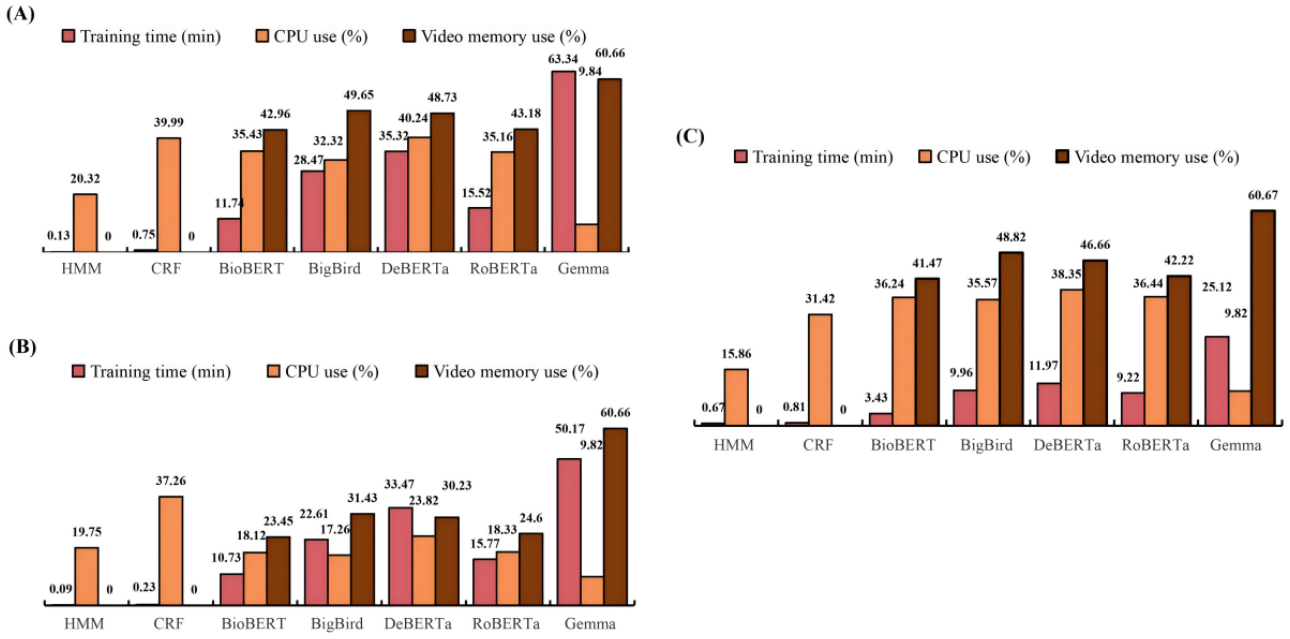
[c]AnatEM: Anatomical Entity Mention.

Model Resource Use Results

Overview

In evaluating prediction accuracy among the HMM, CRF, RoBERTa, BigBird, DeBERTa, BioBERT, and Gemma models, we documented their training time, CPU use, and GPU memory consumption performance, as shown in Figure 8.

Figure 8. Training time and central processing unit (CPU) and graphics processing unit (GPU) uses of the models. (A) The figure shows the resource use results on the Revised Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) dataset. (B) The figure shows the resource use results on the BioCreative V CDR (BC5CDR) dataset. (C) The figure shows the resource use results on the Anatomical Entity Mention (AnatEM) dataset. BERT: Bidirectional Encoder Representations from Transformers; BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; CRF: conditional random fields; DeBERTa: Decoding-enhanced BERT with Disentangled Attention; HMM: hidden Markov model; RoBERTa: Robustly Optimized BERT Pretraining Approach; BigBird: Big Transformer Models for Efficient Long-Sequence Attention.



Training Time

Gemma consistently required the most extended training times across all datasets, reaching a peak of 63.34 minutes in the Revised JNLPBA dataset. In stark contrast, HMM was notably more efficient, completing training in only 0.13 minutes in the same dataset and a mere 0.09 minutes in the BC5CDR dataset.

CPU Use

DeBERTa recorded the highest CPU use—40.24% and 38.35% in the Revised JNLPBA and AnatEM datasets, respectively. In the BC5CDR dataset, CRF had the highest CPU use at 37.26%. In contrast, Gemma demonstrated minimal CPU requirements, with use rates consistently at approximately 10% across all 3 datasets.

GPU Memory Consumption

Gemma had the highest GPU use, with rates consistently at approximately 61% across all 3 datasets. In contrast, the HMM and CRF traditional machine learning models recorded zero GPU use. This is because these models primarily rely on the CPU for their computations and do not leverage the parallel processing capabilities of GPUs, which are designed to accelerate deep learning tasks. Among the BERT-based models, BioBERT exhibited the lowest GPU use, consuming 42.96% and 41.47% in the Revised JNLPBA and AnatEM datasets, respectively, and only 23.45% in the BC5CDR dataset. This lower GPU consumption indicates that BioBERT, while still using GPU resources, does so more efficiently than Gemma, potentially due to its optimized architecture and more efficient use of GPU memory for processing.

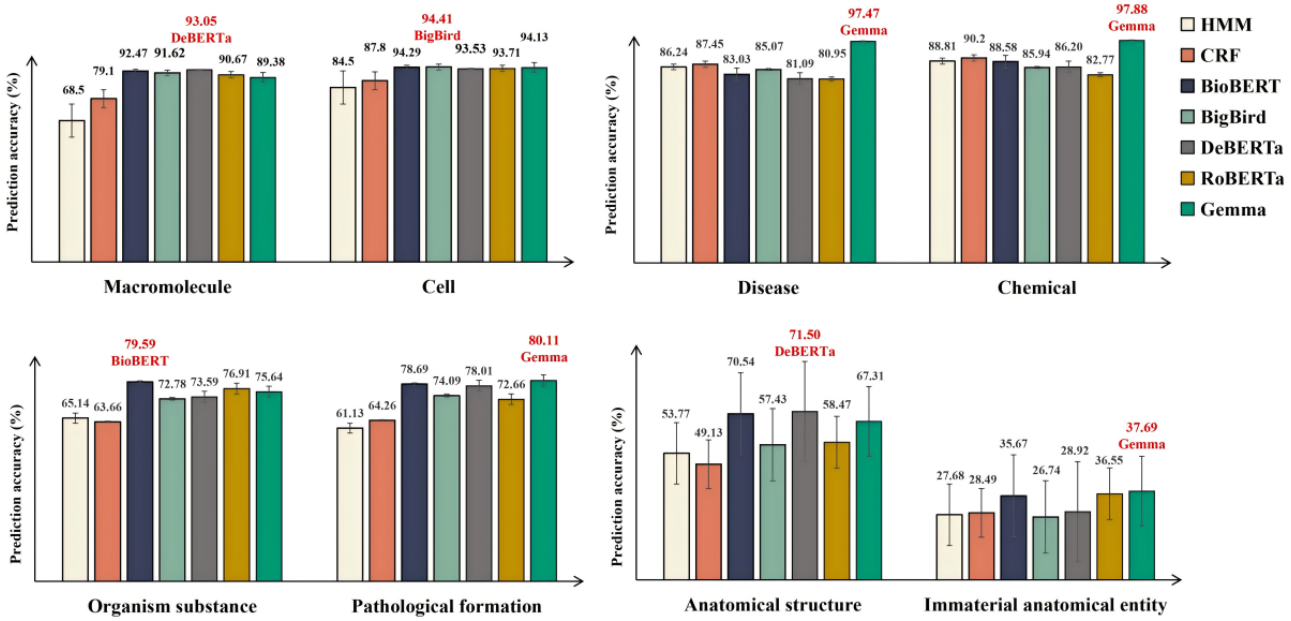
Overall

Gemma required significant resources and was characterized by the highest GPU use across all datasets and the most extended training times. This computational intensity may enhance accuracy but comes at the cost of increased operational resources. In contrast, BioBERT demonstrated high prediction accuracy and lower resource consumption among the deep learning models, indicating its efficiency and suitability for environments with strict resource constraints.

Entity Type Prediction Accuracy Results

On the basis of the data presented in Figure 9, we identified specific strengths in different models using a categorical relaxation method to merge entity types. Gemma achieved the highest prediction accuracy in the “chemical,” “disease,” “pathological formation,” and “immaterial anatomical entity” types, whereas BioBERT excelled in the “organism substance” entity type. DeBERTa performed best in the “macromolecule” and “anatomical structure” entity types, demonstrating its strengths. In addition, BigBird showed a superior performance in the “cell” type.

Figure 9. Relationship between the models' prediction accuracy and merged entity types. BERT: Bidirectional Encoder Representations from Transformers; BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; CRF: conditional random fields; DeBERTa: Decoding-enhanced BERT with Disentangled Attention; HMM: hidden Markov model; RoBERTa: Robustly Optimized BERT Pretraining Approach; BigBird: Big Transformer Models for Efficient Long-Sequence Attention.

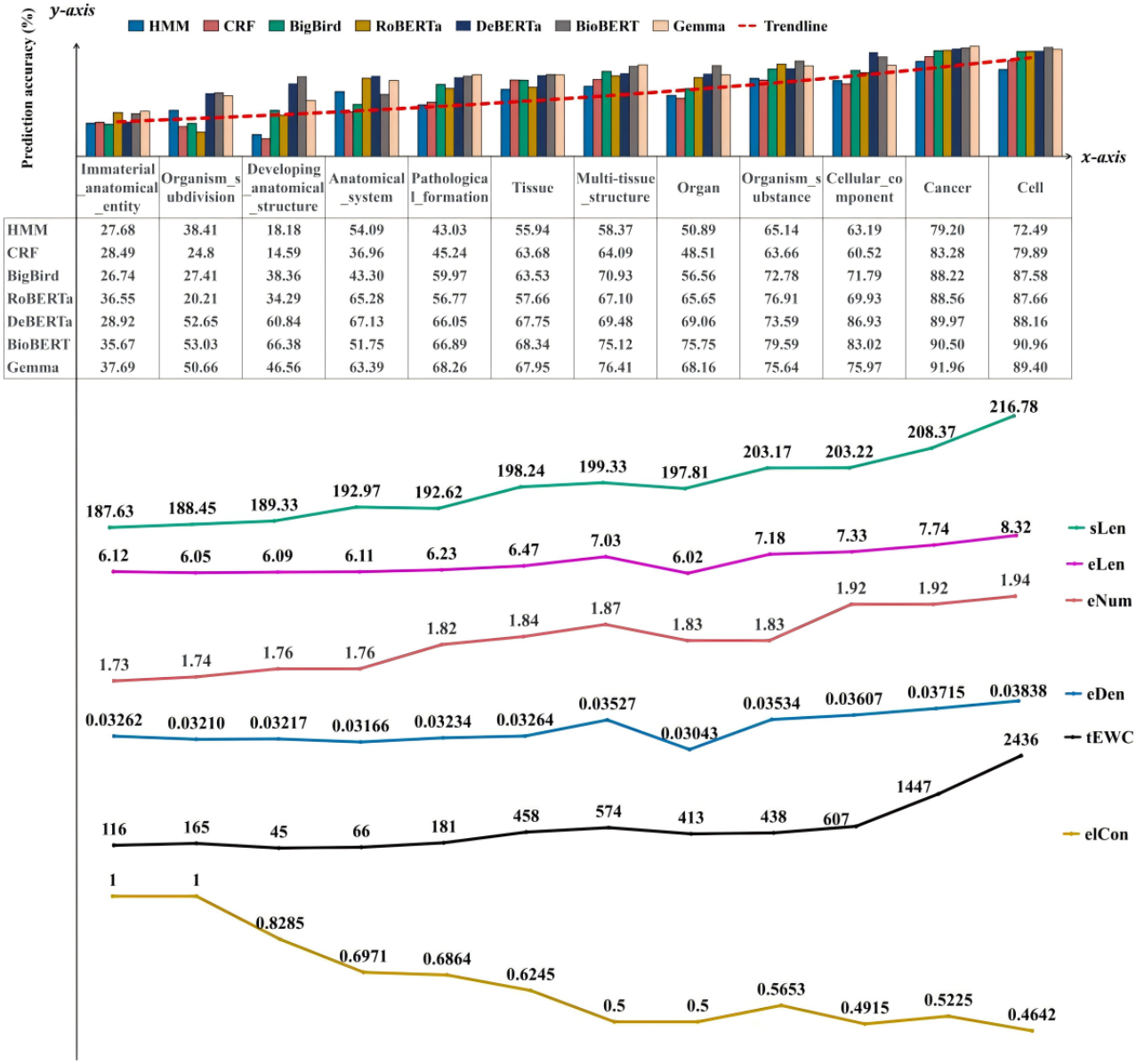


Macrofactor Trends Impacting Model Prediction Accuracy

The results of this section primarily involve the trend relationships between the prediction accuracy of 7 NER models and 6 macrofactor metrics, focusing on the average values of these metrics across various datasets.

In the AnatEM dataset, as depicted in Figure 10, there was a distinct correlation—entity types that yielded higher prediction accuracies corresponded to increased values in *sLen*, *eLen*, *eNum*, *eDen*, and *tEWC*. This pattern indicates that the model's ability to accurately predict entities is enhanced with the rising complexity and volume of data associated with those entity types. Conversely, an inverse relationship was observed with *elCon*, which decreased as the other metrics increased. For instance, BioBERT recorded a high prediction accuracy of 90.96% in the "cell" entity type. This outstanding performance correlated with the highest metrics observed in the dataset—*sLen* peaked at 216.78, *eLen* peaked at 8.32, *eNum* peaked at 1.94, and *eDen* peaked at 0.03838. Then, the "cell" entity type's *tEWC* was notably high at 2436, demonstrating the BioBERT's ability to process extensive textual data effectively. However, *elCon* was significantly lower at 0.4642. This indicates that, although this model is adept at managing complex and voluminous data, it does not consistently ensure accurate entity labeling, suggesting a trend in which higher accuracy and complexity metrics do not correspond with label consistency.

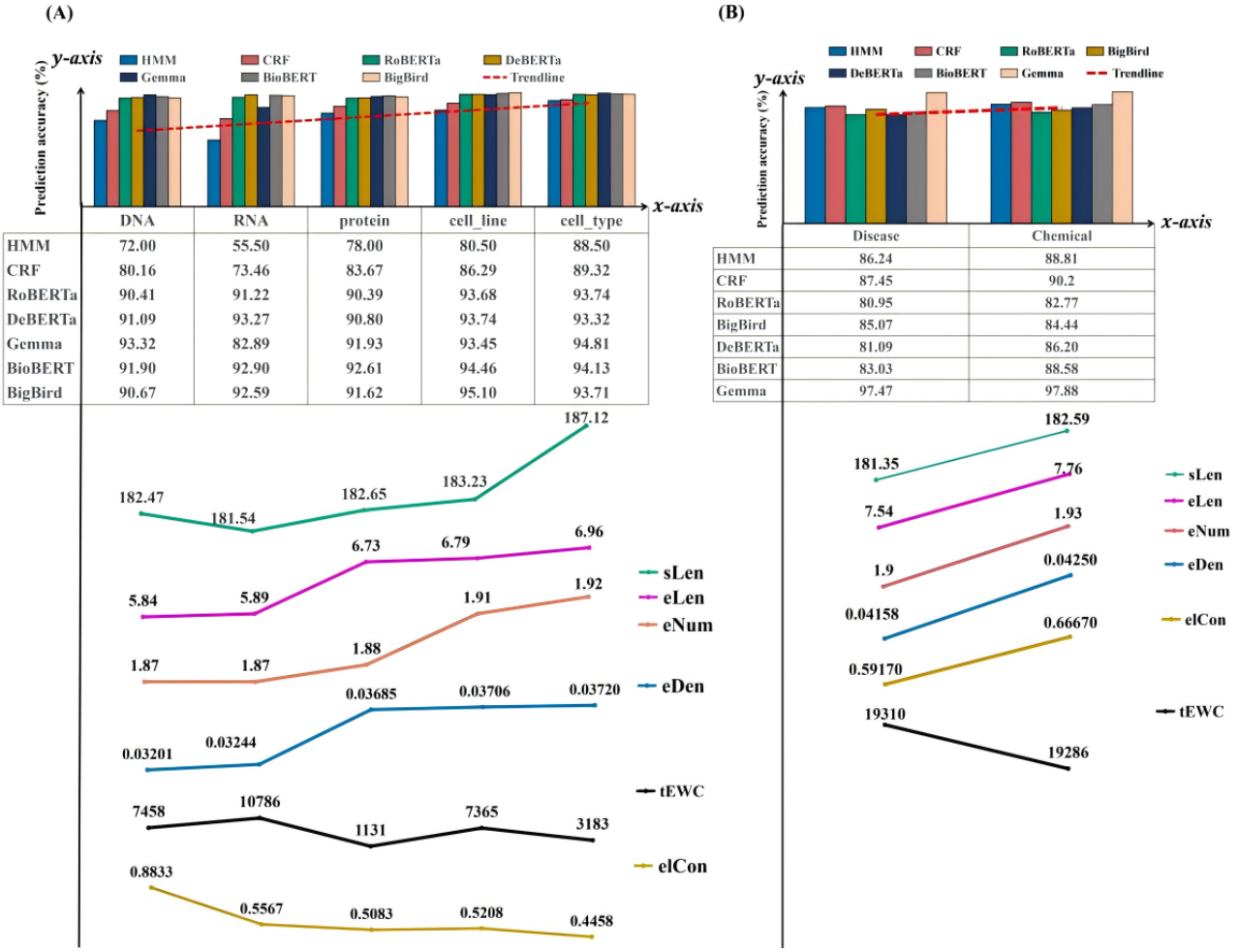
Figure 10. Relationship between macrofactors and model prediction accuracy in the Anatomical Entity Mention (AnatEM) dataset. BERT: Bidirectional Encoder Representations from Transformers; BigBird: Big Transformer Models for Efficient Long-Sequence Attention; BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; CRF: conditional random fields; DeBERTa: Decoding-enhanced BERT with Disentangled Attention; eDen: entity density; elCon: entity label consistency; eLen: entity phrase length; eNum: number of entity words in each entity phrase; HMM: hidden Markov model; RoBERTa: Robustly Optimized BERT Pretraining Approach; sLen: sentence length; tEWC: total entity word count in each entity type.



	Immaterial_anatomical_entity	Organism_subdivision	Developing_anatomical_structure	Anatomical_system	Pathological_formation	Tissue	Multi-tissue_structure	Organ	Organism_substance	Cellular_component	Cancer	Cell
HMM	27.68	38.41	18.18	54.09	43.03	55.94	58.37	50.89	65.14	63.19	79.20	72.49
CRF	28.49	24.8	14.59	36.96	45.24	63.68	64.09	48.51	63.66	60.52	83.28	79.89
BigBird	26.74	27.41	38.36	43.30	59.97	63.53	70.93	56.56	72.78	71.79	88.22	87.58
RoBERTa	36.55	20.21	34.29	65.28	56.77	57.66	67.10	65.65	76.91	69.93	88.56	87.66
DeBERTa	28.92	52.65	60.84	67.13	66.05	67.75	69.48	69.06	73.59	86.93	89.97	88.16
BioBERT	35.67	53.03	66.38	51.75	66.89	68.34	75.12	75.75	79.59	83.02	90.50	90.96
Gemma	37.69	50.66	46.56	63.39	68.26	67.95	76.41	68.16	75.64	75.97	91.96	89.40

The Revised JNLPBA dataset substantiated these observations, as shown in Figure 11. The "cell_type" entity type generally yielded a higher accuracy among the 7 models, showing elevated values for *sLen*, *eLen*, *eNum*, and *eDen*, with average values recorded at 187.12 (SD 10.52), 7.76 (SD 1.24), 1.91 (SD 0.32), and 0.0372 (SD 0.005), respectively. This indicated that the models achieved high accuracy and effectively handled denser entity distributions. Moreover, there seemed to be a negative correlation between higher values of *elCon* and *tEWC* and these accuracies. In addition, a comparable trend emerged in the BC5CDR dataset, where the "chemical" entity type yielded the highest accuracy across the RoBERTa, DeBERTa, and BioBERT models, aligning with the highest measurements of *sLen*, *eLen*, *eNum*, *eDen*, and *tEWC*, coupled with the lowest value of *elCon*.

Figure 11. Relationship between macrofactors and model prediction accuracy in two datasets. (A) The figure shows the relationship in the Revised Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) dataset. (B) The figure shows the relationship in the BioCreative V CDR (BC5CDR) dataset. BERT: Bidirectional Encoder Representations from Transformers; BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining; CRF: conditional random fields; DeBERTa: Decoding-enhanced BERT with Disentangled Attention; eDen: entity density; elCon: entity label consistency; eLen: entity phrase length; eNum: number of entity words in each entity phrase; HMM: hidden Markov model; RoBERTa:Robustly Optimized BERT Pretraining Approach; sLen: sentence length; tEWC: total entity word count in each entity type; BigBird: Big Transformer Models for Efficient Long-Sequence Attention.



Macrofactor Selection Results

Examining Tables 5-7, the results from the MFE algorithm reveal the importance of different macrofactors across various models and datasets. In the CRF model, *eNum* was identified as the most influential macrofactor for prediction accuracy in the Revised JNLPBA and AnatEM datasets. In contrast, *eLen* was found to be more influential in the BC5CDR dataset. This finding indicates that the *eNum* and *eLen* play varying roles depending on the dataset's characteristics. The Gemma model exhibited results similar to those of the CRF model, with *eNum* significantly impacting prediction accuracy for the Revised JNLPBA and AnatEM datasets. At the same time, *eLen* significantly influenced prediction accuracy for the BC5CDR dataset.

In the BioBERT model, *eNum* consistently emerged as the most critical macrofactor across all datasets (Revised JNLPBA, BC5CDR, and AnatEM); despite varying macrofactor combinations in the initial 2 layers, *eNum* was consistently chosen for the final layer. This indicates that, among the 6 macrofactors, *eNum* significantly influences BioBERT's prediction accuracy.

Table 5. Macrofactors selected by the multilevel factor elimination (MFE) algorithm in the conditional random fields model across different datasets.

MFE algorithm	On the basis of the Revised JNLPBA[a] dataset	On the basis of the BC5CDR[b] dataset	On the basis of the AnatEM[c] dataset
Input	sLen[d], eLen[e], eNum[f], eDen[g], elCon[h], and tEWC[i]	sLen, eLen, eNum, eDen, elCon, and tEWC	sLen, eLen, eNum, eDen, elCon, and tEWC
Layer 1	sLen, eLen, eNum, and elCon	eLen, eNum, eDen, and tEWC	sLen, eLen, eNum, and eDen
Layer 2	sLen and eNum	eLen and eNum	eNum and eDen
Layer 3	eNum	eLen	eNum

[a]JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

[b]BC5CDR: BioCreative V CDR.

[c]AnatEM: Anatomical Entity Mention.

[d]sLen: sentence length.

[e]eLen: entity phrase length.

[f]eNum: number of entity words in each entity phrase.

[g]eDen: entity density.

[h]elCon: entity label consistency.

[i]tEWC: total entity word count in each entity type.

Table 6. Macrofactors selected by the multilevel factor elimination (MFE) algorithm in the Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) model across different datasets.

MFE algorithm	On the basis of the Revised JNLPBA[a] dataset	On the basis of the BC5CDR[b] dataset	On the basis of the AnatEM[c] dataset
Input	sLen[d], eLen[e], eNum[f], eDen[g], elCon[h], and tEWC[i]	sLen, eLen, eNum, eDen, elCon, and tEWC	sLen, eLen, eNum, eDen, elCon, and tEWC
Layer 1	eLen, eNum, elCon, and tEWC	eLen, eNum, eDen, and tEWC	sLen, eLen, eNum, and eDen
Layer 2	eLen and eNum	eNum and tEWC	eLen and eNum
Layer 3	eNum	eNum	eNum

[a]JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

[b]BC5CDR: BioCreative V CDR.

[c]AnatEM: Anatomical Entity Mention.

[d]sLen: sentence length.

[e]eLen: entity phrase length.

[f]eNum: number of entity words in each entity phrase.

[g]eDen: entity density.

[h]elCon: entity label consistency.

[i]tEWC: total entity word count in each entity type.

Table 7. Macrofactors selected by the multilevel factor elimination (MFE) algorithm in the Gemma model across different datasets.

MFE algorithm	On the basis of the Revised JNLPBA[a] dataset	On the basis of the BC5CDR[b] dataset	On the basis of the AnatEM[c] dataset
Input	sLen[d], eLen[e], eNum[f], eDen[g], elCon[h], and tEWC[i]	sLen, eLen, eNum, eDen, elCon, and tEWC	sLen, eLen, eNum, eDen, elCon, and tEWC
Layer 1	sLen, eNum, elCon, and tEWC	eLen, eNum, eDen, and elCon	sLen, eLen, eNum, and tEWC
Layer 2	sLen and eNum	eLen and eNum	eLen and eNum
Layer 3	eNum	eLen	eNum

[a]JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications.

[b]BC5CDR: BioCreative V CDR.

[c]AnatEM: Anatomical Entity Mention.

[d]sLen: sentence length.

[e]eLen: entity phrase length.

[f]eNum: number of entity words in each entity phrase.

[g]eDen: entity density.

[h]elCon: entity label consistency.

[i]tEWC: total entity word count in each entity type.

Discussion

Principal Findings

This study comprehensively evaluated various NER models in medical informatics, focusing on prediction accuracy, resource use, and the impact of macrofactors and hyperparameters. The primary findings indicate that BERT-based models (BioBERT, RoBERTa, BigBird, and DeBERTa) exhibited generally higher accuracy than traditional statistical models (HMM and CRF). These BERT-based models, fine-tuned using the ADAM optimizer with a consistent learning rate of 0.0001, demonstrated outstanding performance. Among them, BioBERT excelled due to its specialized pretraining on extensive medical literature. The Gemma LLM, fine-tuned using the LoRA technique, achieved the highest accuracy on the BC5CDR dataset but showed variability across the other datasets, highlighting the need for further optimization. Macrofactors such as "entity phrase length (*eLen*)" and "the number of entity words in each entity phrase (*eNum*)" significantly influenced model performance, with the MFE algorithm effectively filtering these macrofactors. Computational resource use revealed that, while Gemma required substantial resources, BioBERT was a balanced NER model with high prediction accuracy and lower computational demands, making it suitable for resource-constrained environments. These findings underscore the importance of continuous refinement and dataset-specific optimization to advance NER capabilities in medical informatics.

Evaluation of Predictive Performance Across NER Models

Evaluating various NER models on medical datasets provides significant insights into their strengths and limitations, highlighting the complexity of accurately identifying and categorizing medical entities. The performance disparity among statistical machine learning models, deep learning models based on BERT architecture, and the Gemma LLM is evident, with the latter 2 consistently demonstrating superior accuracy and robustness.

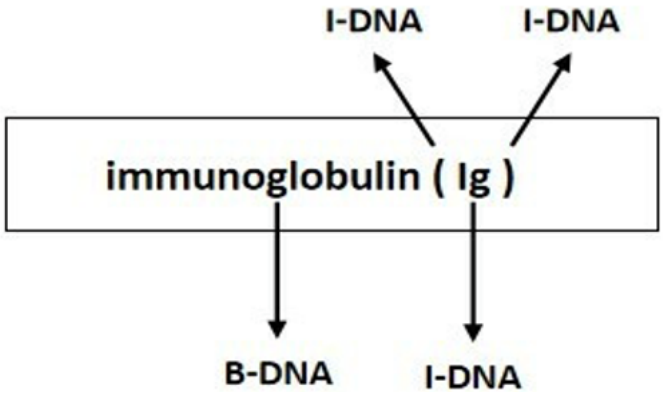
Statistical machine learning models such as HMM and CRF exhibited a moderate performance in medical NER tasks. The HMM's relatively lower scores across all datasets, with AVG_MICRO scores of 0.7265, 0.825, and 0.6112 for the Revised JNLPBA, BC5CDR, and AnatEM datasets, respectively, reflected its limited capacity to capture the nuanced patterns inherent in medical texts. While CRF outperformed HMM with higher AVG_MICRO scores of 0.8476, 0.9019, and 0.743 for the same datasets, respectively, it still fell short compared to the deep learning models. This indicates inherent limitations in CRF's feature engineering and sequence modeling capabilities, which are less adept at handling the complexity and variability of medical entities.

In contrast, the BERT-based models (BioBERT, RoBERTa, BigBird, and DeBERTa) consistently demonstrated a strong performance across multiple datasets. BioBERT's superior performance on the Revised JNLPBA and AnatEM datasets, with AVG_MICRO scores of 0.932 and 0.8494 and AVG_MACRO scores of 0.9298 and 0.6975, respectively, can be attributed to its specialized pretraining on large-scale medical literature. This pretraining allows BioBERT to capture domain-specific nuances, enhancing its ability to recognize and classify complex medical entities accurately. In the Revised JNLPBA dataset, BioBERT achieved the highest F_1-scores in categories such as "B-DNA" (0.9057) and "I-cell_type" (0.9550), showcasing its robust generalization capabilities. However, the slightly lower performance of these models on the BC5CDR dataset, where AVG_MICRO and AVG_MACRO scores were generally below 0.8726 and 0.858, respectively, suggests potential overfitting or sensitivity to the specific characteristics of this dataset. This variability underscores the necessity for further fine-tuning and incorporating additional domain-specific knowledge to improve the models' generalization capabilities across diverse datasets.

The Gemma LLM demonstrated exceptional performance on the BC5CDR dataset, achieving the highest AVG_MICRO score of 0.9962 and an AVG_MACRO score of 0.981. In specific entity types within this dataset, Gemma attained F_1-scores of 0.9806 in "B-Disease" and 0.9926 in "B-Chemical," indicating its strong capability to recognize these entities. This suggests that Gemma's architecture and training regimen are particularly well suited for this dataset, which contains only 2 entity types (disease and chemical). The narrow focus of the BC5CDR dataset may have allowed Gemma to optimize its performance more effectively than on more complex datasets with diverse entity types. However, Gemma's less pronounced results on the Revised JNLPBA and AnatEM datasets, with AVG_MICRO scores of 0.9088 and 0.8029 and AVG_MACRO scores of 0.8298 and 0.6496, respectively, indicate that, while it excels in binary entity recognition, it may require further adjustments to handle more complex, multi-entity datasets effectively. In the Revised JNLPBA dataset, for instance, Gemma showed a high performance in categories such as "B-protein" (0.9368) and "I-DNA" (0.9503) but struggled in "B-RNA" (0.7925) compared to BioBERT's 0.9177, highlighting areas for potential improvement.

Furthermore, we found significant variability in prediction accuracy across different entity types for models such as BioBERT, Gemma, and CRF, as evidenced by contrasting F_1-scores for entities like 'B-Developing_anatomical_structure' and 'I-Immaterial_anatomical_entity.' This inconsistency, illustrated in Figures 4 and 5, is primarily attributed to 2 interrelated factors: uneven data distribution and the complexities inherent in medical terminologies. First, uneven data distribution and noise in datasets such as the Revised JNLPBA, BC5CDR, and AnatEM impeded the models' ability to fully understand and accurately predict underrepresented entity types, affecting their overall performance. For example, the "protein" and "cell_type" entities in the Revised JNLPBA dataset are significantly more abundant (5256 and 2070 entity words, respectively) compared to "cell_line" and "RNA" (404 and 161 entity words, respectively). Second, multiple entity types for a single entity word complicate prediction. For instance, in the Revised JNLPBA dataset, certain symbols function as entity words within entity phrases. As depicted in Figure 12, the parentheses in the entity phrase "immunoglobulin (Ig)" from the training set are categorized as the "I-DNA" entity type. However, when predicting the entity types of parentheses in the test set, the models occasionally misclassify them as "I-DNA" even though they do not belong to this entity type in some contexts. Specifically, the phrase "(PCBA)" in the test set is not recognized as an entity phrase. However, the models may erroneously assign the parentheses to the "I-DNA" entity type, leading to prediction errors.

Figure 12. Entity type division.



These findings underscore the importance of selecting and optimizing models based on the specific characteristics of the datasets to which they are applied. The consistent performance of the deep learning models, particularly those based on BERT architecture, and the promising results from LLMs such as Gemma highlight the potential for advanced NER models to improve medical entity recognition significantly. However, achieving optimal results across diverse medical contexts will require continuous refinement and adaptation of these models to the unique nuances of each dataset.

Impact of Hyperparameters and Optimizer on the Fine-Tuning of the BERT-Based Models

The fine-tuning results of the BERT-based models (BioBERT, BigBird, DeBERTa, and RoBERTa) across the Revised JNLPBA, BC5CDR, and AnatEM datasets reveal the significant role of hyperparameter tuning and the effectiveness of the ADAM optimizer in achieving optimal prediction accuracy in medical NER tasks.

The consistent use of the ADAM optimizer with a learning rate of 0.0001 across all models demonstrated its efficacy in fine-tuning BERT-based architectures. The ADAM optimizer's adaptive learning rate mechanism, which adjusts based on the first and second moments of the gradients, ensures efficient and effective convergence. This balance is crucial for fine-tuning complex models such as BERT and its variants as it mitigates the risk of overshooting the optimal point or converging too slowly. The success of this learning rate underscores the importance of selecting a rate that balances convergence speed and accuracy.

Beyond the learning rate, hyperparameters such as batch size, epochs, and dropout rate were critical in determining model performance. Optimal configurations for these parameters varied significantly across models and datasets, emphasizing the need for dataset-specific and model-specific tuning. For instance, BioBERT achieved an AVG_MICRO of 0.932 on the Revised JNLPBA dataset with a batch size of 50, a dropout rate of 0.5,

and 5 epochs. In contrast, a smaller batch size of 10 was necessary for optimal prediction accuracy on the BC5CDR dataset. This variation illustrates how dataset characteristics heavily influence batch size selection; larger batch sizes can improve computational efficiency by leveraging the parallel processing capabilities of GPUs, whereas smaller batch sizes may allow for more frequent updates and better convergence.

The number of epochs required to achieve optimal prediction accuracy also showed variability, indicating the need for tailored training durations based on dataset complexity. BioBERT needed 5 epochs for the Revised JNLPBA dataset but 10 epochs for the BC5CDR dataset, suggesting that more training iterations are essential for specific datasets to fully capture their complexity and variability. Different datasets may demand varied training durations to achieve optimal performance.

The dropout rate, a critical parameter to prevent overfitting, exhibited significant variation across models and datasets. BioBERT, BigBird, and RoBERTa demonstrated better prediction accuracy with a lower dropout rate of 0.1 on the BC5CDR dataset, whereas a higher dropout rate of 0.5 was optimal for the Revised JNLPBA dataset. This variation underscores the differing overfitting risks and regularization needs across datasets. Higher dropout rates can prevent overfitting in more straightforward datasets but may hinder performance in more complex datasets requiring nuanced learning.

These findings emphasize the critical role of hyperparameter tuning in fine-tuning BERT-based models. While the ADAM optimizer with a learning rate of 0.0001 proved adequate, batch size, epochs, and dropout rate required careful adjustment to the specific characteristics of each dataset and model architecture. This variability highlights the necessity of a tailored approach in hyperparameter optimization to achieve optimal prediction accuracy. The consistent performance of deep learning models based on BERT architecture underscores their potential for significantly improving medical NER tasks. However, achieving optimal results across diverse medical contexts will require continuous refinement and adaptation of these models to the unique nuances of each dataset.

Impact of Hyperparameters and Optimizers on the Fine-Tuning of the Gemma Model

The fine-tuning of the Gemma model using the LoRA technique led to significant improvements in its performance metrics across various medical NER tasks, as shown in Table 4. The consistent increases in AVG_MICRO metrics across 3 datasets demonstrate the effectiveness of LoRA in enhancing model accuracy. Although the AVG_MACRO metric declined in one dataset, it still improved in the other two, reaffirming LoRA's overall positive impact on model performance.

LoRA involves freezing the pretrained model weights and introducing trainable low-rank decomposition matrices in each transformer layer, significantly reducing the number of trainable parameters and GPU memory requirements. By optimizing these low-rank matrices, LoRA adapts the model to specific tasks without the computational burden of retraining the entire model. This technique focuses on adjusting the weights of hyperparameters rather than changing their values, allowing for more precise and efficient fine-tuning. The learning rate, managed using a cosine learning rate scheduler with a warm-up ratio of 0.1, ensures stable and efficient convergence. The consistent use of a learning rate of 0.0001 across various datasets demonstrated its effectiveness in balancing convergence speed and accuracy.

Hyperparameters such as batch size, epochs, and dropout rate were critical in determining model performance. Although their values remained constant during the LoRA fine-tuning process, the adjustments made by LoRA allowed the model to adapt effectively to the specific characteristics of each dataset. For instance, the Revised JNLPBA dataset benefited from a larger batch size and fewer epochs. In comparison, the BC5CDR dataset required a smaller batch size and more epochs to capture its complexities adequately. This variability underscores the importance of dataset-specific configurations in hyperparameter tuning to maximize model performance.

The dropout rate, crucial for preventing overfitting, also showed a significant impact depending on the dataset's structure. For example, the Revised JNLPBA dataset, with its more straightforward structure, benefited from its original higher dropout rate, providing more robust regularization. Conversely, the BC5CDR and AnatEM datasets, with their more complex structures, yielded a better performance with their original lower dropout rates, allowing the model to retain more information during training. This finding highlights the necessity of understanding the specific regularization needs of each dataset to optimize model performance effectively.

The consistent improvement in AVG_MICRO and AVG_MACRO metrics demonstrates the enhanced generalization capability of the Gemma model after fine-tuning. These improvements are particularly significant for medical NER tasks, where precise entity recognition is critical for extracting meaningful information from complex texts. The results suggest that LoRA enables Gemma to capture nuanced patterns within the datasets more effectively, leading to better prediction accuracy and balanced performance across different entity types. Furthermore, the fine-tuning of the Gemma model using the LoRA technique highlights the critical role of weight adjustment in hyperparameter tuning and advanced optimization strategies in improving model performance. The significant improvements observed emphasize the potential of LoRA in efficiently adapting large-scale models to specialized tasks, making them more practical for deployment in various medical contexts.

Evaluation of Resource Use Across NER Models

Gemma consistently required the most extended training times across all datasets, peaking at 63.34 minutes in the Revised JNLPBA dataset. This extended training duration reflects Gemma's complex architecture, including numerous layers and parameters. The model's intricate design requires significant processing power and time to adjust its parameters accurately through extensive training iterations.

Regarding CPU use, DeBERTa recorded the highest use in the Revised JNLPBA and AnatEM datasets, with use rates of

40.24% and 38.35%, respectively. DeBERTa's high CPU use indicates its sophisticated architecture, which includes enhanced attention mechanisms and deeper layers compared to other models. These features demand substantial computational resources, leading to increased CPU consumption. The model's reliance on complex attention mechanisms to capture fine-grained dependencies within the text likely contributes to its higher computational intensity, impacting its efficiency in resource-constrained environments.

In the BC5CDR dataset, CRF exhibited the highest CPU use, reaching 37.26%. This high CPU use can be attributed to the intensive computations required for feature extraction and sequence labeling in a dataset with relatively more straightforward entity types. While less complex than the deep learning models, the CRF model still requires significant CPU resources for processing and optimizing the conditional dependencies between the labels, especially when the dataset has clear, well-defined entity types that increase the computational workload.

In contrast, Gemma demonstrated minimal CPU requirements, maintaining use rates of approximately 10% across all datasets. This low CPU use suggests that Gemma offloads most of its computational workload to the GPU, leveraging its parallel processing capabilities to handle the model's complex computations more efficiently. Gemma's architecture is designed to exploit GPU parallelism, which allows for faster processing of large batches of data, thereby reducing the strain on the CPU. Thus, Gemma had the highest GPU use, consistently at approximately 61% across all datasets. This high GPU use underscores Gemma's dependency on GPU resources to manage its computationally intensive tasks. The model's architecture, characterized by numerous layers and a high parameter count, requires substantial GPU resources for the parallel processing needed during training and inference. The reliance on GPUs is due to their ability to handle multiple operations simultaneously, which is essential for deep learning models with large-scale parameters and complex computations.

BioBERT demonstrated the lowest GPU use among the BERT-based models, consuming 42.96% and 41.47% on the Revised JNLPBA and AnatEM datasets, respectively, and only 23.45% on the BC5CDR dataset. BioBERT's lower GPU consumption can be attributed to its optimized architecture and efficient memory management, allowing it to balance computational demands with performance effectively. BioBERT's design leverages GPU resources efficiently, ensuring that it can achieve high accuracy without excessively taxing computational resources.

Overall, Gemma's significant resource requirements, evidenced by the fact that it had the highest GPU use and longest training times, highlight that its enhanced accuracy comes at a substantial cost in operational resources. The model's advanced architecture necessitates extensive computational power to manage the large-scale parallel processing required for its numerous parameters and layers. This makes Gemma suitable for environments in which accuracy and abundant resources are paramount. In contrast, BioBERT's combination of high prediction accuracy and lower resource consumption underscores its efficiency and suitability for environments with strict resource constraints. BioBERT's ability to balance performance with resource use makes it versatile for various medical NER applications. These findings emphasize the importance of selecting NER models based on the deployment environment's specific resource availability and operational constraints. Future research should focus on optimizing the computational efficiency of NER models without compromising their performance, ensuring that they can be effectively deployed across various medical contexts.

Variability in Prediction Accuracy Across Entity Types in NER Models

Gemma demonstrated a superior performance in the entity categories of "chemical," "disease," "pathological formation," and "immaterial anatomical entity," likely due to its advanced design in feature extraction and semantic representation. Gemma uses optimized embedding techniques and attention mechanisms specifically tailored to manage the complexities inherent in these categories. For example, Gemma may use molecular property–based embedding representations in chemical entity recognition, enabling it to capture critical features of chemical substances. Gemma integrates rich medical ontologies and knowledge graphs in the context of diseases and pathological formations, enhancing its understanding of medical terminology and complex pathological descriptions. In addition, Gemma's attention mechanisms focus on critical segments of the text, improving performance in recognizing immaterial anatomical entities.

BioBERT's outstanding performance in the "organism substance" entity category can be attributed to its extensive pretraining in medical literature, particularly that covering organism substances. On the basis of the BERT bidirectional transformer architecture, BioBERT captures long-range dependencies in complex medical texts. The intricate language and contextual information in organism substance literature are effectively parsed by BioBERT's architecture, leading to superior performance in this category. However, BioBERT's performance in the "chemical" category was not as strong as Gemma's, possibly because its pretraining data focus more on organism-related content than on chemical substances' specific features.

DeBERTa's superior performance in the "macromolecule" and "anatomical structure" entity categories was due to its innovative attention mechanisms and position encoding methods. Macromolecules and anatomical structures often exhibit high complexity and a hierarchical nature. DeBERTa's disentangled attention mechanism allows for a fine-grained capture of internal relationships and hierarchical information within these complex structures. Its enhanced position encoding method effectively represents these complex medical entities, leading to an outstanding performance in these categories.

BigBird's excellent performance in the "cell" entity category reflects its ability to handle long texts. Texts in cell biology are typically very detailed and information dense. BigBird's architecture allows for more oversized context windows, enabling effective processing and analysis of these extensive texts. BigBird maintains efficiency and accuracy through its

sparse attention mechanism, resulting in significant advantages in precisely classifying cell types.

This analysis reveals the necessity for a refined approach to improving medical NER models, emphasizing that, while extensive dataset training can improve the overall accuracy, architectural adjustments and targeted training are essential to mitigate disparities in model performance across diverse entity types. Therefore, enhancing the accuracy and robustness of NER models involves increasing the variety and volume of training data and optimizing model architectures to address the specific challenges posed by less represented or more intricate entity types.

Influence of Macrofactors on Prediction Accuracy in Medical NER

In analyzing how macrofactor metrics influence prediction accuracy across various medical entity types within 7 NER models, we observed significant nuances that reflect the complexity of modeling in medical NER tasks. Notably, entities characterized by higher values in *sLen*, *eLen*, *eNum*, and *eDen* generally yielded better prediction accuracy. This correlation suggests that entities with more extensive and detailed textual representations tend to be predicted more accurately, highlighting the models' capacity to handle intricate data structures effectively. However, a notable exception arises with *elCon*, which inversely correlated with these metrics, indicating a potential trade-off between detailed data processing and consistent label accuracy.

This phenomenon is further complicated by the varying impact of the *tEWC* on prediction accuracy across different datasets. For example, in the Revised JNLPBA dataset, lower *tEWC* values were associated with higher accuracies, suggesting that models perform better with more concise entity representations. In contrast, other datasets showed that higher *tEWC* values, indicative of richer contextual data, enhanced model performance. This inconsistency underscores the complex influence of data characteristics on model effectiveness and suggests that the optimal balance of data quantity and quality varies significantly across datasets. Therefore, tailored model training and data preparation strategies are essential to optimize prediction accuracy according to the unique characteristics of each dataset.

These observations necessitate a strategic approach to model training and data preparation that considers the unique demands of each dataset. While handling more extensive datasets can lead to better entity recognition in some contexts, balancing this with the need for precision and consistency in entity labeling is crucial. As NER technologies evolve, it becomes imperative to refine model architectures and training methodologies to ensure that models can manage the dual challenges of complexity and volume without sacrificing accuracy.

Refining Macrofactor Sensitivity to Improve NER Model Precision

Using the MFE algorithm for hierarchical macrofactor screening, *eNum* or *eLen* in each entity phrase had the most significant impact on the prediction accuracy of NER models. Unlike broader textual metrics such as *sLen* that provide general context, *eNum* directly measures the complexity of entities, whereas *eLen* captures the length of entity phrases. These factors significantly influence how models process and interpret dense information. A higher *eNum* generally indicates semantically rich entities that are potentially more challenging to analyze. At the same time, a higher *eLen* suggests that entity phrases contain more detailed and extended descriptions requiring careful parsing.

The consistent selection of *eNum* or *eLen* in the final layer of the screening process across various datasets underscores their pivotal role in enhancing the precision of entity recognition. The impact of *eNum* on model accuracy suggests that entities with a higher density of words require sophisticated model capabilities to discern and categorize detailed information accurately. Similarly, entities with longer phrases (*eLen*) present a challenge as they involve more complex syntactic and semantic structures that must be interpreted correctly.

Therefore, refining models' ability to analyze entities with higher *eNum* or longer *eLen* values could be a strategic approach to advancing NER technologies, particularly in domains such as medicine, where precise and reliable entity recognition is crucial. Enhancing how models manage and use the detailed information encapsulated in *eNum* and *eLen* will improve the robustness and effectiveness of medical NER models, ensuring that they meet the complex demands of varied and extensive datasets.

Conclusions

Medical NER is a crucial component of medical informatics, essential for identifying and categorizing named entities within unstructured medical text data. A proficient NER model significantly enhances various downstream applications, such as medical text classification, question answering, and information retrieval. Developing a high-performance NER model requires a meticulous approach that includes selecting relevant macrofactors, designing the model's architecture, and curating specialized training data tailored to medical contexts.

Our evaluation method for NER models extends beyond general metrics such as accuracy, recall, and F_1-score by incorporating an extensive analysis of macrofactors relevant to medical entities. This comprehensive approach enables a multidimensional evaluation of the models, providing insights into how different entity types, attributes, and contextual factors influence performance. For example, our findings indicate that, while "disease" frequently occurs in medical texts and requires high accuracy, entities such as "Immaterial_anatomical_entity" may not require the same precision. This discrepancy highlights the need for targeted optimization strategies for different entity types, which is crucial for advancing medical NER models.

In addition, our study explored the characteristics of entities to improve the creation of high-quality medical NER datasets and documents. This focus enhances the NER models' ability to identify entities accurately and addresses the specific needs of medical texts. While our analysis extensively covered macrofactors and their impact, it did not delve into misclassifications of entity labels or the fine-grained interactions between entity words. These areas could further refine our

understanding of model accuracy. Moreover, examining hardware performance illuminates these models' internal efficiency and resource use, which is crucial for their deployment in real-world scenarios.

In conclusion, evaluating medical NER models is essential for developing effective and precise NLP applications in health care. It gives medical researchers the insights to select and refine NER models suited to various medical scenarios, ultimately improving these systems' accuracy, robustness, and reliability. This foundational work sets the stage for future research that could explore the intricate relationships within NER systems, further enhancing the capabilities of medical informatics.

Acknowledgments

This work was supported by the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences program (grant 2021-I2M-1-057).

Conflicts of Interest

None declared.

References

1. Li J, Wei Q, Ghiasvand O, Chen M, Lobanov V, Weng C, et al. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. BMC Med Inform Decis Mak 2022 Sep 06;22(Suppl 3):235 [FREE Full text] [doi: 10.1186/s12911-022-01967-7] [Medline: 36068551]
2. Gridach M. Character-level neural network for biomedical named entity recognition. J Biomed Inform 2017 Jun;70:85-91. [doi: 10.1016/j.jbi.2017.05.002] [Medline: 28502909]
3. Yang T, He Y, Yang N. Named entity recognition of medical text based on the deep neural network. J Healthc Eng 2022;2022:3990563 [FREE Full text] [doi: 10.1155/2022/3990563] [Medline: 35295179]
4. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: challenges and opportunities. In: Proceedings of the 2016 IEEE International Conference on Big Data. 2016 Presented at: Big Data '16; December 5-8, 2016; Washington, DC p. 1937-1945 URL: https://ieeexplore.ieee.org/document/7840814 [doi: 10.1109/bigdata.2016.7840814]
5. Durango MC, Torres-Silva EA, Orozco-Duque A. Named entity recognition in electronic health records: a methodological review. Healthc Inform Res 2023 Oct;29(4):286-300 [FREE Full text] [doi: 10.4258/hir.2023.29.4.286] [Medline: 37964451]
6. Chen P, Zhang M, Yu X, Li S. Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT. BMC Med Inform Decis Mak 2022 Dec 01;22(1):315 [FREE Full text] [doi: 10.1186/s12911-022-02059-2] [Medline: 36457119]
7. Wu H, Ji J, Tian H, Chen Y, Ge W, Zhang H, et al. Chinese-named entity recognition from adverse drug event records: radical embedding-combined dynamic embedding-based BERT in a bidirectional long short-term conditional random field (Bi-LSTM-CRF) model. JMIR Med Inform 2021 Dec 01;9(12):e26407 [FREE Full text] [doi: 10.2196/26407] [Medline: 34855616]
8. Cheng M, Xiong S, Li F, Liang P, Gao J. Multi-task learning for Chinese clinical named entity recognition with external knowledge. BMC Med Inform Decis Mak 2021 Dec 31;21(1):372 [FREE Full text] [doi: 10.1186/s12911-021-01717-1] [Medline: 34972505]
9. Ao Y, Zhang Y, Tang H. Marine shellfish entity recognition based on BiLSTM-CRF model. In: Proceedings of the 6th International Conference on Pattern Recognition and Artificial Intelligence. 2023 Presented at: PRAI '23; August 18-20, 2023; Haikou, China p. 217-222 URL: https://ieeexplore.ieee.org/document/10332114 [doi: 10.1109/prai59366.2023.10332114]
10. Wang W, Li X, Ren H, Gao D, Fang A. Chinese clinical named entity recognition from electronic medical records based on multisemantic features by using robustly optimized bidirectional encoder representation from transformers pretraining approach whole word masking and convolutional neural networks: model development and validation. JMIR Med Inform 2023 May 10;11:e44597 [FREE Full text] [doi: 10.2196/44597] [Medline: 37163343]
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online October 11, 2018 [FREE Full text] [doi: 10.5260/chara.21.2.8]
12. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv Preprint posted online February 3, 2018 [FREE Full text] [doi: 10.18653/v1/n18-1202]
13. Liu Y, Ott M, Goyal N, DuBois J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv Preprint posted online July 26, 2019 [FREE Full text] [doi: 10.5260/chara.21.2.8]
14. Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, et al. Big Bird: Transformers for Longer Sequences. arXiv Preprint posted online July 28, 2020 [FREE Full text] [doi: 10.48550/arXiv.2007.14062]
15. He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. arXiv Preprint posted online June 5, 2020 [FREE Full text] [doi: 10.48550/arXiv.2006.03654]

16. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
17. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023 Nov 22;25(1):bbad493. [doi: 10.1093/bib/bbad493] [Medline: 38168838]
18. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv Preprint posted online March 31, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.18223]
19. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform Assoc 2024 Sep 01;31(9):1812-1820. [doi: 10.1093/jamia/ocad259] [Medline: 38281112]
20. Zheng Y, Zhang R, Zhang J, Ye Y, Luo Z, Ma Y. LlamaFactory: unified efficient fine-tuning of 100+ language models. arXiv Preprint posted online March 20, 2024 [FREE Full text] [doi: 10.18653/v1/2024.acl-demos.38]
21. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The Llama 3 herd of models. arXiv Preprint posted online August 15, 2024 [FREE Full text] [doi: 10.48550/arXiv.2407.21783]
22. Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, et al. Opt: open pre-trained transformer language models. arXiv Preprint posted online May 21, 2022 [FREE Full text] [doi: 10.48550/arXiv.2205.01068]
23. Team G, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, et al. Gemma: open models based on gemini research and technology. arXiv Preprint posted online March 13, 2024 [FREE Full text] [doi: 10.48550/arXiv.2403.08295]
24. Ohno Y, Kato R, Ishikawa H, Nishiyama T, Isawa M, Mochizuki M, et al. Using the natural language processing system medical named entity recognition-Japanese to analyze pharmaceutical care records: natural language processing analysis. JMIR Form Res 2024 Jun 04;8:e55798 [FREE Full text] [doi: 10.2196/55798] [Medline: 38833694]
25. Freund F, Tamla P, Hemmje M. Towards improving clinical practice guidelines through named entity recognition: model development and evaluation. In: Proceedings of the 31st Irish Conference on Artificial Intelligence and Cognitive Science. 2023 Presented at: AICS '23; December 7-8, 2023; Letterkenny, Ireland p. 1-8 URL: https://ieeexplore.ieee.org/document/10470480 [doi: 10.1109/aics60730.2023.10470480]
26. Ahmad PN, Shah AM, Lee K. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. Healthcare (Basel) 2023 Apr 28;11(9):1268 [FREE Full text] [doi: 10.3390/healthcare11091268] [Medline: 37174810]
27. Yoon W, So CH, Lee J, Kang J. CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. BMC Bioinformatics 2019 May 29;20(Suppl 10):249 [FREE Full text] [doi: 10.1186/s12859-019-2813-6] [Medline: 31138109]
28. Yu H, Mao XL, Chi Z, Wei W, Huang H. A robust and domain-adaptive approach for low-resource named entity recognition. In: Proceedings of the 2020 IEEE International Conference on Knowledge Graph. 2020 Presented at: ICKG '20; August 9-11, 2020; Nanjing, China p. 9-11 URL: https://ieeexplore.ieee.org/document/9194550 [doi: 10.1109/icbk50248.2020.00050]
29. Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models. arXiv Preprint posted online October 25, 2019 [FREE Full text] [doi: 10.5260/chara.21.2.8]
30. Erdmann A, Wrisley DJ, Allen B, Brown C, Cohen-Bodénès S, Elsner M, et al. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL '19; 2019; Minneapolis, MN p. 2223-2234 URL: https://aclanthology.org/N19-1231/ [doi: 10.18653/v1/n19-1231]
31. Usha MS, Smrity AM, Das S. Named entity recognition using transfer learning with the fusion of pre-trained SciBERT language model and bi-directional long short term memory. In: Proceedings of the 25th International Conference on Computer and Information Technology. 2022 Presented at: ICCIT '22; December 17-19, 2022; Cox's Bazar, Bangladesh p. 460-465 URL: https://ieeexplore.ieee.org/document/10055784 [doi: 10.1109/iccit57492.2022.10055784]
32. Nagaraj P, Dass MV, Mahender E, Kumar KR. Breast cancer risk detection using XGB classification machine learning technique. In: Proceedings of the 2022 IEEE International Conference on Current Development in Engineering and Technology. 2022 Presented at: CCET '22; December 23-24, 2022; Bhopal, India p. 1-5 URL: https://ieeexplore.ieee.org/document/10080076 [doi: 10.1109/ccet56606.2022.10080076]
33. Ozcelik O, Toraman C. Named entity recognition in Turkish: a comparative study with detailed error analysis. Inf Process Manage 2022 Nov;59(6):103065. [doi: 10.1016/j.ipm.2022.103065]
34. Akhtyamova L. Named entity recognition in Spanish biomedical literature: short review and Bert model. In: Proceedings of the 26th Conference of Open Innovations Association. 2020 Presented at: FRUCT '20; April 20-24, 2020; Yaroslavl, Russia p. 20-24 URL: https://ieeexplore.ieee.org/document/9087359 [doi: 10.23919/fruct48808.2020.9087359]
35. Fu J, Liu P, Neubig G. Interpretable multi-dataset evaluation for named entity recognition. arXiv Preprint posted online November 13, 2020 [FREE Full text] [doi: 10.18653/v1/2020.emnlp-main.489]
36. Zhou X, Ma H, Gu J, Chen H, Deng W. Parameter adaptation-based ant colony optimization with dynamic hybrid mechanism. Eng Appl Artif Intell 2022 Sep;114:105139. [doi: 10.1016/j.engappai.2022.105139]

37. Yao R, Guo C, Deng W, Zhao H. A novel mathematical morphology spectrum entropy based on scale-adaptive techniques. ISA Trans 2022 Jul;126:691-702. [doi: 10.1016/j.isatra.2021.07.017] [Medline: 34446283]
38. Huang MS, Lai PT, Lin PY, You YT, Tsai RT, Hsu W. Biomedical named entity recognition and linking datasets: survey and our recent development. Brief Bioinform 2020 Dec 01;21(6):2219-2238. [doi: 10.1093/bib/bbaa054] [Medline: 32602538]
39. Li J, Sun Y, Johnson RJ, Sciaky D, Wei CH, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford) 2016 May 09;2016:baw068 [FREE Full text] [doi: 10.1093/database/baw068] [Medline: 27161011]
40. Pyysalo S, Ananiadou S. Anatomical entity mention recognition at literature scale. Bioinformatics 2014 Mar 15;30(6):868-875 [FREE Full text] [doi: 10.1093/bioinformatics/btt580] [Medline: 24162468]
41. Zhou Y, Cahya S, Combs SA, Nicolaou CA, Wang J, Desai PV, et al. Exploring tunable hyperparameters for deep neural networks with industrial ADME data sets. J Chem Inf Model 2019 Mar 25;59(3):1005-1016. [doi: 10.1021/acs.jcim.8b00671] [Medline: 30586300]
42. Florea AC, Andonie R. Weighted random search for hyperparameter optimization. Int J Comput Commun 2019 Apr 14;14(2):154-169. [doi: 10.15837/ijccc.2019.2.3514]
43. Wen C, Chen T, Jia X, Zhu J. Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary. Data Intel 2021;3(3):402-417. [doi: 10.1162/dint_a_00105]
44. Wu Y, Liu L. Selecting and composing learning rate policies for deep neural networks. ACM Trans Intell Syst Technol 2023 Feb 16;14(2):1-25. [doi: 10.1145/3570508]
45. Marchisio A, Massa A, Mrazek V, Bussolino B, Martina M, Shafique M. NASCaps: a framework for neural architecture search to optimize the accuracy and hardware efficiency of convolutional capsule networks. In: Proceedings of the 39th International Conference on Computer-Aided Design. 2020 Presented at: ICCAD '20; November 2-5, 2020; Virtual Event p. 1-9 URL: https://dl.acm.org/doi/10.1145/3400302.3415731 [doi: 10.1145/3400302.3415731]
46. Jabir B, Falih N. Dropout, a basic and effective regularization method for a deep learning model: a case study. Indones J Electr Eng Comput Sci 2021 Nov 01;24(2):1009-1016. [doi: 10.11591/ijeecs.v24.i2.pp1009-1016]
47. Zhalgasbayev A, Khauazkhan A, Sarsenova Z. Fine-tuning the gemma model for Kaggle assistant. In: Proceedings of the 2024 IEEE AITU Digital Generation Conference. 2024 Presented at: IEEE AITU '24; April 3-4, 2024; Astana, Kazakhstan p. 104-109 URL: https://ieeexplore.ieee.org/document/10585529 [doi: 10.1109/ieeeconf61558.2024.10585529]
48. Tsai RT, Wu SH, Chou WC, Lin YC, He D, Hsiang J, et al. Various criteria in the evaluation of biomedical named entity recognition. BMC Bioinformatics 2006 Feb 24;7:92 [FREE Full text] [doi: 10.1186/1471-2105-7-92] [Medline: 16504116]
49. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet 2018 Sep 17;19(Suppl 1):65 [FREE Full text] [doi: 10.1186/s12863-018-0633-8] [Medline: 30255764]
50. Nasir IM, Khan MA, Yasmin M, Shah JH, Gabryel M, Scherer R, et al. Pearson correlation-based feature selection for document classification using balanced training. Sensors (Basel) 2020 Nov 27;20(23):6793 [FREE Full text] [doi: 10.3390/s20236793] [Medline: 33261136]
51. Zhou JY, Song LW, Yuan R, Lu XP, Wang GQ. Prediction of hepatic inflammation in chronic hepatitis B patients with a random forest-backward feature elimination algorithm. World J Gastroenterol 2021 Jun 07;27(21):2910-2920 [FREE Full text] [doi: 10.3748/wjg.v27.i21.2910] [Medline: 34135561]

Abbreviations

ADAM: Adaptive Moment Estimation
AnatEM: Anatomical Entity Mention
AVG_MACRO: macroaverage
AVG_MICRO: microaverage
BERT: Bidirectional Encoder Representations From Transformers
BigBird: Big Transformer Models for Efficient Long-Sequence Attention
CPU: central processing unit
CRF: conditional random field
DeBERTa: Decoding-enhanced Bidirectional Encoder Representations From Transformers with Disentangled Attention
eDen: entity density
elCon: entity label consistency
eLen: entity phrase length
eNum: number of entity words in each entity phrase
GPU: graphics processing unit
HMM: hidden Markov model
JNLPBA: Joint Workshop on Natural Language Processing in Biomedicine and its Applications
LLM: large language model

LoRA: low-rank adaptation
MFE: multilevel factor elimination
NER: named entity recognition
NLP: natural language processing
RoBERTa: Robustly Optimized BERT Pretraining Approach
sLen: sentence length
tEWC: total entity word count in each entity type

Edited by C Lovis; submitted 23.04.24; peer-reviewed by MO Khursheed, I Rida, S Mao; comments to author 08.06.24; revised version received 09.08.24; accepted 15.09.24; published 17.10.24.

Please cite as:
Liu S, Wang A, Xiu X, Zhong M, Wu S
Evaluating Medical Entity Recognition in Health Care: Entity Model Quantitative Study
JMIR Med Inform 2024;12:e59782
URL: https://medinform.jmir.org/2024/1/e59782
doi:10.2196/59782
PMID:

Original Paper

A New Natural Language Processing–Inspired Methodology (Detection, Initial Characterization, and Semantic Characterization) to Investigate Temporal Shifts (Drifts) in Health Care Data: Quantitative Study

Bruno Paiva[1], MD; Marcos André Gonçalves[1], PhD; Leonardo Chaves Dutra da Rocha[2], PhD; Milena Soriano Marcolino[3], PhD; Fernanda Cristina Barbosa Lana[3], PhD; Maira Viana Rego Souza-Silva[3], MD; Jussara M Almeida[1], PhD; Polianna Delfino Pereira[3], PhD; Claudio Moisés Valiense de Andrade[1], PhD; Angélica Gomides dos Reis Gomes[4], MD; Maria Angélica Pires Ferreira[5], PhD; Frederico Bartolazzi[6], MD; Manuela Furtado Sacioto[7], MD; Ana Paula Boscato[8], MD; Milton Henriques Guimarães-Júnior[9], MD; Priscilla Pereira dos Reis[10], MD; Felício Roberto Costa[3], MD; Alzira de Oliveira Jorge[11], PhD; Laryssa Reis Coelho[12], MD; Marcelo Carneiro[13], PhD; Thaís Lorenna Souza Sales[1], MD; Silvia Ferreira Araújo[14], MD; Daniel Vitório Silveira[15], MD; Karen Brasil Ruschel[1], PhD; Fernanda Caldeira Veloso Santos[16], MSc; Evelin Paola de Almeida Cenci[17], MSc; Luanna Silva Monteiro Menezes[1], MSc, MD; Fernando Anschau[18], MSc, MD; Maria Aparecida Camargos Bicalho[19], MD; Euler Roberto Fernandes Manenti[20], PhD; Renan Goulart Finger[21], MD; Daniela Ponce[22], PhD; Filipe Carrilho de Aguiar[23], MD; Luiza Margoto Marques[7]; Luís César de Castro[24], PhD; Giovanna Grünewald Vietta[25], PhD; Mariana Frizzo de Godoy[6], MD; Mariana do Nascimento Vilaça[26], MD; Vivian Costa Morais[7]

[1]Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, Belo Horizonte, Brazil
[2]Computer Science Department, Universidade Federal de São João del-Rei, Brazil, São João del-Rei, Brazil
[3]Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, Belo Horizonte, Brazil
[4]Hospitais da Rede Mater Dei, Belo Horizonte, Brazil
[5]Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
[6]Hospital Santo Antônio, Curvelo, Brazil
[7]Faculdade Ciências Médicas de Minas Gerais, Belo Horizonte, Brazil
[8]Hospital Tacchini, Bento Gonçalves, Brazil
[9]Hospital Márcio Cunha, Ipatinga, Brazil
[10]Hospital Metropolitano Doutor Célio de Castro, Belo Horizonte, Brazil
[11]Hospital Risoleta Tolentino Neves, Belo Horizonte, Brazil
[12]Faculdade de Medicina, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Teófilo Otoni, Brazil
[13]Hospital Santa Cruz, Santa Cruz do Sul, Brazil
[14]Hospital Semper, Belo Horizonte, Brazil
[15]Hospital Unimed BH, Belo Horizonte, Brazil
[16]Hospital Universitário de Santa Maria, Santa Maria, Brazil
[17]Hospital Moinhos de Vento, Porto Alegre, Brazil
[18]Hospital Nossa Senhora da Conceição, Porto Alegre, Brazil
[19]Fundação Hospitalar do Estado de Minas Gerais, Belo Horizonte, Brazil
[20]Hospital Mãe de Deus, Porto Alegre, Brazil
[21]Hospital Regional do Oeste, Chapecó, Brazil
[22]Faculdade de Medicina de Botucatu, Universidade Estadual Paulista Júlio de Mesquita Filho, Botucatu, Brazil
[23]Hospital das Clínicas, Universidade Federal de Pernambuco, Recife, Brazil
[24]Hospital Bruno Born, Lajeado, Brazil
[25]Hospital SOS Cárdio, Florianópolis, Brazil
[26]Hospital Metropolitano Odilon Behrens, Belo Horizonte, Brazil

Corresponding Author:
Bruno Paiva, MD

Computer Science Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Street Daniel de Carvalho
1846, apto 201
Belo Horizonte, 30431310
Brazil
Phone: 55 31999710134
Email: angelfire7@gmail.com

Abstract

Background: Proper analysis and interpretation of health care data can significantly improve patient outcomes by enhancing services and revealing the impacts of new technologies and treatments. Understanding the substantial impact of temporal shifts in these data is crucial. For example, COVID-19 vaccination initially lowered the mean age of at-risk patients and later changed the characteristics of those who died. This highlights the importance of understanding these shifts for assessing factors that affect patient outcomes.

Objective: This study aims to propose detection, initial characterization, and semantic characterization (DIS), a new methodology for analyzing changes in health outcomes and variables over time while discovering contextual changes for outcomes in large volumes of data.

Methods: The DIS methodology involves 3 steps: detection, initial characterization, and semantic characterization. Detection uses metrics such as Jensen-Shannon divergence to identify significant data drifts. Initial characterization offers a global analysis of changes in data distribution and predictive feature significance over time. Semantic characterization uses natural language processing–inspired techniques to understand the local context of these changes, helping identify factors driving changes in patient outcomes. By integrating the outcomes from these 3 steps, our results can identify specific factors (eg, interventions and modifications in health care practices) that drive changes in patient outcomes. DIS was applied to the Brazilian COVID-19 Registry and the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) data sets.

Results: Our approach allowed us to (1) identify drifts effectively, especially using metrics such as the Jensen-Shannon divergence, and (2) uncover reasons for the decline in overall mortality in both the COVID-19 and MIMIC-IV data sets, as well as changes in the cooccurrence between different diseases and this particular outcome. Factors such as vaccination during the COVID-19 pandemic and reduced iatrogenic events and cancer-related deaths in MIMIC-IV were highlighted. The methodology also pinpointed shifts in patient demographics and disease patterns, providing insights into the evolving health care landscape during the study period.

Conclusions: We developed a novel methodology combining machine learning and natural language processing techniques to detect, characterize, and understand temporal shifts in health care data. This understanding can enhance predictive algorithms, improve patient outcomes, and optimize health care resource allocation, ultimately improving the effectiveness of machine learning predictive algorithms applied to health care data. Our methodology can be applied to a variety of scenarios beyond those discussed in this paper.

(JMIR Med Inform 2024;12:e54246) doi:10.2196/54246

KEYWORDS

health care; machine learning; data drifts; temporal drifts

Introduction

Overview

Health care data are a critical resource that can be used to improve patient outcomes and the financial performance of health care institutions [1,2]. By analyzing patient data, health care providers can gain insights into patients' health status, identify trends, and make informed decisions about treatment plans. Properly collected, managed, treated, and interpreted health care data can help providers improve operational efficiency and reduce costs, thereby improving financial results [3].

One of the primary ways health care data can be used to enhance medical decisions and potentially improve patient outcomes is through predictive analysis. This technique uses historical data to identify patterns and predict future outcomes, thereby enabling the recognition of high-risk patients, the simulation of different therapeutic approaches, and the personalization of patient care. However, relying on historical data has its caveats, as the predictive capacity of different variables is not fixed over time. Ignoring these aspects of temporal data may lead to prediction errors and learning instabilities. These variations in performance are part of what is known as temporal data shifts [4-7].

A temporal data shift refers to a change in the statistical properties of a data set over time, which can degrade model accuracy. In health care, this may occur due to various reasons, including changes in data collection practices, software updates or replacements, changes in patient behavior or lifestyle habits,

and the introduction of new therapeutic technologies. These temporal events may lead to inconsistencies and discrepancies in the data, which may affect both the accuracy and reliability of the data and models trained on them. The impacts can be significant [4,7], as they can lead to incorrect diagnoses, inappropriate treatment plans, and poor patient outcome predictions. This highlights the importance of managing, characterizing, and mitigating these temporal effects [8].

We are particularly interested in how temporal data drifts can be used to analyze the effectiveness of new patient treatment options. Changes in predictive capacity can provide insights into the impact of new treatments on patient outcomes. For instance, by comparing data collected before and after introducing a new treatment, we can identify any shifts that may indicate improved patient outcomes. If the data drift analysis indicates a positive impact of the new treatment, health care providers may choose to continue to monitor the data to ensure that the positive effects are sustained while maintaining the use of the new therapeutic option [9].

A notable example of a condition that experienced an important data drift over time is the HIV infection. In the 1980s, HIV infection was a strong predictor of early death. However, it has now become more of a chronic condition, such as diabetes mellitus or systemic hypertension. In the same manner, advancements in breast cancer treatment have significantly increased survivorship over the years [10].

Similarly, several infectious diseases, such as poliomyelitis or measles, have been nearly eradicated in most parts of the world, making them unlikely hypotheses for new diagnoses [11,12]. In the case of COVID-19, vaccination has dramatically changed the profile of hospitalizations and deaths [4,13], initially decreasing the mean age of patients at risk and creating a clear distinction between the periods before and after vaccination.

Our Main Contribution: The Detection, Initial Characterization, and Semantic Characterization Methodology

Building upon the idea of analyzing data drifts to obtain insights into how and whether new technologies or treatments have impacted patient outcomes, this paper proposes a novel, 3-step health care temporal analysis methodology, called detection, initial characterization, and semantic characterization (DIS). The proposed DIS methodology is summarized in Figure 1. It consists of three main steps, (1) detection, (2) initial characterization, and (3) semantic characterization, which are described in the following sections.

In summary, we exploited various drift detection metrics in the detection step to identify any significant instances of data drift. Some of the metrics we explored in this step include Jensen-Shannon divergence [14], autoencoder reconstruction error [15], and centroid distances [16]. If changes were detected, we proceeded to the initial characterization step, where we obtained a global (data set–level) descriptive analysis of what changed and how the discriminative and predictive power of each feature and the distribution of labels evolved over time.

Additionally, we introduced the concept of temporal granularity in the data drift domain, which holds particular significance in health care data drifts and influences the instantiation of our third and final step. High temporal granularity is observed when a data set allows the visualization of numerous events over time for individual patients, with a clear understanding of the chronological order among these events. Conversely, low temporal granularity is observed when each patient is considered a singular event in time, lacking clarity regarding the precedence or sequence of different attributes.

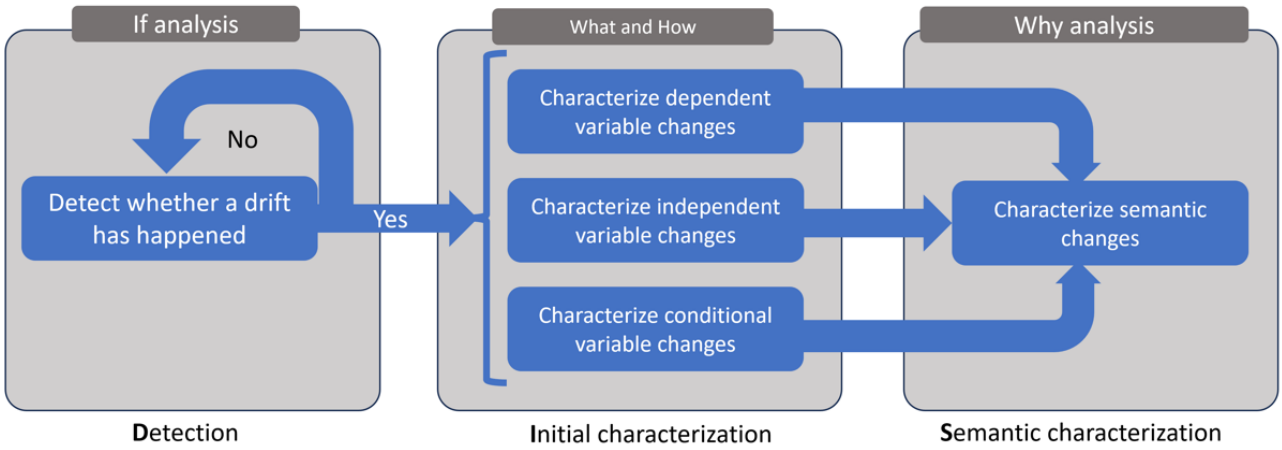
Finally, guided by these principles, we proceeded to the third, semantic characterization step, which exploits concepts popularized in the natural language processing (NLP) domain to provide a localized (instance-level) perspective of why certain shifts occurred. To achieve this, we exploited vector embeddings derived from health care events, such as sequences of the *International Classification of Diseases (ICD)* codes, vital data measurements, and consumption items. Each of these semantic units (*ICD* codes, measurements, consumed items, etc) was treated as an "event" or, in NLP terminology, a "token." By using NLP-inspired techniques to create semantic embeddings for these entities, we aimed to uncover insights into the changing context and its impact on the outcomes of interest over time.

Before delving into the details of each step in our methodology, it is crucial to emphasize that our DIS approach differs significantly from common practices. While conventional methods usually involve an ad hoc combination of techniques for data collection, qualitative data processing and extraction, and data analysis, our DIS methodology offers a planned and structured procedure, as illustrated in Figure 1. This procedure delineates the required steps to understand data drift in health care data. As we will demonstrate and discuss, these steps can be tailored to various case studies by applying different techniques depending on specific data characteristics. We also offer guidance for selecting one particular approach for a given scenario. Furthermore, we discuss how the results of each step can inform the execution of the following ones and how the combined results of all steps can support our understanding of the drift.

More broadly, to the best of our knowledge, this is the first study to examine data drifts in health care from a technology incorporation standpoint. Rather than solely focusing on enhancing the robustness of machine learning (ML) models, we delved into the underlying factors driving temporal shifts in patient outcomes. Our aim was to study the impact of emerging technologies such as new drugs, patient care policies, or vaccines. In the following sections, we detail the steps of our DIS methodology and illustrate its application in 2 case studies with distinct characteristics in terms of temporal data shifts: the Brazilian COVID-19 Registry data set [17] and the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) data set [18]. By doing so, we illustrate how DIS can obtain insights into the reasons behind some real-life data drifts, as well as their potential impacts, both positive and negative, from a health care perspective.

The main contributions of this paper are summarized in Textbox 1.

Figure 1. Overview of the detection, initial characterization, and semantic characterization (DIS) methodology.



Textbox 1. Main contributions of our study.

Contributions

1. The proposal of a new data drift characterization and analysis methodology, detection, initial characterization, and semantic characterization (DIS), that is flexible enough to work on different scenarios. DIS encapsulates and cohesively organizes a sequence of necessary steps for data drift analysis.
2. A new semantic analysis step based on natural language processing embeddings for temporal understanding, which focuses on comprehending the context of relevant outcomes by examining changes in their embedding vectors over time. By incorporating such semantic techniques, DIS provides deeper insights into the reasons behind temporal changes, especially when combined with domain-specific knowledge. This approach allows for a more nuanced analysis of data evolution over time, capturing complex patterns and relationships that may not be apparent with traditional methods such as cluster analysis.
3. The application of the DIS methodology to 2 different case studies with very different temporal granularity profiles illustrates the possibility of conducting insightful analyses using the methodology. We also offer guidelines to aid practitioners in making informed decisions about which methods to use in each step of our methodology, based on particular characteristics of the data. This demonstrates the generalizability and applicability of DIS across different scenarios.

Methods

A Detailed Description of the DIS Methodology

Detection Step

In step 1 (detection), the main focus is on assessing whether the data have relevant temporal variations. Monitoring and detecting such data drifts are crucial for upholding the accuracy and reliability of ML models and for identifying beneficial and detrimental changes in health care caused by interventions, such as the introduction of new treatments or drugs. From the perspective of a health care service or company, this step identifies whether changes are occurring, potentially prompting further investigations that could enhance service efficiency over time.

For the *detection step*, we recommend splitting the data into temporal chunks and then comparing the data distributions in consecutive chunks. A drift is detected whenever the distributions of distinct chunks exhibit significant differences. Various metrics to compare empirical distributions are available in the literature. These metrics have different characteristics and underlying principles, which may lead to relevant differences in their effectiveness in detecting temporal data drifts. In this work, we considered the following metrics: centroid cosine distance [16], Jensen-Shannon divergence [14], autoencoder reconstruction error [15], classifier error (in separating 2-time chunks) [19], and principal component analysis (PCA) reconstruction error [20] metrics.

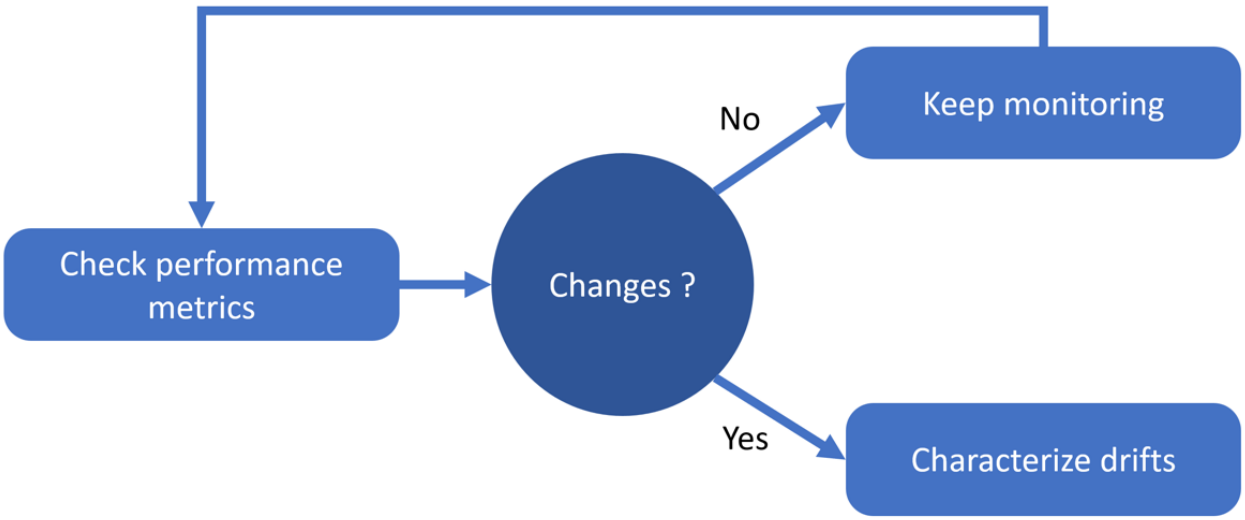
The centroid cosine distance metric assesses changes in the central points of data clusters over time and is sensitive to numeric outliers, particularly in heavy-tailed distributions where extremes can be multiple orders of magnitude larger than typical values. The PCA reconstruction error captures variations in data structure by quantifying the difference between original and reconstructed data. Similarly, autoencoder reconstruction error focuses on reconstruction accuracy. Both metrics measure the “novelty” of a data point and are sensitive to numerical outliers. By contrast, the classifier error evaluates a model’s ability to distinguish past from future data, providing insights into how drift affects predictive capabilities. Finally, the Jensen-Shannon divergence quantifies distributional changes, offering a broader perspective on underlying data distribution shifts over time. While reconstruction errors and centroids excel at detecting local outliers and structural changes, the Jensen-Shannon divergence and classifier error provide a more comprehensive view of distributional shifts, making them valuable for modeling the impact of temporal drifts on data distributions.

As an example, our prior analysis of the Brazilian COVID-19 Registry [17] revealed a data drift that significantly impacted the death prediction task, suggesting that vaccination had a pivotal role in the profiles of hospitalized and deceased patients

during the COVID-19 pandemic [4]. Although this is an interesting finding, the previous study did not present a proper structure to detect, monitor, and interpret such drifts generically, nor did it propose mechanisms to detect semantic information associated with specific outcomes.

The drift caused by vaccination can be initially hypothesized by comparing the data distributions of consecutive chunks (eg, near future vs recent past) using a classification approach. This involves monitoring the prediction model's performance over time using metrics such as accuracy, precision, and recall. Alternatively, the distribution of different features over time can be tracked using metrics such as Jensen-Shannon divergence or autoencoder reconstruction errors. If the model's performance drops (or changes) significantly over time or if the differences between the metrics exceed a certain threshold, it may indicate a data drift. A summary of this monitoring loop is illustrated in Figure 2.

Figure 2. The temporal drift monitoring loop. We usually observe temporal shifts as important variations in model effectiveness over time.



Initial Characterization Step

Once a drift has been detected, we proceed to step 2 (initial characterization), where we begin to understand, from a global perspective (all data), *how* the data have changed (Table 1 [21]). This stage focuses on developing a general (global) comprehension of the *whats* and *hows* contributing to the changes observed in the data collection. Specifically, we are interested in characterizing variations in both *dependent* $P(y)$ and independent $P(x)$ variables, as well as the conditional probability of the dependent variables given the independent variables $P(y|x)$. To reach these goals, we examine how $P(y)$ has changed by plotting its frequency over time; the same is valid for $P(x)$. For $P(y|x)$, we can explore different complementary techniques that can help understand the drifts globally. We can analyze how the different correlation metrics between the top independent variables and the dependent variable change over time, for instance, with Pearson [22] or Spearman [23] correlations, or analyze the feature importance of tree-based learners or entropy-based measures such as information gain or chi-square over time [24]. Another possibility is to exploit explainability metrics based on game theory, such as Shapley additive explanations values [25].

"Sudden drift" describes a situation where changes are abrupt and usually caused by a single event, such as a change in data collection practices, where an attribute stops being collected. "Incremental drift" describes gradual and directional changes in a data distribution, such as the observed increase in the populations with overweight and obesity over the past years. "Gradual drift" is similar but does not imply directional changes. Instead, it encompasses other gradual changes, such as the slow change in the hospital admission profile over many years. Finally, "reoccurring drift" refers to a drift pattern that repeats over time, such as the seasonal increase in emergency services admitting patients with influenza during predictable seasons of the year.

This type of analysis facilitates understanding how the relationship between predictive variables and the outcome of interest has evolved from a global perspective. Additionally, it is helpful to check the rate of change for each selected outcome by using similarity metrics and comparing the different groups of patients over time. At this stage, it is feasible to answer valuable research and business questions. For instance, we may observe a decreased likelihood of the "death" outcome in a given population, such as patients with COVID-19 or patients with breast cancer. We may also spot changes in the profiles of the patients who had adverse outcomes. Following these initial insights, the subsequent task is to understand *why* such changes happened, the goal of *step 3*.

Table 1. Drift types concerning the passing of time, according to Moreno-Torres et al [21][a].

Data drift type	Description
Sudden drift	Abrupt and unexpected changes in the data
Incremental drift	Gradual and continuous changes over time
Gradual drift	Slow and steady changes in the data distribution
Reoccurring drift	Periodic or repetitive shifts in the data

Semantic Characterization Step

In step 3, the main focus is to learn *why* the changes we observed in step 2 happened. This step integrates fundamental research and business value into our methodology and is heavily dependent on the temporal granularity of the data under evaluation. To the best of our knowledge, this is the first study to examine data drifts in health care from a technology incorporation standpoint. For instance, as mentioned earlier, we may have already learned, as a result of step 1, that a given disease or condition, such as COVID-19, had a decreased lethality over a specific time period. Given this information, what will add value to health care services is the discovery of which repeatable interventions within this time frame can be consistently beneficial.

We begin step 3 by proposing a novel NLP-inspired technique based on token embedding techniques, such as Word2Vec [26], to detect local or individual changes in outcome contexts over time. We opt for NLP-inspired techniques because they effectively model and comprehend "semantics" and "contexts." In this context, we treat each patient as a "document" and any temporally discrete health care event or information, such as disease codes or items used during a hospital stay, as a "token" (ie, the equivalent of a "word" or a "subword" in NLP). For instance, the underlying premise is that a patient's semantics can be understood by examining their diseases and consumption history. On the basis of on this representation, we characterize which entities or outcome groups have undergone the most significant changes regarding their defining characteristics in comparison to a baseline or initial time chunk. This assessment assumes a setting where we have an outcome y and the task of predicting this outcome using independent variables X. This characterization can be achieved by comparing the distance of each class's centroid to a reference centroid, where a "centroid" represents the arithmetic mean of each patient's features.

The procedure to compute each of these *centroids* is explained in Multimedia Appendix 1. In this figure, we show a simplified view of 2 groups of patients in 2D and how the centroids are calculated to be at the spatial "center" of the groups by averaging their attributes. We can compare different centroids using either a cosine distance or a cosine similarity (equation 1). This type of analysis can guide our research toward a specific hypothesis, filtering down to the pattern changes in specific outcomes, such as death or the need for mechanical ventilation during a hospital stay.

(1)

In equation 1, the cosine distance is simply 1 –cosine similarity.

The centroid of each class in the first (time) chunk will be analyzed over time, providing insights into which outcomes (eg, death vs nondeath or hospitalization vs nonhospitalization) underwent the most significant changes. From this observation, we can focus our analysis on the interest group. This approach, which will be further illustrated in our experiments, allows us to compute semantic distances among patients, between patients and outcomes, and among different outcomes.

To apply step 3 to a data set, we need to remember that health care data come in different temporal and semantic granularities. For instance, data sets such as the Brazilian COVID-19 Registry (details presented in the DIS Instantiation for the Brazilian COVID-19 Registry Data Set section) treat each patient as a single data point, characterized by atomic temporal granularity, where temporal effects are observed only at a populational level. In data sets with such low temporal granularity, it is as if all events happened simultaneously at the patient level, and we know only the relationship between those events and the patients. In these cases, modeling the relationships between entities and their resulting semantic vectors may require techniques such as graph vectorization.

On the other extreme, data sets with high temporal granularity, such as MIMIC-IV (details presented in the DIS Instantiation for the MIMIC-IV Data Set section), present patients existing within their own timelines, as well as at the populational level. Furthermore, MIMIC-IV has different levels of semantic detail, such as sequential disease codes that could be aggregated into broader groups based on their chapters (eg, both "prostate cancer" and "breast cancer" could be grouped under the "neoplasms" disease code chapter).

In both cases, we would first refer to step 2 to identify suitable candidates for the NLP-inspired modeling. In the case of MIMIC-IV, as demonstrated later, the data show a gradual and trending shift over time, with in-hospital mortality consistently decreasing over the years. Given this pattern and the granularity available in these data, we create sequences of discrete information tokens to elucidate the observed variations for each patient, such as ordinal disease codes or chapters, if a more compact set of possible semantic units is desired.

Finally, we can append "artificial tokens" at the appropriate positions on each patient's sequence, such as a "death" token at the end of the sequences of deceased patients or an "ICU" token when the patient is transferred to the intensive care unit (ICU), if applicable. With these sequences, we can obtain semantic vectors representing diseases, patients, or outcomes. Following this process on discrete temporal chunks, such as

years or months, we obtain distinct outcome tokens for each temporal chunk (eg, "death 2020" and "death 2021," effectively separating the same outcome over 2 years). With this, it is possible to compare the tokens, examining their relative distance and semantic similarity to each other and other tokens. This allows the identification of what has become more or less similar to the analyzed outcome over time.

Next, we will illustrate the application of our methodology to the 2 aforementioned case studies, with different temporal granularities. The 2 cases are very different in terms of their temporal granularity, volume, and nature of data, demonstrating the generalization capability of DIS.

DIS Instantiation

We illustrate the application of DIS to analyze temporal shifts by using the MIMIC-IV [18] and the Brazilian COVID-19 Registry data sets [17].

The MIMIC-IV data set is a comprehensive, open-access, and deidentified in-hospital patient record containing sequential diagnosis data; consumption items; vital data records; unstructured eHealth data (text data); and clinical notes for approximately 40,000 ICU patients from 2008 to 2019, designed for research in health care and medical science [18]. In this data set, age is reported in age groups, which is a requirement for deidentification.

The Brazilian COVID-19 Registry is a multicenter retrospective cohort of 10,897 patients with a confirmed diagnosis of COVID-19 admitted between March 2020 and December 2021 from 41 different Brazilian hospitals. For the purpose of the present analysis, variables collected at hospital presentation and at patient discharge were used. The data set consists of >200 features, including known comorbidities, patient's age and sex, laboratory tests (such as complete blood count, C-reactive protein, and arterial blood gas analysis), vital signs at hospital presentation (ie, arterial blood pressure, respiratory rate, and heart rate), and clinical outcomes [17].

As mentioned earlier, we chose these 2 case studies, as they illustrate scenarios where the available data have very different temporal granularity characteristics, meaning the patient's timeline can be reconstructed from the data at either a local (individual) or a populational level.

Ethical Considerations

This study was approved by the Ethics and Research Committee of the Federal University of Minas Gerais (CAAE 70801523.7.1001.5149).

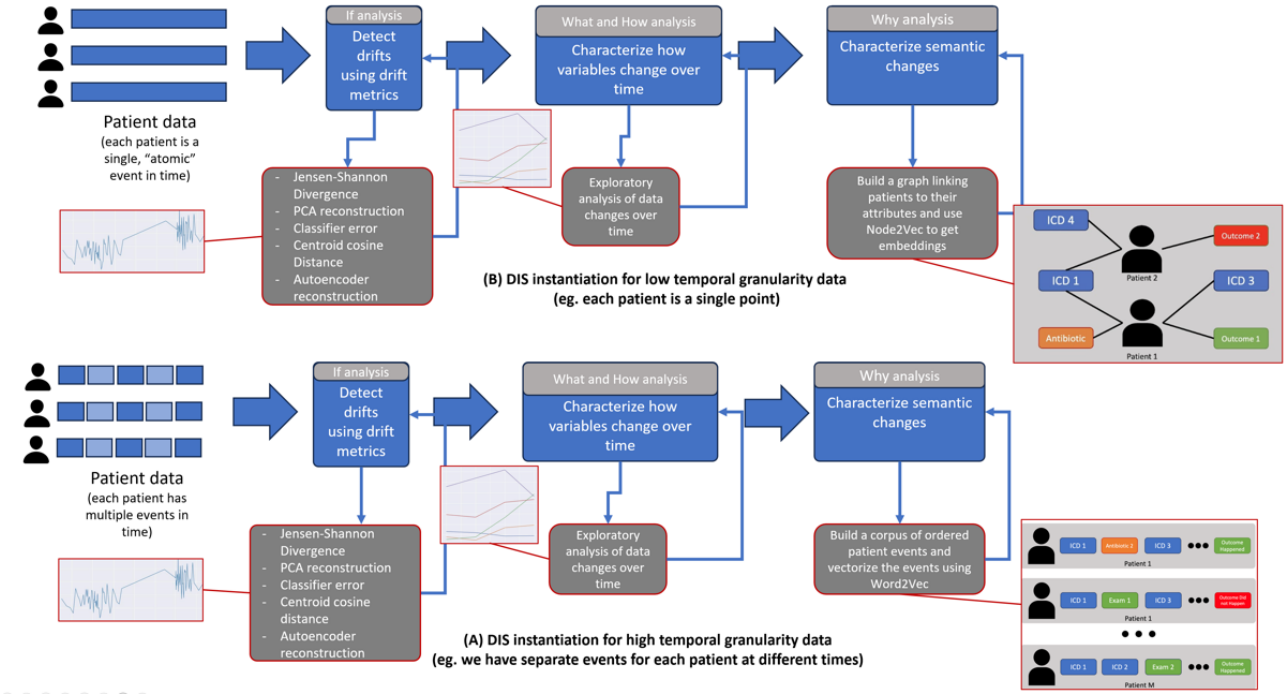
Results

Overview

The MIMIC-IV data set comprised 299,712 patients (median age 48, IQR 29-65 years), while the Brazilian COVID-19 Registry data set comprised 10,898 patients (median age 60, IQR 48-71 years).

Figure 3 illustrates how the DIS methodology is instantiated concerning the data's temporal granularity for each scenario. As explained, DIS consists of 3 steps (detection, initial characterization, and semantic characterization). The temporal granularity of the available data affects specifically the last step (semantic characterization). The figure also shows that several methods can be applied for the detection step. In our experiments, we tested and compared 5 different methods regarding their capability of accurately identifying temporal drifts in the detection step. In the second step, different exploratory techniques that measure the relationship between the dependent $P(y)$ and independent $P(x)$ variables over time can be used. We exploited multiple alternative techniques, such as feature importance and Pearson correlation. Finally, in the last step, our aim was to generate semantic embeddings for outcomes and other health care events over time and to derive insights from comparing these embeddings. We tested two different alternatives for producing such insights: (1) using our semantic embedding modeling and (2) using traditional clustering techniques over the untreated (original) data without the semantic treatment. The goal of using these 2 techniques was to illustrate insights that can be obtained with the semantic layer, which would be difficult to obtain otherwise.

Figure 3. Overview of the instantiation of the detection, initial characterization, and semantic characterization (DIS) methodology to 2 scenarios with different temporal granularities. (A) Medical Information Mart for Intensive Care, version IV (MIMIC-IV) DIS instantiation and (B) Brazilian COVID-19 Registry DIS instantiation. ICD: International Classification of Diseases; PCA: principal component analysis.



DIS Instantiation for the MIMIC-IV Data Set

A notable characteristic of this data set is its high temporal granularity, enabling the tracking of time progression within each individual's hospital stay. High temporal granularity means we know the sequence of health care events at the individual level. This facilitates obtaining invaluable insights into the relationships between such events, much like it helps us learn about the semantics of words in NLP. It has been consistently shown that the order of precedence between words and how often they appear with other words are representative of those words' semantics [26]. We claim that the order of precedence and cooccurrence between health care events can also contain the "semantics" of those events. A distributed representation built from these relationships could cluster similar health care events, such as the representation of different types of diabetes or hypertension and their associated complications, in close proximity in the space. Although all dates in the data set are anonymized for privacy reasons, we can track each individual's sequence of events using the provided masking of dates. This date masking is consistent in a manner that allows for time tracking during each patient's hospital stay, and it contains a special attribute that allows for the association of patients with the yearly interval during which they were hospitalized. These yearly interval data allow us to compare how patients in each year group behaved as a group, meaning we can also measure temporal effects at the populational level. The period covered by these data set ranges from 2008 to 2019.

In other words, the data set offers temporal granularity at both the population and individual levels. However, breaking this data set into arbitrary temporal chunks is challenging because the dates are masked. Despite this, the data set contains a nonmasked anchor year group that assigns each patient to an actual year interval during which they were hospitalized. Multimedia Appendix 2 explains how this variable works. Essentially, a random time delta is fixed for each patient and added to all relevant dates, effectively masking them while preserving the relative time intervals for that patient. Consequently, direct comparison of dates between 2 different patients is not feasible, except for their "anchor_year_group" variables. For instance, a patient hospitalized in 2015 may have (through the added random time delta) dates that appear later than those of a patient hospitalized in 2020. We can only directly compare dates within the context of each patient. The real year interval during which each patient was hospitalized is preserved in their "anchor_year_group" variable, which we use in all chunking for this data set henceforth.

DIS: Detection Step (MIMIC-IV)

As described, the temporal chunks in MIMIC-IV were given by the "anchor_year_group" variable. We used this variable to separate patients into the 4 groups provided within the data set. We then used alternative drift detection approaches, namely Jensen-Shannon divergence, autoencoder reconstruction error, PCA reconstruction error, centroid distances, and classifier prediction error in separating time chunks plot for this data set considering in-hospital ICD diagnosis. The Jensen-Shannon divergence formula is shown in equation 2, where *KL* is the Kullback-Leibler (KL) divergence [27], and *P* and *Q* are the 2 variables being compared.

We started step 1 of DIS with the *drift detection* substep. As previously described, the temporal chunks in MIMIC-IV were identified through the "anchor_year_group" variable. We used this variable to separate patients into the 4 groups provided within the data set. Figure 4 shows the Jensen-Shannon divergence plot for this data set considering in-hospital *ICD*

diagnosis. The Jensen-Shannon divergence is shown in equation (2), in which *KL* is the KL divergence, *P* and *Q* are the 2 variable distributions being compared, and we compute na average of each possible KL divergence combination between the two distributions. Since the KL divergence is asymmetric, the calculation described can be interpreted as a symmetric divergence between the two distributions. This metric was tracked to evaluate whether the data distributions changed over time, how fast they changed, and whether the data shift was temporary.

$$JSD(P||Q) = 1/2\ KL(P||M) + 1/2\ \mathrm{KL}(Q||M) \quad \textbf{(2)}$$

In equation 2, *KL* is the KL divergence, *M* is 1/(*P* + *Q*), and *P* and *Q* are the distributions of the variables we compared.

Figure 4 presents the results of our drift detection metrics, applied to the various "anchor_year_groups" in the MIMIC-IV data set. The figure depicts the normalized magnitude of the drift signal calculated per "anchor_year_group." The drift signals were normalized in the 0 range for visualization, as shown in equation 2. The results for the Jensen-Shannon divergence, PCA reconstruction error, and centroid cosine distances revealed a trend toward increasing distance between the variable distributions over time, which did not revert to prior levels, suggesting a gradual temporal shift. As seen in Multimedia Appendix 2, this drift occurred gradually over several years, with a more pronounced change between the first 2 temporal chunks.

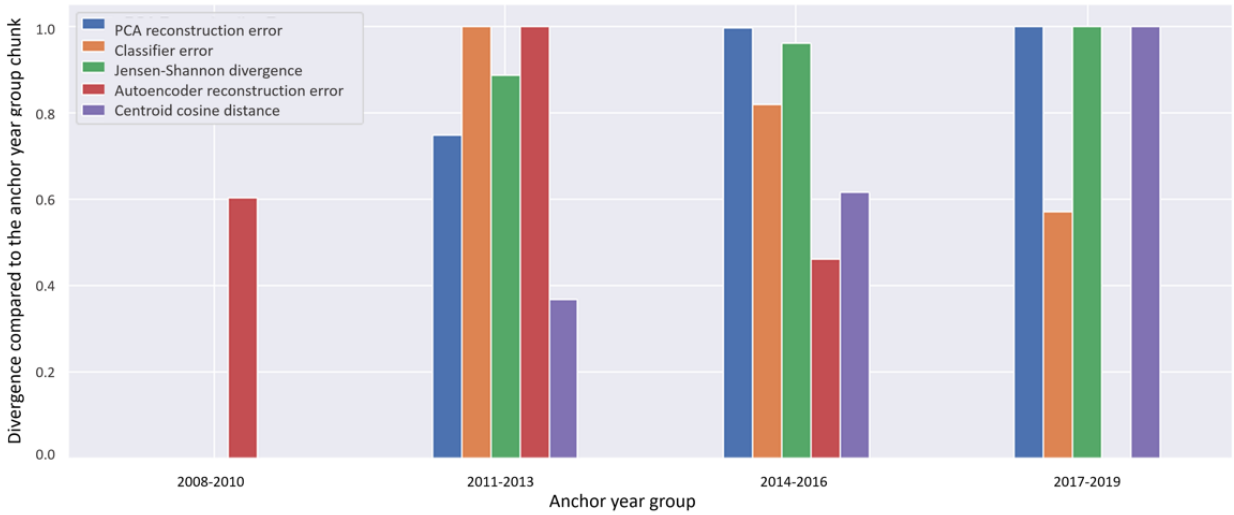
By contrast, when examining the autoencoder reconstruction error and classifier error metrics, a peak divergence was observed in the second time chunk (2011-2013), which gradually trended toward the baseline. As models with more parameters, these 2 drift metrics were sensitive to a combination of the data distribution, novel data points (ie, rare diseases or diseases not present in the reference time slice), and numerical outliers in the case of the autoencoder reconstruction error. For example, the disease codes appearing in the second chunk had the smallest intersection with the reference chunk, meaning they had the fewest diseases occurring concurrently in both chunks. This likely explains why the autoencoder reconstruction error and classifier error metrics exhibited their highest peaks in this slice.

In summary, the Jensen-Shannon divergence metric yielded more robust drift signals in our tests. It is important to note that the best metric depends on the most relevant type of drift for the data collection being analyzed. The Jensen-Shannon divergence is robust at detecting distribution changes, just as robust as the classifier error metric. If we are interested in detecting the occurrence of outliers or novel samples not seen before, the reconstruction errors might result in better detection. The choice of metric must be informed by the characteristics of the metrics themselves as well as the characteristics of the data stream being monitored.

$$\mathrm{NormalizedSignal} = (X - \min[X])/(\max[X] - \min[X]) \quad \textbf{(3)}$$

As show in equation 3, normalization is used to calculate the normalized magnitude of the drift signal.

Figure 4. Different drift detection metrics over time on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) data set, considering in-hospital International Classification of Diseases (ICD) diagnosis. PCA: principal component analysis.



DIS: Initial Characterization Step (MIMIC-IV)

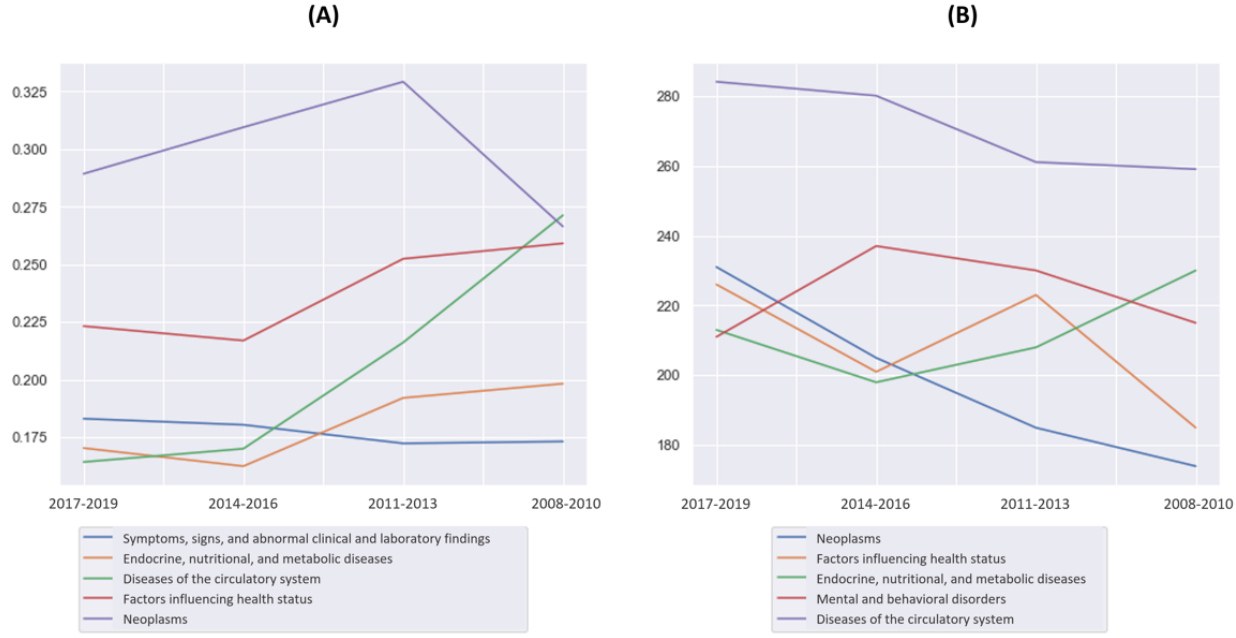
After establishing that a drift has indeed occurred, especially based on the results of the most accurate method, the Jensen-Shannon divergence, we proceeded to *step 2*. In this step, we strived to understand how the independent variables (*P*(*X*)) affect the outcome, which is our *dependent variable (P(y|X))*, and how the relationship between dependent and independent variables changes over time. This analysis can be accomplished by examining changes in correlations and feature importance over time, as well as by characterizing the distribution of different features over time. For instance, in Multimedia Appendix 3, we show how the relative distribution of the "death" outcome has changed over time in this data set. This means that our data exhibit a consistent trend toward in-hospital mortality reduction over time, which indicates a change in the relative distribution of the 2 possible categories (deceased × not deceased) for this outcome.

In Figure 5, we show the correlations and feature importance variations of the top 5 most correlated and the top 5 most

predictive *ICD* chapters (according to *ICD-10*) and the "death" outcome (according to feature importance). For instance, Figure 5A shows some expressive variations, such as how circulatory system diseases seem to grow more correlated with death over time, Figure 5B shows how neoplasms seem to become less predictive of death over time.

In Multimedia Appendix 4, we show how the different outcome groups behave over time from given baseline, in particular the patterns of independent variables given the outcome categories P(y|X) observed in the first temporal chunk. To obtain this result, we computed the arithmetic mean of each class's features in each "anchor_year_group" and calculated the cosine distance between these means over time, taking the first chunk as a reference to compare all other chunks against it. In this particular figure, we represent each patient as a "corpus" containing all their health care events (such as diseases and medications used during the hospital stay), then encode each feature as a 1-hot sparse matrix (each event can have the value "0," if it did not happen for a particular patient, or "1," if it did), and subsequently average these features. Notably, this representation treats each patient as a "bag of health care events," disregarding the order of precedence between those events, unlike what we did in our semantic characterization step. In the specific case, we show how the "death" outcome exhibits greater temporal drifts over the available time chunks in both data sets compared to the overall hospitalized patient population.

Figure 5. (A) Pearson correlations between the top 5 International Classification of Diseases (ICD) chapters (according to ICD-10) most correlated with the death outcome over time. (B) Feature importance among the top 5 ICD chapters most predictive of the death outcome over time.



DIS: Semantic Characterization (MIMIC-IV)

In Table 2, we show the top 5 *ICD-10* chapters that have become more and less similar to the "death" outcome over time. Notably, certain diseases, such as neoplasms, have become less similar, while others, such as malformations and circulatory system diseases, have become more similar. That is consistent with the findings in step 2, and over the next few paragraphs, we describe the procedure to obtain this similarity score. We explain the token-level vectorization process for both dependent and independent variables in Figure 6. First, we compiled a temporally ordered list of patient data, consisting of discrete data points such as items consumed during hospital stay (antibiotics, anti-inflammatories, etc), disease codes (using *ICD*), and procedures. At the end of each patient's sequence, we appended the outcome category for that patient. To classify the outcome, we divided binary outcomes into distinct tokens, such as "deceased" and "not deceased," and used the corresponding token to generate our training corpus. Continuous outcomes and dependent variables could be binarized using a simple histogram binarization scheme, as demonstrated in the next analysis. Following the corpus generation, we used it to train token embeddings with Word2Vec [26]. This method produced embedding vectors for both dependent and independent variables, allowing semantic comparisons between different entities, such as the "death" outcome and different disease codes. We created 1 outcome token for each outcome category and temporal chunk in our data set. This allowed us to evaluate how an outcome such as "death" may have drifted closer to or farther from certain diseases or procedures over time.

In Multimedia Appendix 5, we show the top 5 conditions that became more similar to the "death" token and the top 5 conditions that became less similar when comparing the first and last time chunks. Since every entity is a "token," we could evaluate similarities between diseases and disease chapters, between patients and diseases they have not yet been diagnosed with, and between outcomes and diseases (Multimedia Appendix 5). In particular, in Multimedia Appendix 6, we demonstrate changes in similarity for the "dysphagia following stroke" *ICD* code within the MIMIC-IV data set [18]. Our analysis revealed a rise in the simultaneous appearance of ICD codes related to obesity between the periods of 2011 to 2013 and 2017 to 2019.

This trend aligns with broader observations indicating an uptick in obesity rates across the United States. Importantly, it is essential to recognize that this method does not permit the establishment of causal relationships; rather, it emphasizes changes in correlation and cooccurrence.

The *step 3* analysis can also be conducted at different levels of granularity to gain a deeper understanding of the observed changes. From step *2*, it can be inferred that mortality has been decreasing and has some relationship with particular disease groups. If *step 3* is performed at the disease code level, as shown in Multimedia Appendix 6, chapters that had considerable shifts in their similarity to the "death" outcome, either increasing or decreasing similarity, can be identified. For instance, the findings confirm what is illustrated in Figure 6, where "cancer" shows a decreasing similarity to the outcome, while the variable "circulatory diseases" exhibits an increasing similarity to the outcome. This observation is further supported by the results shown in Multimedia Appendix 7, where an absolute increase in the number of patients with cancer over time is shown, associated with a relative decrease in in-hospital cancer-related deaths between 2008 and 2019.

To further illustrate how the proposed DIS semantic analysis based on embedding distances among entities of interest can help in better comprehending the reasons for the drifts, we contrasted the previous analyses of our third step with a traditional clustering analysis for the MIMIC-IV data set. This analysis used a syntactically oriented term frequency–inverse document frequency (TF-IDF) [28] representation for the entities, built from the same corpus of clinical entities. In a TF-IDF representation, each dimension corresponds to a unique term (word) in the document corpus. The value in each dimension reflects the importance of that term in a specific document, calculated by multiplying the term's frequency in the document (term frequency) by the inverse frequency of the term across all documents (inverse document frequency). In our case, each "document" was a patient, and each "word" was a health care event, such as the identification of a novel disease. We applied a spectral clustering [29] procedure to the TF-IDF representation of the entities to create the clusters. The results are shown in Figure S8. To obtain the 4 clusters displayed in Multimedia Appendix 8, we used a silhouette analysis using 2 to 15 clusters.

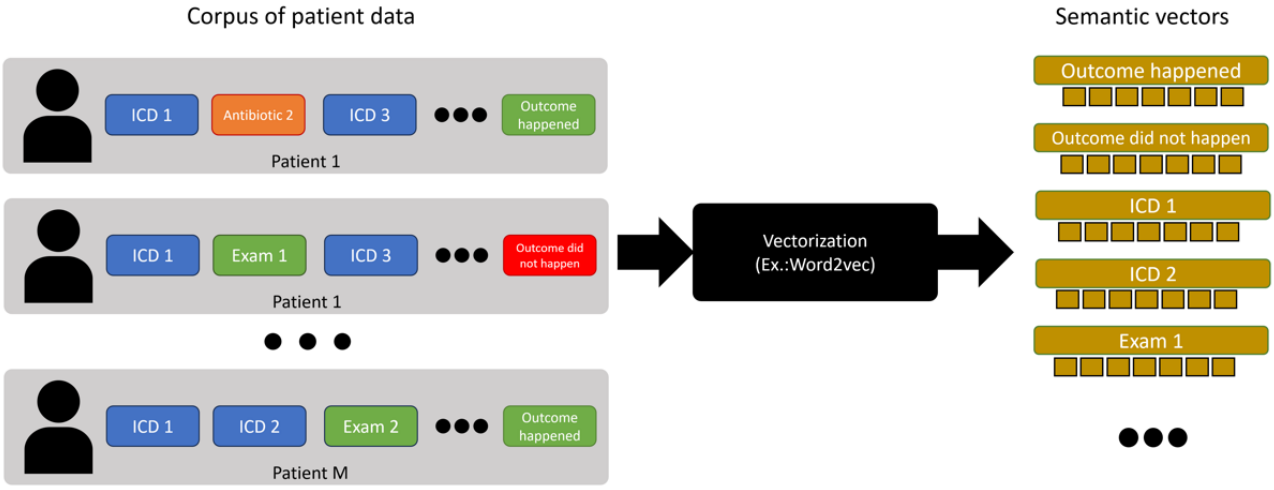
Multimedia Appendix 8 shows the top 5 most frequent diseases for each of the 4 clusters (y-axis). On the x-axis, we present the index of each cluster. Multimedia Appendix 8 shows how the relative frequency of each cluster changed over each "anchor_year_group." A few points stood out from the clustering analysis illustrated in Multimedia Appendix 8. As it can be observed, the cluster analysis using syntactically oriented vectors made it harder to interpret the drivers of a data drift when compared to DIS. For instance, some semantically similar diseases, such as "other and unspecified hyperlipidemia" and "hyperlipidemia, unspecified," may have very distinct profiles in different clusters, such as in clusters 0 and 2, each having a high concentration of patients with either one of these diseases. The main problem of this particular cluster analysis based on syntactically oriented representation is the separation of semantically similar entities into distinct clusters. In DIS, similar entities will be represented similarly and thus analyzed in conjunction.

Table 2. Change in similarity by ICD[a] chapter.

ICD chapter	Change in similarity	Direction
Diseases of the nervous system	–0.14	Less similar
Diseases of the musculoskeletal system	–0.12	Less similar
External causes of morbidity and mortality	–0.10	Less similar
Diseases of the digestive system	–0.08	Less similar
Neoplasms	–0.02	Less similar
Congenital malformations	+0.40	More similar
Diseases of the circulatory system	+0.35	More similar
Diseases of the genitourinary system	+0.30	More similar
Endocrine, nutritional, and metabolic diseases	+0.25	More similar
Diseases of the skin and subcutaneous tissue	+0.20	More similar

[a]ICD: International Classification of Diseases.

Figure 6. How to generate semantic vectors? We start by generating a corpus of temporally ordered patient discrete data points. Then, we vectorize the tokens of this corpus using Word2Vec to obtain semantic vectors for dependent and independent variables. ICD: International Classification of Diseases.



DIS Instantiation for the Brazilian COVID-19 Registry Data Set

The median age was 60 (IQR 48-71) years, and 21.72% (2367/10,898) were women (5012 patients). In this data set, 21% of registered patients died, yielding an unbalanced classification problem when predicting future deaths. The data set has low temporal granularity, with only 1 data point per patient, which precludes time tracking during hospital stays. Consequently, we could measure time only at the populational level. In other words, unlike the previous case study, there was a single "snapshot" for each patient, with no temporal evolution at the individual level.

DIS: Detection Step (Brazilian COVID-19 Registry Data Set)

As in the previous case study scenario, we evaluated the same 5 alternative techniques, namely the PCA reconstruction error, autoencoder reconstruction error, classifier error (in separating past vs future), and Jensen-Shannon divergence. All these metrics measure the drift compared to a reference temporal slice and do not require setting a specific outcome or using labeled data.

The outcomes of this procedure are illustrated in Figure 7, where the divergence sharply increases starting from the final quarter of 2020, based on the Jensen-Shannon divergence metric. Numerically, a drift is indicated in this interval as the divergence levels surpass a user-defined threshold, such as a fixed threshold of 2 SDs or a threshold informed by domain expertise. As depicted in the figure, the PCA reconstruction error, autoencoder reconstruction error, and centroid cosine distances indicate positive drift signals in the quarter starting from April 2020. During this semester, the Brazilian COVID-19 Registry data set exhibited a small number of numeric outliers, which were identified by these methods. Conversely, the Jensen-Shannon method signaled a data drift in the quarter starting from October 2020, which aligns with the "official" start of the second wave in Brazil in November 2020. Meanwhile, the classifier error method indicated a drift in July 2020, which falls between the identification of numeric outliers and the actual distribution change from the first wave to the second wave. Both the Jensen-Shannon method and the classifier error method signaled drift closer to known actual changes, while the other, more reconstruction-based methods were more sensitive to numeric shifts, which were not necessarily associated with changes in the underlying distributions.

Figure 7. Different drift detection metrics over time in the Brazilian COVID-19 Registry data set. PCA: principal component analysis.

DIS: Initial Characterization Step (COVID-19)

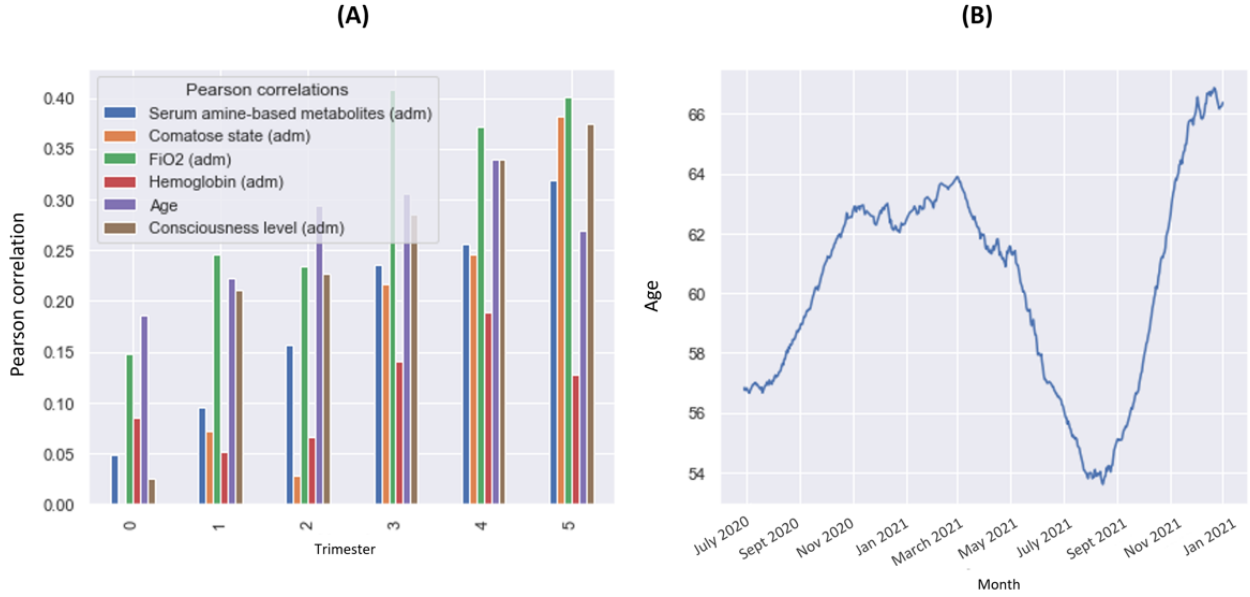
Once a drift was detected, we proceeded with the second DIS step, initial characterization. This step aims to understand the main drivers ("what") of changes during the considered period and "how" they affect the underlying outcomes. At a high level, this begins with the characterization of the changes in the outcome (the independent variable) over time. In Figure 7, we illustrate this upon evaluating the variation in COVID-19–related mortality in our data set. This example displays a trend toward a reduction in the death outcome over time. At the initial characterization step, it is expedient to examine the distribution of the outcome of interest (death, ICU admissions, etc) as well as those of the most predictive independent variables (eg, those with the highest correlation with the desired outcomes or higher feature importance in a tree-based classifier).

To guide the next steps, it is helpful to check how much each outcome category's properties (such as the mean age of the deceased patient population or the prevalence of hypertension) have changed over time. In particular, focusing on which outcomes have changed the most helps target specific subsets of the data that could better explain the observed phenomena. We show an example in Multimedia Appendix 9, where we analyzed such variations in the Brazilian COVID-19 Registry data set. To build the graphs in this figure, we split our data set into time chunks. For each chunk, we separated all patients into classes according to their outcome (eg, dividing the population into deceased and nondeceased and then representing the chunk by averaging all of the patient's features in each category). For each subgroup of patients within the same time chunk and sharing the same outcome, we computed the centroid of that class (the arithmetic mean of all attributes). We then took the first chunk as a reference and compared each class's chunk arithmetic mean to the reference using a cosine distance. Multimedia Appendix 9 shows how much the deceased patients' characteristics changed more than those of the overall population during the same period.

A better comprehension of the drift drivers during the COVID-19 pandemic emerges from Figure 8. As shown in Figure 8A, we observed how the overall best predictors of death changed over time through Pearson correlation analysis conducted each trimester on the data set. At the beginning of the pandemic, age was the single best predictor of death, in trimesters 0 and 2. As the vaccination campaign started, older adults were prioritized and received immunization first. This led to a progressive deterioration of the predictive value of age, as well as an overall decrease in mortality (Multimedia Appendix 10), culminating in the latest trimester, where age was the worst predictor among the top 5 variables. In Figure 8B, it can be seen that the median age of the deceased patient population over time.

In summary, the second step revealed that the COVID-19 data showed a progressive decrease in patient mortality (Multimedia Appendix 10), with a more pronounced change in the group of deceased patients (Multimedia Appendix 9). It was also possible to notice that the overall characteristics of the patients who were dying changed abruptly (Figure 8). From the remaining characterization steps in Figure 8, we can see that age lost its predictive capacity (Figure 8A) over time, while clinical features such as the patients' fraction of inspired oxygen (FiO_2) became better predictors of death. Concurrently, there was a reduction in the median age of patients who were dying (Figure 8B).

Figure 8. (A) Pearson correlations over time for the overall top 6 most predictive variables in the Brazilian COVID-19 Registry data set. (B) Median age of hospitalized patients dying from COVID-19.



DIS: Semantic Characterization (COVID-19)

Following the conclusions from the previous step, we moved further into the semantic characterization step. As the Brazilian COVID-19 Registry data have low temporal granularity and most of their features are continuous, what requires data categorization to enable the use of NLP techniques to treat words and other semantic units.

Subsequently, due to low granularity at the individual level, we needed to model relationships between these now-discrete entities. In more detail, we assumed that the temporal precedence between events imposes a relationship between them and that this relationship can be learned and embedded into a distributional representation. The issue with low temporal granularity data is that the order of precedence is not known; hence, it is not possible to model it directly. Therefore, we modeled all health events (from the perspective of a single individual) as if they happened simultaneously. Therefore, in this setting, we modeled the passing of time only from the perspective of the population and not from the perspective of the individual. This means that we only knew, for instance, that a given patient had events (such as new diseases or use of medications) 1, 2, and 3, but we did not know the order of precedence between these attributes, something that was explicit in the MIMIC-IV data due to high temporal granularity. We began by discretizing the continuous features with a histogram discretizer, which essentially breaks the data intervals into "equal width segments" and then assigns a "token" (ie, a string or integer value) that is unique to patients having that attribute in that specific range of values.

After that, we created a graph with patients, discretized continuous attributes, discrete attributes, and outcomes, such as the one in Figure 9. To build this graph, we connected each patient to their attribute tokens and outcomes while creating one outcome token for each time chunk under analysis. Finally, we embedded the graph using a node embedding algorithm such as Node2Vec [30]. We contrasted this procedure with the one adopted to characterize the MIMIC-IV data set (Figure 10). As discussed before, in MIMIC-IV, the temporal order is defined at the individual level, with entity relationships determined by the timeline. By contrast, the Brazilian COVID-19 Registry data set presents events as occurring "simultaneously" at the patient level, limiting the understanding of relationships between events and patients. In this case, to derive semantic vectors representing entity relationships, we approached it as a graph vectorization problem.

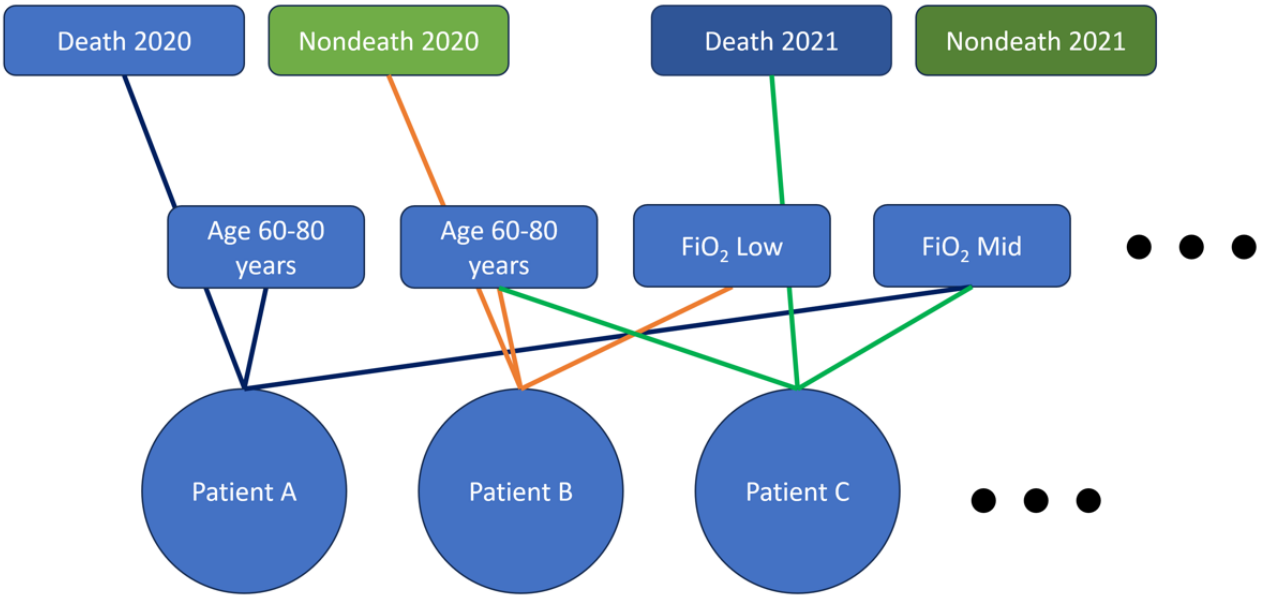
To analyze the resulting model, we compared the outcome embedding vectors to evaluate their similarity to each other and to other patient attributes. We show the results of this procedure in Multimedia Appendix 11. From that, it is evident that the 2021 death outcome token increased in similarity to lower age groups, such as age groups 18 to 39 years and 40 to 61 years, while decreasing in similarity to older age groups, such as age groups 62 to 83 years and 84 to 105 years. This observation further validates the previous findings and introduces new elements not captured in earlier steps. We could also see an increase in similarity to lower admission heart rates and lower admission serum sodium values, as well as lower FiO_2 at admission, showing a shift in disease severity markers over this time frame.

As mentioned earlier, for comparative purposes, to emphasize the semantic capabilities of the proposed DIS procedures, we compared our semantic-step results with those obtained through traditional clustering analysis for the Brazilian COVID-19 Registry data set (Multimedia Appendix 12). In this analysis, entities were represented by a syntactically oriented TF-IDF representation. In Multimedia Appendix 12, we show the top 5 highest-value features for each of the 6 clusters selected using the silhouette analysis, as was done for the MIMIC-IV case. In Multimedia Appendix 12, we show how the relative frequency of each cluster changed over time in each trimester.

Similar to the MIMIC-IV case, the clustering analysis of the COVID-19 data was not as straightforward to interpret as the

DIS analysis when searching for the drivers of data drift. For example, we identified a cluster of patients who underwent a "transplant" and another cluster of patients with diabetes mellitus type 2, but the reasons why these particular clusters were selected and the reasons for the drifts could not be easily derived through a straightforward analysis of these syntactically oriented clusters.

Figure 9. Example of how to create a patient graph with tokenized dependent variables and temporal outcome tokens.



Discussion

Comparison With Prior Work

Multiple studies have analyzed variations observed over time in class distributions and model effectiveness and their overall impacts. Studies such as Salles et al [31] and Mouro et al [32], for instance, performed a detailed characterization of such effects in textual data sets of documents organized into topics. Health care data, however, are quite different from simple text data [31,32]. To begin with, this type of data is multimodal, including tabular and sequential information in the form of vital measurements, disease code diagnosis, and items consumed during a hospital stay, as well as common text, images, wave forms, and sometimes even sound waves. Furthermore, the data may experience sudden and specific drifts driven by new medications, vaccines, surgeries, and public policies [9]. For example, an effective vaccine may cause the eradication of a disease, resulting in a subsequent data drift [33]. While most studies on health care data focus on either drift detection or drift adaptation [33,34], our work is unique in that it focused on drift detection, monitoring, and characterization. We advanced the existing literature by leveraging these 3 steps to pursue explanations for health care data drifts.

Concerning terminologies and problem-setting definitions, Gama et al [35] defined data changes as being related to the distribution of the independent variables P(X) and dependent variables P(y) or the conditional probability of dependent variables for given independent variables P(y|X). Works unified and consolidated some of the underlying terminologies [21,36]. As defined by Lu et al [36], data and concept drifts can be categorized based on how they behave over time, being (1) sudden (ie, 1 event permanently changes the "meaning" of a concept), (2) incremental (ie, 1 event incrementally generates gradual changes to the "meaning" of a concept), (3) gradual (ie, the concepts interchange gradually until the complete shift occurs), or (4) reoccurring (ie, a transient concept drift).

Approaches to detect and learn in the presence of concept drifts do exist. However, in most contexts, naively monitoring data drifts may be expensive, as it often requires data labeling. As an alternative approach, Haque et al [37] used an ensemble of classifiers to report their prediction confidences and monitor changes in their confidence distribution to detect when a concept drift occurred. In the data sets used in this paper, however, deaths are readily available labeled data, which means that our main issue was related to learning in the presence of a data drift.

A common approach to drift detection is monitoring model outputs, as in the study by Sahiner et al [38]. These "model monitoring" approaches are not always possible or desirable; for instance, Tiwari and Agarwal [39] argued that labels are a resource that is not always available and suggested exploring other options, such as detecting drifts by monitoring changes in the underlying data distributions. Following this idea, we propose a *drift monitoring* procedure that is independent of labels and focuses on distribution changes over time. Additionally, Tiwari and Agarwal [39] provide a comprehensive review of useful health care data type classification and data drift management strategies in data streaming scenarios. Textbox 2 details the categorization of health care data proposed by these authors.

In addition to the categorization mentioned in Textbox 2, Tiwari and Agarwal [39] discussed the use of sampling in diverse forms to handle data streams and drifts. In health care data, it is common to encounter massive data sets encompassing multiple years and thousands of patients. For such cases, sampling may

be a viable option. Given the size and nature of our data sources, we opted to work with the complete data set available instead of using sampling. The decision to use sampling should be evaluated depending on the type of ML algorithm used, the available computing capabilities, and the data set size.

Drift detection has multiple beneficial impacts on health care. Once detected and treated, a drift can be used to help maintain and enhance model effectiveness. Additionally, it can be useful to detect whether a new treatment is changing the outcomes of a disease in a meaningful manner or even understand populational trends to derive health policies. A recent example is the COVID-19 pandemic. This topic was explored in the studies by Jung et al [41] and Jassat et al [42], which showed differences in hospitalized patient profiles as new COVID-19 waves spread. Another study has explored how the death prediction task evolved throughout the pandemic, showing that factors such as vaccination changed the profile of patients who were severely ill [4]. These characterizations can help in the detection of important pandemic events, such as the impacts of vaccination, the emergence of new COVID-19 strains, and the emergence of new viral strains resistant to current therapies. In this context, we focused our characterization efforts on technology evaluation through the lens of data drifts in a health care setting.

Some solutions have been reported in the literature to address the challenge of learning in the presence of data drifts, and most of these solutions focused on sample selection or sample weighting, with variations on how they derive the final weighting or sampling. Klinkenberg [43], for instance, tackled the problem by using support vector machines for both sample selection and sample weighting, using an iterative process that sequentially trains support vector machines to find the training instances that constitute the model's support vectors [43]. Kolter and Maloof [44] used a special weighted ensemble to learn in the presence of such drifts. Salles et al [6,31,45] used a temporal weighting function that can be automatically learned to select relevant samples for each training window. Finally, Rocha et al [7] tackled the problem using temporal contexts. The authors analyzed document collections that evolved over time and defined a temporal context as portions of documents that minimize the temporal effects of class distribution, term distribution, and class similarity over time. This method is used to devise a greedy strategy to optimize the trade-off between undersampling and temporal effects. We were inspired by this latter work in our methodology. Most of these approaches, however, are not applied to the health care setting, focusing mostly on common text data.

Another relevant setting is detecting drifts in data streams. This is potentially relevant to some health care data, especially sensor data, which are most commonly obtained from hospitalized patients but also streamed from personal health devices such as smartwatches and heart rate sensors. Zliobaite et al [46], for instance, proposed a continuous loop of labeling new samples under a labeling budget and used active learning to detect data drifts.

Class imbalance is another important aspect of detecting data drifts in health care data. Disease occurrence is naturally unbalanced, with common diseases such as diabetes or hypertension affecting between 5% and 30% of the population [47,48]. Rare diseases, by contrast, have a prevalence in the order of <10 patients per 100,000 or 1,000,000 inhabitants, with combined prevalence among all rare diseases being estimated to be between 3.5% and 5% [49]. Most approaches to handling such class imbalances in the data drift literature focus on oversampling, undersampling, or a combination of both. Gao et al [50], for instance, proposed oversampling the minority class over multiple time slices while undersampling the majority class using only the most recent slice. Ditzler and Polikar [51], by contrast, focused on using incremental learning combined with the synthetic minority oversampling technique [52] to learn a classification ensemble that can deal with both the class imbalance and concept drifts in streamed data. In particular, the combination of models and data sets used in our work was robust to such class imbalance issues and did not require using these types of techniques, as discussed in the following sections.

Textbox 2. Categorization of health care data.

Categories

1. Clinical data, such as the records in Medical Information Mart for Intensive Care, version IV (MIMIC-IV) [18] and the Brazilian COVID-19 Registry [17], are desirable if the goal is to describe data drifts related to the impact of specific interventions, such as the introduction of a new drug or therapy.
2. Self-administered data, obtained from questionnaires, usually investigate lifestyle variables, such as smoking or alcohol consumption habits.
3. Biological data, usually obtained by measuring parameters in biological samples such as blood and urine, are often the result of a laboratory study.
4. Molecular data are the kind of data encoded in protein databases such as UniProt [40], genome databases, or even drug-to-molecule interaction databases.
5. Exposure data encode patients' exposure to given events, drugs, or interventions.
6. Modeling data are data generated from models, including estimated risks given the patient's exposure.

Summary of the Main Results of Applying DIS to the MIMIC-IV Data Set

The instantiation of the drift detection step using several distribution comparison metrics showed the flexibility of the methodology. It also demonstrated that, for the purpose of separating the temporal chunks in this particular scenario, metrics such as the Jensen-Shannon divergence or the classifier errors capture the underlying distributions better than particular outliers or novel samples. Higher values in these metrics imply more significant "populational" changes, such as a gradual shift in the composition of the in-hospital population's disease burden.

As seen in the drift detection step (Table 2), there is a gradual but persistent pattern in MIMIC-IV, happening over several years. This gradual change may be caused by various factors, such as an increased tendency for patients who are terminally ill to receive end-of-life care at home or advancements in therapeutic techniques for certain diseases. The nature of the expected data change can be hypothesized based on characteristics such as the suddenness or gradualness of the drift, its persistence, and its duration, along with the results from the next analytical steps in DIS. This difference becomes evident when comparing the MIMIC-IV and the Brazilian COVID-19 Registry data sets.

The initial characterization step (Figure 5) revealed a trend toward a decrease in overall mortality over time, and this is the "context" in which we interpreted subsequent findings. Additionally, Figure 6 indicates that the overall characteristics of the deceased patients changed more than those of the overall in-hospital population over the observed time frame. This means that the reduction in overall mortality is due to changes in the characteristics of the patients who died. The findings in Figure 5 show how different diseases impacted mortality predictions over time. Figure 5 shows that 2 *ICD-10* chapters, "diseases of the circulatory system" and "cancer," had important changes during this period. By associating the findings of step 1 with those of *step 2*, we can begin to understand the factors contributing to decreased mortality over time, but it does not provide the "full picture."

The DIS semantic characterization step, which measures how the contexts of the independent variables relate to those of the dependent variables over time at a more semantic level, yields interesting results that complement the previous ones. Multimedia Appendix 6 shows an example of such a result, that is, changes in similarity for the "dysphagia following stroke" ICD-10 code within the MIMIC-IV data set [18]. There has been an increase in the cooccurrence of many obesity-related ICD codes between the 2011 and 2013 and 2017 and 2019 time slices. This is aligned with general observations of the increase in obesity prevalence in the overall US population. It is worth noting that this technique does not allow us to draw causal conclusions but instead focuses on the correlation and cooccurrence changes. The cooccurrence of death and "cancer," as well as the presence of "external causes," has decreased over the period, possibly indicating a reduction in iatrogenic events, improved cancer treatment leading to lower lethality, or that patients with cancer are receiving more end-of-life care at home. This may be an explanation as to why overall in-hospital mortality has decreased in this data set.

As overall mortality decreases, patterns affecting the decrease of similarities between entities, such as the lethality of circulatory diseases, unchanged. This means that increases in similarity with the outcome may be simply due to the decrease in the lethality of other groups. To investigate this, we filtered the data only for cancer disease codes, as in Multimedia Appendix 5. The figure reveals important decreases in mortality in mostly severe and hard-to-treat cancers, such as brain, colon, lung, and secondary (metastatic) tumors.

It is also possible that the observed patterns may be attributed to multiple factors at the same time. For instance, recent policy changes favoring home care for patients who are terminally ill may influence who dies in the hospital. If these patients are more likely to die at home, we might have a "survivorship bias," where mostly the ones who did not die received hospital care and the patients who were terminally ill were sent back home. Over this time frame, there were important advances in immunobiological therapies for tumors, such as lung cancer, as well as early diagnostic techniques that have made it possible to cure some early cases when the tumor is still resectable. Combining these factors yields a lower lethality, which has decreased over time despite an increase in the total number of patients with neoplasm, as shown in Figure 6.

In summary, the application of the DIS methodology to the MIMIC-IV data set allowed us to determine important trends that help understand certain phenomena observed in the data. Moreover, it facilitates the formulation of interesting hypotheses, which are harder to validate based only on the data themselves. Nevertheless, in a real-world scenario, such hypotheses could be the subject of further investigation using other data sources, such as official policy implementation records, country-wide demographic records, or even published literature.

Summary of the Main Results of Applying DIS to the Brazilian COVID-19 Registry Data Set

The *drift detection* step, especially using the Jensen-Shannon divergence, revealed important data drifts in the Brazilian COVID-19 Registry data set, which commenced approximately at the same time interval as the vaccination rollout in Brazil, between late 2020 and early 2021 [14]. The initial characterization revealed a trend toward decreasing mortality over time, with the steepest decrease closely matching our drift detection. This means that thus far, there has been an important variable distribution shift as well as a change in the distribution of the outcome itself.

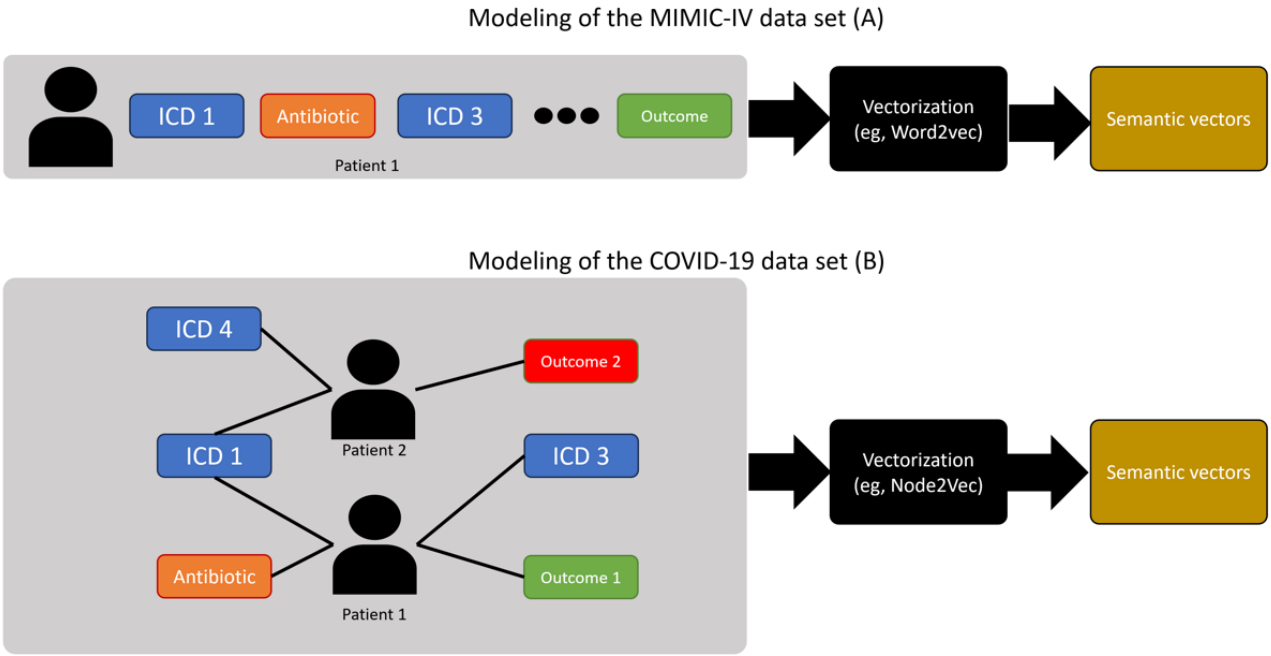
We analyzed how the top 5 highest Pearson correlation variables behaved over time (Figure 8). Figure 8A shows how the relative ranking and correlation of the best predictors of death changed over the course of the pandemic, with features such as "age" being the strongest predictors at the early stages and gradually becoming less predictive over time. Figure 8A also shows how patient severity markers, such as "FiO_2" and "altered level of consciousness," gradually became more important predictors over time, hinting at the change from "older patients dying from

COVID-19" to "patients who were severely ill at admission dying from COVID-19." From our analysis, the patient's age is shown to be a consistently robust predictor of COVID-19–related hospitalization and death. In Figure 8B, we show the median age of the patients who died from COVID-19. This shows how one of the most predictive features in this data set has changed over time, with the median dying age decreasing from approximately 63 years at the peak of the pandemic to approximately55 years in a time frame coinciding with the start of the vaccination campaign in Brazil [5]. However, the median age starts to rise again, possibly relating to another drift, such as the emergence of new viral strains that can disproportionately affect the older adult population. This fluctuation in the median age of deceased patients leads to the aforementioned deterioration of the correlation scores. Furthermore, this pattern with the age variable decreasing over time is consistent with how the vaccines were rolled out to the public, with older age groups being prioritized for vaccination [46]. If these groups received vaccines earlier and consequently reduced their probability of death, this would likely reduce the median and mean deceased patients' ages.

The main results of the semantic characterization step (Multimedia Appendices 10-12 and Figures 8-10), where we compared the semantic vectors for the "death" outcome in 2020 and 2021, validate several findings from the initial characterization step and introduce new findings. For instance, the results show a decrease in similarity between the outcome and older groups (eg, the age groups "84-105" years vs "62-83" years) with an increase in similarity between the outcome and younger groups. This validates the findings in Figure 8A, where median age declines steadily up until roughly September 2021. Figure 8A also shows how the "death" outcome had an increase in similarity to several disease severity markers, such as lower admission serum sodium, lower admission arterial blood pressure, fewer comorbidities, and lower FiO_2. This potentially indicates that, when compared to 2020, patients who died in 2021 were more severely ill at admission, had fewer comorbidities, and were younger (presumably unvaccinated). This is a significant pattern change, especially compared to the bulk of deceased patients in the initial chunk, who were mostly older adults with lower severity at admission. This change in pattern implies that, at the analyzed time frame, patients who were young and severely ill at admission were more common among patients who were dying. However, this should be analyzed in conjunction with the previous findings from the other steps. For instance, we know that the overall mortality has decreased, and this patient profile (young and severely ill at admission) could also be present in the first temporal chunk. What possibly happened was the removal of a significant portion of older patients who were dying from the population through events such as vaccination, as evidenced by the reduced mortality and diminished predictive power of age.

To conclude, the DIS analysis hints at the central role of vaccination in the COVID-19 pandemic, which reduced the odds of older patients dying from the disease following the rollout of the vaccines. This hypothesis was raised by the alignment between the detected data drift and mortality reduction during the vaccination period. Additionally, the observed decrease in the median age of the patients who were dying corresponded to the age-stratified vaccination strategy. Furthermore, the shift of mortality burden to patients who were young and severely ill upon admission, who were likely unvaccinated, demonstrates how they possibly kept dying while this process unfolded.

Figure 10. (A) Modeling of the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) data set as an ordered sequence of patient tokens. (B) Modeling of the Brazilian COVID-19 Registry data set as a graph connecting multiple patients through their common token. ICD: International Classification of Diseases.



Limitations

We have proposed a methodology to discover and interpret temporal shifts in health care data. While our approach provides valuable insights by uncovering many correlations and semantic connections, DIS still cannot establish causal relationships outcomes and semantic units. The causal part is only hypothesized and inferred, but the methodology does not go so far as to return causal links for arbitrary outcomes. Furthermore, we have not applied the methodology to certain relevant health care domains, such as images (eg, x-rays, computed tomography, or ultrasound) and wavelets (eg, electrocardiograms or electroencephalograms).

That said, here, we offer some insights into how we could apply DIS to handle temporal shifts in nonquantitative data or raw magnetic resonance imaging data. For this, we would first need to obtain a distributed representation of the data in such a manner that samples from similar patients have similar embedding vectors. For instance, we could use DINOv2 embeddings or contrastive language-image pretraining embeddings in images. This type of pretrained neural network exists for multiple data types, which facilitates its application to multiple domains. From the embeddings, we can apply the first step of our methodology as applied to tabular data, computing Jensen-Shannon divergence (or autoencoder errors, classifier errors, etc) to detect whether a drift exists in the data. Exploring these data in the second step presents some challenges, as it might involve exploring both the embedding and raw data spaces. For instance, we can use clustering and centroid analysis (applied to the embeddings) to find samples where the drift is particularly pronounced. Then, we can go back to the raw data and analyze the samples to check for patterns. In essence, the third step remains similar in nature. The idea is to train a neural network model such that the embeddings of the samples closely resemble the embeddings of the outcomes experienced by those patients over time. One such way to obtain these embeddings, starting from pretrained ones, is to use losses such as the triplet loss to approximate patient sample embeddings from outcome embeddings. The interpretation of the triplet loss, as presented in our paper, will change according to the temporal granularity of the samples. If the data have high temporal granularity, the positive pairs (which the loss will learn to represent more closely in space) will obey an ordered sequence of events. For instance, 2 magnetic resonance imaging tests will be proximate if they belong to the same patient and happen close to each other in time and if they are visually and semantically similar. Conversely, if the data have low temporal granularity, the embeddings should be learned to align patient samples to their outcome embeddings. Then, for the analysis of such embeddings, we would have to analyze the raw data samples closer to the outcome embeddings. If one splits the time, say, in 2 years and is working with the "death" outcome, one would be expected to have 1 such outcome for each year. Then, analyzing the samples closer to each of the outcome embeddings should help build an understanding of the relevant changes in a more generalized setting, and this might require some domain expertise. We intend to explore these ideas in future work.

Finally, we cannot claim that our 3 steps (encompassing the "if," "what," and "why" of a data drift) are a comprehensive list of all possible steps to analyze a temporal shift. Instead, we believe our steps to be a minimum required subset. While it is possible that these steps might not cover all possible situations, they allowed us to obtain interesting insights from the 2 data sets presented in our work, as discussed earlier. We and other researchers plan to continue to study, extend, and adapt this methodology in future work to test the limits of our approach and whether new steps or a refinement of the ones proposed at the fiber granularity level is necessary.

We intend to explore methods for enhancing models' resilience to data drifts, as well as examine different health care–relevant data types, such as images, wavelets, and multimodal data.

Conclusions

We have proposed DIS, a temporal data drift methodology for analyzing the changes in health outcomes and variables over time while discovering contextual changes for outcomes in large volumes of data. We applied DIS to 2 very different case studies and demonstrated how it can provide valuable insights into changing patterns in the data and the underlying reasons driving such changes.

The DIS methodology goes beyond simple detection; it comprehensively characterizes temporal data drifts. By analyzing the underlying causes, patterns, and magnitudes of drifts, health care stakeholders can gain a deeper understanding of the factors influencing data changes over time. This deeper understanding has practical implications for health care organizations, allowing them to improve patient care, optimize resource allocation, and enhance operational efficiency by leveraging the insights gained from monitoring and characterizing temporal data drifts.

The practical implications of our methodology are far-reaching. Early detection of data drifts can trigger timely interventions, enabling proactive adjustments to treatment plans, health care policies, and quality improvement initiatives. Our methodology empowers health care practitioners and data analysts to effectively monitor and manage temporal data drifts, ultimately leading to better health care outcomes and informed decision-making processes.

Acknowledgments

The authors would like to thank the hospitals and staff for their support in this project. This study was supported by the Minas Gerais State Agency for Research and Development (Fundação de Amparo à Pesquisa do Estado de Minas Gerais [FAPEMIG]; grants APQ-01154-21 and APQ-00262-22), National Institute of Science and Technology for Health Technology Assessment (Instituto de Avaliação de Tecnologias em Saúde [IATS])/National Council for Scientific and Technological Development

(Conselho Nacional de Desenvolvimento Científico e Tecnológico [CNPq]; grant465518/2014-1), and National Council for Scientific and Technological Development (grants 421773/2022-7, 403184/2021-5, and 401898/2022-9). This study was also partially financed with resources from the Center for Innovation and Artificial Intelligence for Health (CI-IA Saúde), the São Paulo State Research Support Foundation (FAPESP; process number 2020/09866-4), the Foundation of Minas Gerais Research Support (FAPEMIG; process number PPE-00030-21), and UNIMED Belo Horizonte. MSM was supported in part by CNPq (grant number 310561/2021-3). MAG was supported in part by CNPq (310538/2020-3). The funding bodies played no role in the design of the study; collection, analysis, or interpretation of the data; or writing the manuscript.

Authors' Contributions

BP, LCDdR, JMA, MSM, CMVdA, FCBL, MVRS-S, PDP, and MAG made substantial contributions to the conception or design of the work and drafted the work. All the authors made substantial contributions to the acquisition, analysis, and interpretation of data for the work; revised the manuscript critically for important intellectual content; and gave final approval of the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Modeling of the centroids as the arithmetic mean of the features in each outcome group. Co is the centroid of cluster O, XCO is the matrix of attributes including all patients in the outcome O, and |CO| is the number of patients in the outcome group O.
[PNG File , 55 KB - medinform_v12i1e54246_app1.png]

Multimedia Appendix 2
The “anchor_year_group” variable on the Medical Information Mart for Intensive Care, version IV (MIMIC-IV) data set. Within each “anchor_year_group,” the actual dates are masked, making it possible to have only a rough estimate of when the patient was at the hospital.
[PNG File , 81 KB - medinform_v12i1e54246_app2.png]

Multimedia Appendix 3
Lethality over time in the Medical Information Mart for Intensive Care, version IV data set.
[PNG File , 63 KB - medinform_v12i1e54246_app3.png]

Multimedia Appendix 4
Drift of the arithmetic mean of each outcome class over time, as measured by cosine distances between each class’s means when compared to the mean of the first “anchor_year_group” in the Medical Information Mart for Intensive Care, version IV data set.
[PNG File , 214 KB - medinform_v12i1e54246_app4.png]

Multimedia Appendix 5
Evaluation of the drivers of lethality data drift in the Medical Information Mart for Intensive Care, version IV data set.
[PNG File , 319 KB - medinform_v12i1e54246_app5.png]

Multimedia Appendix 6
Changes in co-occurrence for the “dysphagia following stroke” International Classification of Diseases in the Medical Information Mart for Intensive Care, version IV data set.
[PNG File , 242 KB - medinform_v12i1e54246_app6.png]

Multimedia Appendix 7
Validation of the data drift in cancer patients. On the left, we show the increase in the absolute number of cancer patients, while on the right, we show the overall lethality reduction for this disease group.
[PNG File , 65 KB - medinform_v12i1e54246_app7.png]

Multimedia Appendix 8
Cluster analysis of the Medical Information Mart for Intensive Care, version IV data set. (A) Top 5 highest-valued features per cluster. (B) Relative frequency of each cluster over time.
[PNG File , 138 KB - medinform_v12i1e54246_app8.png]

Multimedia Appendix 9
Drift of the arithmetic means of the dying patients versus the overall population over time, as measured by cosine distances between each class's means on each time chunk over time, in the Brazilian COVID-19 Registry data set.
[PNG File , 227 KB - medinform_v12i1e54246_app9.png]

Multimedia Appendix 10
Lethality over time in the Brazilian COVID-19 Registry data set.
[PNG File , 80 KB - medinform_v12i1e54246_app10.png]

Multimedia Appendix 11
Top 15 largest increases and decreases in similarity between the "death" tokens for 2021 and 2020 in the Brazilian COVID-19 Registry data set.
[PNG File , 373 KB - medinform_v12i1e54246_app11.png]

Multimedia Appendix 12
Cluster analysis of the Brazilian COVID-19 Registry data set. (A) Top 5 highest-valued features per cluster. (B) Relative frequency of each cluster over time.
[PNG File , 106 KB - medinform_v12i1e54246_app12.png]

References

1. Vayena E, Dzenowagis J, Brownstein JS, Sheikh A. Policy implications of big data in the health sector. Bull World Health Organ 2018 Jan 01;96(1):66-68 [FREE Full text] [doi: 10.2471/BLT.17.197426] [Medline: 29403102]
2. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. Eur J Public Health 2019 Oct 01;29(Supplement_3):23-27 [FREE Full text] [doi: 10.1093/eurpub/ckz168] [Medline: 31738444]
3. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood) 2014 Jul;33(7):1115-1122. [doi: 10.1377/hlthaff.2014.0147] [Medline: 25006136]
4. de Paiva BB, Delfino-Pereira P, Gomes VM, Souza-Silva MV, Valiense C, Marcolino MS, et al. Characterizing and understanding temporal effects in COVID-19 data. In: Proceedings of the 1st Workshop on Healthcare AI and COVID-19. 2022 Presented at: ICML 2022; July 22, 2022; Baltimore, MD URL: https://proceedings.mlr.press/v184/paiva22a/paiva22a.pdf
5. Moura EC, Cortez-Escalante J, Cavalcante FV, Barreto IC, Sanchez MN, Santos LM. Covid-19: temporal evolution and immunization in the three epidemiological waves, Brazil, 2020-2022. Rev Saude Publica 2022 Nov 18;56:105 [FREE Full text] [doi: 10.11606/s1518-8787.2022056004907] [Medline: 36515307]
6. Salles T, Rocha L, Pappa GL, Mourã F, Meira WJ, Gonçalves M. Temporally-aware algorithms for document classification. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010 Presented at: SIGIR '10; July 19-23, 2010; Geneva, Switzerland. [doi: 10.1145/1835449.1835502]
7. Rocha L, Mourão F, Pereira A, Gonçalves MA, Meira WJ. Exploiting temporal contexts in text classification. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008 Presented at: CIKM '08; October 26-30, 2008; Napa Valley, CA. [doi: 10.1145/1458082.1458117]
8. Lahza H, Lahza HF. A survey on detecting healthcare concept drift in AI/ML models from a finance perspective. Front Artif Intell 2022 Apr 17;5:955314 [FREE Full text] [doi: 10.3389/frai.2022.955314] [Medline: 37139355]
9. McLean C, Capurro D. Concept drift detection to assess the diffusion of process innovations in healthcare. AMIA Annu Symp Proc 2022;2022:746-755 [FREE Full text] [Medline: 37128394]
10. Sundquist M, Brudin L, Tejler G. Improved survival in metastatic breast cancer 1985-2016. Breast 2017 Feb;31:46-50. [doi: 10.1016/j.breast.2016.10.005] [Medline: 27810699]
11. Lima ES, Romero EC, Granato CF. Current polio status in the world. J Bras Patol Med Lab 2021;57:1-6. [doi: 10.5935/1676-2444.20210022]
12. Dabbagh A, Patel MK, Dumolard L, Gacic-Dobo M, Mulders MN, Okwo-Bele JM, et al. Progress toward regional measles elimination - worldwide, 2000-2016. MMWR Morb Mortal Wkly Rep 2017 Oct 27;66(42):1148-1153 [FREE Full text] [doi: 10.15585/mmwr.mm6642a6] [Medline: 29073125]
13. Graña C, Ghosn L, Evrenoglou T, Jarde A, Minozzi S, Bergman H, et al. Efficacy and safety of COVID-19 vaccines. Cochrane Database Syst Rev 2022 Dec 07;12(12):CD015477 [FREE Full text] [doi: 10.1002/14651858.CD015477] [Medline: 36473651]
14. Menéndez ML, Pardo JA, Pardo L, Pardo MC. The Jensen-Shannon divergence. J Frankl Inst 1997 Mar;334(2):307-318. [doi: 10.1016/S0016-0032(96)00063-4]

15. Menon AG, Gressel G. Concept drift detection in phishing using autoencoders. In: Proceedings of the Machine Learning and Metaheuristics Algorithms, and Applications. 2020 Presented at: SoMMA 2020; October 14-17, 2020; Chennai, India. [doi: 10.1007/978-981-16-0419-5_17]
16. Deng Z, Li C, Song R, Liu X, Qian R, Chen X. Centroid-guided domain incremental learning for EEG-based seizure prediction. IEEE Trans Instrum Meas 2024;73:1-13. [doi: 10.1109/TIM.2023.3334330]
17. Marcolino MS, Pires MC, Ramos LE, Silva RT, Oliveira LM, Carvalho RL, et al. ABC2-SPH risk score for in-hospital mortality in COVID-19 patients: development, external validation and comparison with other available scores. Int J Infect Dis 2021 Sep;110:281-308 [FREE Full text] [doi: 10.1016/j.ijid.2021.07.049] [Medline: 34311100]
18. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV. PhysioNet. 2021 Mar 16. URL: https://physionet.org/content/mimiciv/1.0/ [accessed 2023-10-18]
19. Wang H, Fan W, Yu PS, Han J. Mining concept-drifting data streams using ensemble classifiers. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003 Presented at: KDD '03; August 24-27, 2003; Washington, DC. [doi: 10.1145/956750.956778]
20. Parmar H, Nutter B, Mitra SD, Long LR, Antani SK. Automated signal drift and global fluctuation removal from 4D fMRI data based on principal component analysis as a major preprocessing step for fMRI data analysis. In: Proceedings of the Biomedical Applications in Molecular, Structural, and Functional Imaging. 2019 Presented at: SPIE Medical Imaging 2019; February 16-21, 2019; San Diego, CA. [doi: 10.1117/12.2512968]
21. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit 2012 Jan;45(1):521-530. [doi: 10.1016/j.patcog.2011.06.019]
22. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Berlin, Germany: Springer; 2009.
23. Myers L, Sirois MJ. Spearman correlation coefficients, differences between. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B, editors. Encyclopedia of Statistical Sciences. Hoboken, NJ: John Wiley & Sons; 2004.
24. Kazemitabar SJ, Amini AA, Bloniarz A, Talwalkar A. Variable importance using decision trees. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA.
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA.
26. Mnih A, Kavukcuoglu K. Learning word embeddings efficiently with noise-contrastive estimation. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, NV.
27. Kullback–Leibler divergence. Wikipedia. URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence [accessed 2023-10-18]
28. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc 1972;28(1):11-21. [doi: 10.1108/eb026526]
29. Hamad D, Biela P. Introduction to spectral clustering. In: Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications. 2008 Presented at: ICTTA 2008; April 7-11, 2008; Damascus, Syria. [doi: 10.1109/ictta.2008.4529994]
30. Grover A, Leskovec J. node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939754]
31. Salles T, Rocha L, Gonçalves MA, Almeida JM, Mourão F, Meira WJ, et al. A quantitative analysis of the temporal effects on automatic text classification. J Assoc Inf Sci Technol 2015 Aug 07;67(7):1639-1667. [doi: 10.1002/asi.23452]
32. Mouro F, Rocha L, Arajo R, Couto T, Gonçalves M, Meira WJ. Understanding temporal aspects in document classification. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008 Presented at: WSDM '08; February 11-12, 2008; Palo Alto, CA. [doi: 10.1145/1341531.1341554]
33. Rotalinti Y, Tucker A, Lonergan M, Myles P, Branson R. Detecting drift in healthcare AI models based on data availability. In: Proceedings of the International Workshops on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. 2022 Presented at: ECML PKDD 2022; September 19-23, 2022; Grenoble, France. [doi: 10.1007/978-3-031-23633-4_17]
34. Nirmala CR, Aljohani M, Sreenivasa BR, M S AR. A novel technique for detecting sudden concept drift in healthcare data using multi-linear artificial intelligence techniques. Front Artif Intell 2022 Aug 31;5:950659 [FREE Full text] [doi: 10.3389/frai.2022.950659] [Medline: 36117781]
35. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv 2014 Mar 01;46(4):1-37. [doi: 10.1145/2523813]
36. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. IEEE Trans Knowl Data Eng 2019 Dec 01;31(12):2346-2363. [doi: 10.1109/tkde.2018.2876857]

37. Haque A, Chandra S, Khan L, Hamlen K, Aggarwal C. Efficient multistream classification using direct density ratio estimation. In: Proceedings of the IEEE 33rd International Conference on Data Engineering. 2017 Presented at: ICDE 2017; April 19-22, 2017; San Diego, CA. [doi: 10.1109/icde.2017.63]
38. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. Br J Radiol 2023 Oct;96(1150):20220878. [doi: 10.1259/bjr.20220878] [Medline: 36971405]
39. Tiwari S, Agarwal S. Data stream management for CPS-based healthcare: a contemporary review. IETE Tech Rev 2021 Jul 20;39(5):987-1010. [doi: 10.1080/02564602.2021.1950578]
40. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res 2015 Jan;43(Database issue):D204-D212. [doi: 10.1093/nar/gku989] [Medline: 25348405]
41. Jung C, Excoffier JB, Raphaël-Rousseau M, Salaün-Penquer N, Ortala M, Chouaid C. Evolution of hospitalized patient characteristics through the first three COVID-19 waves in Paris area using machine learning analysis. PLoS One 2022 Feb 22;17(2):e0263266 [FREE Full text] [doi: 10.1371/journal.pone.0263266] [Medline: 35192649]
42. Jassat W, Mudara C, Ozougwu L, Tempia S, Blumberg L, Davies MA, et al. Difference in mortality among individuals admitted to hospital with COVID-19 during the first and second waves in South Africa: a cohort study. Lancet Glob Health 2021 Sep;9(9):e1216-e1225 [FREE Full text] [doi: 10.1016/S2214-109X(21)00289-8] [Medline: 34252381]
43. Klinkenberg R. Learning drifting concepts: example selection vs. example weighting. Intell Data Anal 2004 Aug 13;8(3):281-300. [doi: 10.3233/IDA-2004-8305]
44. Kolter JZ, Maloof MA. Dynamic weighted majority: an ensemble method for drifting concepts. J Mach Learn Res 2007;8(91):2755-2790.
45. Salles T, Rocha L, Mourão F, Gonçalves M, Viegas F, Meira WJ. A two-stage machine learning approach for temporally-robust text classification. Inf Syst 2017 Sep;69:40-58. [doi: 10.1016/j.is.2017.04.004]
46. Zliobaite I, Bifet A, Pfahringer B, Holmes G. Active learning with drifting streaming data. IEEE Trans Neural Netw Learning Syst 2014 Jan;25(1):27-39. [doi: 10.1109/tnnls.2012.2236570]
47. Hypertension. World Health Organization. 2023 Mar 16. URL: https://www.who.int/news-room/fact-sheets/detail/hypertension [accessed 2023-06-11]
48. About diabetes. American Diabetes Association. URL: https://diabetes.org/about-diabetes [accessed 2023-06-11]
49. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2020 Feb;28(2):165-173 [FREE Full text] [doi: 10.1038/s41431-019-0508-0] [Medline: 31527858]
50. Gao J, Ding B, Fan W, Han J, Yu PS. Classifying data streams with skewed class distributions and concept drifts. IEEE Internet Comput 2008;12(6):37-49. [doi: 10.1109/mic.2008.119]
51. Ditzler G, Polikar R. Incremental learning of concept drift from streaming imbalanced data. IEEE Trans Knowl Data Eng 2013 Oct;25(10):2283-2301. [doi: 10.1109/TKDE.2012.136]
52. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]

Abbreviations

DIS: detection, initial characterization, and semantic characterization
FiO2: fraction of inspired oxygen
ICD: International Classification of Diseases
ICU: intensive care unit
KL: Kullback-Leibler
MIMIC-IV: Medical Information Mart for Intensive Care, version IV
ML: machine learning
NLP: natural language processing
PCA: principal component analysis
TF-IDF: term frequency–inverse document frequency

Edited by C Lovis; submitted 02.11.23; peer-reviewed by C Wu, L Guo; comments to author 10.02.24; revised version received 30.05.24; accepted 07.07.24; published 28.10.24.

Please cite as:
Paiva B, Gonçalves MA, da Rocha LCD, Marcolino MS, Lana FCB, Souza-Silva MVR, Almeida JM, Pereira PD, de Andrade CMV, Gomes AGDR, Ferreira MAP, Bartolazzi F, Sacioto MF, Boscato AP, Guimarães-Júnior MH, dos Reis PP, Costa FR, Jorge ADO, Coelho LR, Carneiro M, Sales TLS, Araújo SF, Silveira DV, Ruschel KB, Santos FCV, Cenci EPDA, Menezes LSM, Anschau F, Bicalho MAC, Manenti ERF, Finger RG, Ponce D, de Aguiar FC, Marques LM, de Castro LC, Vietta GG, Godoy MFD, Vilaça MDN, Morais VC
A New Natural Language Processing–Inspired Methodology (Detection, Initial Characterization, and Semantic Characterization) to Investigate Temporal Shifts (Drifts) in Health Care Data: Quantitative Study
JMIR Med Inform 2024;12:e54246
URL: https://medinform.jmir.org/2024/1/e54246
doi:10.2196/54246
PMID:

Original Paper

Data-Driven Identification of Factors That Influence the Quality of Adverse Event Reports: 15-Year Interpretable Machine Learning and Time-Series Analyses of VigiBase and QUEST

Sim Mei Choo[1,2], MSc; Daniele Sartori[3], MSc; Sing Chet Lee[1], MSc; Hsuan-Chia Yang[2,4,5,6*], PhD; Shabbir Syed-Abdul[2,4,7*], MD, PhD

[1]Centre of Compliance & Quality Control, National Pharmaceutical Regulatory Agency, Petaling Jaya, Malaysia

[2]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan

[3]Uppsala Monitoring Centre, Uppsala, Sweden

[4]International Center for Health Information Technology, Taipei Medical University, Taipei, Taiwan

[5]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan

[6]Research Center of Big Data and Meta-Analysis, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

[7]School of Gerontology and Long-Term Care, Taipei Medical University, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Shabbir Syed-Abdul, MD, PhD
Graduate Institute of Biomedical Informatics
Taipei Medical University
301 Yuantong Rd
Taipei, 235
Taiwan
Phone: 886 66202589 ext 10930
Email: drshabbir@tmu.edu.tw

Abstract

Background: The completeness of adverse event (AE) reports, crucial for assessing putative causal relationships, is measured using the vigiGrade completeness score in VigiBase, the World Health Organization global database of reported potential AEs. Malaysian reports have surpassed the global average score (approximately 0.44), achieving a 5-year average of 0.79 (SD 0.23) as of 2019 and approaching the benchmark for well-documented reports (0.80). However, the contributing factors to this relatively high report completeness score remain unexplored.

Objective: This study aims to explore the main drivers influencing the completeness of Malaysian AE reports in VigiBase over a 15-year period using vigiGrade. A secondary objective was to understand the strategic measures taken by the Malaysian authorities leading to enhanced report completeness across different time frames.

Methods: We analyzed 132,738 Malaysian reports (2005-2019) recorded in VigiBase up to February 2021 split into historical International Drug Information System (INTDIS; n=63,943, 48.17% in 2005-2016) and newer E2B (n=68,795, 51.83% in 2015-2019) format subsets. For machine learning analyses, we performed a 2-stage feature selection followed by a random forest classifier to identify the top features predicting well-documented reports. We subsequently applied tree Shapley additive explanations to examine the magnitude, prevalence, and direction of feature effects. In addition, we conducted time-series analyses to evaluate chronological trends and potential influences of key interventions on reporting quality.

Results: Among the analyzed reports, 42.84% (56,877/132,738) were well documented, with an increase of 65.37% (53,929/82,497) since 2015. Over two-thirds (46,186/68,795, 67.14%) of the Malaysian E2B reports were well documented compared to INTDIS reports at 16.72% (10,691/63,943). For INTDIS reports, higher pharmacovigilance center staffing was the primary feature positively associated with being well documented. In recent E2B reports, the top positive features included reaction abated upon drug dechallenge, reaction onset or drug use duration of <1 week, dosing interval of <1 day, reports from public specialist hospitals, reports by pharmacists, and reaction duration between 1 and 6 days. In contrast, reports from product registration holders and other health care professionals and reactions involving product substitution issues negatively affected the quality of E2B reports. Multifaceted strategies and interventions comprising policy changes, continuity of education, and

human resource development laid the groundwork for AE reporting in Malaysia, whereas advancements in technological infrastructure, pharmacovigilance databases, and reporting tools concurred with increases in both the quantity and quality of AE reports.

Conclusions: Through interpretable machine learning and time-series analyses, this study identified key features that positively or negatively influence the completeness of Malaysian AE reports and unveiled how Malaysia has developed its pharmacovigilance capacity via multifaceted strategies and interventions. These findings will guide future work in enhancing pharmacovigilance and public health.

(JMIR Med Inform 2024;12:e49643) doi:10.2196/49643

KEYWORDS

pharmacovigilance; medication safety; big data analysis; feature selection; interpretable machine learning

Introduction

Background

Pharmacovigilance (PV) is the science and activities related to the detection, assessment, understanding, and prevention of adverse effects or any other possible drug-related problems [1]. Individual case safety reports (ICSRs) of suspected adverse drug reactions and adverse events following immunization (hereafter collectively referred to as adverse events [AEs]) collected in spontaneous reporting systems (SRSs) remain the cornerstone of postmarketing drug safety surveillance [2,3] (see Multimedia Appendix 1 for a list of definitions [4-10]).

Over 170 participating countries in the World Health Organization (WHO) Programme for International Drug Monitoring (PIDM) share reports of suspected AEs and collaborate worldwide in monitoring and identifying signals of AEs [4]. The WHO PIDM signal detection process is anchored on data recorded in the WHO global ICSR database, VigiBase, developed and maintained by the WHO Collaborating Centre for International Drug Monitoring, Uppsala Monitoring Centre (UMC), Sweden. Common technical specifications for report transmission and standard terminologies for drugs and reactions have evolved over the years to facilitate global information sharing and efficient analysis [4,5]. Currently, VigiBase accepts 3 standard formats: the original International Drug Information System (INTDIS) and 2 revisions of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) Guidelines for the Electronic Transmission of ICSRs, namely E2B(R2), and the latest E2B(R3), with all data being transformed to a format most closely resembling E2B(R2) in VigiBase [4].

The participating members are characterized by diverse contexts—sociocultural, political, and clinical—that affect the measures in which reports are collected and processed, as well as their quality [4]. Ideally, a robust PV system should consider all data quality parameters, including accuracy, completeness, conformity, consistency, currency, duplication, integrity, precision, relevance, and understandability. Among all these parameters, problems associated with completeness (ie, missing data) have long been regarded as critical factors hampering the usefulness of existing reports [5,6]. Influxes of poorly documented reports could increase operational burdens, upsurge a system's resources, and even mask or delay the detection of drug safety signals [11].

While only 4 elements (identifiable patient, identifiable reporter, medicinal product, and AE) are required for a valid report, they are often insufficient for productive analyses of potential causal relationships between medicinal products and AEs [6]. In 2014, the UMC developed the vigiGrade completeness score, an automated multidimensional tool that measures the amount of clinically relevant information in reports essential for causality assessment, replacing the 4-grade WHO documentation grading scheme since the 1990s [6,12]. The vigiGrade score quantifies report completeness based on a selection of ICH-E2B fields: time to onset, indication, event outcome, patient age and sex, dose information, country of origin, reporter, type of report, and free-text fields. The vigiGrade score can be used to pinpoint trends in report quality over time and reflect systematic data quality issues in collections of reports from member countries. For instance, vigiGrade uncovered miscoded age units in US reports and missing AE outcomes in Italian reports [6]. The score may also guide reviewers in judging whether the information in a report suffices for a problem to be investigated [4]. Notably, vigiGrade has proven to be an indicator of a true signal and is part of the data-driven predictive model used by the UMC, vigiRank, for signal detection [13].

The PV System in Malaysia

PV activities in Malaysia began in the 1980s with the establishment of the Malaysian Adverse Drug Reactions Advisory Committee (MADRAC) under the Drug Control Authority (DCA) [7]. The Malaysian national PV center is based within the National Pharmaceutical Regulatory Agency (NPRA) under the Pharmaceutical Services Programme of the Ministry of Health (MOH). Malaysia became a member of the WHO PIDM in 1990 and is regarded as an established PV center, receiving >30,000 reports annually, which is well above the WHO criteria of 200 reports per million inhabitants per year since 2009. Every AE report recorded in the national PV database (QUEST; see Multimedia Appendix 1 [7] for a detailed description) is carefully processed and assessed by trained pharmacists at the national center and subsequently reviewed by the MADRAC before submission to the UMC for inclusion in VigiBase (Figure S1 in Multimedia Appendix 2).

Problem Statement and Research Benefits

Previous studies have not evaluated the quality of Malaysian AE reports, and little is known about the underlying factors affecting their vigiGrade completeness scores. However, identifying and validating factors associated with report quality

was made difficult by the large number and variety of potentially correlated characteristics of a spontaneous report—at the reaction, drug, patient, reporter, sender, or regulator level (see the literature review on AE report quality in Multimedia Appendix 3 [6,14-38]). As of 2019, Malaysian reports demonstrated a 5-year average completeness score of 0.79, surpassing the global average of approximately 0.44 in VigiBase and approaching the benchmark for well-documented reports (0.80). Therefore, this study primarily aimed to use a hypothesis-free, data-driven approach to explore the main drivers influencing the completeness of Malaysian reports in VigiBase over a 15-year period using vigiGrade. A secondary objective was to understand the strategic measures taken by the Malaysian authorities that preceded the relatively high completeness score across different time frames. A better understanding of the drivers of AE report completeness may be helpful for the NPRA and regulators worldwide.

Methods

Data-Driven Framework for Identifying Factors Associated With AE Report Quality

Overview

Our study used big data analysis approaches incorporating machine learning (ML) methods, which are becoming increasingly prevalent in clinical and epidemiological research [8,39]. These approaches aimed to overcome limitations inherent in traditional approaches in handling complex interactions among variables (eg, multicollinearity and nonlinearity). Importantly, ML methods focus on identifying patterns and associations within complex data rather than on establishing causal inference [39]. For clarity, we have outlined the similarities in concepts and nomenclatures between ML and traditional medical statistics in Multimedia Appendix 1.

Hybrid Feature Selection

Our study leveraged ML methods for feature selection, mitigating human bias in analyzing extensive report characteristics, which might be overlooked by traditional hypothesis-driven approaches that are prone to high selection biases [9]. We combined statistical filtering and ML algorithms to preselect features, reducing overfitting risks (often arising from redundancy and multicollinearity) and computational costs [40,41]. Notably, Stevens et al [9] used random forest (RF)–based feature selection to identify potential risk factors associated with cardiovascular diseases. In almost all domains, incorporating domain expertise remains vital for developing meaningful and effective models [8]. Multimedia Appendix 4 [8,9,39-48] provides detailed explanations.

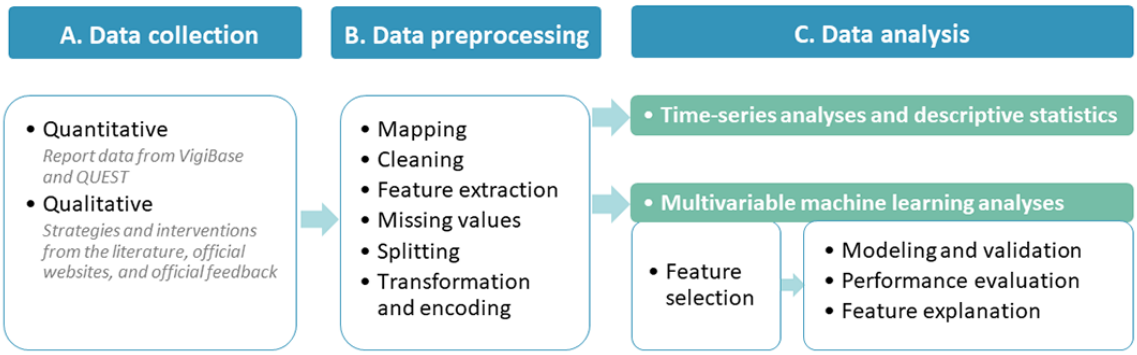
Interpretable ML

Post hoc explanation methods such as Shapley additive explanations (SHAP) have been increasingly used to provide interpretability for complex black-box models such as RF [49]. Van den Bosch et al [50] used regression coefficients and SHAP value analyses to identify risk factors associated with 30-day mortality among patients undergoing colorectal cancer surgery. Gong et al [51] also developed an ML framework for acute kidney injury prediction and interpretation using SHAP values to assess feature contributions and identify specific patient impacts. Multimedia Appendix 5 [49-57] provides detailed explanations.

Study Design

This observational study used interpretable ML and descriptive time-series analyses of PV database data. Fundamentally, it was an exploratory data analysis assessing a large number of report characteristics without prespecified hypotheses [42] aiming to identify factors influencing report quality. The main steps of our methodological workflow are illustrated in Figure 1.

Figure 1. Overview of study workflow.



Data Collection

AE report data were obtained from VigiBase in CSV format. Supplementary information, such as means of reporting, sender type, and sender region, was retrieved from the NPRA QUEST3+ (latest version) database in CSV format. For the secondary objective of understanding the key drivers of AE report completeness, strategies and interventions implemented in Malaysia were collected from the literature, official websites, and feedback from the NPRA. We included all reports recorded in VigiBase as of February 2021; reported in Malaysia; and received by the NPRA from January 1, 2005, to December 31, 2019. This 15-year range was chosen for its relevance to current PV needs and to enable the timely identification of necessary improvements. We excluded reports that (1) were suspected duplicates identified by the UMC's vigiMatch [4] (see Multimedia Appendix 1 for operational definitions), (2) were not sourced from Malaysia, (3) had null average completeness scores, and (4) lacked drugs marked as suspected or interacting.

Study Variables

Dependent Variables or Outcomes

The vigiGrade completeness score (*C*; ranges from 0.07 to 1) was classified as well documented ($C>0.8$) or not well documented ($C\leq0.8$; see Multimedia Appendix 1 for operational definitions).

Independent Variables or Explanatory Features

The variables related to administrative, sender, reporter, patient, drug, and reaction characteristics are presented in Table S1 in Multimedia Appendix 2.

Data Preprocessing

Data Mapping

Supplementary data from QUEST3+ were mapped to the primary data set from VigiBase using the primary identifier. Given the distinct differences in reporting elements of INTDIS and E2B formats (input values vary in certain data fields), we divided the data set for separate analysis.

Data Cleaning and Feature Extraction

We cleaned and engineered the features from the available data based on the literature, domain knowledge, and previous experience. Information about the WHO Anatomical Therapeutic Chemical (ATC) classification system codes was provided by the UMC in the data set. If an active ingredient was linked to more than one code, an ATC level-2 code was manually assigned based on indication, route of administration, dosage, product information, and clinical narratives. Reporting qualifications in INTDIS format were harmonized with the E2B format with reference to supplementary data from the NPRA. We calculated the number of suspected or interacting drugs, concomitant drugs, and reactions for each report. We also included the annual staffing level of the national PV center in Malaysia and the means of reporting (based on report identifier).

Missing Values

Continuous variables consisting of null values were converted to categorical variables based on data distribution and domain knowledge. Missing values for categorical variables were grouped as a *null* category.

Data Splitting

To ensure a consistent distribution of target classes, we applied stratified random sampling. We allocated 90% of the data for training, which underwent 10-fold cross-validation, and reserved 10% for testing to gauge model performance on unseen samples. Our approach prioritized the extraction of insights from the current data set rather than overgeneralizations on future data.

Transformation and Encoding

To overcome data complexity and maximize interpretability, data at the drug event level were transformed to the case (report) level. Observations related to concomitant drugs were excluded as the vigiGrade scoring method is restricted to drugs listed as suspected or interacting [6]. We took the average value of a case for continuous variables whereby, for categorical variables, we examined the presence (or absence) of a particular drug- or event-related characteristic. Continuous variables were standardized. Binary categorical variables such as patient sex were integer encoded. One-hot encoding was performed on the remaining categorical variables, including ordinal variables and categories labelled as *null* or *unknown*. In the following sections, we distinguish variables from features, where the latter correspond to the processed variables in a binary fashion for the ML model input [43,56].

Multivariable ML Analysis

Feature Selection

We performed hybrid feature selection to eliminate redundant or less informative features before data mining using the ML algorithm. To avoid data leakage and the corresponding model overfitting, we conducted a 2-stage feature selection solely based on training data [8,44]. We first applied the univariable filter method to independently assess and preselect the features and subsequently selected the top-ranked features using RF-based recursive feature elimination coupled with multicollinearity assessment. The detailed processes are provided in Multimedia Appendix 4.

Modeling and Validation

We applied a supervised ML method to identify key features relevant to the reports classified as well documented. Specifically, the RF classifier was selected for its robustness to nonparametric distributions, nonlinearity, and outliers [58] and its out-of-the-box performance. Its built-in feature importance metrics allowed us to assess the relative attribution of a feature to the classification task. The more a feature is used to make key decisions with the forest of decision trees, the higher its relative importance. To mitigate class imbalance in the INTDIS data set, we used RandomUnderSampler with a 0.25 ratio that achieved optimal balanced performance of prediction and recall. We chose undersampling over synthetic sampling methods to preserve the real-world data characteristics. For the imbalanced INTDIS data set, we adjusted the class_weight parameter in the RF classifier to *balanced*. We evaluated the RF classification models using 10-fold cross-validation.

Performance Evaluations

Classification performance was measured using the area under the receiver operating characteristic curve, accuracy, recall (sensitivity), and precision (positive predictive values). For the imbalanced INTDIS data set, F_1-scores (harmonic average of precision and recall) were reported.

Feature Explanations

To mitigate the issue of black-box predictions**,** we used TreeExplainer [52] to generate SHAP summary plots that succinctly display the magnitude, prevalence, and direction of a feature's effect by measuring each feature's attributions to the classification. In SHAP, the feature effect is a measure of how much the value of a specific feature influences the prediction made by the model.

Software and Packages

All ML analyses were developed in Jupyter Notebook (Project Jupyter) using Python (version 3.7.9; Python Software

Foundation). Statistical tests were performed using *pandas* (version 1.2.4) and *statsmodels* (version 0.12.2) [59]. ML analysis was completed using the *scikit-learn* package (version 0.24.2) [60]. SHAP values were calculated using TreeExplainer [52].

Time-Series and Descriptive Statistical Analysis

We used time-series analysis and descriptive statistics to evaluate the trends in report quality and the characteristics associated with well-documented reports over different time frames. One-way ANOVA or 2-tailed Student *t* tests were conducted on continuous variables, whereas the chi-square or Fisher exact test was used to compare categorical variables, as appropriate. A *P* value of <.05 was considered statistically significant. All analyses were conducted using the SAS software (version 9.4; SAS Institute).

Ethical Considerations

This study was registered with the Malaysian National Medical Research Register (NMRR-20-983-53984 [Investigator Initiated Research]) and received ethics approval from the Medical Review and Ethics Committee, MOH, Malaysia (reference: KKM/NIHSEC/P20–1144(4)).

Results

Overview

We analyzed the completeness of Malaysian AE reports in VigiBase received by the NPRA over 15 years. A total of 132,738 reports were included in the analysis following the predefined inclusion and exclusion criteria (Figure S2 in Multimedia Appendix 2). Table S1 in Multimedia Appendix 2 summarizes the characteristics of the INTDIS and E2B reports included in this study concerning administration, reporter, patient, drug, and reaction by status of being well documented. Among the included reports, 48.17% (63,943/132,738) were in the INTDIS format, and 51.83% (68,795/132,738) were in the E2B format. Over two-thirds (46,186/68,795, 67.14%) of E2B reports were well documented compared to 16.72% (10,691/63,943) of INTDIS reports.

Multivariable ML Analysis

Selected Features

For the INTDIS subsets, 90 features were preselected using univariate filter methods and further narrowed down to 33 features following RF-based recursive feature elimination ranking and multicollinearity assessment. For the E2B subsets, 90 features were preselected and subsequently reduced to 40.

Classification Performance

The performance of the RF models in classifying reports as well or not well documented is presented in Table 1.

Table 1. Classification performance of random forest model for the training (10-fold cross-validation) and test set.

	Recall (%)	Precision (%)	Accuracy (%)	AUROC[a] (%)	F_1-score (%)
INTDIS[b]					
Training, mean (SD)	99.6 (0.03)	95.4 (0.13)	99.0 (0.03)	99.8 (0.01)	97.4 (0.06)
Validation, mean (SD)	73.7 (1.90)	77.0 (1.02)	90.3 (0.39)	95.0 (0.33)	75.3 (1.16)
Test	74.9	74.3	91.5	95.1	74.6
E2B					
Training, mean (SD)	99.7 (0.02)	99.7 (0.02)	99.6 (0.01)	99.9 (0.001)	99.7 (0.01)
Validation, mean (SD)	96.9 (0.27)	91.6 (0.34)	92.0 (0.24)	94.9 (0.29)	94.2 (0.20)
Test	96.9	90.9	91.4	95.1	93.8

[a]AUROC: area under the receiver operating characteristic curve.

[b]INTDIS: International Drug Information System.

Top Factors Predicting Status of Malaysian Reports Being Well Documented

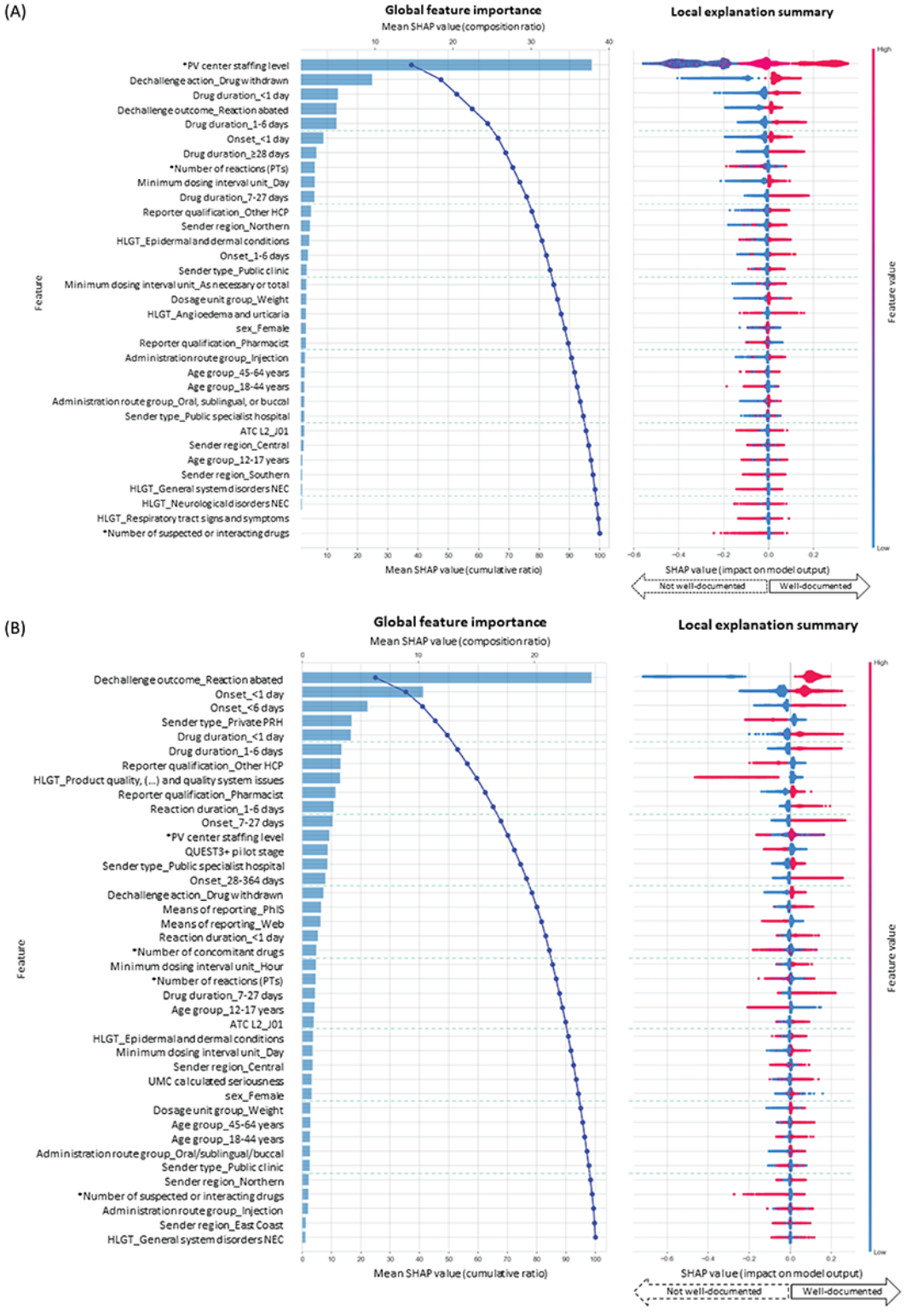
RF ML Model

Figure S3 in Multimedia Appendix 2 reveals the top-ranked features that contributed to the status of INTDIS and E2B reports being well documented derived from the RF ML model's built-in feature importance metrics. However, the directions of their contribution were not known due to the black-box nature of the RF model. For INTDIS reports received between 2005 and 2016, PV center staffing was identified as the most important factor in predicting their status of being well documented. Other important factors included suspected drug withdrawal, the number of reactions reported, reaction abated upon drug dechallenge, and patient sex. Reaction abated upon drug dechallenge, on the other hand, appeared to be the most important factor predicting whether an E2B report received between 2015 and 2019 was well documented. Reports from other health care professionals (HCPs), reactions occurring <1 day, reports submitted by product registration holders (PRHs), and the number of concomitant drugs were also among the top 5 important factors.

SHAP Interpretation Method

The SHAP post hoc interpretations provided us with an understanding of the magnitude, prevalence, and direction of feature effects. Figure 2 depicts rich summaries of individual attributions for all features, allowing us to discover key factors that influence well-documented reports. Features with higher global importance have a greater influence on the model's predictions. A feature with predominantly red dots to the right (eg, reaction abated upon drug dechallenge in Figure 2B) implies a positive contribution to well-documented reports, whereas a negative contribution is indicated if the direction is to the left (eg, reports submitted by PRHs in Figure 2B). Features with lower global importance but a long tail stretching in one direction indicate a rare but high-magnitude effect [52,56]. As mean SHAP values are calculated across all cases, a feature with a lower impact but higher prevalence may have a higher SHAP value [50]. For the least globally important features, we observed that their feature effects were not constant across cases, with blue and red dots dispersed in both directions. This variation may arise from interactions with other features that modulate their importance in different cases [52].

Figure 2. (A) Top 33 features for the International Drug Information System (INTDIS) subset from 2005 to 2016; (B) top 40 features for the E2B subset during the years 2015 to 2019. The Shapley additive explanation (SHAP) bar plot illustrates global feature importances based on mean absolute SHAP values, highlighting the impact of each feature on the model's predictions. Higher values represent greater influence. The waterfall plot indicates the cumulative contribution of features to the model. The SHAP summary plot of local explanations displays each observation as a dot, with its position on the x-axis (SHAP value) indicating the impact of a feature on the model's classification for that observation. Continuous features (marked with an asterisk) range from low (blue) to high (red) values, whereas categorical features of a binary nature are either absent (blue) or present (red). The distribution of dots indicates the magnitude and prevalence of a feature effect. Features are ordered by global importance. ATC: Anatomical Therapeutic Chemical; HCP: health care professional; HLGT: Medical Dictionary for Regulatory Activities High-Level Group Term; NEC: not elsewhere classified; PhIS: pharmacy hospital information system; PRH: product registration holder; PT: Medical Dictionary for Regulatory Activities Preferred Term; PV: pharmacovigilance; UMC: Uppsala Monitoring Centre; UMC calculated seriousness: serious cases classified automatically by a UMC-developed algorithm.



Regarding INTDIS reports, in earlier years, PV center staffing was the primary factor driving the Malaysian rate of reports being well documented, with this factor alone accounting for >35% of the model's explainability. The next most important factor favoring well-documented reports was drug withdrawal, followed by a duration of drug use of <1 week, reaction abated upon drug dechallenge, and reaction occurring <1 day. In contrast, an increased number of reported reactions and reports from pharmacists predicted not well-documented INTDIS reports.

In more recent years, the most important factor favoring well-documented E2B reports from Malaysia was reaction abated upon drug dechallenge, which alone was responsible for >25% of the model's explainability. Among the top 25 features, which provided 90% of the model's interpretation on classifying status of being well documented, 6 (24%) were found to be negatively associated with well-documented reports: reports submitted by PRHs; reports made by other HCPs; reactions under the Medical Dictionary for Regulatory Activities (MedDRA) High-Level Group Terms (HLGTs) *product quality, supply, distribution, manufacturing,* and *quality system issues*; reports received during the QUEST3+ pilot stage, reports received via web reporting, and adolescent patients (aged 12-17 years). E2B reports that involved reactions with a shorter time to onset and duration of drug use were more likely to be well documented. Other identified key drivers of Malaysian well-documented E2B reports were reports made by pharmacists, reports submitted from public specialist hospitals, pharmacy hospital information system (PhIS)–integrated reporting, and the involvement of systemic antimicrobials (ATC code J01).

Time-Series and Descriptive Statistical Analysis

Trend Analysis of Malaysian AE Report Quality (2005-2019)

Figure 3 depicts the time trends in AE reporting in Malaysia, illustrating how both the quantity and completeness scores of AE reports received by the NPRA grew over a 15-year period from 2005 to 2019. In Tables 2 and 3, we summarize the trends divided into 5-year subperiods, further stratified by sender type and reporter qualification. Of the total 132,738 reports received, 56,877 (42.84%) were well documented. Before 2014, the average completeness score consistently fell short of 0.5 but was slightly above the global average of 0.44. The volume of reports surged by 121% from 2013 to 2014, whereas the proportion of well-documented reports rose from practically 0% to 18.93% (2843/15,013). Since 2015, more than half (53,929/82,497, 65.37%) of the Malaysian reports were well documented, averaging 0.79 (SD 0.23) over the last 5 years, with a new high of 0.82 in 2019.

Figure 3. Distribution of average completeness scores and counts of Malaysian reports by status of being well documented in VigiBase over the study period.

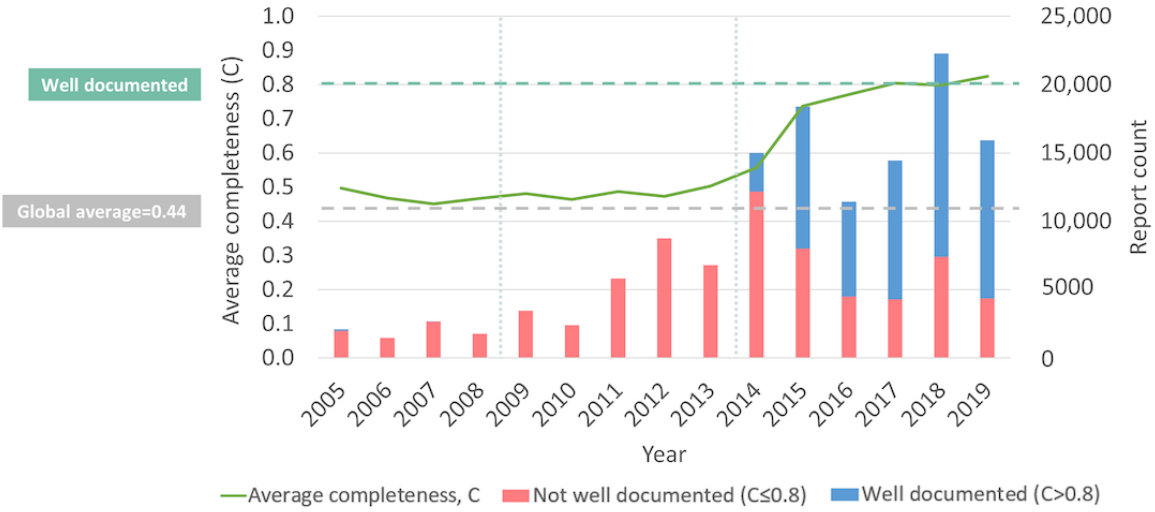


Table 2. The 5-year stratified summary statistics of overall Malaysian report quality.

	Total	2005-2009	2010-2014	2015-2019	*P* value[a]
Overall reports, n (%)	132,738 (100)	11,458 (8.6)	38,783 (29.2)	82,497 (62.2)	N/A[b]
Completeness, mean (SD)	0.68 (0.25)	0.47 (0.15)	0.51 (0.18)	0.79 (0.23)	<.001
Well-documented reports (C^c>0.8), n (%)	56,877 (42.8)	105 (0.9)	2843 (7.3)	53,929 (65.4)	<.001

[a]*P* value based on ANOVA.

[b]N/A: not applicable.

[c]vigiGrade completeness score.

Table 3. The 5-year stratified summary statistics of well-documented Malaysian reports (N=56,877).

	Total, n (%)	2005-2009 (n=105), n (%)	2010-2014 (n=2843), n (%)	2015-2019 (n=53,929), n (%)	*P* value[a]
Sender type					<.001
Public specialist hospital	31,503 (55.4)	72 (68.6)	1542 (54.2)	29,889 (55.4)	
Public nonspecialist hospital	4954 (8.7)	3 (2.9)	217 (7.6)	4734 (8.8)	
Public clinic	15,609 (27.4)	4 (3.8)	866 (30.5)	14,739 (27.3)	
Other public services	3 (0)	0 (0)	1 (0)	2 (0)	
University hospital	496 (0.9)	9 (8.6)	25 (0.9)	462 (0.9)	
Private PRH[b]	939 (1.7)	11 (10.5)	58 (2)	870 (1.6)	
Private hospital or clinic	2114 (3.7)	6 (5.7)	106 (3.7)	2002 (3.7)	
Private community pharmacy	45 (0.1)	0 (0)	0 (0)	45 (0.1)	
Consumer	30 (0.1)	0 (0)	0 (0)	30 (0.1)	
Unknown	1184 (2.1)	0 (0)	28 (1)	1156 (2.1)	
Reporter qualification					<.001
Physician	11,446 (20.1)	53 (50.5)	488 (17.2)	10,905 (20.2)	
Pharmacist	40,295 (70.8)	16 (15.2)	1850 (65.1)	38,429 (71.3)	
Other HCP[c]	4084 (7.2)	36 (34.3)	500 (17.6)	3548 (6.6)	
Consumer	78 (0.1)	0 (0)	0 (0)	78 (0.1)	
Unknown	974 (1.7)	0 (0)	5 (0.2)	969 (1.8)	

[a]*P* value based on the Fisher exact test.

[b]PRH: product registration holder.

[c]HCP: health care professional.

Over the 15 years, most well-documented reports in Malaysia came from public health facilities, with public specialist hospitals contributing more than half (31,503/56,877, 55.38%). Public clinics emerged as key contributors in later stages, with well-documented reports increasing considerably from 3.8% (4/105) in the period from 2005 to 2009 to 30.46% (866/2843) in the following 5 years. Compared to public services, the private sector consistently demonstrated a marginal contribution to quality AE reporting in Malaysia. In the earlier years, physicians contributed approximately half (53/105, 50.5%) of the well-documented reports. In the subsequent periods, reports from pharmacists showed a rise in quantity and average completeness, yielding the highest overall rate of being well documented (1850/2843, 65.07% to 38,429/53,929, 71.26%) among all reporter types from 2010 to 2019.

Key Strategies and Interventions Implemented in Malaysia (2005-2019)

In Malaysia, various strategies and interventions were implemented over the 15 years with the intent of improving AE reporting, as summarized by 5-year period in Figure 4 [7,61-65]. While the impacts of most interventions are usually multifaceted at the national level and challenging to measure with limited quantitative information, there is a particular interest in understanding the influence of staffing levels at the PV center, the introduction of a new PV database, and enhancements to reporting tools on reporting quality at different time points.

Figure 4. Key strategies and interventions implemented to improve adverse event (AE) reporting in Malaysia between 2005 and 2019. CPD: continuing professional development; DIS: drug information service; HCP: health care professional; PhIS: pharmacy hospital information system; PRH: product registration holder; PRP: provisionally registered pharmacist; PV: pharmacovigilance; RiMUP: *risalah maklumat ubat untuk pengguna*.

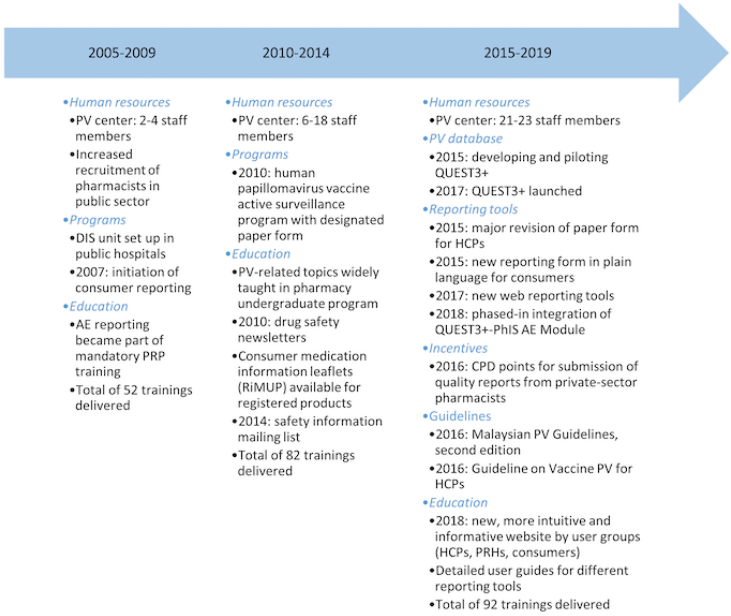


Figure 5 shows the annual trends in PV center staffing in relation to rates of reports being well documented. Figure 6 depicts how the transition in report submission format (from INTDIS to E2B) and reporting means correlated with report quantity and average completeness. In Figure 7, we focus on the rates of reports being well documented before and after the implementation of the new PV database (QUEST3+) and key enhancements to reporting tools since 2015. Information about reporting means was not available for INTDIS reports collected from the historical QUEST2 database. We further examined the influence and popularity of different reporting means among various reporters following the official launch of QUEST3+ and new web reporting tools in the first quarter of 2017 (Figure S4 in Multimedia Appendix 2).

Figure 5. Annual trends in pharmacovigilance (PV) center staffing levels and rate of reports being well documented (2005-2019).

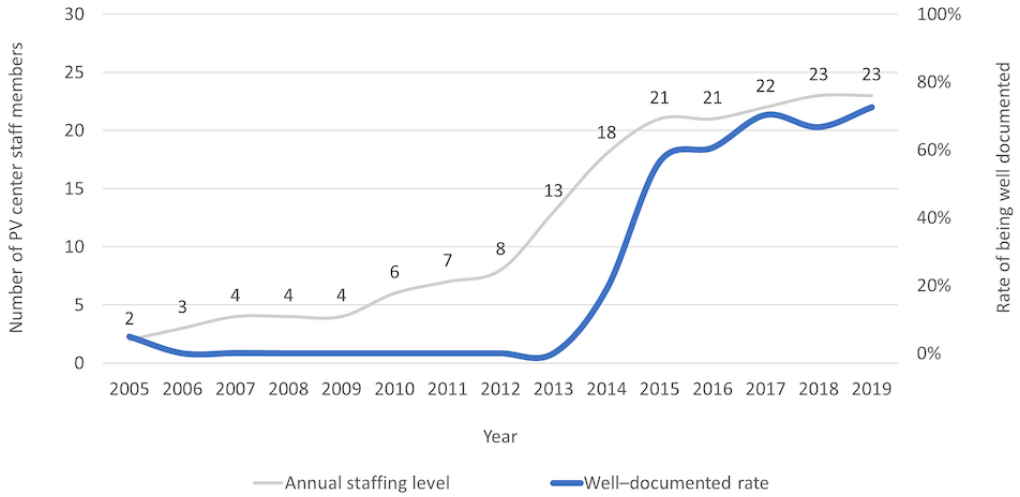


Figure 6. Mean completeness score and report count by submission format, pharmacovigilance database, and means of reporting yearly from 2013 to 2019. The size of the bubble corresponds to the report count. INTDIS: International Drug Information System; PhIS: pharmacy hospital information system.

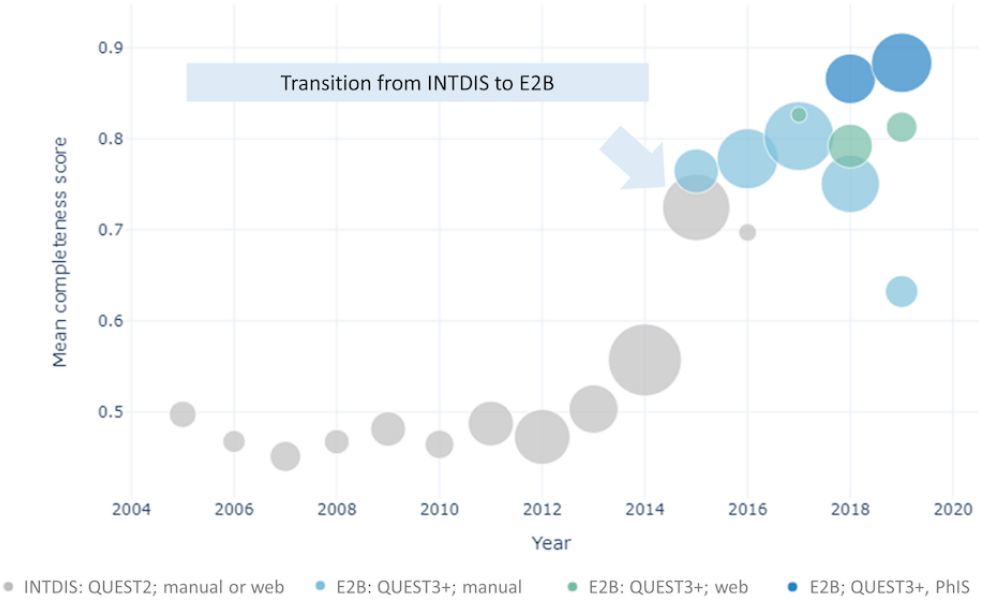
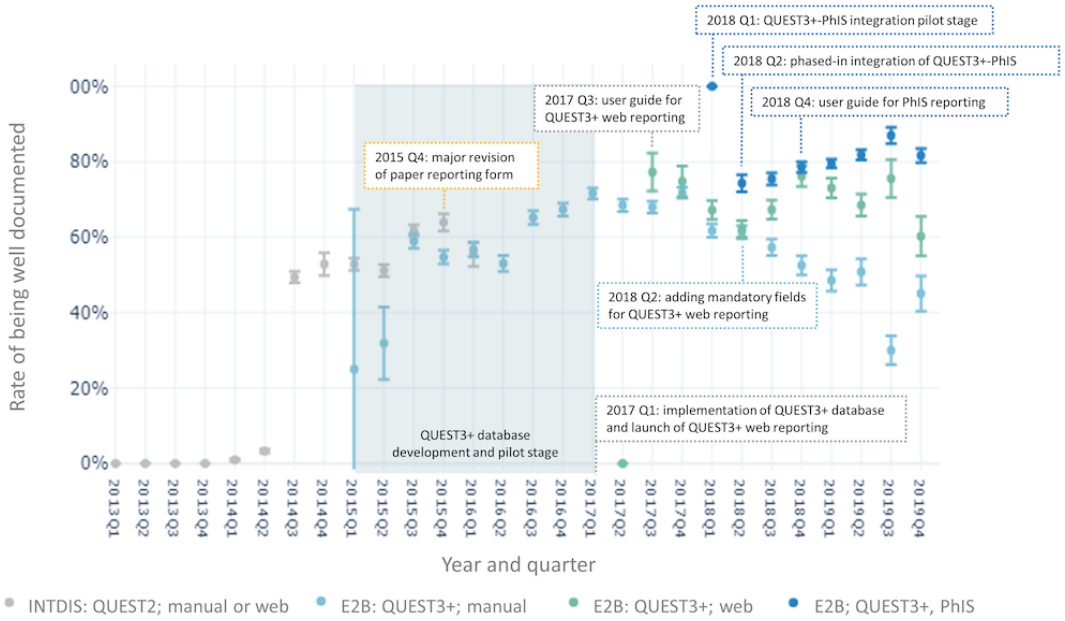


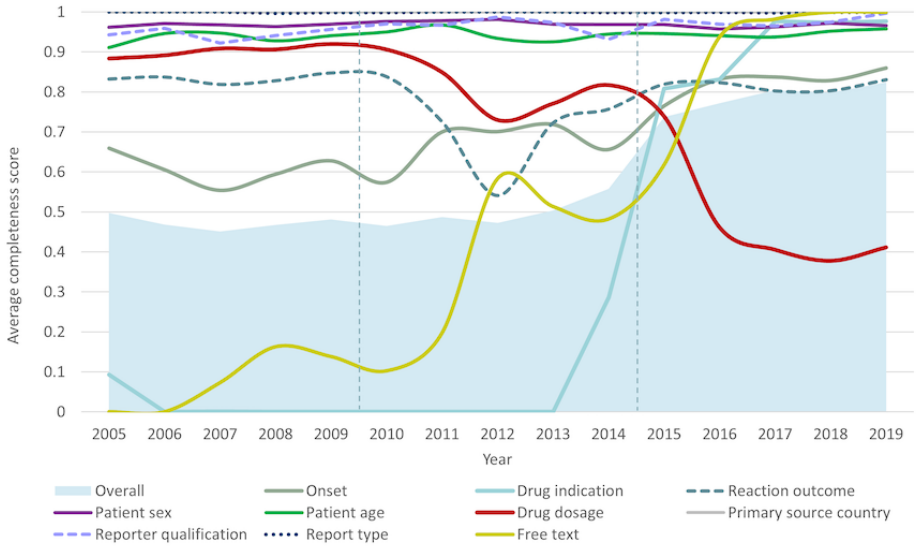
Figure 7. Rate of reports being well documented by submission format, pharmacovigilance database, and means of reporting quarterly from 2013 to 2019. 95% CI error bars (equivalent to 1.96×SE) were constructed. INTDIS: International Drug Information System; PhIS: pharmacy hospital information system; Q: quarter.



Average Completeness of Individual Dimensions for Malaysian AE Reports

Figure 8 illustrates the trends in the average completeness of the individual dimensions for Malaysian reports in VigiBase. The dimensions for report type, primary source country, reporter qualification, and patient age and sex were consistently the most completed. Completeness of free text and drug indication improved from zero in the earlier period of 2005 to 2009 to >0.9 in recent years. An uptrend in improvements was also observed for reaction onset. Completeness for reaction outcome dipped in 2011 to 2012 but subsequently rebounded to >0.8. Unexpectedly, we observed a noteworthy drop in drug dosage completeness since 2010. The average completeness of each individual dimension for vigiGrade was evaluated for E2B subsets (Figure S5 in Multimedia Appendix 2).

Figure 8. Trends in the distribution of average completeness scores for Malaysian reports in VigiBase over the study period. The shaded area indicates overall completeness; the line represents the individual dimension.



Discussion

Overview of Principal Findings

In our study, we used a comprehensive approach to examine the factors influencing the quality of Malaysian AE reports and decipher the temporal trends and interventions that shaped the Malaysian PV landscape from 2005 to 2019. In the first part of our study, by harnessing a data-driven approach that encompassed ML-based feature selection and analysis, we identified the key features that predict the status of Malaysian reports in VigiBase being well documented. Our hybrid feature selection helped in mitigating the risks of overfitting and unstable interpretability that are commonly associated with high variance [8,9,40] and resulted in a robust and valid RF model, as evidenced by the excellent classification performance (>90%) across all training, validation, and test sets for the E2B subset that reflects recent patterns of Malaysian reports (Table 1). While the model for the highly imbalanced INTDIS subset (containing only 10,691/63,943, 16.72% of well-documented reports) demonstrated satisfactory performance, with recall, precision, and F_1-scores of >70%, the model faced overfitting. This issue, evident from the decreased validation and test performance, has been acknowledged as a limitation in our study. The black-box RF model was made interpretable using SHAP values that, in agreement with human intuition, did not contradict the findings from the time-series analyses. To the best of our knowledge, this is the first study to use state-of-the-art interpretable ML methods to obtain insights concerning the key factors contributing to AE report quality in the PV field. Supplemented with insights drawn from our time-series and descriptive analyses, we summarized the identified features associated with well-documented Malaysian reports under 3 main themes: administrative; sender and reporter; and reaction, drug, and patient.

In the second part of our study, our extensive time-series and descriptive analyses illuminated the notable progress Malaysia has made in both the quantity and quality of AE reports over the years. We delved further into the chronological trends and characteristics of Malaysian AE reports, outlined the key strategies and interventions implemented at 5-year intervals, and tracked the influence of interventions of interest. These findings offer valuable insights into the multifaceted strategies and interventions that have driven enhancements in Malaysian AE reporting quality. Finally, by focusing on individual dimensions for vigiGrade completeness scoring, we not only gained additional insights into the specific characteristics and aspects of report completeness within the Malaysian context but also pinpointed systematic data quality issues that warrant further attention and improvement work.

Factors and Characteristics Associated With Well-Documented Malaysian Reports

Administrative

Our ML analysis distinguished PV center staffing level as the most important factor positively associated with well-documented INTDIS reports over 12 years between 2005 and 2016 (Figure S3 in Multimedia Appendix 2 and Figure 2A). Previous studies [66-68] have also highlighted manpower at national PV centers as a barrier to functional and sustainable PV systems. During the initial stage (2005-2009), only 2 to 4 staff members handled all PV operations in Malaysia (Figure 5). As the number of reports grew rapidly, the center struggled with a backlog of reports. The center swiftly grew from 8 to 21 staff members within 3 years (2012-2015) alongside the expansion of PV functions and task specializations, which coincided with a notable improvement in the completeness of reports received (Figures 3 and 5). Compared to the period from 2005 to 2009, more training workshops were delivered from 2010 to 2019 (Figure 4) to enhance staff and reporter competencies. As the staffing level and rate of reports being well documented became relatively stable afterward (Figure 5), the influence of PV staffing levels appeared less distinctive for E2B reports (Figure 2B). This could be due to some staff members focusing on other PV duties such as signal detection and assessment, risk management, and risk communication

rather than AE processing. It is worth noting that the transition to E2B may have also obfuscated the influence of increased staffing on report completeness. While less distinctive in our model, we cannot exclude that other centers with E2B reports and low staffing may still face overall low report completeness.

The capabilities and scalability of PV databases are essential for effective data collection and management [5]. Integrating electronic reporting into hospital information systems has been reported to reduce duplicate work and effectively improve AE reporting [69]. In Malaysia, the new QUEST3+ database was later integrated with the AE reporting module of a centralized PhIS implemented at Malaysian public hospitals and clinics, enabling the automatic input of reporter and sender information and improved reporter accessibility to patient, drug use, and regulatory information (eg, product registration number and batch number). Our ML analysis of E2B subsets (Figure 2B) revealed that PhIS-integrated reporting positively contributed to well-documented Malaysian reports, whereas web reporting had an unexpected negative association. The volume of PhIS reports surpassed other means in 2019 (Figure 7 and Figure S4 in Multimedia Appendix 2) and maintained the highest reporting quality. Web reporting was initially introduced in 2000, but due to unstable systems and slowly developing IT infrastructures at some public health facilities in the last decade, most reports before 2015 were submitted manually via postage, fax, or email. After mandatory fields were added, web reporting recorded a rate of >70% of reports being well documented, but it later declined to 60%. A similar but larger declining trend was also observed with manual reporting (Figure 5). These declines corresponded to the shift to PhIS reporting by public-sector pharmacists and were likely associated with other HCPs and consumers (Figure S4 in Multimedia Appendix 2). Our findings highlight that electronic reporting tools serve as ad hoc support for well-trained reporters but IT infrastructure maturity and widespread user acceptance are required for success. This is in line with a recent realist review [70] asserting that technological interventions alone, without capacity building, have little or no impact on health data quality.

Sender and Reporter

Our study revealed that the greatest proportion of well-documented Malaysian reports, amounting to 91.54% (52,066/56,877), originated from public health facilities overseen by the MOH (Table 3). Among the well-documented reports, 98.15% (55,825/56,877) were submitted by HCPs. This finding aligns with those of previous studies [71-73] conducted across different regions of Malaysia, which consistently highlight that most HCPs recognize AE reporting as part of their professional obligations. In Malaysia, consumer-generated reports constituted only 0.14% (78/56,877) of the well-documented reports, which stands in contrast to countries such as Denmark and Norway where three-fifths of well-documented reports came from consumers or non-HCPs [6].

Our observations revealed that Malaysian pharmacists generated 70.85% (40,295/56,877) of well-documented reports (Table 3), with most serving the public health sector [63]. Specifically, we identified public specialist hospitals and pharmacists as key features that positively contributed to the well-documented Malaysian E2B reports (Figure 2B). It is noteworthy that, during the earlier stage (2005-2009), pharmacists' reports exhibited the poorest quality and were recognized as a key factor predicting not well-documented INTDIS reports from 2005 to 2016 (Figure 2A). Nonetheless, over the years, following increased recruitment in public health services [74,75] and multifaceted initiatives aimed at strengthening pharmacists' roles and skills (Figure 4), pharmacists have become an integral part of AE monitoring in Malaysia. Almost 9 in 10 public hospital pharmacists in Malaysia have reported at least one AE in the past [76].

Malaysia presents a unique scenario in which pharmacists play a leading role in PV activities, distinguishing it from global trends, in which physicians contribute nearly two-thirds of well-documented reports [6]. The Malaysian context aligns with findings from a Spanish study in which pharmacists reported a great majority of the AEs due to the integration of PV into routine hospital pharmacy practices [77]. Similar observations were made in a pharmacist-led AE monitoring and management model in China, where pharmacists provided higher-quality reports among all HCPs [14]. Within Malaysian public hospitals, the pharmacist-led Drug Information Service (DIS) unit is responsible for facility-level PV activities, including responding to queries related to AEs; disseminating safety information; and compiling, verifying, and submitting AE reports to the NPRA [65,78]. In addition to direct detection and reporting by pharmacists, a collaborative mechanism exists within public health facilities where physicians and other HCPs are aware of the role of pharmacists in monitoring and reporting the AEs detected during clinical rounds or discussions. Moreover, we observed that over half (19,188/33,559, 57.18%) of the well-documented E2B reports made by pharmacists came from public specialist hospitals. These reports were believed to have benefited from the input of specialist physicians, suggesting a positive contribution of collaborative efforts among HCPs in enhancing the quality of AE reports.

In contrast, reports from PRHs and other HCPs (including regulatory affairs officers, clinical trial associates, nurses, and medical assistants) were flagged as the key features negatively associated with Malaysian report completeness (Figure 2B). These findings are consistent with the features observed in the United States [15], Brazil [16], Spain [17], South Korea [18], and Japan [19,20]. Of note, the NPRA classified the reports from PRHs as reported by other HCPs when the primary reporter was unknown. Among 7833 E2B reports from PRHs, only 778 (9.93%) reports were well documented (Table S1 in Multimedia Appendix 2), with an overall completeness score of 0.39 (Figure S5 in Multimedia Appendix 2). Information regarding drug dosage, reaction onset, reaction outcome, and patient age was most incomplete in Malaysian reports from PRHs. This could be attributed to the lack of a robust PV culture among PRHs. Existing literature [6,79] suggests that PRHs might prioritize submitting a report to fulfill pharmaceutical legislation [80] that mandates that PRHs report any suspected AEs within strict timelines even when minimal information is available. There could also be instances in which the primary reporter did not provide consent for follow-up. Conversely, it is conceivable that pharmacists and physicians serving at health facilities were

more motivated to make a clinically meaningful report even on a voluntary basis [17,79,81].

Reaction, Drug, and Patient

As AE reporting is highly dependent on individual motivation [5], we were interested in understanding whether the nature of drugs and reactions affects the quality of reports. While a report may involve more than one drug or reaction, most studies on AE report quality have not assessed all drugs and reactions reported in a case. Toki and Ono [21] examined only the primary suspected drug in a multivariable logistic regression model, whereas Araujo et al [22] and Masuka and Khoza [23] evaluated a specific drug group using simple univariable analysis. Other studies have evaluated only case-level information, such as case seriousness [14,18,24-26], fatal outcome [15], and causality [18]. As we converted drug-reaction pairs to case-level data, we evaluated the influence of drug- and reaction-related factors on overall report completeness for all reported suspected and interacting drugs.

Our ML analysis of the E2B subset (Figure 2B) revealed that a case where the reaction abated following a drug dechallenge (ie, positive dechallenge) was the primary key feature associated with a well-documented report. While information on drug dechallenge was unknown in most cases (Table S1 in Multimedia Appendix 2), it is possible that a positive dechallenge may have strengthened the reporter's confidence in the drug-reaction causal relationship and, thus, the motivation to construct a clinically meaningful report. While our findings suggest that positive dechallenge may have motivated more complete reports, there is no supporting study on this. It is important to note that reports that are well documented by vigiGrade standards might also tend to have a positive dechallenge. In other words, it may be that vigiGrade tends to flag those reports that have a positive dechallenge as well documented.

As expected, cases containing information on time to onset were more likely to be well documented as the onset dimension incurs the highest penalty of 50% for missing data and 30% if the uncertainty exceeds 1 month [6]. Our findings suggest that cases with a shorter (ie, <1 day) time to onset, dosing interval, and duration of drug use were most likely to be well documented. This observation might be attributable to better recall and description of events occurring within a shorter time frame following drug use or to greater reporter confidence in the drug-reaction relationship due to stronger temporal association and a lower likelihood of confounding factors. On the other hand, reports that involved reactions lasting 1 to 6 days tended to carry more information compared to those that involved reactions lasting <1 day or >6 days. A competing hypothesis is that reactions occurring within this time frame allow for sufficient time for more observation or data gathering while still being easily observed and described by patients and HCPs. Nevertheless, further research is needed to determine the specific factors contributing to the observed differences in report quality for cases with varying time to onset, dosing intervals, and durations of drug use.

Antibiotics for systemic use (ATC code J01) emerged as a key feature favorably contributing to well-documented E2B reports. First, it could be attributed to the baseline reporting patterns, where systemic antibiotics were the most commonly reported drug group in Malaysia, with over one-fifth of AE reports involving at least one systemic antibiotic (Table S1 in Multimedia Appendix 2). Second, this observation might suggest that Malaysian HCPs exercise heightened caution when using anti-infectives, leading to a higher likelihood of detecting AEs related to anti-infectives with higher report completeness. According to a study from a Malaysian infectious disease hospital [65], most inquiries (37.8%) received by the DIS unit concerned anti-infective drugs (ATC code J), which included other β-lactam antibacterials (ATC code J01D), direct-acting antivirals (ATC code J05A), and penicillins (ATC code J01C), with the largest proportion of the inquiries pertaining to their AEs and pediatric dosage adjustments. Although this observation could be expected in an infectious disease hospital, it also highlights the role of pharmacist-led DIS in AE monitoring and reinforces our previous discussion regarding pharmacists working in public health facilities tending to submit more complete reports. Trainings by the NPRA often prioritized pharmacists working in DIS units, who then conducted echo training for HCPs in their respective health facilities.

In addition, our analysis revealed a positive association between reports marked as serious and well-documented Malaysian reports (Figure 2B), consistent with previous studies from France [25,26], China [14,24], and South Korea [18] that highlighted higher completeness for serious reports. Previous research has also indicated that Malaysian HCPs prioritize reporting serious AEs [71]. The heightened gravity and potential consequences of these cases might prompt reporters to exercise greater diligence in ensuring reporting quality, including PRHs who are subjected to stricter reporting timelines for serious cases. Fatal outcomes were not flagged as the key feature contributing to Malaysian reports being well documented, likely due to their low prevalence (1048/68,795, 1.52%; Table S1 in Multimedia Appendix 2). Nonetheless, Malaysian fatal reports had a low overall completeness of 0.55 compared to 0.80 for nonfatal reports (Figure S5 in Multimedia Appendix 2). Another observational study evaluating reports submitted to the US Food and Drug Administration [15] also found that cases of patient deaths had the lowest completeness scores across reporting sources. This could be attributed to the absence of medical terminology describing the cause of death or indicate an investigation into a potential drug involvement.

Reactions under the HLGTs *product quality, supply, distribution, manufacturing,* and *quality system issues*, of which 97.78% (2242/2293) were related to product substitution issues and 91.41% (2096/2293) were sourced from public health facilities primarily by pharmacists, were captured as the key feature that negatively contributed to E2B reports being well documented. Subsequent investigations revealed that product substitution issues were most prevalent in 2018 (1355/2242, 60.44%) and primarily involved brand switching between 2 generic products: amlodipine (1012/2242, 45.14%) and perindopril (291/2242, 12.98%). Among these, only 24.8% (251/1012) and 32.6% (95/291) of the reports involving amlodipine and perindopril, respectively, were well documented. In the Malaysian public health care sector, drugs are procured

through 3 distinct mechanisms: a national concessionaire, national tenders, and direct purchases by health facilities for items not covered by the former 2 mechanisms [82]. However, in situations in which a product substitution issue is suspected and public health facilities need to directly procure alternative products for items already listed in the former 2 mechanisms, AEs or product complaints must be submitted as justification. It is believed that the reporters might submit reports containing only minimal information solely to comply with the drug procurement procedure.

Another key negative feature identified in the E2B subset was the presence of adolescent patients (aged 12-17 years). In comparison, reports involving adult patients (aged 18-44 years) and midlife adult patients (aged 45-64 years), which comprised the largest proportion of Malaysian reports, tended to be well documented. In South Korea, overall reports involving children and adolescents (aged 0-19 years) were negatively associated with being well documented in comparison to the older adult group (aged ≥65 years), whereas reports involving adults (aged 19-65 years) had a positive association [18].

Trends in Malaysian AE Reporting Quality Between 2005 and 2019

Building on the preceding discussion on the factors and characteristics associated with well-documented Malaysian AE reports, we expanded our scope to the chronological progression of AE reporting in Malaysia from 2005 to 2019. Our analyses underline that policy changes, continuity of education, and human resource development laid the foundation for a functional and sustainable SRS in Malaysia. Meanwhile, advancements in technological infrastructure, PV databases, and reporting tools contributed to the observed increase in both the quantity and quality of AE reports. These findings echo the expert-recommended 4-tier hierarchy of needs to achieve systemic capacity building for PV [83]—progressing from structures, systems, and roles to staff and infrastructure to skills to tools.

Malaysia, with its SRS governed within an established legal and regulatory framework, has historically struggled with challenges of underreporting and poorly reported AEs, as evidenced in Figure 3. In an effort to establish a functional and sustainable PV system, Malaysia placed early priorities on cultivating a reporting culture among HCPs and strengthening national PV capacities through collaborative efforts involving multiple stakeholders (Figure 4). Among them were policy changes to strengthen pharmacists' role in AE monitoring, increased recruitment of public-sector pharmacists and PV staff, active surveillance programs for targeted medicinal products, public awareness campaigns, and continuity of PV education to HCPs from undergraduate and preservice to at-service levels [7,61-63,65,74,75]. These initiatives were consistent with the existing literature, which emphasizes that multifaceted strategies and interventions work more synergistically to improve AE reporting than a single intervention [79,84,85]. Notably, our findings from the ML analysis suggest a positive association between higher staffing levels at the PV center and well-documented INTDIS reports, which could underscore the potential need for capacity building in the early phase of PV implementation.

As PV activities in Malaysia attained a higher level of maturation, the NPRA began to put greater emphasis on improving report quality. Comparative studies examining reporting forms from various countries have consistently highlighted that the Malaysian paper reporting form captures the most comprehensive information [86,87]. In response to the influx of reports observed in 2014 (Figure 3), the NPRA set their efforts on enhancing AE reporting tools and processing capabilities (Figure 4). Enhancements were made to the paper reporting form in 2015, including the addition of structured checkboxes and a reporting guide to ensure that more complete and harmonized clinical information could be obtained for subsequent causality assessment [27,88]. Concurrently, the NPRA began developing and piloting QUEST3+, an upgraded regulatory database system that marked a new submission format to the UMC—transitioning from INTDIS to E2B (Figure 6). The official launch of the QUEST3+ database took place in January 2017, replacing the historical QUEST2 database. Alongside these paradigm shifts, the NPRA also revamped and relaunched its web reporting tool for HCPs and introduced a new plain-language web reporting tool specifically for consumers (ConSERF). With the maturation of IT infrastructures, in 2018, the QUEST3+ database was integrated in phases with the centralized PhIS across Malaysian public health facilities. Reporting guides, drop-down lists, and validation alerts were also added to web and PhIS-integrated reporting tools to enhance the completeness and consistency of the collected data. Interestingly, as previously discussed, our comparative findings regarding PhIS-integrated tools used by well-trained pharmacists and new web reporting tools likely used by other HCPs and consumers highlight the complementary role of electronic reporting tools as ad hoc aids for well-trained reporters, whereas the effectiveness of these tools in improving AE reporting also relies on the maturity of IT infrastructures and their acceptance by users.

As a consequence of continuous efforts to strengthen PV capacities and technological advancements, Malaysia has seen considerable improvements not only in the quantity but also in the quality of reports. From 2015 to 2019, approximately two-thirds of Malaysian reports were well documented compared to approximately 1 in 5 reports from the rest of the world [28]. It is worth noting that, while overall completeness improved after the transition to the E2B submission format in 2015, our investigations revealed that low completeness in drug dosage (Figure 8) was systematically confounded by miscoding errors during report conversion to E2B-XML files before report transmission to VigiBase and, thus, was comparatively lowest in all subsets (Figure S5 in Multimedia Appendix 2). As a consequence of missing "number of unit in the interval" and miscoded "number of separate dosages," the drug dosage dimension for a Malaysian report in E2B format was penalized when the total daily dose for a case could not be calculated from the specified fields [29,89]. Similar to global reports [6], Malaysian E2B reports carried more administrative information, such as report type followed by reporter qualification and patient characteristics (ie, sex and age), but less drug- and

reaction-related information, such as drug dosage (despite the aforementioned confounding), reaction onset, and reaction outcome. In contrast to global reports [6], the inclusion of mandatory fields in electronic reporting tools led to a higher completeness of drug indication and free-text narratives in Malaysian reports. Reports from the literature and other sources, made by other HCPs, and submitted by PRHs had the lowest overall completeness scores (<0.5). Fatal reports and those from community pharmacies or other public services also tended to contain less information (<0.6).

Limitations

Our study is constrained by the limitations inherent to cross-sectional observational data and ML analysis, where causal reasoning and statistical inference cannot be determined [54]. The features identified in our study should be understood as predictors associated with well-documented reports but not as causal factors. Owing to the assumptions that multifaceted interventions often work synergistically and that control groups are frequently absent in nationwide implementation [79], the exact impact of individual interventions on reporting quality cannot be determined. As such, our study serves as an exploratory analysis, and the highlighted features offer a starting point for further in-depth review.

Our study faced challenges with the class imbalance inherent to the INTDIS data set, which heightened the risk of model overfitting. While undersampling improved the balanced performance of precision and recall, it could introduce new biases. Given that the INTDIS format is now obsolete in Malaysia, our focus is shifting toward the more recent E2B features.

Our models did not include causality information for several reasons. AE reports received by the NPRA and VigiBase come from a variety of sources, and the probability that the suspected AE is drug related is not the same in all cases [90]. Reporters and senders might use different methods for assessing causality, such as Naranjo probability scores and WHO-UMC causality categories, which were not available. In addition, it is important to note that causality may change as knowledge expands, and the UMC does not validate the causality assessments of the received reports.

Our data set is also constrained by the timeliness of report submission from QUEST to VigiBase and did not include all the reports received by the NPRA by December 31, 2019. As the systematic data quality issues uncovered in our study have already been communicated to the NPRA, follow-up work is underway to address these issues. Therefore, the findings of this study will not be representative of the future completeness scores of Malaysian reports in VigiBase. This also implies that the key features identified in our study were subject to multiple systematic biases, which are typically encountered when using real-world data [90,91].

Conclusions and Future Work

By using a data-driven approach and the vigiGrade method, we pinpointed the trends and milestones of the Malaysian AE reporting system and demonstrated how the country has striven to contribute large numbers of high-quality reports to global PV. Our work also highlights the vigiGrade method by the UMC as an effective tool for monitoring the quality of AE reports and aiding countries in evaluating to enhance reporting. Our multidimensional perspective on AE reporting trends and strategies in Malaysia, informed by data-driven insights, underlines the complexity and evolving nature of the SRS and the importance of continual improvement for global PV.

Using interpretable ML methods, we identified specific features that were positively associated with Malaysian AE reports being well documented. Notable factors include higher PV center staffing for INTDIS reports, reaction abated upon drug dechallenge, reaction onset or drug use duration of <1 week, dosing interval of <1 day, reports from public specialist hospitals, reports by pharmacists, and reaction duration between 1 and 6 days for recent E2B reports. Conversely, reports from PRHs and other HCPs indicated areas for potential improvement in the quality of Malaysian reports. These identified features could potentially serve as a basis for future research and strategies aimed at improving PV practices, thus improving drug safety surveillance and, ultimately, public health outcomes.

Furthermore, our time-series analysis showcased how Malaysia has built up and strengthened its PV capacity via multifaceted strategies and interventions to enhance both the quantity and quality of AE reports. Policy changes, continuity of education, and human resource development have all contributed to the foundation for a functional and sustainable SRS in Malaysia, whereas advancements in technological infrastructure, PV databases, and reporting tools concurred with the rise in both the quantity and quality of AE reports. These findings resonate with the expert-recommended 4-tier hierarchy of needs for systemic PV capacity building—from structures, systems, and roles to staff and infrastructure to skills to tools [83,92].

Building on our findings on Malaysia's progress in AE reporting and factors identified for report quality, we propose several areas for future work. To understand how and in what measure the findings from the time-series analysis contributed to the completeness of Malaysian reports, viewing the interventions set up by the NPRA as complex [93]—targeting multiple individuals or a wide range of behaviors and involving multiple interacting components—could be instrumental. Future evaluations may use this newly updated framework for complex intervention research [94].

Our findings revealed that the private health sector, including PRHs, private hospitals, private clinics, and community pharmacies, exhibited suboptimal contributions. This highlights persistent challenges pertaining to underreporting and unsatisfactory report quality in these sectors, necessitating further research into understanding behavioral or organizational barriers for developing targeted interventions [95,96]. Considering that preservice and in-service trainings often do not adequately prepare HCPs for data-related tasks [96], stronger stakeholder coordination and collaboration are imperative for continuous competency-based training and fostering an effective data use culture across health systems [95]. Regular feedback on reporting performance could be considered to facilitate self-monitoring among all senders.

Sustainable improvement in surveillance data quality and use requires a whole-systems approach encompassing governance, people, tools, and processes [95]. Given that data quality is highly reliant on their collection at health facilities, future work can prioritize people and environments essential for functional information systems as well as validation upon data entry to ensure completeness, accuracy, and consistency [88]. Our identification of systematic data quality issues highlighted a gap in data-driven continuous quality improvement [95], underscoring the need for internal quality assurance procedures for AE data management and transmission, including routine systematic checks and periodic in-depth reviews [88,97].

Looking ahead, as Malaysian reports currently use the E2B(R2) format, future efforts can navigate toward transitioning to the E2B(R3)-compliant database and reporting tools as the inclusion of null flavors in the E2B(R3) format helps address missing information by explaining data absence. Future work on data governance could explore leveraging automation, ML, and natural language processing to improve the overall efficiency and quality of AE data collection, processing, and management [98,99].

Acknowledgments

This paper is based on the master's thesis completed by SMC at the Graduate Institute of Biomedical Informatics, Taipei Medical University, under the supervision of SS-A and the joint supervision of DS and Jim Barrett from the Uppsala Monitoring Centre (UMC), as well as SCL from the National Pharmaceutical Regulatory Agency (NPRA). No generative artificial intelligence tools were used to create the original content for publication. The views expressed in this paper do not represent the opinions of the NPRA, the UMC, or the World Health Organization. The authors thank the Director-General of Health Malaysia for his permission to publish this study and are indebted to all stakeholders who contributed reports to QUEST and VigiBase. Special thanks to Usman Iqbal, Ekansh Gayakwad, Suo-Chen Chien, and Yu-Chin Chu from Taipei Medical University and Wai Lam Hoo from the University of Malaya for their valuable assistance and advice. The authors are also grateful for the unwavering support provided by Azuana Ramli, Norleen Mohamed Ali, Kobu Thiruvanackan, Nora Ashikin Mohd Ali, Mohd Ghazli Ismail, and Wee Kee Wo from the NPRA. The publication fund for this research was sponsored in part by the National Science and Technology Council (NSTC) Taiwan under grant NSTC 110-2320-B-038-029-MY3.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to the data protection policies of the Uppsala Monitoring Centre and the National Pharmaceutical Regulatory Agency. Data requests should be made directly to the institutions.

Authors' Contributions

SMC contributed to conceptualization, methodology, data curation, software, formal analysis, and original draft writing. DS and SCL contributed equally to conceptualization, methodology, formal analysis, and supervision. SS-A contributed to conceptualization, methodology, supervision, and funding acquisition. HCY contributed to methodology, formal analysis, and funding acquisition. All authors have reviewed, edited, and approved the final manuscript. SS-A and HCY are co-corresponding authors on this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of definitions or operational definitions.
[PDF File (Adobe PDF File), 243 KB - medinform_v12i1e49643_app1.pdf]

Multimedia Appendix 2
Supplementary figures and tables.
[PDF File (Adobe PDF File), 1866 KB - medinform_v12i1e49643_app2.pdf]

Multimedia Appendix 3
Literature review on quality of reports in the spontaneous reporting system.
[PDF File (Adobe PDF File), 101 KB - medinform_v12i1e49643_app3.pdf]

Multimedia Appendix 4
Feature selection.
[PDF File (Adobe PDF File), 206 KB - medinform_v12i1e49643_app4.pdf]

Multimedia Appendix 5
Tree-based machine learning models and interpretable machine learning.
[PDF File (Adobe PDF File), 79 KB - medinform_v12i1e49643_app5.pdf]

References

1. The importance of pharmacovigilance. World Health Organization. 2002. URL: https://apps.who.int/iris/handle/10665/42493 [accessed 2021-05-22]
2. Natsiavas P, Malousi A, Bousquet C, Jaulent MC, Koutkias V. Computational advances in drug safety: systematic and mapping review of knowledge engineering based approaches. Front Pharmacol 2019 May 17;10:415 [FREE Full text] [doi: 10.3389/fphar.2019.00415] [Medline: 31156424]
3. Bate A, Hornbuckle K, Juhaeri J, Motsko SP, Reynolds RF. Hypothesis-free signal detection in healthcare databases: finding its value for pharmacovigilance. Ther Adv Drug Saf 2019 Aug 5;10:2042098619864744 [FREE Full text] [doi: 10.1177/2042098619864744] [Medline: 31428307]
4. Uppsala Monitoring Centre. URL: https://www.who-umc.org/ [accessed 2021-05-22]
5. García CH, Pinheiro L, Maciá MA, Stroe R, Georgescu A, Dondera R, et al. Spontaneous adverse drug reactions: subgroup report. Heads of Medicines Agencies, European Medicines Agency. URL: https://tinyurl.com/4n76snn5 [accessed 2021-05-22]
6. Bergvall T, Norén GN, Lindquist M. vigiGrade: a tool to identify well-documented individual case reports and highlight systematic data quality issues. Drug Saf 2013 Dec 17;37(1):65-77. [doi: 10.1007/s40264-013-0131-x]
7. National Pharmaceutical Regulatory Agency, Ministry of Health Malaysia. URL: https://npra.gov.my/index.php/en/ [accessed 2021-02-22]
8. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health 2020 Apr 02;41:21-36 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094437] [Medline: 31577910]
9. Stevens LM, Linstead E, Hall JL, Kao DP. Association between coffee intake and incident heart failure risk. Circ Heart Failure 2021 Feb;14(2):e006799. [doi: 10.1161/circheartfailure.119.006799]
10. Merriam-Webster, Inc. Merriam-Webster Dictionary. Merriam-Webster, Inc. URL: https://www.merriam-webster.com/ [accessed 2024-03-25]
11. Klein K, Scholl JH, De Bruin ML, van Puijenbroek EP, Leufkens HG, Stolk P. When more is less: an exploratory study of the precautionary reporting bias and its impact on safety signal detection. Clin Pharma Therapeutics 2017 Oct 25;103(2):296-303. [doi: 10.1002/cpt.879]
12. Edwards IR, Lindquist M, Wiholm BE, Napke E. Quality criteria for early signals of possible adverse drug reactions. The Lancet 1990 Jul;336(8708):156-158 [FREE Full text] [doi: 10.1016/0140-6736(90)91669-2]
13. Caster O, Juhlin K, Watson S, Norén GN. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank. Drug Saf 2014 Jul 23;37(8):617-628. [doi: 10.1007/s40264-014-0204-5]
14. Chen Y, Niu R, Xiang Y, Wang N, Bai J, Feng B. The quality of spontaneous adverse drug reaction reports in China: a descriptive study. Biol Pharml Bull 2019;42(12):2083-2088. [doi: 10.1248/bpb.b19-00637]
15. Moore TJ, Furberg CD, Mattison DR, Cohen MR. Completeness of serious adverse drug event reports received by the US Food and Drug Administration in 2014. Pharmacoepidemiol Drug Safety 2016 Feb 10;25(6):713-718. [doi: 10.1002/pds.3979]
16. Ribeiro A, Lima S, Zampieri ME, Peinado M, Figueras A. Filling quality of the reports of adverse drug reactions received at the Pharmacovigilance Centre of São Paulo (Brazil): missing information hinders the analysis of suspected associations. Expert Opin Drug Saf 2017 Aug 23;16(12):1329-1334. [doi: 10.1080/14740338.2017.1369525]
17. Plessis L, Gómez A, García N, Cereza G, Figueras A. Lack of essential information in spontaneous reports of adverse drug reactions in Catalonia—a restraint to the potentiality for signal detection. Eur J Clin Pharmacol 2017 Mar 1;73(6):751-758. [doi: 10.1007/s00228-017-2223-5]
18. Oh IS, Baek YH, Kim HJ, Lee M, Shin JY. Differential completeness of spontaneous adverse event reports among hospitals/clinics, pharmacies, consumers, and pharmaceutical companies in South Korea. PLoS ONE 2019 Feb 14;14(2):e0212336. [doi: 10.1371/journal.pone.0212336]
19. Tsuchiya M, Obara T, Miyazaki M, Noda A, Sakai T, Funakoshi R, et al. High-quality reports and their characteristics in the Japanese Adverse Drug Event Report database (JADER). J Pharm Pharm Sci 2021 Apr 08;24:161-173. [doi: 10.18433/jpps31417]
20. Tsuchiya M, Obara T, Sakai T, Nomura K, Takamura C, Mano N. Quality evaluation of the Japanese Adverse Drug Event Report database (JADER). Pharmacoepidemiol Drug 2019 Dec 10;29(2):173-181. [doi: 10.1002/pds.4944]
21. Toki T, Ono S. Assessment of factors associated with completeness of spontaneous adverse event reporting in the United States: a comparison between consumer reports and healthcare professional reports. J Clin Pharm Ther 2019 Nov 25;45(3):462-469. [doi: 10.1111/jcpt.13086]
22. Araujo AG, Lucchetta RC, Tonin FS, Pontarolo R, Borba HH, Wiens A. Analysis of completeness for spontaneous reporting of disease-modifying therapies in multiple sclerosis. Expert Opin Drug Saf 2021 Mar 11;20(6):735-740. [doi: 10.1080/14740338.2021.1897566]
23. Masuka JT, Khoza S. An analysis of the trends, characteristics, scope, and performance of the Zimbabwean pharmacovigilance reporting scheme. Pharmacol Res Perspect 2020 Sep 15;8(5):e00657. [doi: 10.1002/prp2.657]

24. Niu R, Xiang Y, Wu T, Zhang Z, Chen Y, Feng B. The quality of spontaneous adverse drug reaction reports from the pharmacovigilance centre in western China. Expert Opin Drug Saf 2019 Jan;18(1):51-58. [doi: 10.1080/14740338.2019.1559812] [Medline: 30574811]
25. Humbert X, Jacquot J, Alexandre J, Sassier M, Robin N, Pageot C, et al. Completeness of pharmacovigilance reporting in general medicine in France. Sante Publique 2019;31(4):561-566. [doi: 10.3917/spub.194.0561]
26. Durrieu G, Jacquot J, Mège M, Bondon-Guitton E, Rousseau V, Montastruc F, et al. Completeness of spontaneous adverse drug reaction reports sent by general practitioners to a regional pharmacovigilance centre: a descriptive study. Drug Saf 2016 Sep 29;39(12):1189-1195. [doi: 10.1007/s40264-016-0463-4]
27. Bahk CY, Goshgarian M, Donahue K, Freifeld CC, Menone CM, Pierce CE, et al. Increasing patient engagement in pharmacovigilance through online community outreach and mobile reporting applications: an analysis of adverse event reporting for the Essure device in the US. Pharmaceut Med 2015;29(6):331-340 [FREE Full text] [doi: 10.1007/s40290-015-0106-6] [Medline: 26635479]
28. Wakao R, Taavola H, Sandberg L, Iwasa E, Soejima S, Chandler R, et al. Data-driven identification of adverse event reporting patterns for Japan in VigiBase, the WHO global database of individual case safety reports. Drug Saf 2019 Sep 26;42(12):1487-1498. [doi: 10.1007/s40264-019-00861-y]
29. Technical description of vigiGrade: completeness score. Uppsala Monitoring Centre. URL: https://tinyurl.com/2b55m5k8 [accessed 2024-03-11]
30. Kheloufi F, Default A, Rouby F, Laugier-Castellan D, Boyer M, Rodrigues B, et al. Informativeness of patient initial reports of adverse drug reactions. Can it be improved by a pharmacovigilance centre? Eur J Clin Pharmacol 2017 Aug;73(8):1009-1018. [doi: 10.1007/s00228-017-2254-y] [Medline: 28391408]
31. Muñoz MA, Delcher C, Dal Pan GJ, Kortepeter CM, Wu E, Wei YJ, et al. Impact of a new consumer form on the quantity and quality of adverse event reports submitted to the United States Food and Drug Administration. Pharmacotherapy 2019 Nov;39(11):1042-1052. [doi: 10.1002/phar.2325] [Medline: 31479525]
32. Fernandez-Fernandez C, Lázaro-Bengoa E, Fernández-Antón E, Quiroga-González L, Montero Corominas D. Quantity is not enough: completeness of suspected adverse drug reaction reports in Spain-differences between regional pharmacovigilance centres and pharmaceutical industry. Eur J Clin Pharmacol 2020 Aug;76(8):1175-1181. [doi: 10.1007/s00228-020-02894-0] [Medline: 32447435]
33. Tsuchiya M, Obara T, Miyazaki M, Noda A, Takamura C, Mano N. The quality assessment of the Japanese Adverse Drug Event Report database using vigiGrade. Int J Clin Pharm 2020 Apr;42(2):728-736. [doi: 10.1007/s11096-020-00969-7] [Medline: 32020439]
34. Rolfes L, van Hunsel F, van der Linden L, Taxis K, van Puijenbroek E. The quality of clinical information in adverse drug reaction reports by patients and healthcare professionals: a retrospective comparative analysis. Drug Saf 2017 Jul;40(7):607-614 [FREE Full text] [doi: 10.1007/s40264-017-0530-5] [Medline: 28405899]
35. Masuka JT, Khoza S. Adverse events following immunisation (AEFI) reports from the Zimbabwe expanded programme on immunisation (ZEPI): an analysis of spontaneous reports in Vigibase from 1997 to 2017. BMC Public Health 2019 Aug 27;19(1):1166 [FREE Full text] [doi: 10.1186/s12889-019-7482-x] [Medline: 31455314]
36. Uppsala reports 68. Uppsala Monitoring Centre. 2015 Jan. URL: https://who-umc.org/media/164371/ur68_final_2_gb.pdf [accessed 2024-03-12]
37. Oosterhuis I, Taavola H, Tregunno PM, Mas P, Gama S, Newbould V, et al. Characteristics, quality and contribution to signal detection of spontaneous reports of adverse drug reactions via the WEB-RADR mobile application: a descriptive cross-sectional study. Drug Saf 2018 Oct;41(10):969-978 [FREE Full text] [doi: 10.1007/s40264-018-0679-6] [Medline: 29761281]
38. Jokinen J, Bertin D, Donzanti B, Hormbrey J, Simmons V, Li H, et al. Industry assessment of the contribution of patient support programs, market research programs, and social media to patient safety. Ther Innov Regul Sci 2019 Nov;53(6):736-745. [doi: 10.1177/2168479019877384] [Medline: 31684774]
39. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract 2017 Mar 31;36(1):3-11. [doi: 10.23876/j.krcp.2017.36.1.3]
40. Kuhn M, Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Models. Boca Raton, FL: CRC Press; 2019.
41. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007 Oct 01;23(19):2507-2517. [doi: 10.1093/bioinformatics/btm344] [Medline: 17720704]
42. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020 Oct;13(10) [FREE Full text] [doi: 10.1161/circoutcomes.120.006556]
43. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003 Mar 1;3:1157-1182 [FREE Full text] [doi: 10.1162/153244303322753616]
44. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870]

45. Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods 2018 Jun;15(6):399-400. [doi: 10.1038/s41592-018-0019-x] [Medline: 29855577]
46. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods 2016 Aug 30;13(9):703-704. [doi: 10.1038/nmeth.3968]
47. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997 Dec;97(1-2):273-324. [doi: 10.1016/S0004-3702(97)00043-X]
48. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform 2018 Aug;116:10-17 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.05.006] [Medline: 29887230]
49. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA.
50. van den Bosch T, Warps AL, de Nerée tot Babberich MP, Stamm C, Geerts BF, Vermeulen L, et al. Predictors of 30-day mortality among Dutch patients undergoing colorectal cancer surgery, 2011-2016. JAMA Netw Open 2021 Apr 26;4(4):e217737. [doi: 10.1001/jamanetworkopen.2021.7737]
51. Gong K, Lee HK, Yu K, Xie X, Li J. A prediction and interpretation framework of acute kidney injury in critical care. J Biomed Inform 2021 Jan;113:103653 [FREE Full text] [doi: 10.1016/j.jbi.2020.103653]
52. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
53. Ariza-Garzon MJ, Arroyo J, Caparrini A, Segovia-Vargas MJ. Explainability of a machine learning granting scoring model in peer-to-peer lending. IEEE Access 2020;8:64873-64890. [doi: 10.1109/access.2020.2984412]
54. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A 2019 Oct 16;116(44):22071-22080. [doi: 10.1073/pnas.1900654116]
55. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Victoria, CA: Leanpub; 2020.
56. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health 2020 Apr;2(4):e179-e191. [doi: 10.1016/s2589-7500(20)30018-2]
57. Shapley L. A value for n-person games. In: Kuhn H, Tucker A, editors. Contributions to the Theory of Games II. Princeton, NJ: Princeton University Press; 1953:307-317.
58. Breiman L. Random forests. Mach Learn 2001;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
59. Seabold S, Perktold J. Statsmodels econometric and modeling with python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy 2010; June 28-July 3, 2010; Austin, TX. [doi: 10.25080/majora-92bf1922-011]
60. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
61. National centre for adverse drug reactions monitoring annual report. National Pharmaceutical Regulatory Agency (NPRA), Ministry of Health Malaysia. 2021 Nov 29. URL: https://tinyurl.com/44ayxzwk [accessed 2022-02-22]
62. Elkalmi RM, Hassali MA, Al-lela OQ, Jamshed SQ. The teaching of subjects related to pharmacovigilance in Malaysian pharmacy undergraduate programs. J Pharmacovigilance 2013;1(2):1-5. [doi: 10.4172/2329-6887.1000106]
63. Annual Report and Statistics of the Pharmaceutical Services Programme. Pharmaceutical Services Programme, Ministry of Health Malaysia. 2019. URL: https://tinyurl.com/y4h7ctn4 [accessed 2021-05-22]
64. Rosli R, Dali AF, Aziz NA, Ming LC, Manan MM. Reported adverse drug reactions in infants: a nationwide analysis in Malaysia. Front Pharmacol 2017 Feb 10;8:30 [FREE Full text] [doi: 10.3389/fphar.2017.00030] [Medline: 28239351]
65. Ali A, Mohd Yusoff S, Mohd Joffry S, Wahab MS. Drug information service awareness program and its impact on characteristics of inquiries at DIS unit in Malaysian public hospital. Arch Pharma Pract 2013;4(1):9. [doi: 10.4103/2045-080x.111576]
66. Suwankesawong W, Dhippayom T, Tan-Koi WC, Kongkaew C. Pharmacovigilance activities in ASEAN countries. Pharmacoepidemiol Drug Saf 2016 Sep 12;25(9):1061-1069. [doi: 10.1002/pds.4023] [Medline: 27174034]
67. Ampadu HH, Hoekman J, Arhinful D, Amoama-Dapaah M, Leufkens HG, Dodoo AN. Organizational capacities of national pharmacovigilance centres in Africa: assessment of resource elements associated with successful and unsuccessful pharmacovigilance experiences. Global Health 2018 Nov 16;14(1):109 [FREE Full text] [doi: 10.1186/s12992-018-0431-0] [Medline: 30445979]
68. Olsson S, Pal SN, Stergachis A, Couper M. Pharmacovigilance activities in 55 low- and middle-income countries: a questionnaire-based analysis. Drug Saf 2010 Aug 01;33(8):689-703. [doi: 10.2165/11536390-000000000-00000] [Medline: 20635827]
69. Ortega A, Aguinagalde A, Lacasa C, Aquerreta I, Fernández-Benítez M, Fernández LM. Efficacy of an adverse drug reaction electronic reporting system integrated into a hospital information system. Ann Pharmacother 2008 Aug 26;42(10):1491-1496. [doi: 10.1345/aph.1l130]

70. A realist review of what works to improve data use for immunization: evidence from low- and middle-income countries. Pan American Health Organization, World Health Organization. 2019. URL: https://tinyurl.com/3p5dveue [accessed 2024-03-04]
71. Balan S. Knowledge, attitude and practice of Malaysian healthcare professionals toward adverse drug reaction reporting: a systematic review. Int J Pharm Pract 2021 Aug 11;29(4):308-320. [doi: 10.1093/ijpp/riab030] [Medline: 34289016]
72. Ali RS, Ismail WI. Adverse drug reactions reporting: knowledge, attitude and practice among healthcare providers at a tertiary hospital in northern region of Malaysia. Asian J Med Health Sci 2021 Oct;4(Supplement 1):214-227 [FREE Full text]
73. Kirthikaa GK. Evaluation of knowledge, attitude, and practice towards adverse drug reaction reporting and reason for underreporting among the private and public medical practitioners of Kuala Lumpur and Selangor. International Medical University Central Digital Repository. 2022. URL: https://rep.imu.edu.my/xmlui/handle/1234.56789/2945?show=full [accessed 2024-03-04]
74. Rosli R, Ming LC, Abd Aziz N, Manan MM. A retrospective analysis of spontaneous adverse drug reactions reports relating to paediatric patients. PLoS ONE 2016 Jun 1;11(6):e0155385. [doi: 10.1371/journal.pone.0155385]
75. Hadi MA, Ming LC. Impact of pharmacist recruitment on ADR reporting: Malaysian experience. South Med Rev 2011 Dec;4(2):102-103 [FREE Full text] [doi: 10.5655/smr.v4i2.1009] [Medline: 23093890]
76. Hadi MA, Helwani R, Long CM. Facilitators and barriers towards adverse drug reaction reporting: perspective of Malaysian hospital pharmacists. J Pharm Health Serv Res 2013 May 24;4(3):155-158. [doi: 10.1111/jphs.12022]
77. Pérez-Ricart A, Gea-Rodríguez E, Roca-Montañana A, Gil-Máñez E, Pérez-Feliu A. Integrating pharmacovigilance into the routine of pharmacy department: experience of nine years. Farm Hosp 2019 Jul 01;43(4):128-133. [doi: 10.7399/fh.11169] [Medline: 31276445]
78. Malayasian adverse drug reactions newsletter. National Pharmaceutical Control Bureau, Ministry of Health Malaysia. 2013 Apr. URL: https://tinyurl.com/3ywrmb5b [accessed 2021-05-22]
79. Li R, Zaidi ST, Chen T, Castelino R. Effectiveness of interventions to improve adverse drug reaction reporting by healthcare professionals over the last decade: a systematic review. Pharmacoepidemiol Drug Saf 2020 Jan;29(1):1-8 [FREE Full text] [doi: 10.1002/pds.4906] [Medline: 31724270]
80. Malaysian guidelines on Good Pharmacovigilance Pratices (GVP) for product registration holders. National Pharmaceutical Regulatory Agency (NPRA), Ministry of Health Malaysia. 2021 Sep 30. URL: https://tinyurl.com/mpass52p [accessed 2022-02-22]
81. Ribeiro-Vaz I, Silva AM, Costa Santos C, Cruz-Correia R. How to promote adverse drug reaction reports using information systems - a systematic review and meta-analysis. BMC Med Inform Decis Mak 2016 Mar 01;16:27 [FREE Full text] [doi: 10.1186/s12911-016-0265-8] [Medline: 26926375]
82. Hamzah NM, Perera PN, Rannan-Eliya RP. How well does Malaysia achieve value for money in public sector purchasing of medicines? Evidence from medicines procurement prices from 2010 to 2014. BMC Health Serv Res 2020 Jun 05;20:509. [doi: 10.1186/s12913-020-05362-8]
83. Potter C. Systemic capacity building: a hierarchy of needs. Health Policy Plan 2004 Sep 01;19(5):336-345. [doi: 10.1093/heapol/czh038]
84. Khalili M, Mesgarpour B, Sharifi H, Daneshvar Dehnavi S, Haghdoost AA. Interventions to improve adverse drug reaction reporting: a scoping review. Pharmacoepidemiol Drug 2020 May 19;29(9):965-992 [FREE Full text] [doi: 10.1002/pds.4966]
85. Gonzalez-Gonzalez C, Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Strategies to improve adverse drug reaction reporting: a critical and systematic review. Drug Saf 2013 May;36(5):317-328. [doi: 10.1007/s40264-013-0058-2] [Medline: 23640659]
86. Singh A, Bhatt P. Comparative evaluation of adverse drug reaction reporting forms for introduction of a spontaneous generic ADR form. J Pharmacol Pharmacotherapeutics 2022 Apr 11;3(3):228-232. [doi: 10.4103/0976-500x.99417]
87. Bandekar MS, Anwikar SR, Kshirsagar NA. Quality check of spontaneous adverse drug reaction reporting forms of different countries. Pharmacoepidemiol Drug 2010 Sep 15;19(11):1181-1185. [doi: 10.1002/pds.2004]
88. Increasing Adverse Event Reporting (IAER) subproject: survey report. International Coalition of Medicines Regulatory Authorities. URL: https://tinyurl.com/2ksfrm6z [accessed 2021-05-22]
89. Electronic transmission of individual case safety reports message specification. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. 2001. URL: https://admin.ich.org/sites/default/files/inline-files/ICH_ICSR_Specification_V2-3.pdf [accessed 2024-03-12]
90. Guideline for using VigiBase data in studies. Uppsala Monitoring Centre (UMC). 2021 Mar 15. URL: https://who-umc.org/media/05kldqpj/guidelineusingvigibaseinstudies.pdf [accessed 2021-05-22]
91. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest 2020 Feb 03;130(2):565-574 [FREE Full text] [doi: 10.1172/JCI129197] [Medline: 32011317]
92. Indicator-based pharmacovigilance assessment tool: manual for conducting assessments in developing countries. United States Agency for International Development. 2009 Dec. URL: https://pdf.usaid.gov/pdf_docs/pnads167.pdf [accessed 2021-05-22]
93. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655. [doi: 10.1136/bmj.a1655]

94. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061]
95. Scobie HM, Edelstein M, Nicol E, Morice A, Rahimi N, MacDonald NE, et al. Improving the quality and use of immunization and surveillance data: summary report of the Working Group of the Strategic Advisory Group of Experts on Immunization. Vaccine 2020 Oct;38(46):7183-7197. [doi: 10.1016/j.vaccine.2020.09.017]
96. Nicol E, Turawa E, Bonsu G. Pre- and in-service training of health care workers on immunization data management in LMICs: a scoping review. Hum Resour Health 2019 Dec 02;17:92. [doi: 10.1186/s12960-019-0437-6]
97. Radecka A, Loughlin L, Foy M, de Ferraz Guimaraes MV, Sarinic VM, Di Giusti MD, et al. Enhancing pharmacovigilance capabilities in the EU regulatory network: the SCOPE joint action. Drug Saf 2018 Aug 21;41(12):1285-1302 [FREE Full text] [doi: 10.1007/s40264-018-0708-5] [Medline: 30128638]
98. Ghosh R, Kempf D, Pufko A, Barrios Martinez LF, Davis CM, Sethi S. Automation opportunities in pharmacovigilance: an industry survey. Pharm Med 2020 Feb 08;34(1):7-18. [doi: 10.1007/s40290-019-00320-0]
99. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. Clin Pharma Ther 2018 Dec 11;105(4):954-961. [doi: 10.1002/cpt.1255]

Abbreviations

AE: adverse event
ATC: Anatomical Therapeutic Chemical
DCA: Drug Control Authority
DIS: Drug Information Service
HCP: health care professional
HLGTs: High-Level Group Terms
ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICSR: individual case safety report
INTDIS: International Drug Information System
MADRAC: Malaysian Adverse Drug Reactions Advisory Committee
MedDRA: Medical Dictionary for Regulatory Activities
ML: machine learning
MOH: Ministry of Health
NPRA: National Pharmaceutical Regulatory Agency
PhIS: pharmacy hospital information system
PRH: product registration holder
PV: pharmacovigilance
RF: random forest
SHAP: Shapley additive explanations
SRS: spontaneous reporting system
UMC: Uppsala Monitoring Centre
WHO PIDM: World Health Organization Programme for International Drug Monitoring
WHO: World Health Organization

Edited by C Lovis; submitted 05.06.23; peer-reviewed by S Matsuda, C Zhao, I Degen; comments to author 20.08.23; revised version received 10.10.23; accepted 24.02.24; published 03.04.24.

Please cite as:
Choo SM, Sartori D, Lee SC, Yang HC, Syed-Abdul S
Data-Driven Identification of Factors That Influence the Quality of Adverse Event Reports: 15-Year Interpretable Machine Learning and Time-Series Analyses of VigiBase and QUEST
JMIR Med Inform 2024;12:e49643
URL: https://medinform.jmir.org/2024/1/e49643
doi:10.2196/49643
PMID:38568722

Original Paper

Construction of a Multi-Label Classifier for Extracting Multiple Incident Factors From Medication Incident Reports in Residential Care Facilities: Natural Language Processing Approach

Hayato Kizaki[1], MSc; Hiroki Satoh[2,3], PhD; Sayaka Ebara[1], BSc; Satoshi Watabe[1], BSc; Yasufumi Sawada[2], PhD; Shungo Imai[1], PhD; Satoko Hori[1], PhD

[1]Division of Drug Informatics, Keio University Faculty of Pharmacy, Tokyo, Japan

[2]Graduate School of Pharmaceutical Sciences, The University of Tokyo, Tokyo, Japan

[3]Interfaculty Initiative in Information Studies, The University of Tokyo, Tokyo, Japan

Corresponding Author:
Hayato Kizaki, MSc
Division of Drug Informatics
Keio University Faculty of Pharmacy
1-5-30 Shibakoen
Minato-ku
Tokyo
Japan
Phone: 81 354002799
Fax: 81 354002799
Email: hayatokizaki625@keio.jp

Abstract

Background: Medication safety in residential care facilities is a critical concern, particularly when nonmedical staff provide medication assistance. The complex nature of medication-related incidents in these settings, coupled with the psychological impact on health care providers, underscores the need for effective incident analysis and preventive strategies. A thorough understanding of the root causes, typically through incident-report analysis, is essential for mitigating medication-related incidents.

Objective: We aimed to develop and evaluate a multilabel classifier using natural language processing to identify factors contributing to medication-related incidents using incident report descriptions from residential care facilities, with a focus on incidents involving nonmedical staff.

Methods: We analyzed 2143 incident reports, comprising 7121 sentences, from residential care facilities in Japan between April 1, 2015, and March 31, 2016. The incident factors were annotated using sentences based on an established organizational factor model and previous research findings. The following 9 factors were defined: procedure adherence, medicine, resident, resident family, nonmedical staff, medical staff, team, environment, and organizational management. To assess the label criteria, 2 researchers with relevant medical knowledge annotated a subset of 50 reports; the interannotator agreement was measured using Cohen κ. The entire data set was subsequently annotated by 1 researcher. Multiple labels were assigned to each sentence. A multilabel classifier was developed using deep learning models, including 2 Bidirectional Encoder Representations From Transformers (BERT)–type models (Tohoku-BERT and a University of Tokyo Hospital BERT pretrained with Japanese clinical text: UTH-BERT) and an Efficiently Learning Encoder That Classifies Token Replacements Accurately (ELECTRA), pretrained on Japanese text. Both sentence- and report-level training were performed; the performance was evaluated by the F_1-score and exact match accuracy through 5-fold cross-validation.

Results: Among all 7121 sentences, 1167, 694, 2455, 23, 1905, 46, 195, 1104, and 195 included “procedure adherence,” “medicine,” “resident,” “resident family,” “nonmedical staff,” “medical staff,” “team,” “environment,” and “organizational management,” respectively. Owing to limited labels, “resident family” and “medical staff” were omitted from the model development process. The interannotator agreement values were higher than 0.6 for each label. A total of 10, 278, and 1855 reports contained no, 1, and multiple labels, respectively. The models trained using the report data outperformed those trained using sentences, with macro F_1-scores of 0.744, 0.675, and 0.735 for Tohoku-BERT, UTH-BERT, and ELECTRA, respectively. The report-trained models also demonstrated better exact match accuracy, with 0.411, 0.389, and 0.399 for Tohoku-BERT, UTH-BERT, and

ELECTRA, respectively. Notably, the accuracy was consistent even when the analysis was confined to reports containing multiple labels.

Conclusions: The multilabel classifier developed in our study demonstrated potential for identifying various factors associated with medication-related incidents using incident reports from residential care facilities. Thus, this classifier can facilitate prompt analysis of incident factors, thereby contributing to risk management and the development of preventive strategies.

(JMIR Med Inform 2024;12:e58141) doi:10.2196/58141

KEYWORDS

residential facilities; incidents; non-medical staff; natural language processing; risk management

Introduction

The prevention of medication-related incidents and the development of preventive measures are crucial for ensuring medication safety. Heinrich law suggests that for every serious accident, 29 minor accidents and 300 incidents exist [1]. Analysis of these incidents and formulation of countermeasures can help prevent serious medical accidents and enhance patient safety. Moreover, these incidents result in a significant psychological impact on the health care providers [2-5], known as “second victim syndrome” [6-8]. Thus, incident prevention measures are considered vital.

The core of incident prevention is focused on the details of the incident; thus, the creation of incident reports plays a key role. Hospitals have traditionally been the primary sources of such data, resulting in extensive research [9-19] and the development of sophisticated incident prevention strategies. However, residential care facility settings, in which residents live for extended periods, present unique challenges. Unlike hospitals, these facilities serve as communal living spaces for older people and often rely on nonmedical staff (not doctors or nurses) to perform health care–related tasks, including medication assistance. This practice raises significant concerns regarding the potential for medication incidents, underlining the importance of extending incident prevention strategies beyond hospital settings. Moreover, previous studies have highlighted a range of medication incidents in Japanese residential care facilities, including dropped drugs and misdelivery or misuse of medicines [20].

Natural language processing (NLP) technology demonstrates considerable potential for enhancing the analysis of incident reports. NLP is an analytical technique involving deep learning processing of human language for extracting meaningful information. Recently, this technology has been applied to classify various types of text data, including blogs [21,22] and electronic medical records [23], and has been extended to incident report classification in health care settings [24,25]. Specifically, classifiers using NLP were constructed to determine the classification and severity of incidents, with hospital incident reports serving as the training data [24,26,27]. Although the incident reports obtained and their corresponding text data, primarily comprising open-ended descriptions, from residential care facilities are considered suitable for NLP analysis, limited efforts have been made toward using NLP for extracting information from incident reports at these facilities.

Our previous research, which focused on identifying the causes of medication incidents in residential care facilities, demonstrated the complex and multifactorial nature of various elements contributing to these incidents [28]. Therefore, this study aimed to create a multilabel classifier that can extract various factors related to medication-related incidents based on incident reports in residential care facilities.

Methods

Data Set

This study included incidents that occurred in residential care facilities in Japan from April 1, 2015, to March 31, 2016, in 106 long-term residential care facilities operated by a single company. The residential care facilities included in our study are privately run, where residents usually pay monthly fees for housing and various care or support services, including meal provision, assistance with activities of daily living, and recreational opportunities. Notably, the majority of the residents avail medication assistance, which is a crucial component of these services.

We exclusively focused on incidents involving care staff who were not medical professionals. An incident report was completed after each incident, documenting the type of incident (forgetting to take medicines, misdelivery or misuse of medicines, loss of medicines, discovery of dropped drugs, and spitting up or falling while taking medicines), conditions at the time, and factors contributing to the incident. All the reports were written in Japanese. The care staff at the participating facilities were encouraged to record even minor incidents in their reports.

The data set comprised 2143 reports. The free-text descriptions of the factors contributing to the incident in each report were segmented into 7121 sentences for further analysis.

Annotation and Data Analysis

Incident factor labels were established based on the organizational factor model by Reason [29] and findings from our previous study [28], which explored the factors of medication assistance-related incidents in residential care facilities. In our previous study, we interviewed individuals involved in incidents, such as misdelivery or misuse of medications. Our findings indicated that “not following procedures” often resulted in these incidents, and identified 4 key contributing factors, namely individual residents, individual staff, team, and work environment. Considering the diverse nature of incidents in residential care facilities that extend

beyond medication omissions or dropped drugs, a broader range of factor labels is warranted. We developed the following 9 causal labels: procedure adherence, medicine, resident, resident family, nonmedical staff, medical staff, team, environment, and organizational management. To establish the labels, we also consulted James Reason [29] organizational accident model. Reason model posits that incidents, which are often precipitated by unsafe acts attributed to multiple factors, are fundamentally rooted in the culture of the organization. This model provides a framework for understanding the complex interplay of the factors resulting in incidents in our study.

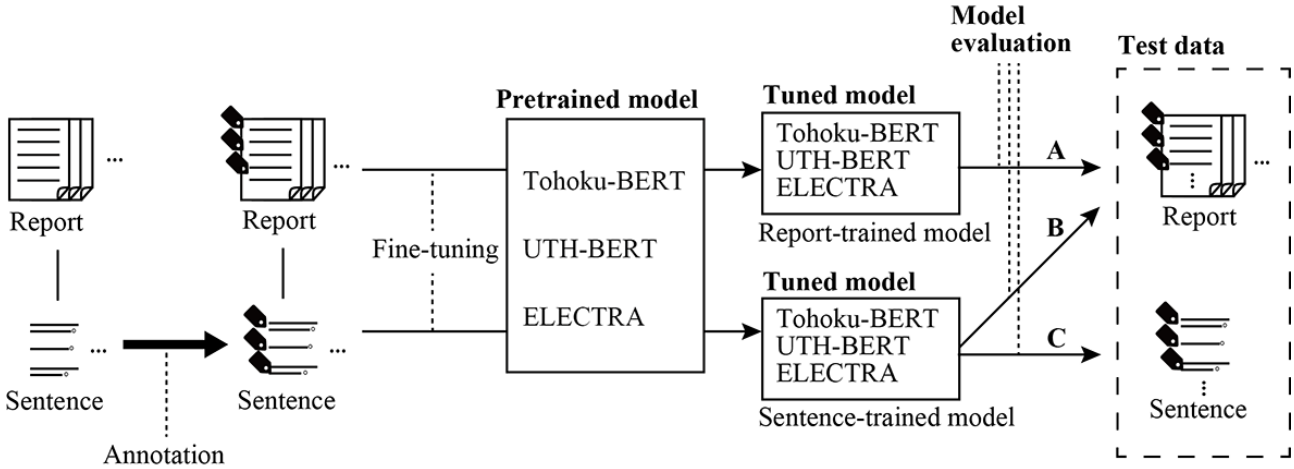
The criteria for annotating the reports were based on our previous study [28] and an analysis of actual medication incident conditions in Japan [20]. To evaluate the reliability of these criteria, we selected a random sample of 50 reports comprising 183 sentences from the data set. These reports were annotated by 2 researchers with relevant medical knowledge (HK and SE). The interannotator agreement (IAA) was assessed using Cohen κ, a statistical measure of agreement. Cohen K values are interpreted as follows: values close to 1 indicate perfect agreement, <0.00, "poor"; 0.00-0.20, "slight"; 0.21-0.40, "fair"; 0.41-0.60, "moderate"; 0.61-0.80, "substantial"; and 0.81-1.00, "almost perfect" [30]. Following this initial assessment, one of the researchers (HK) annotated all the sentences.

The distribution of reports according to the number of labels assigned was analyzed. The average number of labels per incident type was calculated and compared. Incident types were categorized as "forgetting to take medicines," "misdelivery or misuse of medicines," "loss of medicines," "discovery of dropped drugs," and "spitting up or falling while taking medicines," as defined in our previous study [20]. For factor analysis, we used the Student *t* test (2-tailed), applying the Bonferroni correction to account for multiple comparisons. We established the significance criterion at $P<.05$. Furthermore, the percentage of reports that contained each label for every type of incident was calculated, thus providing a detailed view of the label distribution across different incident categories.

Deep Learning Models

Figure 1 shows an overview of the model development. In this study, a multilabel classifier was built from an annotated multilabel data set to manage multiple descriptions of incident factors. Due to the insufficient number of labels, which made the accurate evaluation of the model's performance challenging, the limited labels associated with "resident family" and "medical staff" were excluded from the model development process. Consequently, the refined model development approach enabled the classifier to simultaneously identify 7 distinct labels. The development of this classifier involves fine-tuning existing pretrained models. These models included 2 Bidirectional Encoder Representations From Transformers (BERT) models, each using different pretraining data sources, and an Efficiently Learning Encoder That Classifies Token Replacements Accurately (ELECTRA) model. ELECTRA, while maintaining a foundational structure similar to that of BERT, achieves enhanced performance in NLP tasks through improved pretraining methods. The input of these models was limited to 512 tokens due to the capability of the pretrained model.

Figure 1. Overview of model development and evaluation. (A) Report-trained model evaluated using reports. (B) Sentence-trained model evaluated using reports. (C) Sentence-trained model evaluated using sentences. BERT: Bidirectional Encoder Representations From Transformers; ELECTRA: Encoder That Classifies Token Replacements Accurately; UTH: University of Tokyo Hospital.



Specifically, one of the BERT models we used was developed by the Natural Language Processing Research Group at Tohoku University (Tohoku-BERT), which was pretrained on the Japanese Wikipedia data as of September 1, 2019 [31] (BERT-based model; 12 layers, 768 dimensions of hidden states, and 12 attention heads, tokenizer: MeCab [32]). The other, University of Tokyo Hospital (UTH)-BERT, was developed by the Department of Artificial Intelligence and Digital Twin in Healthcare at the University of Tokyo and pretrained using extensive Japanese clinical text [33] (BERT-based model: 12 layers, 768 dimensions of hidden states, and 12 attention heads, tokenizer: MeCab [32]). The ELECTRA model was developed by the Izumi Laboratory at the University of Tokyo and pretrained with Japanese Wikipedia data as of June 1, 2021 [34] (ELECTRA-based model; 12 layers, 768 dimensions of hidden states, and 12 attention heads, tokenizer: MeCab [32]). No additional preprocessing was conducted beyond what is described.

Our study used 2 distinct models: one based on reports and the other on sentences. In the report-trained model, each report

served as a single training unit, whereas in the sentence-trained model, each sentence served as a training unit.

The hyperparameters that can be adjusted before training are defined in Table S1 in Multimedia Appendix 1.

Task and Metrics

Performance was evaluated in terms of precision, recall, F_1-score, and exact match accuracy. The exact match accuracy specifically measures the percentage of predictions that are correct across all labels. The data set was divided into training and test data at a ratio of 4:1, and the model was evaluated using the average of the 5-fold cross-validation results.

The report-trained model was evaluated using the reports as test data (Figure 1A). The sentence-trained model was evaluated using reports (Figure 1B) and sentences (Figure 1C) as test data.

Generalizability Analysis

We extracted 136 incident reports involving nonmedical staff and 31 reports involving care staff from hospital incident data collected by the Japan Council for Quality Health Care between January 2010 and June 2023 to examine the generalizability of the constructed model. We assessed the ability for extrapolation of the report-trained model derived from the extractor pretrained on Tohoku-BERT using the F_1-score.

Ethical Considerations

All the procedures were performed per the principles of the Declaration of Helsinki. In this study, all data were analyzed anonymously, and informed consent was waived owing to the retrospective observational design of this study. Residents and staff in residential facilities were informed of this study through postings at each facility and were allowed to refuse permission concerning the use of their data. This study was approved by the Research Ethics Review Committee of the Faculty of Pharmaceutical Sciences, University of Tokyo (approved on August 3, 2023) and the Research Ethics Review Committee of the Faculty of Pharmacy, Keio University (approved on July 14, 2023; 230714-1).

Results

Data Set Analysis

The average report length was 62.6 (SD 34.3; median 56, IQR 38-81) tokens, and the average sentence length was 18.2 (SD 9.8; median 16, IQR 11-23) tokens. None of the sentences or reports exceeded 512 tokens. The incidents were categorized as follows: forgetting to take medicines (648 incidents), misdelivery or misuse of medicines (293 incidents), loss of medication (18 incidents), discovery of dropped drugs (1024 incidents), and spitting up or falling while taking medicines (160 incidents).

Annotation and Features of the Incident Factors

An example of this label is presented in Textbox 1. The IAA for each label was calculated, and all the labels achieved an IAA score exceeding 0.6 (Table 1), thereby validating the effectiveness of the developed annotation guidelines. Notably, the κ coefficients for the factors related to “resident family” and “team” were exceptionally high, exceeding 0.9 in all instances. Using these guidelines, the remaining reports were sequentially annotated. The “resident” related factor label was most frequently assigned as shown in Table 1. Conversely, the “resident family” and “medical staff” factors were rarely assigned.

Table 2 presents the distribution of the number of labels assigned per sentence and per report. The most frequent occurrences were 1 label per sentence and 2 labels per report, accounting for 77.5% (5518) sentences and 34.2% (733) reports of the total occurrences, respectively. Table 3 categorizes the number of labels per report according to the incident type. Reports involving forgetting to take medicines and misdelivery or misuse of medicines tended to have a higher number of labels than other incidents (P<.001).

Table 4 shows the percentage of incident reports, with each label categorized by the incident type. Reports describing incidents, such as “spitting up or falling while taking medicines” and “discovery of dropped drugs,” often included “resident” factors and less frequently mentioned “team” factors. In contrast, reports of “forgetting to take medicines” and “misdelivery or misuse of medicines” commonly included “environmental” factors.

Textbox 1. Example of the label.

Procedure adherence

- Care staff did not follow the instructions for double-checking medication assistance.
- Care staff failed to confirm that medications were swallowed until the end.

Medicine

- Due to the concurrent use of herbal medicine with pills, there was a higher risk of dropping them.
- The number of medications to be taken after breakfast was large.

Resident

- The resident was unable to manage their medications.
- Their life rhythm was irregular, with mealtimes being inconsistent.

Resident family

- Family members were assisting with meals, which prevented intervention in medication administration.
- Medication management was being handled by the family.

Nonmedical staff

- Preparation for breakfast was not sufficient, leading to delays in service time and causing staff to rush.
- There was a low awareness that numbness made it difficult for residents to hold medication packets.

Medical staff

- Inexperienced nurses relied on each other, resulting in a lack of necessary checks.
- The nursing notes failed to include the required documentation.

Team

- There was a lack of coordination between meal assistance and medication assistance staff.
- Important information from doctor visits was not properly communicated.

Environment

- The resident was taking medicines during the busiest time for medication assistance.
- The proximity of tables in the restaurant made it impossible to check medication intake.

Organizational management

- Measures against the previous incidents had not been implemented.
- The procedure manual had not been updated.

Table 1. The IAA[a] values and number of sentences for the 9 labels (N=7121).

Label	IAA	Sentences, n
Procedure adherence	0.634	1167
Medicine	0.741	694
Resident	0.898	2455
Resident family	0.954	23
Nonmedical staff	0.638	1905
Medical staff	0.869	46
Team	0.930	195
Environment	0.755	1104
Organizational management	0.692	195

[a]IAA: interannotator agreement.

Table 2. The number of labels per incident report (N=2143) and per sentence (N=7121).

Number of labels	Reports, n (%)	Sentences, n (%)
0	10 (0.5)	508 (7.1)
1	278 (13.0)	5518 (77.5)
2	733 (34.2)	1021 (14.3)
3	689 (32.2)	70 (1.0)
4	348 (16.2)	3 (0.04)
5	78 (3.6)	0 (0)
6	7 (0.3)	0 (0)
≥7	0 (0)	0 (0)

Table 3. The average number of labels per incident contents.

	Labels, mean (SD)
Forgetting to take medicines (n=648)	2.92 (1.10)
Misdelivery or misuse of medicines (n=293)	2.85 (1.12)
Loss of medicines (n=18)	2.61 (1.38)
Discovery of dropped drugs (n=1024)	2.45 (0.94)
Spitting up or falling while taking medicines (n=160)	2.19 (0.93)

Table 4. The percentages of reports that contain each label for every type of incident.

	Procedure adherence	Medicine	Resident	Resident family	Nonmedical staff	Medical staff	Team	Environ-ment	Organizational management
Forgetting to take medicines (n=648)	41.0	33.0	59.0	2.3	72.8	2.8	17.0	53.4	11.0
Misdelivery or misuse of medicines (n=293)	39.6	28.7	59.7	0.7	73.0	4.8	14.0	53.6	11.3
Loss of medicines (n=18)	27.8	33.3	61.1	0	61.1	0	11.1	38.9	27.8
Discovery of dropped drugs (n=1024)	57.0	20.3	84.3	0.3	50.2	0	0.8	25.9	6.2
Spitting up or falling while taking medicines (n=160)	12.5	22.5	81.3	0	73.1	0.6	1.9	24.4	2.5

Model

Table S2 in Multimedia Appendix 1 shows the average label distributions for both the training and test data as part of the 5-fold cross-validation process. Since the labels for "resident family" and "medical staff" are notably fewer than those of other categories, they were excluded from the development of the multilabel classifier.

The performances of the fine-tuned Tohoku-BERT, UTH-BERT, and ELECTRA models were assessed using 3 different approaches: a report-trained model evaluated using reports (Table 5), sentence-trained model evaluated using reports (Table 6), and sentence-trained model evaluated using sentences (Table 7), as summarized in Tables 5-7. The analysis revealed that the report-trained model (Table 5) generally achieved higher F_1-scores than the sentence-trained model evaluated using the report data (Table 6). The performance of the sentence-trained model was better when evaluated using sentences (Table 7) than when evaluated using reports (Table 6).

Table 8 lists the exact match accuracies of these models across the board, specifically for instances involving multiple labels. The sentence-trained model evaluated using sentences exhibited the highest exact match accuracy for the overall test data. When limited to the test data with multiple labels, the report-trained model evaluated using reports demonstrated the highest exact match accuracy.

The extrapolation of the report-trained model, fine-tuned using Tohoku-BERT, revealed that the mean F_1-score (micro F_1-score) was 0.72 for reports involving care staff alone and 0.65 for those involving nonmedical staff (Table S1 in Multimedia Appendix 2).

Table 5. Performance of the model. Report-trained model evaluated using reports.

Class	Tohoku-BERT[a]			UTH[b]-BERT			ELECTRA[c]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Procedure adherence	0.808	0.834	0.820	0.774	0.856	0.816	0.818	0.804	0.804
Medicine	0.734	0.804	0.766	0.734	0.690	0.710	0.718	0.764	0.734
Resident	0.902	0.964	0.932	0.932	0.950	0.942	0.920	0.944	0.930
Nonmedical staff	0.802	0.852	0.820	0.778	0.842	0.806	0.822	0.852	0.832
Team	0.834	0.598	0.674	0.650	0.378	0.446	0.836	0.568	0.662
Environment	0.782	0.840	0.808	0.760	0.768	0.764	0.764	0.836	0.792
Organizational management	0.598	0.322	0.390	0.334	0.218	0.244	0.432	0.402	0.392
Macro F_1-score	—[d]	—	0.744	—	—	0.675	—	—	0.735

[a]BERT: Bidirectional Encoder Representations From Transformers.

[b]UTH: University of Tokyo Hospital.

[c]ELECTRA: Encoder That Classifies Token Replacements Accurately.

[d]Not applicable.

Table 6. Performance of the model. Sentence-trained model evaluated using reports.

Class	Tohoku-BERT[a]			UTH[b]-BERT			ELECTRA[c]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Procedure adherence	0.910	0.376	0.528	0.922	0.382	0.540	0.940	0.444	0.602
Medicine	0.926	0.316	0.466	0.854	0.664	0.746	0.942	0.176	0.296
Resident	0.992	0.444	0.614	0.982	0.726	0.834	0.988	0.494	0.656
Nonmedical staff	0.918	0.686	0.784	0.952	0.382	0.544	0.962	0.468	0.630
Team	1.000	0.274	0.424	0.850	0.148	0.242	0.800	0.084	0.146
Environment	0.976	0.468	0.636	0.964	0.366	0.532	0.930	0.466	0.620
Organizational management	1.000	0.064	0.116	0.600	0.018	0.034	0.884	0.068	0.126
Macro F_1-score	—[d]	—	0.610	—	—	0.496	—	—	0.439

[a]BERT: Bidirectional Encoder Representations From Transformers.

[b]UTH: University of Tokyo Hospital.

[c]ELECTRA: Encoder That Classifies Token Replacements Accurately.

[d]Not applicable.

Table 7. Performance of the model. Sentence-trained model evaluated using sentences.

Class	Tohoku-BERT[a]			UTH[b]-BERT			ELECTRA[c]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
Procedure adherence	0.798	0.798	0.796	0.802	0.750	0.768	0.754	0.844	0.796
Medicine	0.712	0.570	0.628	0.616	0.564	0.584	0.698	0.594	0.640
Resident	0.864	0.862	0.862	0.872	0.808	0.836	0.862	0.844	0.850
Nonmedical staff	0.744	0.686	0.714	0.692	0.670	0.678	0.774	0.674	0.716
Team	0.786	0.596	0.674	0.702	0.534	0.598	0.674	0.592	0.598
Environment	0.756	0.720	0.732	0.688	0.640	0.662	0.754	0.678	0.710
Organizational management	0.574	0.398	0.454	0.482	0.220	0.288	0.522	0.338	0.406
Macro F_1-score	—[d]	—	0.694	—	—	0.631	—	—	0.674

[a]BERT: Bidirectional Encoder Representations From Transformers.

[b]UTH: University of Tokyo Hospital.

[c]ELECTRA: Encoder That Classifies Token Replacements Accurately.

[d]Not applicable.

Table 8. Exact match accuracy.

	Exact match accuracy	Exact match accuracy in test data with only multiple labels
Tohoku-BERT[a]		
Report-trained model evaluated using reports	0.411	0.408
Sentence-trained model evaluated using reports	0.217	0.113
Sentence-trained model evaluated using sentences	0.656	0.318
UTH[b]-BERT		
Report-trained model evaluated using reports	0.389	0.378
Sentence-trained model evaluated using reports	0.202	0.113
Sentence-trained model evaluated using sentences	0.605	0.280
ELECTRA[c]		
Report-trained model evaluated using reports	0.399	0.394
Sentence-trained model evaluated using reports	0.198	0.095
Sentence-trained model evaluated using sentences	0.646	0.303

[a]BERT: Bidirectional Encoder Representations From Transformers.

[b]UTH: University of Tokyo Hospital.

[c]ELECTRA: Encoder That Classifies Token Replacements Accurately.

Discussion

Principal Results

Our study constructed a multilabel classifier that used NLP models, including BERT and ELECTRA, to identify the factors contributing to medication-related incidents by nonmedical staff using incident reports of residential care facilities. Unlike previous studies that mainly focused on classifying incident types and harm severity in hospital settings [24-27], our approach focused on the complex factors involved in medication-related incidents involving nonmedical staff. This complexity often renders accurate classification challenging. To our knowledge, our study is the first to demonstrate the potential contribution of NLP technology in extracting incident factors and formulating measures from incident reports obtained from residential care facilities.

In our study, we identified and annotated 9-factor labels, leading to over 99% (2133) of the reports and more than 92% (6613) of the sentences being assigned these labels. Reports without factor labels merely described the conditions of the incident occurrences, lacking in-depth factor analysis. Hence, the 9-factor labels identified appear to be suitable for representing the contributory factors in medication-related incidents involving nonmedical staff in residential care facilities. In contrast, 2 specific labels, “resident family” and “medical staff,” were relatively limited. In residential care facilities in Japan, where

nonmedical staff primarily provide medication assistance, the involvement of medical staff is limited, and family member participation is irregular. Consequently, these factors are less frequently represented, resulting in a small number of labels.

This study developed 2 types of models: 1 trained on individual sentences and the other on the entire reports. The report-trained model consistently outperformed the sentence-trained model, particularly achieving over 0.1 improvement in the F_1-score for factors involving nonmedical staff. Thus, report-level training, which retains more contextual information than sentence-level training, significantly enhanced the model performance. The exact match accuracy was the highest using the sentence-trained model, exceeding 0.6. However, this accuracy dramatically decreased to approximately 0.3 when the test data were limited to those with multiple labels. This significant reduction underscores the prevalence of sentences with a single label, deeming it unsuitable for evaluation as a multilabel classifier. Therefore, we also evaluated the performance of the sentence-trained model using report units as test data; however, the performance was significantly inferior to that of the report-trained model. Conversely, training on the report data yielded an exact match accuracy of approximately 0.4, which remained stable across tests with multiple labels. These findings demonstrate the successful development of a multilabel classifier that can rather accurately classify multiple labels; nevertheless, the potential for further improvement exists. Further, 1 approach for model improvement involves analyzing label co-occurrences that are prone to errors and creating a data set of these label combinations from an existing data set for upsampling. Additionally, modifications to the model, such as incorporating other pretrained models or applying domain adaptation techniques, could also be effective methods for improving performance.

Within our report-trained multilabel classifier, the macro F_1-scores of Tohoku-BERT and ELECTRA were notably similar, outperforming those of UTH-BERT, which had a lower score. This variation in performance was likely attributed to the characteristics of the pretraining data. Tohoku-BERT and ELECTRA were pretrained using the Japanese Wikipedia data, offering a broad range of general knowledge, whereas UTH-BERT was specifically pretrained on clinical texts. The lower classification performance for UTH-BERT may be attributed to the less specialized terminology included in the incident reports predominantly completed by nonmedical staff in residential care facilities compared to that in clinical texts.

Our analysis of the report-trained model's performance across various labels revealed that, with the notable exceptions of organizational management and team factors, as assessed by UTH-BERT, the F_1-scores consistently exceeded 0.6. Thus, the model accurately classified a broad spectrum of labels, thus demonstrating its effectiveness in automatically identifying incident factors from medication-related reports in residential care settings. However, classifying organizational management has proven to be more challenging. This difficulty can be attributed to the variability in the label assignment and the relatively limited number of labels in this category. Notably, the κ coefficient for the organizational management label was lower than that for the other labels, and the number of labels assigned to this category was also smaller. We assume that the low κ coefficient is partly attributed to the broader range of factors covered by this label, contributing to greater ambiguity compared to other labels. This highlights potential areas for the enhancement of the design and training process of our classifier.

Evaluation of the extrapolation of the constructed report-trained model confirmed that the F_1-score was slightly inferior to that of its initial construction. This reduction was primarily due to the notably few reports used for extrapolation evaluation. Furthermore, the characteristics of the individual who completed the report could have influenced their performance, particularly since the report was from a hospital setting. Moreover, the extrapolation results showed that the model's performance on reports involving care staff alone (F_1-score=0.72) was higher than that on those involving nonmedical staff (F_1-score=0.65). These findings indicate that the model is particularly effective in identifying factors in medication-related incidents involving care staff, suggesting specialization in extracting relevant information from such reports.

Limitations

In total, 1 limitation of this study is the inclusion of data with a limited number of labels. Although 9 labels were assigned at the annotation stage, 2 specific labels, "resident family" and "medical staff," were excluded from the multilabel classifier due to insufficient quantity. When a multilabel classifier that included these 2 labels was constructed, the F_1-score for both labels was nearly zero. The F_1-scores for the other labels remained almost unchanged compared to the case where the multilabel classifier was developed without including these 2 labels. Therefore, the impact of excluding these 2 labels is considered to be minimal. Furthermore, the performance of the model for each label tended to show higher F_1-scores with a higher IAA and a greater number of labels. This problem can be resolved by increasing the number of incident reports and labels.

Future Directions

Our model has the potential to streamline the identification of factors underlying medication-related incidents in residential care settings. This could result in a more effective planning of measures to prevent medication-related incidents. Moreover, it can offer nonmedical staff opportunities for learning and growth through prompt feedback following the occurrence of medication-related incidents.

Conclusions

The multilabel classifier developed in this study can identify various factors associated with medication-related incidents based on incident reports from residential care facilities. This classifier can facilitate prompt analysis of incident factors, thereby contributing to risk management and the development of preventive strategies.

Acknowledgments

We would like to thank all residents and care staff in the residential facilities operated by SOMPO Care and Mr Daisuke Yamamoto, BSc, of the SOMPO Care Corporation for his efforts in providing the data. This study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (JP22K19657).

Data Availability

Permission to provide data to individuals other than this study's investigators was not included under ethical approval and would require a new application for approval. The anonymized analyzed data may be obtained from the corresponding author upon reasonable request.

Authors' Contributions

HK, HS, and SH designed this study. HK, HS, and SE conducted the annotation and created the annotation guidelines. HK performed the data analysis, created the NLP model, and conducted all the experiments. SW supported the development of the NLP model from an NLP technical perspective. SI and YS advised on this study's concepts and processes. HS and SH supervised this study. All the authors have reviewed and approved the final paper. Authors HK (hayatokizaki625@keio.jp) and HS (satoh@mol.f.u-tokyo.ac.jp) are co-corresponding authors for this article.

Conflicts of Interest

HK, SE, SW, SI, and SH declare no conflicts of interest. YS and HS are researchers in a laboratory that received grants from SOMPO Care, Inc.

Multimedia Appendix 1
Hyperparameters of each model and label distribution of training and test data.
[DOCX File , 40 KB - medinform_v12i1e58141_app1.docx]

Multimedia Appendix 2
Extrapolation results of the fine-tuned Tohoku-BERT (Bidirectional Encoder Representations From Transformers).
[DOCX File , 39 KB - medinform_v12i1e58141_app2.docx]

References

1. Heinrich HW. Industrial Accident Prevention: A Scientific Approach. US: McGraw-Hill book Company; 1931.
2. Khammissa RA, Nemutandani S, Shangase SL, Feller G, Lemmer J, Feller L. The burnout construct with reference to healthcare providers: a narrative review. SAGE Open Med 2022;10:20503121221083080 [FREE Full text] [doi: 10.1177/20503121221083080] [Medline: 35646362]
3. Khatri N, Brown GD, Hicks LL. From a blame culture to a just culture in health care. Health Care Manage Rev 2009;34(4):312-322. [doi: 10.1097/HMR.0b013e3181a3b709] [Medline: 19858916]
4. Hamed MMM, Konstantinidis S. Barriers to incident reporting among nurses: a qualitative systematic review. West J Nurs Res 2022;44(5):506-523. [doi: 10.1177/0193945921999449] [Medline: 33729051]
5. Oweidat I, Al-Mugheed K, Alsenany SA, Abdelaliem SMF, Alzoubi MM. Awareness of reporting practices and barriers to incident reporting among nurses. BMC Nurs 2023;22(1):231 [FREE Full text] [doi: 10.1186/s12912-023-01376-9] [Medline: 37400810]
6. Huang H, Chen J, Xiao M, Cao S, Zhao Q. Experiences and responses of nursing students as second victims of patient safety incidents in a clinical setting: a mixed-methods study. J Nurs Manag 2020;28(6):1317-1325. [doi: 10.1111/jonm.13085] [Medline: 32654338]
7. Wu AW. Medical error: the second victim. The doctor who makes the mistake needs help too. BMJ 2000;320(7237):726-727 [FREE Full text] [doi: 10.1136/bmj.320.7237.726] [Medline: 10720336]
8. Naya K, Aikawa G, Ouchi A, Ikeda M, Fukushima A, Yamada S, et al. Second victim syndrome in intensive care unit healthcare workers: a systematic review and meta-analysis on types, prevalence, risk factors, and recovery time. PLoS One 2023;18(10):e0292108 [FREE Full text] [doi: 10.1371/journal.pone.0292108] [Medline: 37788270]
9. Härkänen M, Saano S, Vehviläinen-Julkunen K. Using incident reports to inform the prevention of medication administration errors. J Clin Nurs 2017;26(21-22):3486-3499. [doi: 10.1111/jocn.13713] [Medline: 28042673]
10. Aseeri M, Banasser G, Baduhduh O, Baksh S, Ghalibi N. Evaluation of medication error incident reports at a Tertiary Care Hospital. Pharmacy (Basel) 2020;8(2):69 [FREE Full text] [doi: 10.3390/pharmacy8020069] [Medline: 32325852]
11. Roberts HI, Kinlay M, Debono D, Burke R, Jones A, Baysari M. Nurses' medication administration workarounds when using electronic systems: an analysis of safety incident reports. Stud Health Technol Inform 2023;304:57-61. [doi: 10.3233/SHTI230369] [Medline: 37347569]

12. Cattell M, Hyde K, Bell B, Dawson T, Hills T, Iyen B, et al. Retrospective review of medication-related incidents at a major teaching hospital and the potential mitigation of these incidents with electronic prescribing and medicines administration. Eur J Hosp Pharm 2024;31(4):295-300. [doi: 10.1136/ejhpharm-2022-003515] [Medline: 36868849]
13. Thomas B, Pallivalapila A, El Kassem W, Al Hail M, Paudyal V, McLay J, et al. Investigating the incidence, nature, severity and potential causality of medication errors in hospital settings in Qatar. Int J Clin Pharm 2021;43(1):77-84 [FREE Full text] [doi: 10.1007/s11096-020-01108-y] [Medline: 32767219]
14. Härkänen M, Vehviläinen-Julkunen K, Franklin BD, Murrells T, Rafferty AM. Factors related to medication administration incidents in England and Wales between 2007 and 2016: a retrospective trend analysis. J Patient Saf 2021;17(8):e850-e857. [doi: 10.1097/PTS.0000000000000639] [Medline: 32168268]
15. Takahashi M, Okudera H, Wakasugi M, Sakamoto M, Shimizu H, Wakabayashi T, et al. Describing and quantifying wrong-patient medication errors through a study of incident reports. Drug Healthc Patient Saf 2022;14:135-146. [doi: 10.2147/DHPS.S371574] [Medline: 36039072]
16. Cottell M, Wätterbjörk I, Hälleberg Nyman M. Medication-related incidents at 19 hospitals: a retrospective register study using incident reports. Nurs Open 2020;7(5):1526-1535 [FREE Full text] [doi: 10.1002/nop2.534] [Medline: 32802373]
17. Alqenae FA, Steinke D, Carson-Stevens A, Keers RN. Analysis of the nature and contributory factors of medication safety incidents following hospital discharge using National Reporting and Learning System (NRLS) data from England and Wales: a multi-method study. Ther Adv Drug Saf 2023;14:20420986231154365 [FREE Full text] [doi: 10.1177/20420986231154365] [Medline: 36949766]
18. Howell A, Burns EM, Bouras G, Donaldson LJ, Athanasiou T, Darzi A. Can patient safety incident reports be used to compare hospital safety? Results from a quantitative analysis of the english national reporting and learning system data. PLoS One 2015;10(12):e0144107 [FREE Full text] [doi: 10.1371/journal.pone.0144107] [Medline: 26650823]
19. Scott J, Dawson P, Heavey E, De Brún A, Buttery A, Waring J, et al. Content analysis of patient safety incident reports for older adult patient transfers, handovers, and discharges: do they serve organizations, staff, or patients? J Patient Saf 2021;17(8):e1744-e1758. [doi: 10.1097/PTS.0000000000000654] [Medline: 31790011]
20. Kizaki H, Yamamoto D, Maki H, Masuko K, Konishi Y, Satoh H, et al. Medication incidents associated with the provision of medication assistance by non-medical care staff in residential care facilities. Drug Discov Ther 2024;18(1):54-59 [FREE Full text] [doi: 10.5582/ddt.2023.01073] [Medline: 38417897]
21. Nishioka S, Asano M, Yada S, Aramaki E, Yajima H, Yanagisawa Y, et al. Adverse event signal extraction from cancer patients' narratives focusing on impact on their daily-life activities. Sci Rep 2023;13(1):15516. [doi: 10.1038/s41598-023-42496-1] [Medline: 37726371]
22. Nishioka S, Watanabe T, Asano M, Yamamoto T, Kawakami K, Yada S, et al. Identification of hand-foot syndrome from cancer patients' blog posts: BERT-based deep-learning approach to detect potential adverse drug reaction symptoms. PLoS One 2022;17(5):e0267901 [FREE Full text] [doi: 10.1371/journal.pone.0267901] [Medline: 35507636]
23. Chaichulee S, Promchai C, Kaewkomon T, Kongkamol C, Ingviya T, Sangsupawanich P. Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. PLoS One 2022;17(8):e0270595 [FREE Full text] [doi: 10.1371/journal.pone.0270595] [Medline: 35925971]
24. Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. BMC Med Inform Decis Mak 2017;17(1):84 [FREE Full text] [doi: 10.1186/s12911-017-0483-8] [Medline: 28606174]
25. Mathew F, Wang H, Montgomery L, Kildea J. Natural language processing and machine learning to assist radiation oncology incident learning. J Appl Clin Med Phys 2021;22(11):172-184 [FREE Full text] [doi: 10.1002/acm2.13437] [Medline: 34610206]
26. Nguyen M, Beidler P, Lybarger K, Anderson A, Holmberg O, Kang J, et al. Automatic prediction of severity score of incident learning reports in radiation oncology using natural language processing. Int J Radiat Oncol, Biol, Phys 2022 Nov;114(3):S93-S94. [doi: 10.1016/j.ijrobp.2022.07.510]
27. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Int J Med Inform 2019;132:103971. [doi: 10.1016/j.ijmedinf.2019.103971] [Medline: 31630063]
28. Kizaki H, Yamamoto D, Satoh H, Masuko K, Maki H, Konishi Y, et al. Analysis of contributory factors to incidents related to medication assistance for residents taking medicines in residential care homes for the elderly: a qualitative interview survey with care home staff. BMC Geriatr 2022;22(1):352 [FREE Full text] [doi: 10.1186/s12877-022-03016-4] [Medline: 35459105]
29. Reason J. Managing the risks of organizational accidents. United Kingdom: ASHGATE; 1997.
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174. [Medline: 843571]
31. cl-tohoku/bert-base-japanese-whole-word-masking. Hugging Face. URL: https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking [accessed 2024-02-07]
32. MeCab: yet another part-of-speech and morphological analyzer. URL: https://taku910.github.io/mecab/ [accessed 2024-02-29]

33. UTH-BERT: a BERT pre-trained with Japanese clinical text. URL: https://ai-health.m.u-tokyo.ac.jp/home/research/uth-bert [accessed 2024-02-07]
34. izumi-lab/electra-base-japanese-discriminator. Hugging Face. URL: https://huggingface.co/izumi-lab/electra-base-japanese-discriminator [accessed 2024-02-07]

Abbreviations

IAA: interannotator agreement
BERT: Bidirectional Encoder Representations From Transformers
ELECTRA: Encoder That Classifies Token Replacements Accurately
NLP: natural language processing
UTH: University of Tokyo Hospital

Edited by C Lovis; submitted 07.03.24; peer-reviewed by S Matsuda, T Tachi; comments to author 05.04.24; revised version received 23.05.24; accepted 16.06.24; published 23.07.24.

Please cite as:
Kizaki H, Satoh H, Ebara S, Watabe S, Sawada Y, Imai S, Hori S
Construction of a Multi-Label Classifier for Extracting Multiple Incident Factors From Medication Incident Reports in Residential Care Facilities: Natural Language Processing Approach
JMIR Med Inform 2024;12:e58141
URL: https://medinform.jmir.org/2024/1/e58141
doi:10.2196/58141
PMID:

Predictive Models for Sustained, Uncontrolled Hypertension and Hypertensive Crisis Based on Electronic Health Record Data: Algorithm Development and Validation

Hieu Minh Nguyen[1], MS; William Anderson[2], MS; Shih-Hsiung Chou[3], PhD; Andrew McWilliams[4,5], MD, MPH; Jing Zhao[6], PhD; Nicholas Pajewski[1,7], PhD; Yhenneko Taylor[1,8], PhD

1
2
3
4
5
6
7
8

Corresponding Author:
Hieu Minh Nguyen, MS

Abstract

Background: Assessing disease progression among patients with uncontrolled hypertension is important for identifying opportunities for intervention.

Objective: We aim to develop and validate 2 models, one to predict sustained, uncontrolled hypertension (≥2 blood pressure [BP] readings ≥140/90 mm Hg or ≥1 BP reading ≥180/120 mm Hg) and one to predict hypertensive crisis (≥1 BP reading ≥180/120 mm Hg) within 1 year of an index visit (outpatient or ambulatory encounter in which an uncontrolled BP reading was recorded).

Methods: Data from 142,897 patients with uncontrolled hypertension within Atrium Health Greater Charlotte in 2018 were used. Electronic health record–based predictors were based on the 1-year period before a patient's index visit. The dataset was randomly split (80:20) into a training set and a validation set. In total, 4 machine learning frameworks were considered: L2-regularized logistic regression, multilayer perceptron, gradient boosting machines, and random forest. Model selection was performed with 10-fold cross-validation. The final models were assessed on discrimination (C-statistic), calibration (eg, integrated calibration index), and net benefit (with decision curve analysis). Additionally, internal-external cross-validation was performed at the county level to assess performance with new populations and summarized using random-effect meta-analyses.

Results: In internal validation, the C-statistic and integrated calibration index were 0.72 (95% CI 0.71 - 0.72) and 0.015 (95% CI 0.012 - 0.020) for the sustained, uncontrolled hypertension model, and 0.81 (95% CI 0.79 - 0.82) and 0.009 (95% CI 0.007 - 0.011) for the hypertensive crisis model. The models had higher net benefit than the default policies (ie, treat-all and treat-none) across different decision thresholds. In internal-external cross-validation, the pooled performance was consistent with internal validation results; in particular, the pooled C-statistics were 0.70 (95% CI 0.69 - 0.71) and 0.79 (95% CI 0.78 - 0.81) for the sustained, uncontrolled hypertension model and hypertensive crisis model, respectively.

Conclusions: An electronic health record–based model predicted hypertensive crisis reasonably well in internal and internal-external validations. The model can potentially be used to support population health surveillance and hypertension management. Further studies are needed to improve the ability to predict sustained, uncontrolled hypertension.

(JMIR Med Inform 2024;12:e58732) doi:10.2196/58732

KEYWORDS

machine learning; risk prediction; predictive model; decision support; blood pressure; cardiovascular; electronic health record

Introduction

Hypertension is a major chronic disease affecting nearly half of the adults in the United States, of whom less than half have their blood pressure (BP) under control [1]. Uncontrolled hypertension, defined as BP ≥140/90 mm Hg may lead to major cardiovascular diseases, organ damage, stroke, or even death, if not properly managed over time [2,3]. Effective treatment for uncontrolled hypertension requires proper monitoring so that further disease progression can be detected and prevented.

Numerous studies have examined the risk factors related to hypertension. Surveillance data show that hypertension is more prevalent in men than in women and in older adults than in younger persons [4]. Racial or ethnic disparities in BP control have been described, owing to risk factors that include racism-related stress, and social barriers such as low health literacy, poverty, and limited access to care [5-7]. Prior studies have also revealed other clinical predictors of hypertension such as comorbidities (eg, coronary heart disease and diabetes), laboratory biomarkers (eg, cholesterol levels), and BMI [8-10]. Leveraging the extensive knowledge base about hypertension risk factors, various prediction models, based on statistical and machine learning methods, have been developed to assess the risk of hypertension onset in the general population [11,12]. However, literature searches revealed a lack of research involving risk prediction for clinically important hypertension states, such as future BP measurements that are consistently elevated, also called sustained uncontrolled hypertension, or hypertensive crisis (ie, BP ≥180/120 mm Hg) in patients with uncontrolled hypertension [13,14].

Predictive models of sustained, uncontrolled hypertension and hypertensive crisis within 1-year following an index visit could inform clinical decision support prompting discussions between patients and clinicians about medication intensification. This index visit can be designated as an outpatient or ambulatory clinic appointment in which the patient had an uncontrolled BP reading and did not have a new antihypertensive medication class added. These specifications can identify a targeted population who may benefit from additional consideration to intensity hypertension medications. From a design perspective, the intended use case for the proposed risk models is to either serve as a real-time nudge that informs a shared decision-making conversation between provider and patient at the index visit or to be deployed as a tool for population health surveillance to help guide timely, proactive outreach. For instance, a care manager may reach out to a patient who is at high risk and did not have medication intensification to schedule a follow-up visit sooner, address barriers to medication adherence, or inquire with the care team about enrolling the patient in a hypertension management program. Furthermore, a valid risk score may aid patients in understanding the importance of treatment decisions; thereby helping to address nonadherence to medications, which is a major risk factor for suboptimal BP control [15]. This study presents the development and validation of 2 risk models, one to predict sustained, uncontrolled hypertension, and one to predict hypertensive crisis within 1 year following an index visit.

Methods

Study Population

This study's cohort consisted of patients aged 18 years or older who had an uncontrolled BP reading (systolic BP ≥140 mm Hg or diastolic BP ≥90 mm Hg) during an ambulatory or outpatient encounter in 2018 at a Greater Charlotte facility of Atrium Health, a large hospital network in the southeastern United States. This study's cohort only included active patients in the health system, that is, those having at least one encounter during the following year, 2019. The first ambulatory or outpatient encounter during 2018 showing an uncontrolled BP reading was considered the patient's index visit. Patients were excluded if, on the index visit, they were prescribed a new antihypertensive drug class that was not present in the 1-year historical medication records. Patients on more than 4 drug classes of antihypertensive medication were excluded. Patients were also excluded if they were in hospice care, were pregnant during 2018 or 2019, were diagnosed with end-stage renal disease, received dialysis, had a renal transplant, or died before December 31, 2019.

Study Variables

Sustained, uncontrolled hypertension in a patient was a binary outcome indicating the presence of ≥2 uncontrolled BP readings (≥140/90 mm Hg) or ≥1 particularly high BP reading (≥180/120 mm Hg) within 1 year following an index visit. The hypertensive crisis outcome was a binary indicator showing whether a patient had any BP reading ≥180/120 mm Hg within 1 year following the index visit.

We determined that using a 1-year look-back window prior to index visit would be appropriate to capture recent, clinically relevant predictor data from the health system's electronic health records (EHRs). This decision also relied on previous observations of models predicting 1-year hypertension status demonstrating successes when they were also using 1-year-old data prior to prediction [16,17]. We collected basic data at the index visit including the patient's age, gender, race and ethnicity, medical insurance, and the last systolic and diastolic BP measurements. Based on a patient's primary address, we determined census tract-level neighborhood socioeconomic disadvantage indicators: the Area Deprivation Index (ADI, national-level percentile), based on the 2016 - 2020 American Community Survey, and the Centers for Disease Control and Prevention's Social Vulnerability Index (version 2020; overall score, national-level percentile) [18,19]. Health care–related predictors included the presence of individual Elixhauser Comorbidities, antihypertensive drug classes prescribed, and number of visits within different clinical settings [20]. We considered several biological measurements including total cholesterol, high-density lipoprotein cholesterol, and low-density lipoprotein cholesterol, triglycerides, and creatinine, as well as weight and BMI. BP measurements in the past 1 year of index visit were aggregated into prediction features (eg, count of BP readings ≥140/90 mm Hg). A detailed description of all prediction features can be found in Table S3 in Multimedia Appendix 1.

Statistical Analyses

We calculated the required sample size of a validation dataset so that the 95% CI for validation C-statistics had a width of 0.05 or less [21]. A sample size of 11,520 patients is adequate for an outcome prevalence between 5% - 50% and a C-statistics between 0.6 - 0.8.

We randomly split the dataset into a training set and an internal validation set according to an 80:20 ratio. We performed median imputation for numerical variables and used imputed median values from the training set for subsequent imputation with the

validation set. To handle considerable missingness with laboratory tests, we categorized the variables using standard cutoff values for their normal ranges and applied the "missing" category to the variable when the value was not available. Once categorical variables were 1-hot encoded, we performed data standardization, subtracting the variables by the sample mean then dividing by the sample SD. No variable selection procedure was conducted. Further, 4 machine learning frameworks were considered: (L2) regularized logistic regression, multilayer perceptron (with 1 hidden layer), gradient boosting machines, and random forest. For hyperparameter tuning, we used grid search strategy with discrimination power (C-statistic) as the selection criteria and performed 10-fold cross-validation on the training set. Hyperparameter-tuned models were identified, one for each modeling framework, and further compared on their cross-validation performance for the final model.

For internal validation, in addition to discrimination, we assessed calibration performance of the 2 final models, one for each outcome, using smoothed calibration curves, estimated with generalized additive models. Based on the calibration curves, we computed the average (ie, integrated calibration index [ICI]), median (E50), 90th percentile (E90) of the absolute difference between expected event rate and predicted risk to summarize calibration errors [22]. We computed 95% CIs using standard methods for C-statistics (DeLong method) and calibration metrics (simulation-based inference) [23]. Additionally, we reported sensitivity, specificity, and positive or negative predictive values and 95% CIs (using exact binomial method) with respect to different decision thresholds. We performed decision curve analysis with the validation set to assess the models' net benefit in comparison with the default policies, that is, treating all and treating no patients [24]. We evaluated net benefit within a range of probability decision threshold, which was 50% - 70% for the outcome sustained, uncontrolled hypertension and ≤20% for the outcome hypertensive crisis.

Finally, we carried out internal-external cross-validation (IECV) at the county level to examine the final models' predictive ability in new patient cohorts, using the existing data [25]. With this approach, we used data from patients receiving care within a given county to validate models developed with the data of all other patients. For reliable validation, we only validated the counties yielding sufficient sample size of at least 200 events [26]. Using random-effect meta-analyses, we estimated the pooled performance measures, that is, the C-statistic, and ICI, across validations. We assessed heterogeneity across validations using the SD of the random effect, denoted as τ, and a chi-square test with a significance level of .05.

Model building was performed using Python (Python Software Foundation, eg, "scikit-learn" package). Other analyses were conducted using R (R Foundation, eg, "pROC," "pmcalibration," and "meta" packages).

Ethical Considerations

The Atrium Health institutional review board approved our research protocol. Informed consent was waived as there were no more than minimal risks to study participants. The study data will not be made available publicly to ensure patient confidentiality and privacy.

Results

Patient Characteristics

This study's cohort consisted of 142,897 patients, almost all of whom received care in 13 North Carolina counties and 1 South Carolina county in the year 2018. The patients were 72.33% (n=103,361) White and 22.52% (n=32,174) Black, had a median age of 61 (IQR 49-71) years and a median national-level ADI ranking of 55 (IQR 35-73). All patients were observed with known sustained, uncontrolled hypertension and hypertensive crisis status during follow-up period. The observed prevalence of sustained, uncontrolled hypertension and hypertensive crisis were 41.67% (n=59,547) and 4.53% (n=6,470), respectively. Across the counties, there were notable racial and socioeconomic differences (Table 1). Except for 1 county, the percentage of White patients ranged between 62.64% and 93.13% and the median of ADI ranking ranged between 39 and 88. The observed prevalence of outcomes among the counties ranged between 37.35% and 47.26% for sustained, uncontrolled hypertension, and, except for 1 county, between 3.89% and 6.06% for hypertensive crisis.

Table . Patient characteristics.

Location of care, county (sample size)	ADI[a], median (IQR)	Age, median (IQR)	Female, n (%)	White, n (%)	Black, n (%)	SUHTN[b], n (%)	HC[c], n (%)
Overall (n=142,897)	55 (35-73)	61 (49-71)	85,113 (59.56)	103,361 (72.33)	32,174 (22.52)	59,547 (41.67)	6470 (4.53)
Mecklenburg (n=64,811)	44 (25-65)	60 (49-71)	39,347 (60.71)	40,595 (62.64)	19,746 (30.47)	26,338 (40.64)	2704 (4.17)
Cabarrus (n=31,403)	57 (42-72)	61 (49-71)	18,421 (58.66)	25,175 (80.17)	4891 (15.57)	13,854 (44.12)	1567 (4.99)
Union (n=9882)	54 (43-74)	60 (49-71)	5569 (56.35)	7400 (74.88)	1873 (18.95)	3963 (40.10)	443 (4.48)
York (n=8963)	51 (39-68)	61 (49-72)	5319 (59.34)	6991 (78.00)	1652 (18.43)	3694 (41.21)	384 (4.28)
Cleveland (n=6866)	78 (71-85)	63 (52-73)	4486 (65.33)	5347 (77.88)	1410 (20.54)	2918 (42.50)	416 (6.06)
Lincoln (n=4230)	71 (63-77)	63 (51-72)	2362 (55.84)	3852 (91.06)	266 (6.29)	1999 (47.26)	226 (5.34)
Gaston (n=3756)	67 (50-80)	60 (49-70)	1941 (51.68)	3145 (83.73)	507 (13.50)	1403 (37.35)	146 (3.89)
Stanly (n=2478)	68 (58-78)	64 (51-74)	1524 (61.50)	2148 (86.68)	304 (12.27)	1116 (45.04)	135 (5.45)
Iredell (n=2411)	39 (21-57)	70 (59-77)	1159 (48.07)	2169 (89.96)	148 (6.14)	953 (39.53)	96 (3.98)
Rutherford (n=1922)	82 (66-88)	67 (57-75)	1163 (60.51)	1645 (85.59)	245 (12.75)	771 (40.11)	82 (4.27)
Burke (n=1462)	79 (72-83)	58 (45-69)	1113 (76.13)	1332 (91.11)	91 (6.22)	608 (41.59)	52 (3.57)
Caldwell (n=1324)	84 (72-90)	64 (52-73)	728 (54.98)	1233 (93.13)	72 (5.44)	521 (39.35)	51 (3.85)
Rowan (n=1121)	69 (63-78)	59 (49-69)	635 (56.65)	868 (77.43)	219 (19.54)	500 (44.60)	45 (4.01)
Anson (n=625)	88 (80-96)	58 (50-69)	362 (57.92)	186 (29.76)	425 (68.00)	272 (43.52)	53 (8.48)

[a]ADI: Area Deprivation Index.

[b]SUHTN: sustained, uncontrolled hypertension.

[c]HC: hypertensive crisis.

Final Models

For each prediction problem, the hyperparameter-tuned models from different modeling frameworks achieved practically equivalent 10-fold cross-validated C-statistics of around 0.71 - 0.72 for the outcome sustained, controlled hypertension and 0.79 - 0.80 for hypertensive crisis, respectively (Table 2). Given that the L2-regularized logistic regression (LR) was a simpler and more computationally efficient framework, we selected the hyperparameter-tuned LR models for training and final validations. Additionally, we examined the relative variable importance in a trained LR model via the magnitude of predictor coefficients and investigated each model's top 10 variables (Table S1 in Multimedia Appendix 1). Notably, increases in systolic BP at index visit, the number of encounters with systolic BP≥140 mm Hg and the number of encounters with BP≥140/90 mm Hg in the past 1 year, age, as well as a prior diagnosis of hypertension and higher ADI ranking, predicted both higher risk of sustained, uncontrolled hypertension and hypertensive crisis.

Table . Optimized hyperparameters and 10-fold cross-validated C-statistics of the hyperparameter-tuned models.

Framework and hyperparameter options[a]		Sustained, uncontrolled hypertension		Hypertensive crisis	
		Optimal value	Cross-validated C-stat (SE)	Optimal value	Cross-validated C-stat (SE)
L2 regularized logistic regression			0.713 (0.002)		0.793 (0.002)
	C[b]: 0.001, 0.01, …, 1000	0.001		0.001	
Gradient boosting			0.716 (0.001)		0.799 (0.002)
	n_estimators: 50, 100, 150	100		100	
	learning_rate: 0.05, 0.1, 0.2	0.2		0.2	
	max_depth: 3, 5, 8	5		3	
Multilayer perceptron			0.713 (0.002)		0.794 (0.002)
	hidden_layer_size: 5, 10, 20	5		5	
	learning_rate_init: 0.001, 0.01, 0.1	0.01		0.01	
	alpha[c]: 0.00001, 0.0001, 0.001	0.001		0.001	
Random forest			0.708 (0.002)		0.785 (0.003)
	n_estimators: 50, 100, 150	50		50	
	max_depth: 3, 5, 7	7		7	
	max_features[d]: “sqrt,” “log2,” none	“sqrt”		None	

[a]Hyperpameters in scikit-learn Python package.

[b]C: inverse of regularization strength.

[c]alpha: strength of L2 regularization.

[d]max_features: number of features (function of n_features) to consider when looking for best split.

Internal and Internal-External Validations

In the internal validation dataset, the LR models achieved acceptable discrimination power for predicting sustained, uncontrolled hypertension with a C-statistic of 0.72 (95% CI 0.71 - 0.72), and reasonably good discrimination power for predicting hypertensive crisis with a C-statistic of 0.81 (95% CI 0.79 - 0.82). Both sustained, uncontrolled hypertension and hypertensive crisis models had accurate risk estimates with an ICI of 0.015 (95% CI 0.012 - 0.020) and 0.009 (95% CI 0.007 - 0.011), respectively. Other metrics (E50 and E90) and calibration curves can be examined in Table 3 and Figure 1. Sensitivity, specificity, and predictive values of the final models across potential decision thresholds were also reported in Table S2 in Multimedia Appendix 1. Further, in decision curve analyses, the models demonstrated higher net benefit than treat-all and treat-none policies within the ranges of plausible decision thresholds (Figure 2).

From IECV of the sustained, uncontrolled hypertension model, the pooled estimates were 0.70 (95% CI 0.69 - 0.71) for C-statistic and 0.021 (95% CI 0.016 - 0.026) for ICI; additionally, there was a small or moderate heterogeneity in C-statistic (τ=0.02, 95% CI 0.01 - 0.03, P<.001) and a small heterogeneity in ICI (τ=0.01, 95% CI 0.01 - 0.01, P<.001). From IECV of the hypertensive crisis model, the pooled estimates were 0.79 (95% CI 0.78 - 0.81) for C-statistic and 0.007 (95% CI 0.005 - 0.009) for ICI; across validations, variation in C-statistic was small or moderate (τ=0.01, 95% CI 0.00 - 0.04, P=.004) and variation in ICI was small (τ=0.00, 95% CI 0.00 - 0.01, P=.009).

Table . Discrimination and calibration performance (with 95% CIs) of the final models on training and internal validation datasets.

Model and dataset		Sample size (missing[a]) n	C-stat	ICI[b]	E50	E90
Sustained, uncontrolled hypertension						
	Training	114,317 (7730)	0.71	0.015	0.015	0.030
	Internal validation	28,580 (1959)	0.72 (0.71-0.72)[c]	0.015 (0.011 - 0.020)[c]	0.009 (0.005 - 0.017)[c]	0.040 (0.030 - 0.047)[c]
Hypertensive crisis						
	Training	114,317 (7730)	0.80	0.006	0.005	0.013
	Internal validation	28,580 (1959)	0.81 (0.79-0.82)[c]	0.009 (0.007 - 0.011)[c]	0.006 (0.004 - 0.007)[c]	0.021 (0.014 - 0.028)[c]

[a]Missing: number of observations containing any missing feature.

[b]ICI: integrated calibration index.

[c]95% CI.

Figure 1. Smoothed calibration plots (with 95% CIs; top) and histograms showing distribution of the predicted probability (bottom) for the sustained, uncontrolled hypertension model (left) and the hypertensive crisis model (right) from internal validation.

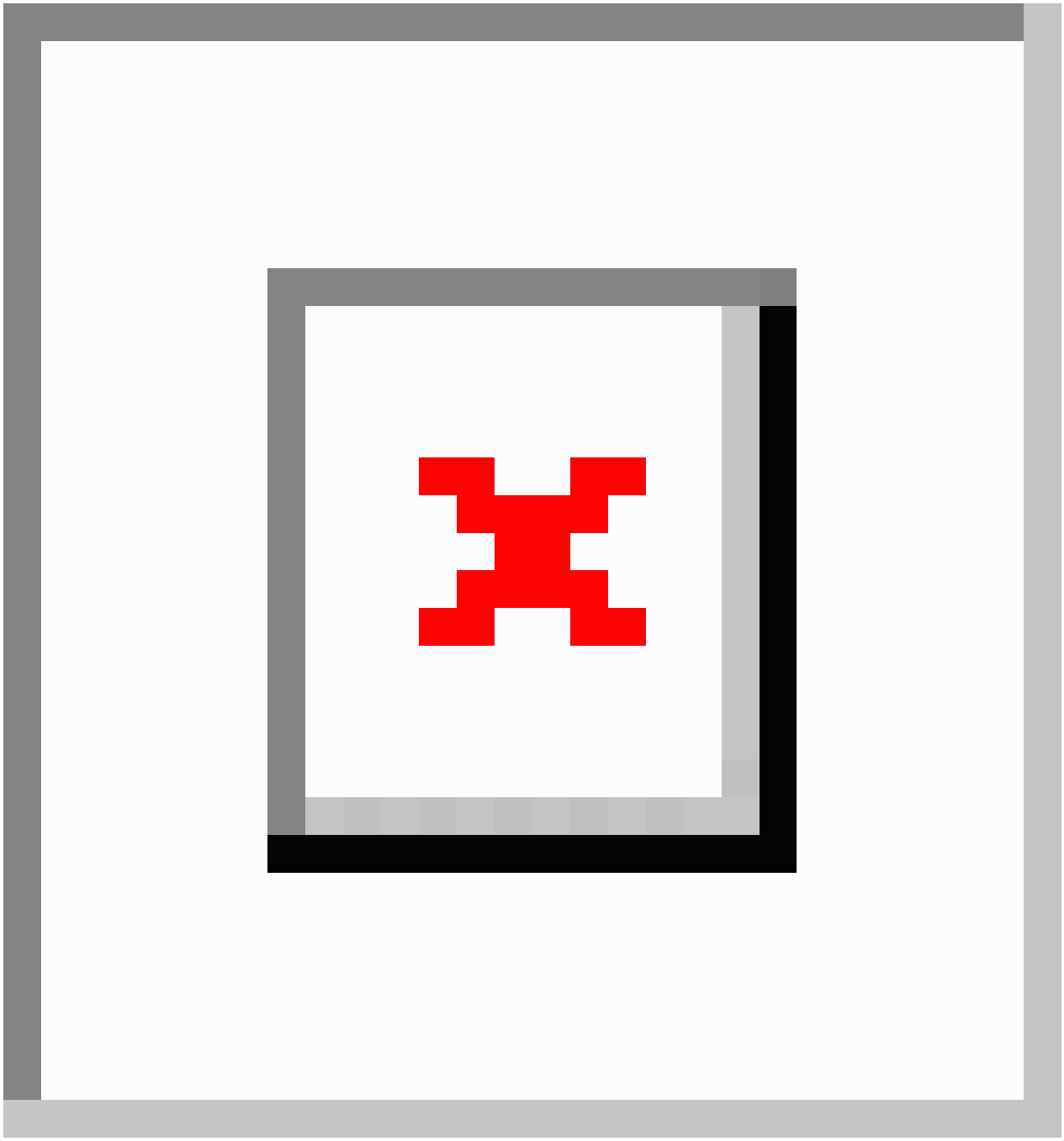
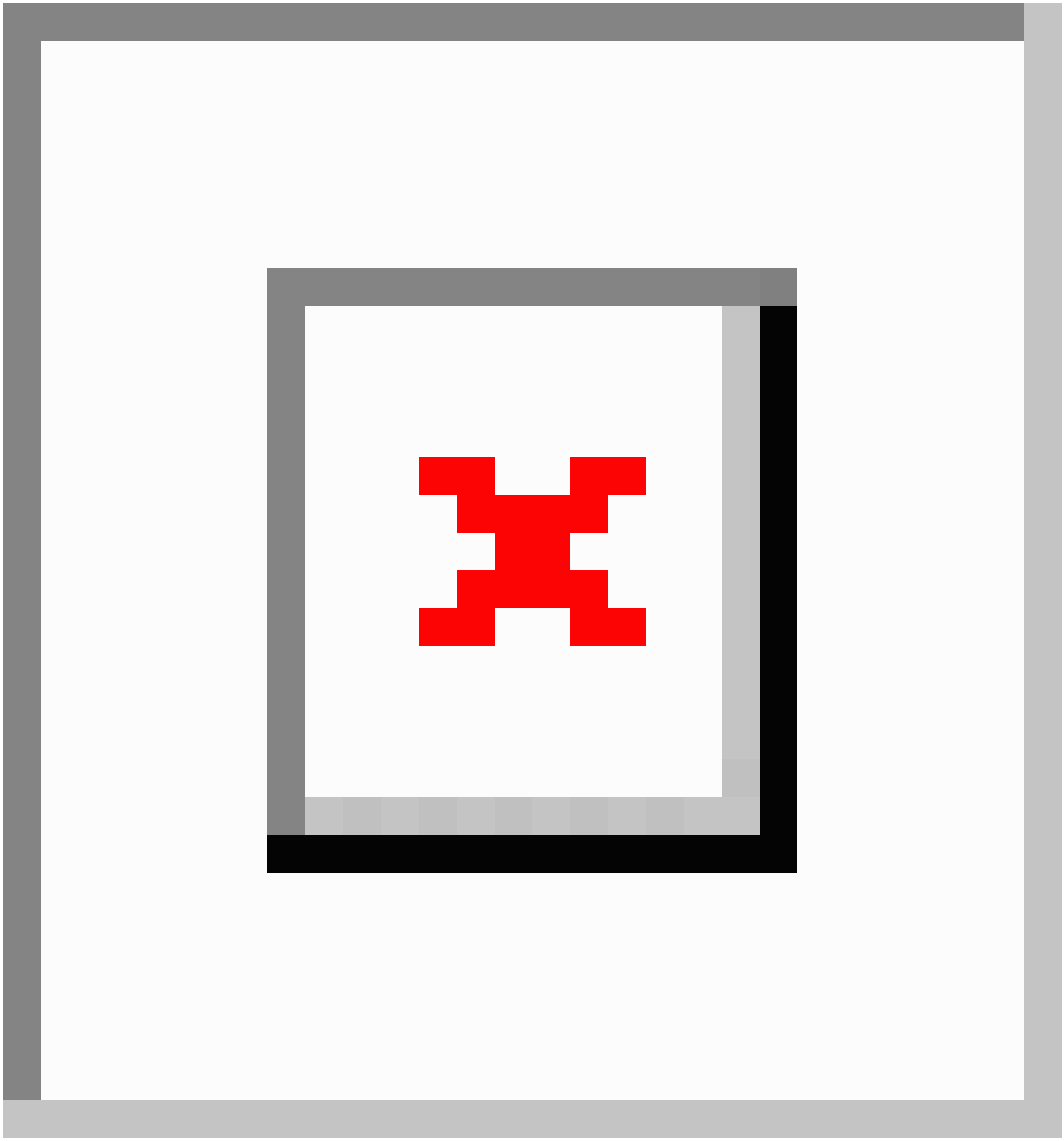


Figure 2. Decision curve analyses of the final models predicting sustained, uncontrolled hypertension (left) and hypertensive crisis (right) from internal validation.



Discussion

Principal Results

Using a large, diverse (in terms of geography, race, and socioeconomic status) patient cohort in the Greater Charlotte area of the United States, we developed and validated risk models to predict sustained, uncontrolled hypertension and hypertensive crisis occurring within 1 year following an index ambulatory or outpatient encounter in which an uncontrolled BP reading was documented. Internal validation showed discrimination performance that was reasonably good for the hypertensive crisis model and lower but acceptable for the sustained, uncontrolled hypertension model. Overall, both models predicted the risks accurately (calibration performance) in the internal validation dataset and further demonstrated clinical utility through decision curve analyses. Using the existing data, IECV assessed the models' predictive ability with new patient cohorts and showed that the models' overall discrimination and calibration performance across validations were consistent with internal validation results. In addition, across validations, we observed small to moderate variation in discrimination performance and a small variation in calibration performance.

Our findings show that the risk of hypertensive crisis in patients with uncontrolled hypertension can be predicted well. The hypertensive crisis model, in particular, showed satisfactory

performance to serve patients within Atrium Health's North and South Carolina markets, and potentially patients from other nearby areas. Based on the model's internal and internal-external validation performance, it also has the potential to be applicable to other health systems. For our final models to be further validated and used elsewhere, the complete specifications of the models can be found in Table S3 in Multimedia Appendix 1.

Therapeutic inertia (TI), that is, the failure of providers to initiate or intensify medication therapy when patients fail to achieve their treatment goals, is a well-known barrier to better clinical outcomes [27]. Among patients with hypertension, TI contributes to worse short- and long-term BP control [28,29]. While the causes of TI are multifactorial, interventions that include provider and patient education and leverage health care data to guide clinical decision-making at the point of care are promising approaches for reducing TI. Within this context, our prediction models can provide useful data to facilitate clinician-patient discussions on the potential need to intensify medications. Our expectation is that the use of these models will ultimately improve medication adherence and BP control in patients with uncontrolled hypertension. Future studies will be needed to assess the models' clinical impacts to support implementation into routine clinical care.

While models such as ours are becoming increasingly more common as clinical decision support tools, several challenges to implementation can be expected. First, there is a need to continuously monitor for potential drifts from the models' expected behaviors, at regular intervals [30]. This may include monitoring for changes in predictive performance, model usefulness, patient population, and predictor data being applied to the models. Substantial efforts may be required for monitoring, investigation of model issues, and model updating. Additionally, based on the 5 rights of clinical decision support framework (ie, right information, right person, right format, right channel, and right time) other challenges can be foreseen [31]. For example, alert fatigue is an issue where a high volume of alerts can overwhelm users and result in total disregard of the information provided. There is also a need to build clinical trust through proper presentation of the models' facts to end users, such as the approved use case, potential risks and benefits of the model, and validation data on performance and clinical utility. All of these factors require careful attention in implementation studies to guide the proper use of the models in practice.

Our study also found that even with a comprehensive set of EHR-based predictors, predicting sustained, uncontrolled hypertension remains a challenging problem. The marginally acceptable discrimination performance results indicate a need to improve further our ability to predict this outcome. As we already attempted complex machine learning methods and a large sample size, more powerful predictors are needed to improve predictive performance. One future direction is to increase our ability to monitor and collect BP data, as BP-related predictors had relatively large impacts on risk scores. Wearable devices, such as fitness trackers, for example, are increasingly popular and can be used to collect BP data and support BP monitoring and management [32].

Limitations

Our study has several limitations. First, our models, while adequate, have room for improvement by adding important predictor variables that were either unavailable or not considered. For example, modifiable risk factors, such as lifestyle behaviors, medication adherence, and medication dosing, are known to be associated with BP control, but these data were either not accessible from the EHR or came from unstructured data sources (eg, clinical notes), which can be challenging to process for prediction. Second, because patients without a BP reading in the prediction window were excluded, and models can only be developed and validated for patients who had some follow-up BP measurements, there is a risk for bias. Fortunately, a relatively small number of patients had no BP reading during the follow-up period, thus, the extent of bias, if any, can be deemed minor. Third, the models were developed for patients on 4 or less antihypertensive drug classes and may not generalize beyond this patient group. Fourth, additional prospective studies are needed to understand how our models would perform when implemented in real-world clinical settings.

Comparison With Prior Work

Our study presented novel applications of predictive modeling to the area of hypertension management. From a design perspective, the prediction outcomes, targeted patient cohort, and intended use case were carefully chosen to optimize models' usefulness for managing hypertension. We noted that while little research was performed about predicting clinically important hypertension states in patients with hypertensiveness, a much larger amount of literature was devoted toward predicting hypertension onset in the general population [11-13]. Comparing with other hypertension prediction studies, ours made use of a relatively diverse set of predictors, including EHR-based data that were not often considered, such as usage of different health services, drug classes prescribed, and social determinants of health [11]. In terms of performance, the C-statistics in our models were similar to those from existing prediction models of hypertension onset, which were between 0.63 - 0.84 according to a meta-analysis [11]. A more recent, published prediction model of uncontrolled hypertension demonstrated a C-statistics of 0.76 [16]. Finally, a large and geographically diverse patient sample allowed us to assess (weak) generalizability through IECV, as well as to assess internal validity with confidence. In contrast, the vast majority of hypertension prediction studies were only internally validated and had smaller sample sizes [11].

Conclusions

We developed and validated risk models for sustained, uncontrolled hypertension, and hypertensive crisis, within 1 year of an index visit showing an uncontrolled BP reading. The hypertensive crisis risk model showed good predictive performance with internal validation and new patient cohorts during IECV. This model could be prospectively validated as a next step. If validity and clinical utility are confirmed, it could then be used within our health system, and potentially elsewhere, as a public health surveillance tool for early detection of hypertensive crises and for supporting physicians with treatment decisions. Further efforts are required to improve the ability to

predict sustained, uncontrolled hypertension, particularly by adding and improving predictor variables.

Authors' Contributions

YT supervised this study. YT, AM, HMN, WA, SHC, JZ, and NP conceptualized this study. HMN, WA, SHC, and JZ developed the statistical analysis plan and performed formal analysis. HMN prepared the first draft of this paper. HMN, WA, SHC, AM, YT, JZ, and NP reviewed and edited this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional tables.
[DOCX File, 41 KB - medinform_v12i1e58732_app1.docx]

References

1. Ritchey MD, Gillespie C, Wozniak G, et al. Potential need for expanded pharmacologic treatment and lifestyle modification services under the 2017 ACC/AHA hypertension guideline. J Clin Hypertens (Greenwich) 2018 Oct;20(10):1377-1391. [doi: 10.1111/jch.13364] [Medline: 30194806]
2. Kjeldsen SE. Hypertension and cardiovascular risk: general aspects. Pharmacol Res 2018 Mar;129:95-99. [doi: 10.1016/j.phrs.2017.11.003] [Medline: 29127059]
3. Papadopoulos DP, Mourouzis I, Thomopoulos C, Makris T, Papademetriou V. Hypertension crisis. Blood Press 2010 Dec;19(6):328-336. [doi: 10.3109/08037051.2010.488052] [Medline: 20504242]
4. Muntner P, Miles MA, Jaeger BC, et al. Blood pressure control among US adults, 2009 to 2012 through 2017 to 2020. Hypertension 2022 Sep;79(9):1971-1980. [doi: 10.1161/HYPERTENSIONAHA.122.19222] [Medline: 35616029]
5. Abrahamowicz AA, Ebinger J, Whelton SP, Commodore-Mensah Y, Yang E. Racial and ethnic disparities in hypertension: barriers and opportunities to improve blood pressure control. Curr Cardiol Rep 2023 Jan;25(1):17-27. [doi: 10.1007/s11886-022-01826-x] [Medline: 36622491]
6. Sulaica EM, Wollen JT, Kotter J, Macaulay TE. A review of hypertension management in black male patients. Mayo Clin Proc 2020 Sep;95(9):1955-1963. [doi: 10.1016/j.mayocp.2020.01.014] [Medline: 32276785]
7. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023 Feb 21;147(8):e93-e621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
8. Wang TJ, Vasan RS. Epidemiology of uncontrolled hypertension in the United States. Circulation 2005 Sep 13;112(11):1651-1662. [doi: 10.1161/CIRCULATIONAHA.104.490599] [Medline: 16157784]
9. Upadhyay RK. Emerging risk biomarkers in cardiovascular diseases and disorders. J Lipids 2015;2015:971453. [doi: 10.1155/2015/971453] [Medline: 25949827]
10. Wang TJ, Gona P, Larson MG, et al. Multiple biomarkers and the risk of incident hypertension. Hypertension 2007 Mar;49(3):432-438. [doi: 10.1161/01.HYP.0000256956.61872.aa] [Medline: 17242302]
11. Chowdhury MZI, Naeem I, Quan H, et al. Prediction of hypertension using traditional regression and machine learning models: a systematic review and meta-analysis. PLoS One 2022;17(4):e0266334. [doi: 10.1371/journal.pone.0266334] [Medline: 35390039]
12. Echouffo-Tcheugui JB, Batty GD, Kivimäki M, Kengne AP. Risk models to predict hypertension: a systematic review. PLoS One 2013;8(7):e67370. [doi: 10.1371/journal.pone.0067370] [Medline: 23861760]
13. Chaikijurajai T, Laffin LJ, Tang WHW. Artificial intelligence and hypertension: recent advances and future outlook. Am J Hypertens 2020 Nov 3;33(11):967-974. [doi: 10.1093/ajh/hpaa102] [Medline: 32615586]
14. Mahabaleshwarkar R, Bond A, Burns R, et al. Prevalence and correlates of uncontrolled hypertension, persistently uncontrolled hypertension, and hypertensive crisis at a healthcare system. Am J Hypertens 2023 Nov 15;36(12):667-676. [doi: 10.1093/ajh/hpad078] [Medline: 37639217]
15. Gupta P, Patel P, Štrauch B, et al. Risk factors for nonadherence to antihypertensive treatment. Hypertension 2017 Jun;69(6):1113-1120. [doi: 10.1161/HYPERTENSIONAHA.116.08729] [Medline: 28461599]
16. Mroz T, Griffin M, Cartabuke R, et al. Predicting hypertension control using machine learning. PLoS One 2024;19(3):e0299932. [doi: 10.1371/journal.pone.0299932] [Medline: 38507433]
17. Ye C, Fu T, Hao S, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. J Med Internet Res 2018 Jan 30;20(1):e22. [doi: 10.2196/jmir.9268] [Medline: 29382633]
18. Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. Am J Public Health 2003 Jul;93(7):1137-1143. [doi: 10.2105/ajph.93.7.1137] [Medline: 12835199]

19. Flanagan BE, Gregory EW, Hallisey EJ, Heitgerd JL, Lewis B. A social vulnerability index for disaster management. J Homel Secur Emerg Manag 2011;8(1). [doi: 10.2202/1547-7355.1792]
20. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser Comorbidity measures into a point system for hospital death using administrative data. Med Care 2009 Jun;47(6):626-633. [doi: 10.1097/MLR.0b013e31819432e5] [Medline: 19433995]
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982 Apr;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]
22. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019 Sep 20;38(21):4051-4065. [doi: 10.1002/sim.8281] [Medline: 31270850]
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [Medline: 3203132]
24. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-574. [doi: 10.1177/0272989X06295361] [Medline: 17099194]
25. Takada T, Nijman S, Denaxas S, et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. J Clin Epidemiol 2021 Sep;137:83-91. [doi: 10.1016/j.jclinepi.2021.03.025] [Medline: 33836256]
26. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016 Jan 30;35(2):214-226. [doi: 10.1002/sim.6787] [Medline: 26553135]
27. Okonofua EC, Simpson KN, Jesri A, Rehman SU, Durkalski VL, Egan BM. Therapeutic inertia is an impediment to achieving the healthy people 2010 blood pressure control goals. Hypertension 2006 Mar;47(3):345-351. [doi: 10.1161/01.HYP.0000200702.76436.4b] [Medline: 16432045]
28. Staessen JA, Thijisq L, Fagard R, et al. Effects of immediate versus delayed antihypertensive therapy on outcome in the systolic hypertension in Europe trial. J Hypertens 2004 Apr;22(4):847-857. [doi: 10.1097/00004872-200404000-00029] [Medline: 15126928]
29. Augustin A, Coutts L, Zanisi L, et al. Impact of therapeutic inertia on long-term blood pressure control: a monte carlo simulation study. Hypertension 2021 Apr;77(4):1350-1359. [doi: 10.1161/HYPERTENSIONAHA.120.15866] [Medline: 33641362]
30. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022 Jan 10;5(1):2. [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
31. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. J Am Med Inform Assoc 2007;14(2):141-145. [doi: 10.1197/jamia.M2334] [Medline: 17213487]
32. Kario K. Management of hypertension in the digital era. Hypertension 2020 Sep;76(3):640-650. [doi: 10.1161/HYPERTENSIONAHA.120.14742] [Medline: 35210178]

Abbreviations

ADI: Area Deprivation Index
BP: blood pressure
EHR: electronic health record
ICI: integrated calibration index
IECV: internal-external cross-validation
LR: logistic regression
TI: therapeutic inertia

Edited by C Lovis; submitted 22.03.24; peer-reviewed by J Speiser, T Wang; revised version received 14.06.24; accepted 30.06.24; published 28.10.24.

Please cite as:
Nguyen HM, Anderson W, Chou SH, McWilliams A, Zhao J, Pajewski N, Taylor Y
Predictive Models for Sustained, Uncontrolled Hypertension and Hypertensive Crisis Based on Electronic Health Record Data: Algorithm Development and Validation
JMIR Med Inform 2024;12:e58732
URL: https://medinform.jmir.org/2024/1/e58732
doi:10.2196/58732

Assessing the Effect of Electronic Health Record Data Quality on Identifying Patients With Type 2 Diabetes: Cross-Sectional Study

Priyanka Dua Sood[1], DrPH; Star Liu[2], MS; Harold Lehmann[1,2], MD, PhD; Hadi Kharrazi[1,2], MD, PhD

[1]

[2]

Corresponding Author:
Priyanka Dua Sood, DrPH

Abstract

Background: Increasing and substantial reliance on electronic health records (EHRs) and data types (ie, diagnosis, medication, and laboratory data) demands assessment of their data quality as a fundamental approach, especially since there is a need to identify appropriate denominator populations with chronic conditions, such as type 2 diabetes (T2D), using commonly available computable phenotype definitions (ie, phenotypes).

Objective: To bridge this gap, our study aims to assess how issues of EHR data quality and variations and robustness (or lack thereof) in phenotypes may have potential impacts in identifying denominator populations.

Methods: Approximately 208,000 patients with T2D were included in our study, which used retrospective EHR data from the Johns Hopkins Medical Institution (JHMI) during 2017 - 2019. Our assessment included 4 published phenotypes and 1 definition from a panel of experts at Hopkins. We conducted descriptive analyses of demographics (ie, age, sex, race, and ethnicity), use of health care (inpatient and emergency room visits), and the average Charlson Comorbidity Index score of each phenotype. We then used different methods to induce or simulate data quality issues of completeness, accuracy, and timeliness separately across each phenotype. For induced data incompleteness, our model randomly *dropped* diagnosis, medication, and laboratory codes independently at increments of 10%; for induced data inaccuracy, our model randomly *replaced* a diagnosis or medication code with another code of the same data type and induced 2% incremental change from −100% to +10% in laboratory result values; and lastly, for timeliness, data were modeled for induced *incremental shift* of date records by 30 days to 365 days.

Results: Less than a quarter (n=47,326, 23%) of the population overlapped across all phenotypes using EHRs. The population identified by each phenotype varied across all combinations of data types. Induced incompleteness identified fewer patients with each increment; for example, at 100% diagnostic incompleteness, the Chronic Conditions Data Warehouse phenotype identified zero patients, as its phenotypic characteristics included only diagnosis codes. Induced inaccuracy and timeliness similarly demonstrated variations in performance of each phenotype, therefore resulting in fewer patients being identified with each incremental change.

Conclusions: We used EHR data with diagnosis, medication, and laboratory data types from a large tertiary hospital system to understand T2D phenotypic differences and performance. We used induced data quality methods to learn how data quality issues may impact identification of the denominator populations upon which clinical (eg, clinical research and trials, population health evaluations) and financial or operational decisions are made. The novel results from our study may inform future approaches to shaping a common T2D computable phenotype definition that can be applied to clinical informatics, managing chronic conditions, and additional industry-wide efforts in health care.

(JMIR Med Inform 2024;12:e56734) doi:10.2196/56734

KEYWORDS

electronic health record; EHR; EHRs; record; records; computable; phenotyping; phenotype; phenotypes; computable phenotypes; data quality; data science; chronic; identify; identification; data types—diagnosis data, medication data, laboratory data; type-2 diabetes; diabetes; diabetic; DM; type 2; hospital system; clinical research and trial; diagnosis; diagnoses; diagnose; diagnostic; diagnostics; phenotypic

Introduction

Type 2 diabetes (T2D) is a common chronic disease affecting more than 11% of the US population [1]. Given the rapidly increasing burden of T2D on health care services and resources, health systems have devised strategies to address such demands by identifying and managing T2D patients across their populations [1]. However, despite clinical guidelines to identify T2D at the point of care, identifying T2D patients in large clinical repositories is still a challenge [2]. Ambiguity of algorithms to identify T2D patients in complex data sets and

data quality issues in routine clinical data sources such as electronic health records (EHRs) are still hindering the development of generalizable approaches to identify T2D patients across different health systems.

Over the last two decades, several approaches have been developed for health care professionals to identify patients with T2D in large clinical data repositories. Multiple T2D phenotype definitions (also known as computational algorithms) are available that define characteristics to identify T2D populations. Some examples of T2D definitions include the Surveillance, Prevention, and Management of Diabetes Mellitus (SUPREME-DM) [3], the Centers for Medicare and Medicaid Services (CMS) Chronic Conditions Data Warehouse (CCW) [4], the Electronic Medical Records and Genomics (eMERGE) Northwestern Group [5], and the Durham Diabetes Coalition (DDC) [6] phenotypes. Each phenotype definition has its own set of inclusion and exclusion criteria using different data types. Data types may include *International Statistical Classification of Diseases, Tenth Edition* (*ICD-10*) diagnosis codes, RxNorm medication codes, and Logical Observation Identifiers Names and Codes (LOINC) laboratory codes, with sequences and frequencies of occurrences of diagnosis, medication and laboratory codes; defined time periods; and care pathways. However, despite these detailed definitions, it is unclear which of the existing T2D phenotypes are prone to inherit data quality issues from clinical data sources such as EHRs. The uncertainty of T2D phenotypes' performance for identifying populations with T2D using real-world data has led to the lack of a universally agreed-upon T2D phenotype.

The concept of data quality in health care varies based on the problems and functional needs of end users. Health care providers, data scientists, or policy makers may differ in their approach to data quality and its significance in practice [7]. Typically, health care data such as captured in EHRs are collected for continuity of clinical care, coding, and billing purposes, and not necessarily to answer specific research questions. Thus, the quality of data collected in EHRs may be sufficient for clinical purposes but may not meet the needs of a researcher or population health intervention. For example, a clinician may identify a patient with T2D at the point of care despite having incomplete data; however, a T2D phenotyping algorithm may miss the same patient in a large EHR data warehouse due to the underlying data quality issues; hence, the patient may inadvertently be excluded from a research study or population health intervention.

Various data quality frameworks have been proposed to measure the quality of health care data. Completeness, accuracy, and timeliness are a few key data quality characteristics that are used across several frameworks [7-9]. The assessment for completeness defines how complete the data are, what the missing elements are, and how usable the data are in their "as is" format. Accuracy is the correctness and consistency of the data elements [8], and timeliness is how recent or current the data are for research and analysis. The assessment of data quality in health care is crucial with the continuous and prominent use of EHRs; however, despite the increasing use of EHR data to identify patients with T2D, the effect of varying levels of key data quality characteristics (ie, completeness, accuracy, and timeliness) on T2D phenotypes is still unknown.

The ongoing challenge of understanding the effect of key data quality issues on T2D phenotypes is further exacerbated by the fact that T2D phenotypes use multiple data types, such as diagnosis codes, medications, and laboratory results. Additionally, given the variability of key data quality issues of these data types across EHRs [10,11], measuring the effect of key existing data quality issues on T2D phenotypes in one EHR may not translate into generalizable findings. For example, one provider's EHR may suffer from incompleteness of diagnosis codes, while another provider's EHR may be affected by inaccurate medication data. Thus, to measure the effect of an EHR's data quality on T2D phenotypes, varying (simulated) levels of key data quality characteristics across all data types (ie, diagnosis, medication, and laboratory) used by T2D phenotypes should be studied. These simulated levels of key data quality issues will in turn help providers to compare their EHR data quality issues with the simulated levels, contrast the potential impact of such data quality issues on identifying T2D patients using various phenotypes, and eventually select the most suitable T2D phenotype for their EHR data.

Currently, evidence is lacking on the effect of data quality issues (eg, completeness, accuracy, and timeliness) and the identification of T2D populations in large clinical data sources such as EHRs. This gap in evidence is further amplified given the variations in characteristics of published T2D phenotypes and potential discrepancies in underlying data types (ie, diagnosis, medication, and laboratory) in EHRs. To address these gaps, our study aimed to assess the impact of varying (simulated) levels of data quality issues across the different data types used by T2D phenotypes. Our study findings can inform health care providers and other stakeholders to select T2D phenotype algorithms that best match their underlying EHR data quality issues.

Methods

Data Source

Our cross-sectional study used retrospective EHR data from the Johns Hopkins Medical Institute (JHMI) data warehouse over a 3-year period from 2017 to 2019. Clinical data included T2D primary diagnostic data (*ICD-10*), laboratory data (LOINC), and medication data (RxNorm). The demographic data included age, sex, race, ethnicity, and the patients' residential state.

Study Population

Our overall study population included approximately 208,000 patients in age groups of 18 to 90 years. This population denominator was identified and extracted in a prior study focusing on T2D patients and funded by the US Food and Drug Administration (5U01FD005942-05). This population denominator, also known as the raw data cut, was identified by the most inclusive query of the JHMI's EHR data warehouse, which also included patients with mentions of diabetes in their clinical notes. Each of the identified T2D phenotypes was applied to the overall study population. At total of approximately

164,000 patients were included in at least one of the T2D phenotypes assessed.

Considering the size of the EHR data set, and given the impracticality of reviewing the individual records of almost a quarter of a million patients, no gold standard population was identified for this research. Indeed, the aim of this research study was not to assess the accuracy of the common T2D phenotypes; instead, we aimed to measure the performance of T2D phenotypes given the underlying data quality issues in EHR data repositories.

T2D Phenotype Definitions

Our assessment included 4 published T2D phenotype definitions, from the CCW, DDC, SUPREME-DM, and eMERGE, as well as 1 unpublished definition (Johns Hopkins University; JHU). Since not all T2D phenotype definitions were defined with the most current *ICD* codes, we converted the DDC, SUPREME-DM, and eMERGE diagnostic phenotypes from *ICD-9* to *ICD-10* [12]. Additionally, not all phenotype definitions using medications had the list of specific RxNorm codes (eg, the DDC and SUPREME-DM phenotypes included only the names of the T2D medication and not the codes). We identified the RxNorm codes for each medication using the National Library of Medicine's RxNav tool [13]. For the purposes of this study, we selected only those RxNorm codes that were associated with the primary ingredient or ingredients of the medication name in the T2D phenotype definition.

Of the 4 published T2D phenotypes, CCW only included diagnosis codes with a reference period of 2 years in the definition. The DDC, SUPREME-DM, and eMERGE phenotypes included a series of detailed care pathways with diagnosis, medication, and laboratory codes and results within specified time periods. However, the unpublished definition from JHU did not have pathways and was the most inclusive definition, as it included all data types (diagnosis, medication, and laboratory) with wide eligibility criteria to identify patients with T2D.

Factors Affecting T2D Phenotyping

The data included demographics, T2D-related data types (ie, diagnosis, medication, and laboratory), and the Charlson Comorbidity Index [14]. The demographic data included age, sex (male, female, and other), race (White, Black, Asian, and other), ethnicity (Hispanic/Latino or non-Hispanic/Latino) and patient location (state of Maryland or other locations). The "other" category for sex, race, and ethnicity was primarily composed of missing data entries. Since the JHMI is in Baltimore, Maryland, the majority (169,215/207,813, 81.4%) of our study population was from Maryland and the remainder (38,598/207,813, 18.6%) was from the surrounding states. The outcome measures included the extent of overlap in identifying patients with T2D and the degree of robustness against data quality issues across phenotypes.

Statistical Analysis

We performed descriptive data analyses across the 5 different phenotypes to identify EHR populations with T2D. We used the χ^2 test for categorical variables and ANOVA for continuous variables. Our analysis included distribution and overlap of the population of interest by diagnosis, medication, and laboratory data types across each of the 5 T2D phenotypes. We introduced methods that simulated or induced data incompleteness, inaccuracy, and lack of timeliness (eg, date shifting) to assess the robustness of each phenotype in capturing the populations of interest. We created unique analytical functions for each data quality issue considering the data types that were applicable across all T2D phenotypes.

To simulate or induce data incompleteness, our procedure randomly dropped codes at 10% increments, from 0% to 100%, for diagnosis, medication, and laboratory codes within each of the T2D phenotypes. Incompleteness was simulated up to 100% as no thresholds of data missingness were known to affect the performance of T2D phenotypes beforehand. Incompleteness was induced for diagnosis, medication, or laboratory codes independently of the other 2 data types. For each increment of incompleteness, T2D phenotypes were reapplied to identify a new (and logically smaller) cohort of patients.

To gauge the impact of inconsistency and inaccuracy in our denominator population for each T2D phenotype, the diagnosis and medication codes were replaced at random at increments of 10% (the same as for incompleteness) from 0% to 100% with another code of the same data type, including T2D and non-T2D codes. We included T2D codes to illustrate expected data quality issues that may impact the phenotypes' performance. For example, 10% of instances of the *ICD-10* diagnosis code E08 (ie, diabetes mellitus due to underlying condition) were replaced at random with the *ICD-10* code E09 (ie, drug- or chemical-induced diabetes mellitus—endocrine, nutritional, or metabolic disease). For laboratory codes, the laboratory values were induced with a 2% incremental change from −100% to 10% in laboratory results. For example, our procedure for simulated or induced laboratory values yielded results of 5.6% to 5.8% for hemoglobin A_{1c} (LOINC code 55454‐3). Additionally, we simulated or induced inaccuracy in units of laboratory data results. We did this by intentionally converting reported laboratory units from US standards to UK standards; in particular, blood glucose level units in mg/dl were converted to mmol/L, and hemoglobin A_{1c} level from percentage to mmol/mol [15].

For timeliness, we simulated or induced date shifts at increments of 30 to 365 days for all phenotypes. For example, our procedure induced a forward shift on December 1, 2019, by 30 days, shifting the date to December 31, 2019, and so on. Lastly, we also induced compounded data quality issues (incompleteness, inaccuracy, and lack of timeliness) to understand each phenotype's resistance to changes that occurred across diagnosis, medication, and laboratory codes simultaneously. Induced compounded incompleteness would mean dropping diagnosis, medication, and laboratory codes randomly at increments of 10% up to 100%. Induced compounded inaccuracy would mean replacing diagnosis and medication codes randomly at increments of 10% up to 100% (laboratory codes were not replaced; their values were manipulated instead). Induced compounded lack of timeliness would mean inducing date shifts

across diagnosis, medication, and laboratory codes at increments of 30 to 365 days.

SQL queries were written for data extraction. All visualizations and statistical analyses were conducted using R (version 4.2.0; R Foundation for Statistical Computing) [16]. The overall findings are showcased as descriptive data tables, Venn diagrams, and line charts depicting the effect of simulated or induced data quality issues on identifying T2D populations using the common T2D phenotype definitions.

Ethical Considerations

This study was reviewed and approved by the IRB committee of the Johns Hopkins School of Public Health (00014440). The deidentified population denominator used in this study is from a prior study funded by the US Food and Drug Administration (5U01FD005942-05).

Results

Characteristics of the Overall and Phenotype-Identified T2D Populations

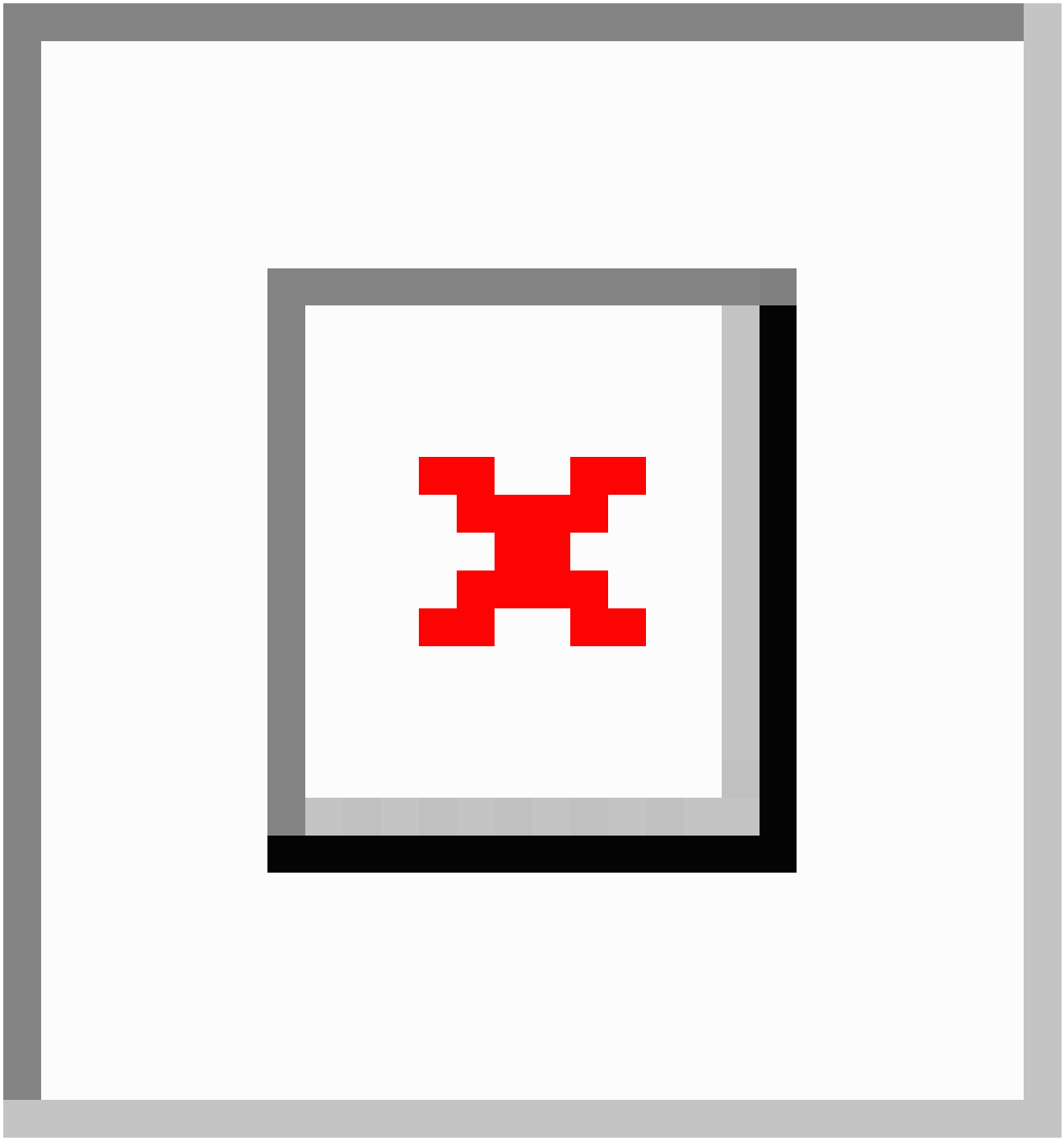
Our overall study population included 207,813 patients with T2D from between 2017 and 2019 in the JHMI EHR data. For the purposes of this analysis, we refer to this population as the raw data cut or the overall study population/denominator. The mean age of the overall study population was 62.4 (SD 15.4) years, with 81.4% (n=169,215) of the population residing in the state of Maryland. Women accounted for 51.3% (n=106,704) of the study population, with 31.8% (n=66,073) Black patients and 89.9% (n=186,785) non-Hispanic/Latino patients. The overall population had a mean number of 0.657 (SD 1.61) inpatient visits and 1.01 (SD 3.61) emergency department visits across the study duration. The mean Charlson Comorbidity Index score was 2.17 (SD 2.25). Table 1 shows the overall characteristics of the study population.

Table . Characteristics of the overall study population (N=207,813).

Characteristics		Values
Age (years)		
	Mean (SD)	62.4 (15.4)
	Median (range)	64.0 (18.0-90.0)
Sex, n (%)		
	Female	106,704 (51.3)
	Male	101,079 (48.6)
Race, n (%)		
	Asian	11,644 (5.6)
	Black	66,073 (31.8)
	White	109,695 (52.8)
	Other	20,401 (9.8)
Ethnicity, n (%)		
	Hispanic/Latino	10,979 (5.3)
	Non-Hispanic/Latino	186,785 (89.9)
	Other	10,049 (4.8)
Charlson Comorbidity Index score		
	Mean (SD)	2.17 (2.25)
	Median (range)	1.33 (0-20.0)
State, n (%)		
	Maryland	169,215 (81.4)
	Other	38,598 (18.6)
Inpatient visits (n)		
	Mean (SD)	0.657 (1.61)
	Median (range)	0 (0-59.0)
Emergency department visits (n)		
	Mean (SD)	1.01 (3.61)
	Median (range)	0 (0-415)

T2D phenotypes were applied to the overall study population. The characteristics of the population identified by each phenotype were notably different due to the phenotypes' varying constraints on medical events and diagnosis, medication, and laboratory codes. The Venn diagram displayed in Figure 1 shows the overlap of the population with T2D across the phenotypes in comparison to the overall study population (ie, the largest, gray circle denotes the raw data cut). These T2D populations were identified using all EHR data types (diagnosis, medication, and laboratory) as needed by the T2D phenotypes. A total of 78% (n=160,030 patients) of the overall study population was identified by at least 1 phenotype, but only 23% (n=47,326) of the overall study population was identified by all T2D phenotypes. DDC identified 139,832 patients with T2D, of which 11,154 (7.98%) were not identified by the other 4 phenotype definitions. Of the 89,772 patients with T2D identified by SUPREME-DM, there were 5911 (3%) patients with T2D that were also identified by the DDC and JHU phenotypes. Additionally, 23,659 (12%) patients were identified by all phenotypes except eMERGE. Additional details of population overlap counts are available in Multimedia Appendix 1, Table S1.

Figure 1. Venn diagram showing overlap of type 2 diabetes populations identified across all phenotype definitions using electronic health record data. CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus; eMERGE: Electronic Medical Records and Genomics.



The distributions of T2D populations identified across phenotype definitions by diagnosis, medication, and laboratory data types were separately calculated (Table 2). The SUPREME-DM phenotype identified the most patients (n=33,268) across all 3 data types, followed by eMERGE with 30,573 patients. Zero patients were observed for some phenotypes when using specific data types. The zero observations were due to the representation of the data types as defined by the criteria of each phenotype definition. For example, the eMERGE phenotype does not have a pathway to identify a patient with T2D based on diagnosis code only, medication code only, or medication and laboratory codes only. Hence, we observed zero patients with diagnosis only, medication only, or medication and laboratory only for eMERGE (Table 2). In the case of CCW, since the phenotype does not include medication or laboratory codes in its definition, zero patients were observed when medication or laboratory codes were required for the identification of T2D patients (Table 2).

We measured the effect of simulated or induced data quality issues of completeness, accuracy, and timeliness for diagnosis, medication, and laboratory data across all T2D phenotypes. The following results include the percentage of patients identified by each phenotype while simulating data quality issues using diagnosis codes. Figures S1 to S3 in Multimedia Appendix 1 show the same diagnosis results but depict the frequency of patients identified by each phenotype. Figures S4 to S17 in Multimedia Appendix 1 show results as the percentages and frequencies of patients identified by each phenotype while simulating data quality issues using medication and laboratory data types.

Table . Distribution of type 2 diabetes (T2D) populations identified by T2D phenotype definitions using different combinations of data types.

Data type	DDC[a] (n=139,832), n	SUPREME-DM[b] (n=89,772), n	eMERGE[c] (n=77,977), n	JHU[d] (n=139,231), n	CCW[e] (n=79,967), n
Diagnosis	40,133	15,983	0	75,027	79,967
Medication	15,511	0	0	5874	0
Laboratory	5973	1907	10,467	2711	0
Diagnosis and medication	26,993	24,328	31,114	21,336	0
Diagnosis and laboratory	21,546	13,237	5823	19,452	0
Medication and laboratory	987	1049	0	83	0
Diagnosis, medication, and laboratory	28,687	33,268	30,573	14,748	0

[a]DDC: Durham Diabetes Coalition.

[b]SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.

[c]eMERGE: Electronic Medical Records and Genomics.

[d]JHU: Johns Hopkins University.

[e]CCW: Chronic Conditions Data Warehouse.

Data Quality Issues

Completeness

To depict the impact of the simulated or induced incompleteness of diagnosis codes on the identified population using the T2D phenotypes, Figure 1 was recreated at increasing levels of data incompleteness (Figure 2). As the percentage of induced missing diagnosis codes increased, the size of the T2D population identified by each phenotype decreased. The CCW phenotype saw the largest decline in the identified population with increasing incompleteness. At 100% induced incompleteness, there were no T2D patients identified by CCW, as the CCW phenotype relies entirely on diagnosis codes for the identification of T2D patients. Lastly, the eMERGE, SUPREME-DM, DDC, and JHU phenotypes continued to identify patients with T2D despite 100% incompleteness of diagnosis codes (Figure 2).

Figure 3 shows the decrease in the percentage of T2D population identified by each phenotype when diagnostic incompleteness was induced from 0% to 100%, that is, diagnosis codes from the *ICD-10* were dropped in increments of 10%. All phenotype definitions showed a similar decrease in the percentage of patients from 0% to approximately 80% of diagnostic incompleteness. The eMERGE and CCW phenotypes saw significant declines in T2D population sizes when the induced incompleteness increased from 80% to 100%. At 100% incompleteness, CCW identified no patients and eMERGE identified only 21% (16,348/77,977) of the patients with T2D, consistent with Table 2 and Figure 2.

Figure 2. Overall population identified by each of the type 2 diabetes phenotype definitions when diagnosis codes were dropped from 20% to 100% to simulate increasing incompleteness of diagnosis codes. CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE: Electronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.

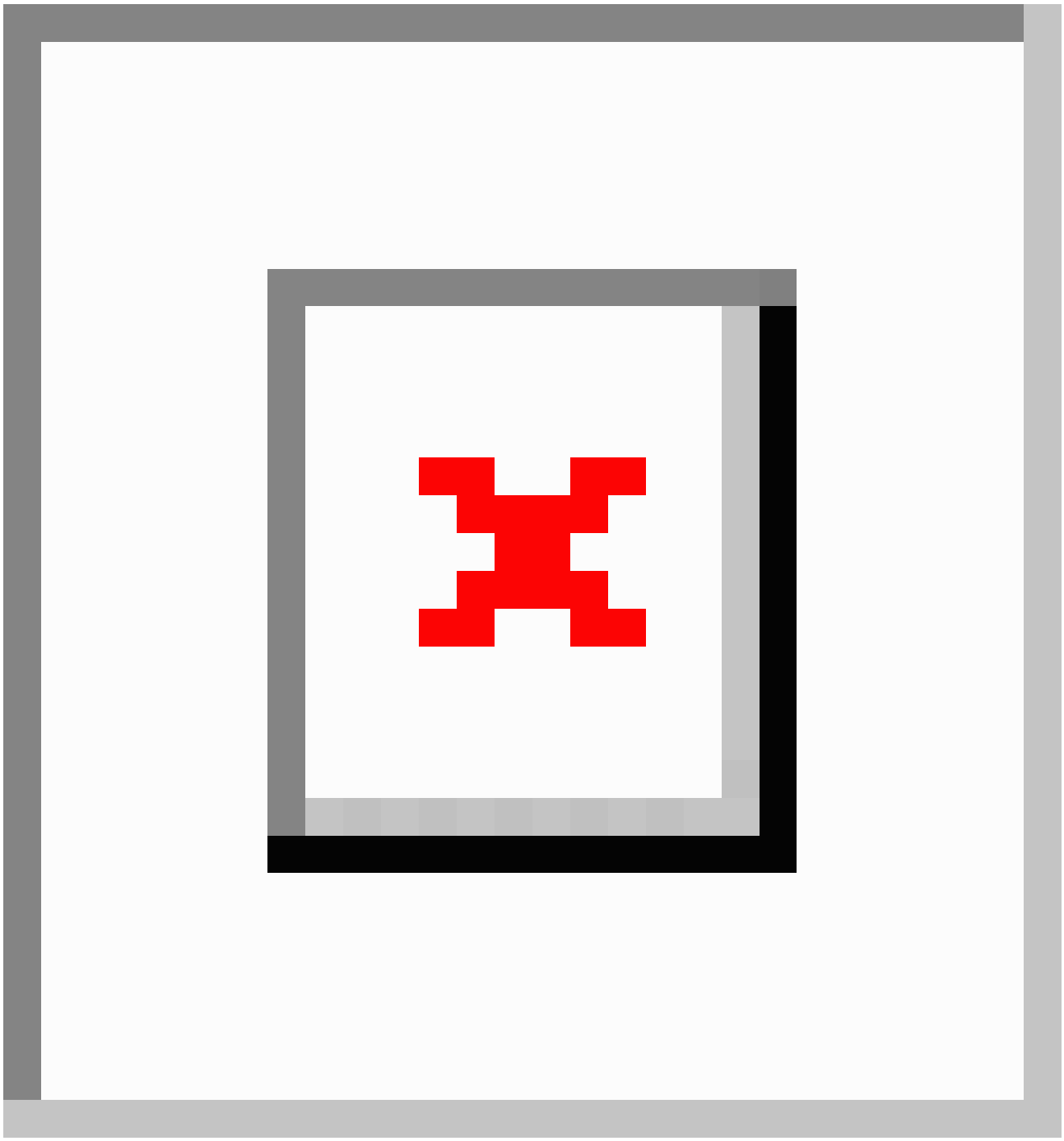
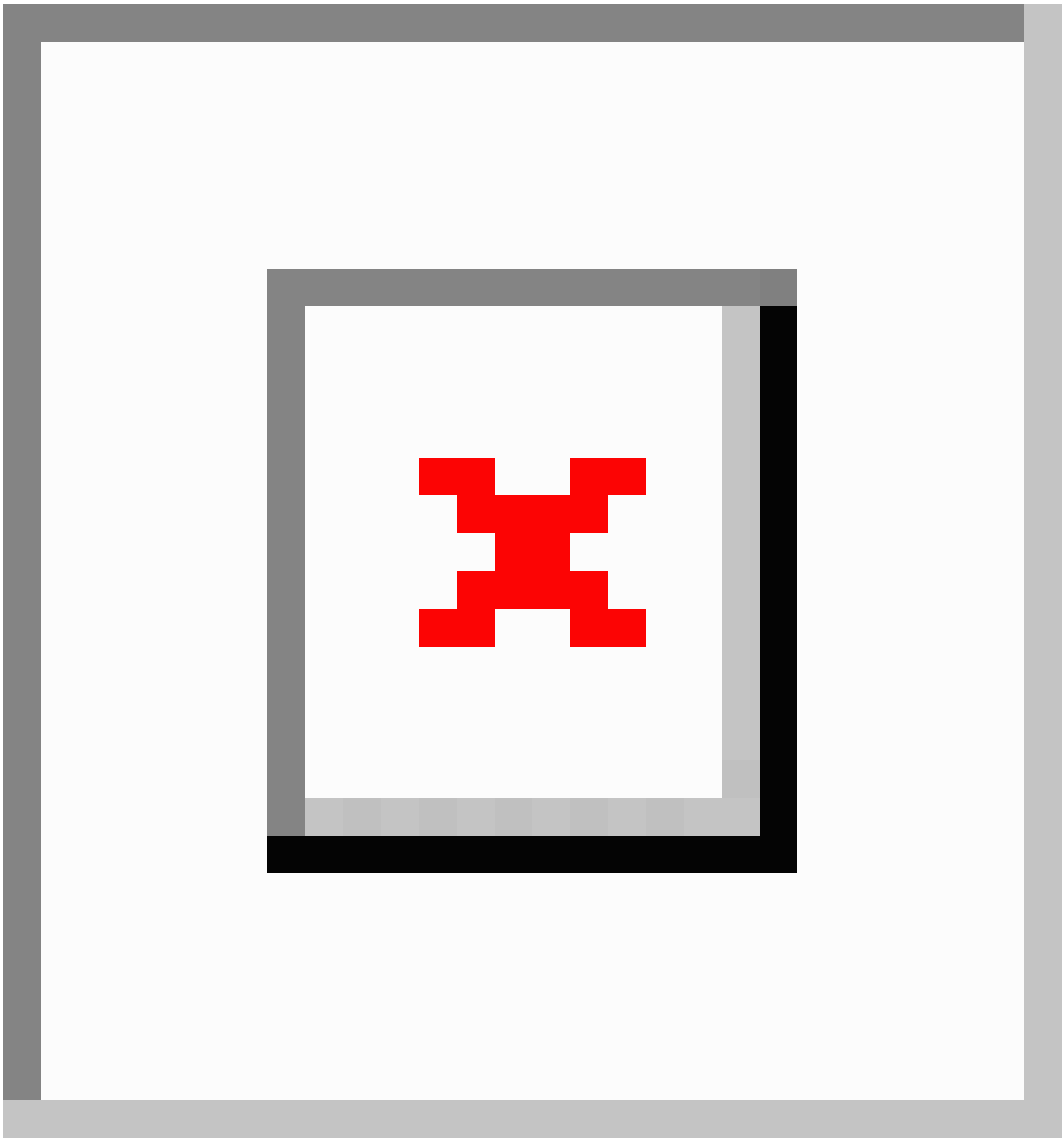


Figure 3. Percent of type 2 diabetes population identified by each type 2 diabetes phenotype definition with increasing incompleteness of diagnosis codes. CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE:lectronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.



Accuracy

Figures 4 and 5 show the overlap and percentages, respectively, of T2D populations identified by each phenotype definition with an increasing percentage of induced replacement of diagnosis codes at random. All phenotype definitions were impacted and continued to identify populations with T2D even with 100% induced inaccuracy despite reductions overall. The CCW, SUPREME-DM, and eMERGE phenotypes showed the greatest decrease when 75% - 100% of the diagnosis codes were replaced, resulting in identification of only 27% - 45% of the patients with T2D.

Figure 4. Overall population identified by each of the type 2 diabetes phenotype definitions when diagnosis codes are replaced at random to simulate increasing diagnostic inaccuracy. CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE: Electronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.

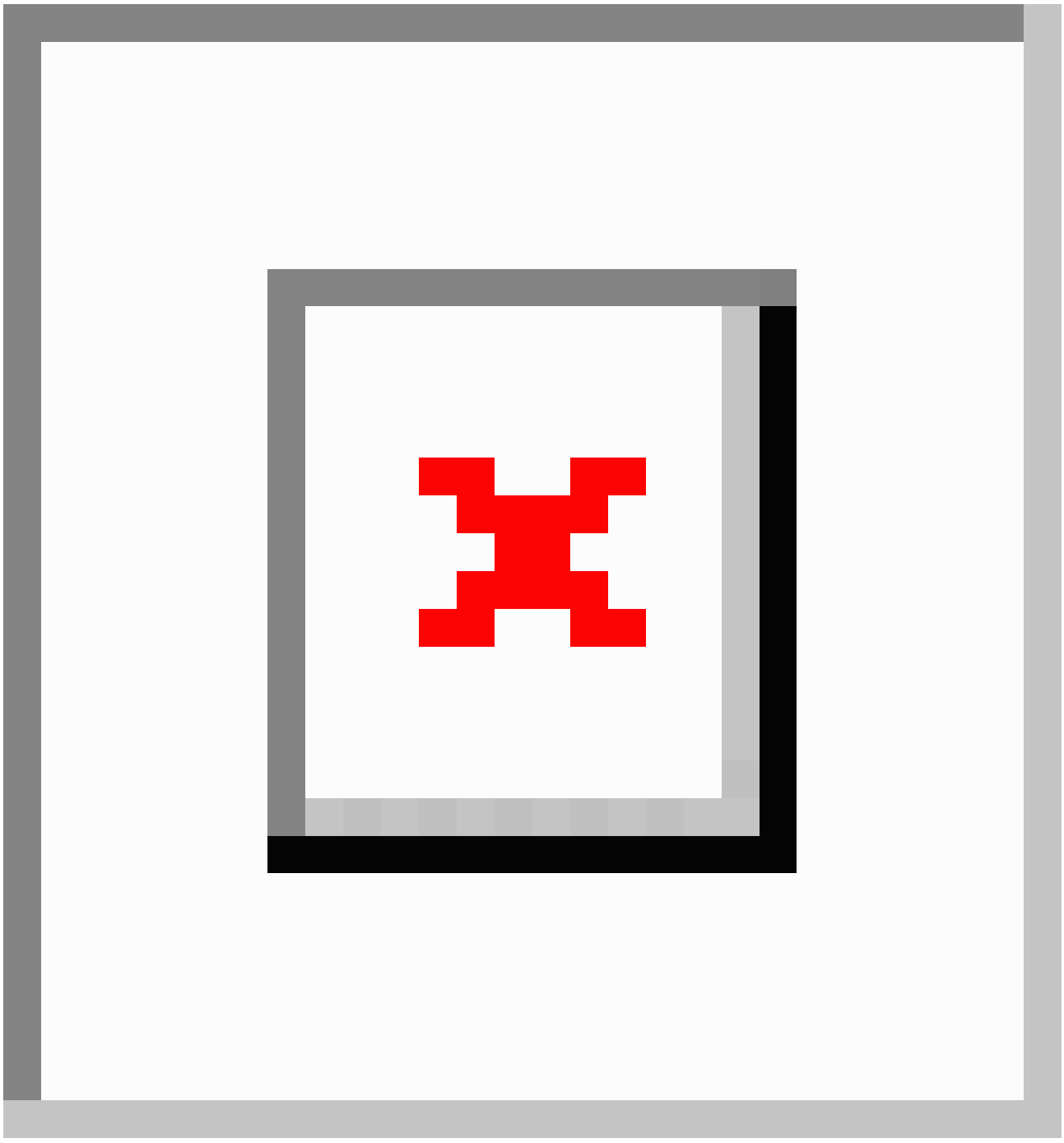
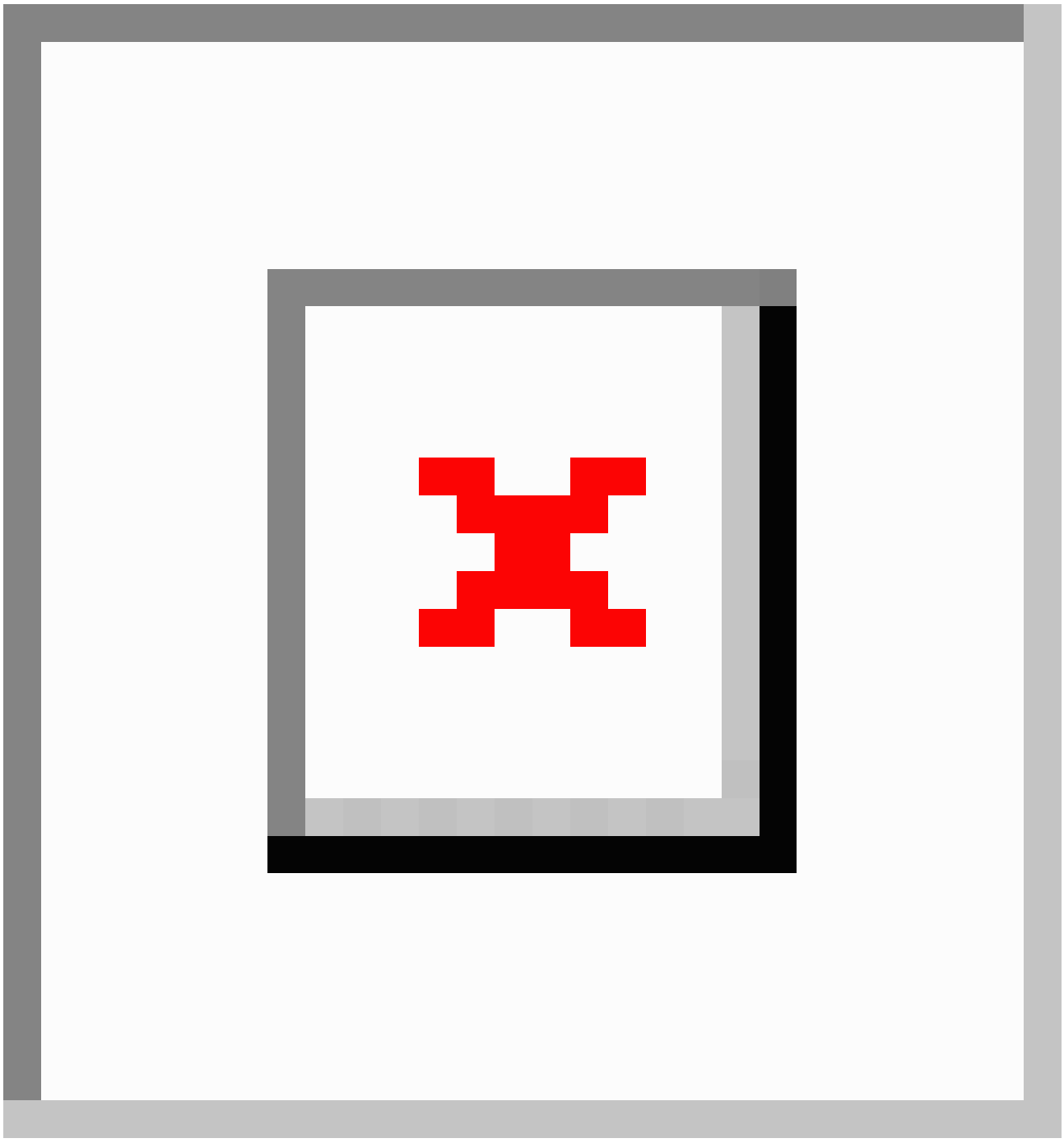


Figure 5. Percentage of type 2 diabetes population identified by each type 2 diabetes phenotype definition with increasing inaccuracy of diagnosis codes. CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE: Electronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.



Timeliness

All T2D phenotypes were similarly impacted by simulated or induced shifts in the date of recorded diagnosis from 30 days to 365 days. The patterns of overlap in population were similar across all phenotypes with each incremental date shift (Figure 6). The percentage of patients identified with T2D showed a decrease over time, however, with at least 85% (176,641/207,813) of the patients with T2D identified during the progression of a year across all phenotypes. The data quality issue of timeliness showed the least impact in the DDC phenotype (Figure 7).

Figure 6. Overall population identified by each of the type 2 diabetes phenotype definitions with shifts in timeliness of diagnostic data ranging from 30 to 365 days. CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE: Electronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.

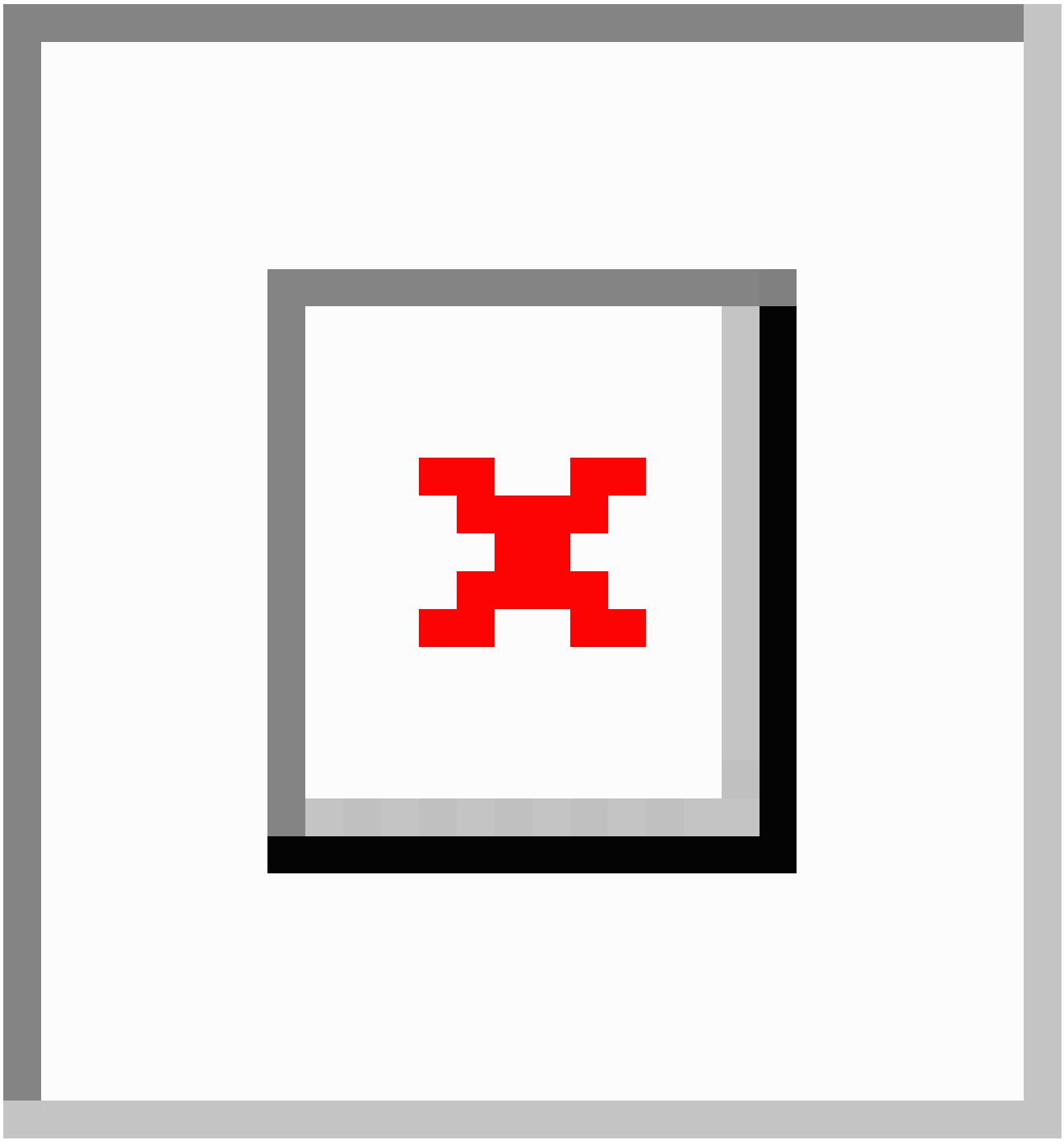
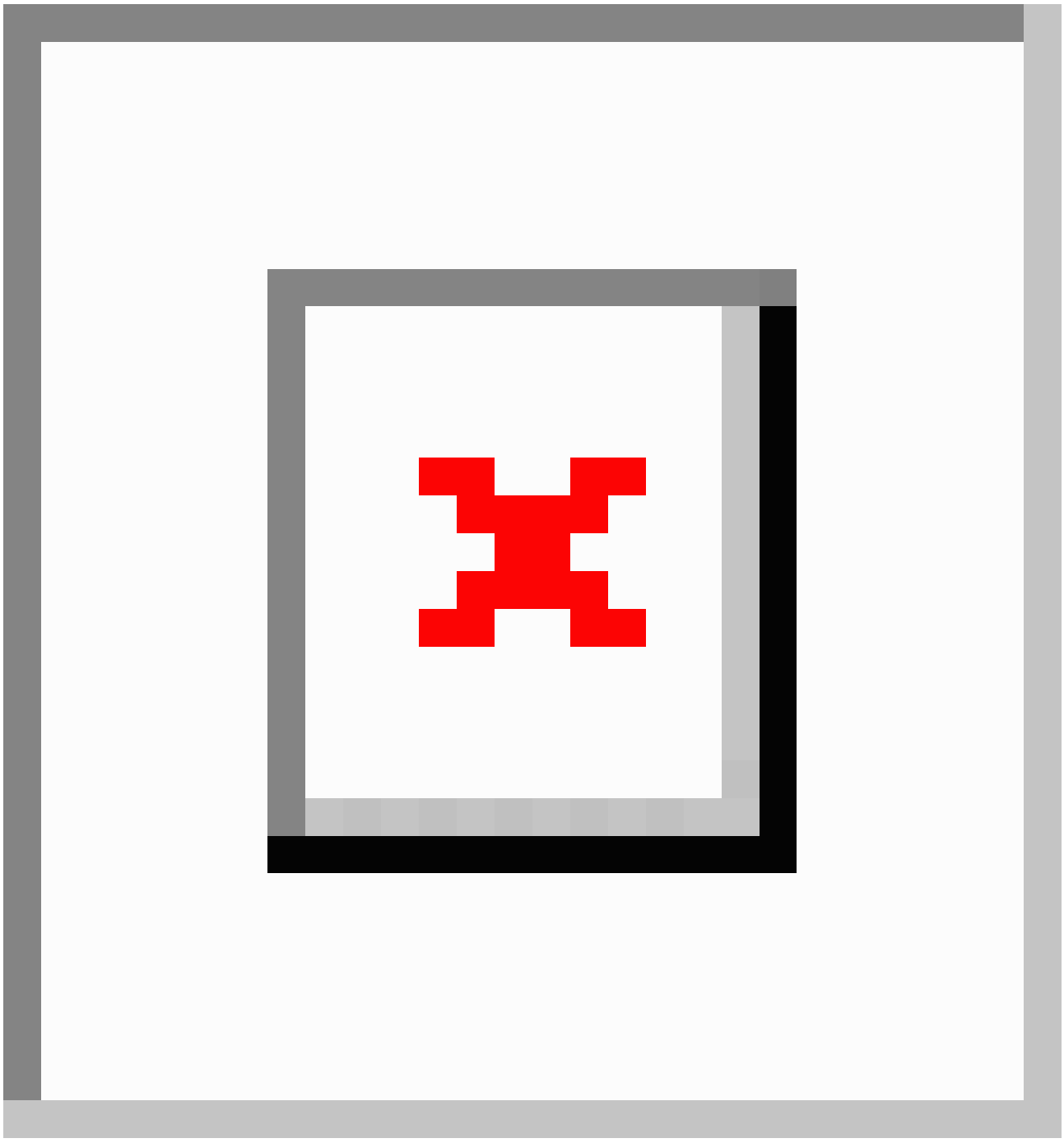


Figure 7. Percentage of type 2 diabetes population identified by each type 2 diabetes phenotype definition with an increasing shift in diagnostic timeliness (ie, number of days shifted). CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE: Electronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.



Compounded Data Quality Issue: Completeness

Figures 8 and 9 show the overlap and percentages, respectively, of T2D populations identified by each phenotype definition with an increasing percentage of induced compounded incompleteness across diagnosis, medication, and laboratory codes. All phenotype definitions were impacted and continued to identify populations with T2D until 100% induced incompleteness, as expected. While all phenotypes exhibited similar rates of decrease with increased compounded incompleteness, CCW was the most robust to incompleteness. SUPREME-DM was the least resistant to induced compounded incompleteness. Figures S18 and S21 in Multimedia Appendix 1 provide results for the percentages and frequencies of patients identified by each phenotype while simulating the compounded data quality issues of inaccuracy and lack of timeliness.

Figure 8. Overall population identified by each of the type 2 diabetes phenotype definitions with compounded increasing incompleteness (diagnostic, medication, and laboratory codes). CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE: Electronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.

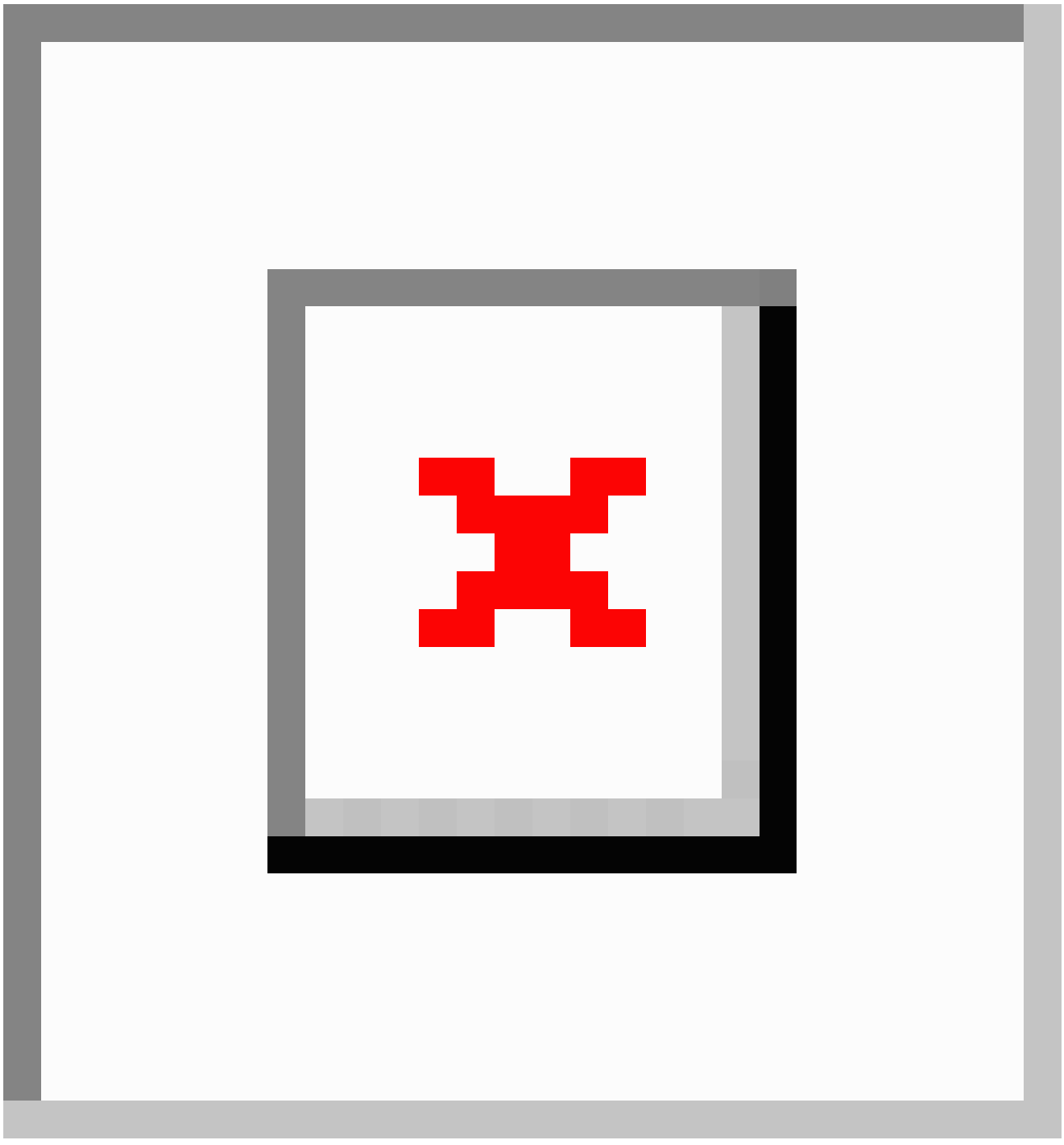
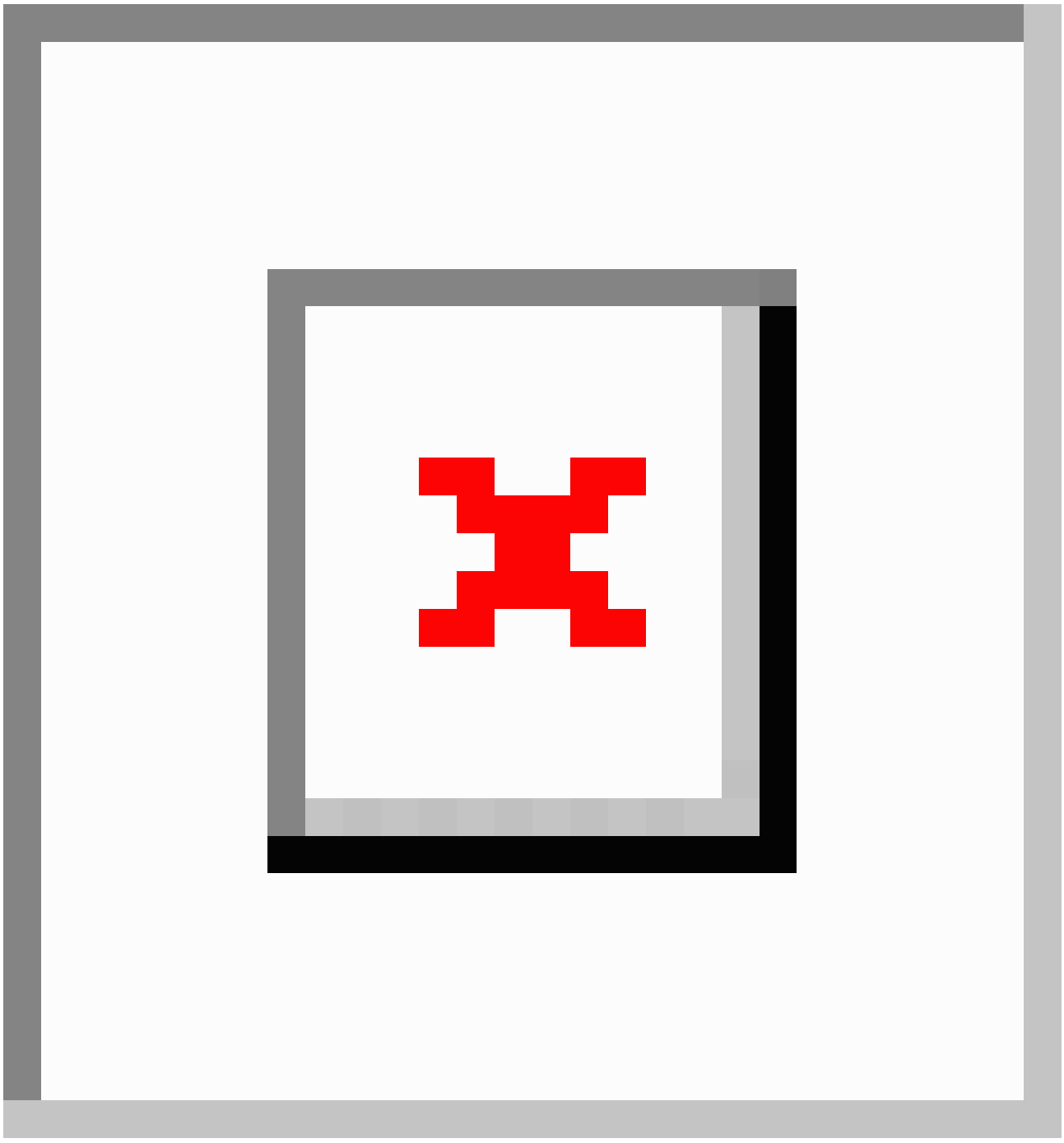


Figure 9. Percentage of type 2 diabetes population identified by each type 2 diabetes phenotype definition with compounded increasing incompleteness (diagnostic, medication, and laboratory codes). CCW: Chronic Conditions Data Warehouse; DDC: Durham Diabetes Coalition; eMERGE: Electronic Medical Records and Genomics; JHU: Johns Hopkins University; SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus.



Discussion

T2D is a common chronic disease with no known universally defined phenotype definition. This is further complicated by lack of assessments that are necessary to understand the quality of data in real-world health care settings against common phenotypes. In our study, we investigated how data quality issues of completeness, accuracy, and timeliness may impact the implementation of existing T2D phenotypes using different data types of diagnosis, medication, and laboratory codes extracted from common clinical data sources such as EHRs. Each data quality issue and phenotype definition may present unique implications at an unpredictable scale in any given health care setting. Collectively, these variables can pose challenges in identifying an eligible denominator population and implementing study population enrollment strategies for epidemiological studies, population health and management, disease surveillance, financial operations, and policies for populations with T2D.

Although each phenotype has distinct characteristics, our results showed considerable overlap in the population identified with T2D across the phenotype definitions (Figure 1 and Table 2). Given the uniqueness of the CCW phenotype, which only includes diagnosis codes, the CCW phenotype identified

approximately 38.4% (79,967/207,813) of the overall study population with T2D, whereas the DDC phenotype, which was inclusive of all data types, identified approximately 67.2% (139,832/207,813) of the total population with T2D. This was similar to the 66.9% (139,231/207,813) of the total population identified by the Hopkins phenotype. Additionally, we observed that the T2D phenotype definitions of SUPREME-DM and eMERGE had similar characteristics and therefore resulted in the identification of approximately 43.2% (89,772/207,813) and 37.5% (77,977/207,813) of the overall study population, respectively. Moreover, when analyzing the distribution of population by data types, the identified populations were significantly different, with no noticeable patterns.

This study revealed significant findings resulting from simulated or induced data quality issues on the overall study population using EHR data across different T2D phenotype definitions. Induced incompleteness of diagnostic data showed the least impact on the DDC phenotype, however, identifying only approximately 60% of the population with T2D at 100% diagnostic incompleteness. The CCW phenotype was the most impacted, as 100% of the diagnosis codes were missing with 100% induced incompleteness given the characteristics of its phenotype definition. With more incomplete data, the uniqueness of the population identified by each phenotype may also shift, thereby showing a decrease in overlap of the population with T2D across all phenotypes. Although trends displayed by phenotype definitions of Hopkins and SUPREME-DM may seem similar, the quantification of the slightest differences in trends can have significant population health implications, resulting in compromised financial and logistical (eg, staffing needs) outcomes. Thus, it is important to understand the resistance of data quality issues across phenotypes given different data types.

The results for induced inaccuracy were similar to those for completeness, but data quality issues for timeliness showed a different trend. The phenotypes of eMERGE, CCW, and SUPREME-DM showed significant decreases in the number of identified patients when the majority of the diagnosis codes were replaced. However, induced timeliness continued to capture significant numbers of patients for all phenotypes, at well above 85%. While the impact of induced inaccuracies of laboratory values (in the negative direction) was mitigated by other pathways of the phenotypes, overall, completeness and accuracy showed a much larger impact on the identification of the population with T2D. The trends for the data quality issues of completeness and accuracy paint a similar picture, which may translate into considerations for population health interventions. The effect of data quality on identifying T2D populations using commonly available T2D phenotypes can affect a variety of interventions and outcomes, such as clinical research, financial analysis, staffing needs, and logistical issues, to name a few.

Lastly, induced compounded (diagnosis, medication, and laboratory) data quality issues showed different trends compared to induced single component (diagnosis, medication, or laboratory) data quality issues. Overall, all phenotypes captured fewer patients when there were compounded data quality issues compared to single-component data quality issues. The CCW phenotype was the most resistant phenotype across compounded incompleteness, inaccuracy, and lack of timeliness. The CCW phenotype's robustness to compounded incompleteness owes to its reliance solely on diagnosis codes; other phenotypes with medication and laboratory components faced steeper drops in patients captured when medication and laboratory codes were both incomplete. When it came to induced compounded inaccuracy, the CCW phenotype demonstrated the greatest resistance until 80% of both diagnosis and medication codes were replaced, at which point the Hopkins phenotype became the most robust (Figure S19 in Multimedia Appendix 1). Similarly, compounded lack of timeliness became a bigger problem for phenotypes that rely on time interval for not only diagnosis, but also medication and laboratory codes. Simulating compounded data quality issues has implications for evaluating the fit of phenotype definitions for various data sources and availabilities of multiple data elements. For instance, if a particular data source only has reliable diagnosis codes, then the CCW phenotype would be the most robust. However, it may come with more false positives due to its sole reliance on diagnosis codes.

Our study highlights the importance of understanding the effects of data quality issues on phenotypes, particularly for common diseases such as T2D. The results from our study are novel and can inform how to better identify denominators for a given purpose, which may be beneficial for both research and operational decisions. Although there is no gold standard data set to compare and analyze baseline thresholds of impact on phenotypes, the results of this study can be used by any researcher using real-world data. Additionally, our study describes methods for assessment of multiple data quality issues that can be applied simultaneously in any clinical and health care research setting.

Our study has some limitations. First, the results discussed are based on the EHR data of JHMI over a period of 3 years; hence, generalizability of the findings may be limited to populations like that of this study population. That said, it may also be not applicable to populations of much smaller sample sizes. Second, our study results should be tested against the epidemiological trends and spread of T2D over time. Third, our study did not measure the embedded issues of data quality problems (ie, only simulated or induced data quality issues); understanding and resolving these issues at the beginning of any study can be vital. Fourth, we did not study other data quality dimensions such as concordance and provenance, the assessment of which can be important. Fifth, all T2D phenotypes used in this study relied on diagnosis, medication, and laboratory codes to identify T2D patients; however, some patients may only have clinically diagnosed information in physician notes (ie, unstructured data or free text) that will be missed using structured diagnosis, medication, or laboratory codes. As a result, there may be some false negatives that may have been excluded from our overall study population despite incorporating the most inclusive criteria for the raw data cut. Sixth, we explored compounded data incompleteness, inaccuracy, and timeliness by dropping multiple data types together at the same increments for each data quality domain. However, there could be alternative versions of compounded simulation where diagnosis, medication, and laboratory data qualities are induced at different levels

simultaneously; this may be of interest for future research. Seventh, we assessed the robustness of the phenotypes as they were designated. However, future research could stratify the analysis by the severity of the condition of interest, as severity could impact diagnosis, medication, and laboratory coding behavior and quality. And lastly, our study did not consider any existing data quality thresholds to assess epidemiological impacts on T2D patients using EHRs (eg, comparing national rates of T2D in a neighborhood vs T2D rates identified using EHR data of patients residing in the same neighborhood).

As a result of these observations, we believe that there is a growing need in the United States for a standardized phenotype definition to identify T2D populations while considering the challenges of data quality issues in real-world data, such as from EHRs. The universal phenotype definition should have the ability to integrate features of EHR data while being resistant to common data quality issues. Such a phenotype definition ought to also consider factors that may introduce racial bias and disparities, which eventually may result in health inequities. And lastly, this fundamental definition must also consider integration and interoperability with other data sources, such as claims data, and alignment with existing data interoperability standards. Our research hopes to inspire T2D subject matter experts, at least, to begin conversations toward creating a universal definition for a disease that is extremely common. That said, there is opportunity for additional research that ties issues of data quality with those of phenotypes, data types, and data sources.

Our research provides novel results to understand the effect of data quality issues in data sources like EHRs to identify T2D population-level groups of interest using commonly available T2D phenotype definitions. The study results can inform research or operational efforts using large clinical data repositories to identify T2D populations. Lastly, the study findings can inform efforts to consolidate T2D phenotypes in the near future.

Acknowledgments

The authors want to thank Dr Hsien Yen Chang and Thomas Richards (Johns Hopkins University Center for Population Health Information Technology) and Dr Martin Bishop (Johns Hopkins Hospital Pharmacy Outpatient) for their perspectives and insights into the data sets and related database challenges. We also thank Dr Eva Tseng and Dr Scott Pilla (Johns Hopkins University School of Medicine) for their knowledge on the International Classification of Disease, 10th Revision codes used for the SUPREME-DM (Surveillance, Prevention, and Management of Diabetes Mellitus) phenotype definition, and Dr. Rita R. Kalyani (Johns Hopkins University School of Medicine) for insights into clinical identifiers and definition of Type 2 Diabetes.

Authors' Contributions

All the authors contributed to the conception and design of the study, including material preparation, data collection, and data analyses. All authors critically reviewed and commented on the manuscript and have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table S1 and Figures S1-S21.
[DOCX File, 2561 KB - medinform_v12i1e56734_app1.docx]

References

1. National Diabetes Statistics Report 2020 estimate of diabetes and its burden in the United States. Diabetes Research Institute. 2022. URL: https://diabetesresearch.org/wp-content/uploads/2022/05/national-diabetes-statistics-report-2020.pdf [accessed 2023-07-13]
2. Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 2013 Dec;20(e2):e319-e326. [doi: 10.1136/amiajnl-2013-001952] [Medline: 24026307]
3. Validating type 1 and type 2 diabetes mellitus in the Minisentinel Distributed Database using the Surveillance, Prevention, and Management of Diabetes Mellitus (SUPREME-DM) datalink. Sentinel Initiative. URL: https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_Methods_Validating-Diabetes-Mellitus_MSDD_Using-SUPREME-DM-DataLink.pdf [accessed 2022-06-20]
4. 27 CCW chronic conditions algorithms. Chronic Conditions Data Warehouse. 2022. URL: https://www2.ccwdata.org/documents/10280/19139608/ccw-cond-algo-diabetes.pdf [accessed 2022-04-13]
5. Pacheco J, Thompson W. Type 2 diabetes mellitus. PheKB. 2012. URL: https://phekb.org/phenotype/18 [accessed 2024-07-09]
6. Type 2 diabetes mellitus phenotype definitions. NIH Collaboratory. URL: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Diabetes_Phenotype%20Definition%20Resources%20and%20Recommendations.pdf [accessed 2022-04-13]
7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]

8. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997;4(5):342-355. [doi: 10.1136/jamia.1997.0040342] [Medline: 9292840]
9. Snyder J. Data cleansing: an omission from data analytics coursework. Inf Syst Educ J 2019;17(6):22-29 [FREE Full text]
10. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19(2):212-218. [doi: 10.1136/amiajnl-2011-000439] [Medline: 22101970]
11. Adhikari P, Devang N, Nandini M. Phenotypical classification of type 2 diabetes mellitus. J Endocrinol Diab 2018 Nov;5(6):1-4. [doi: 10.15226/2374-6890/5/6/001119]
12. ICD-9-CM to and from ICD-10-CM and ICD-10-PCS crosswalk or general equivalence mappings. National Bureau of Economic Research. URL: https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings [accessed 2022-04-13]
13. RxNav. National Library of Medicine. URL: https://mor.nlm.nih.gov/RxNav/ [accessed 2024-07-09]
14. NCI Comorbidity Index overview. National Cancer Institute. URL: https://healthcaredelivery.cancer.gov/seermedicare/considerations/comorbidity.html [accessed 2022-04-13]
15. Guide to HbA1c. Diabetes.co.uk. URL: https://www.diabetes.co.uk/what-is-hba1c.html#:~:text=HbA1c%20can%20be%20expressed%20as,is%20measured%20in%20mmol%2Fl [accessed 2022-04-13]
16. The R project for statistical computing. R Foundation. URL: https://www.r-project.org/ [accessed 2022-04-13]

Abbreviations

CCW: Chronic Conditions Data Warehouse
DDC: Durham Diabetes Coalition
EHR: electronic health record
eMERGE: Electronic Medical Records and Genomics
***ICD*:** *International Classification of Diseases*
IRB: institutional review board
JHMI: Johns Hopkins Medical Institution
JHU: Johns Hopkins University
LOINC: Logical Observation Identifiers Names and Codes
SUPREME-DM: Surveillance, Prevention, and Management of Diabetes Mellitus
T2D: type 2 diabetes

Edited by C Lovis; submitted 24.01.24; peer-reviewed by H Li, S Li, Y Wang; revised version received 07.05.24; accepted 08.06.24; published 27.08.24.

Please cite as:
Sood PD, Liu S, Lehmann H, Kharrazi H
Assessing the Effect of Electronic Health Record Data Quality on Identifying Patients With Type 2 Diabetes: Cross-Sectional Study
JMIR Med Inform 2024;12:e56734
URL: https://medinform.jmir.org/2024/1/e56734
doi:10.2196/56734

Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study

Ippei Akiya[1], MSc; Takuma Ishihara[2], PhD; Keiichi Yamamoto[3], PhD

1

2

3

Corresponding Author:
Ippei Akiya, MSc

Abstract

Background: Synthetic patient data (SPD) generation for survival analysis in oncology trials holds significant potential for accelerating clinical development. Various machine learning methods, including classification and regression trees (CART), random forest (RF), Bayesian network (BN), and conditional tabular generative adversarial network (CTGAN), have been used for this purpose, but their performance in reflecting actual patient survival data remains under investigation.

Objective: The aim of this study was to determine the most suitable SPD generation method for oncology trials, specifically focusing on both progression-free survival (PFS) and overall survival (OS), which are the primary evaluation end points in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods, including CART, RF, BN, and the CTGAN, and the performance of each method was evaluated.

Methods: Using multiple clinical trial data sets, 1000 data sets were generated by using each method for each clinical trial data set and evaluated as follows: (1) median survival time (MST) of PFS and OS; (2) hazard ratio distance (HRD), which indicates the similarity between the actual survival function and a synthetic survival function; and (3) visual analysis of Kaplan-Meier (KM) plots. Each method's ability to mimic the statistical properties of real patient data was evaluated from these multiple angles.

Results: In most simulation cases, CART demonstrated the high percentages of MSTs for synthetic data falling within the 95% CI range of the MST of the actual data. These percentages ranged from 88.8% to 98.0% for PFS and from 60.8% to 96.1% for OS. In the evaluation of HRD, CART revealed that HRD values were concentrated at approximately 0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS. CART demonstrated better similarity than RF, in that CART caused overfitting and RF (a kind of ensemble learning approach) prevented it. In SPD generation, the statistical properties close to the actual data should be the focus, not a well-generalized prediction model. Both the BN and CTGAN methods cannot accurately reflect the statistical properties of the actual data because small data sets are not suitable.

Conclusions: As a method for generating SPD for survival data from small data sets, such as clinical trial data, CART demonstrated to be the most effective method compared to RF, BN, and CTGAN. Additionally, it is possible to improve CART-based generation methods by incorporating feature engineering and other methods in future work.

(JMIR Med Inform 2024;12:e55118) doi:10.2196/55118

KEYWORDS

oncology clinical trial; survival analysis; synthetic patient data; machine learning; SPD; simulation

Introduction

When submitting an application for the approval of a new pharmaceutical product to health authorities, it is imperative to demonstrate its efficacy and safety through multiple clinical trials. However, 86% of these trials encounter difficulties meeting the targeted number of subjects within the designated recruitment period, often leading to extensions of the trial duration or completion of the trial without reaching the target number of subjects [1-3]. The challenge of patient recruitment not only delays the submission of regulatory applications but also hinders the timely provision of effective treatment to patients, which consequently contributes to increased development costs and the escalation of drug prices and potentially exacerbates the strain on health care financing.

In recent years, the use of real-world data (RWD) has emerged as a potential solution for addressing these issues. The Food and Drug Administration has also released draft guidelines [4], garnering attention on the application of RWD as an external control arm in clinical trials [5,6]. Furthermore, it has been reported that it is possible to optimize eligibility using RWD and machine learning, thereby increasing the number of eligible subjects that can be included [7].

In addition to these approaches, we hypothesize that it is possible to generate synthetic patient data (SPD) from control arm data in past clinical trials and use it to establish a control arm for a new clinical trial. The use of SPD, an emerging research approach in the health care research field [8-17], involves the generation of fictitious individual patient-level data from real data, which possess statistical properties similar to those of actual data. This approach is anticipated to facilitate health care research while addressing data privacy concerns [14,18-21].

Regarding its application in clinical trials, concerns have been raised about the feasibility of generating SPDs with statistical properties similar to those of actual data due to the relatively smaller volume of clinical trial data compared to RWD, such as electronic health records or registry data. However, previous studies [22-25] have reported the successful generation of SPDs with statistical properties generally comparable to the actual data, although there are certain limitations. Additionally, with the expansion of clinical trial data-sharing platforms such as ClinicalStudyDataRequest.com, Project Data Sphere, and Vivli, acquiring subject-level clinical trial data has become more accessible. Consequently, advancements in research on the utility of SPD and the expansion of clinical trial data-sharing platforms are expected to have potential applications in clinical trials.

Our focus lies in the application of this technology in oncology clinical trials that evaluate popular efficacy end points such as overall survival (OS) and progression-free survival (PFS)–related survival functions and median survival time (MST) [26]. In previous studies on SPD, there has been a notable emphasis on reporting patient background data and single–time point data [22-25]. However, research focusing specifically on the relationship between SPD and survival data remains relatively insufficient [27].

As the first step in examining our hypothesis that the use of SPD can be beneficial in accelerating health care research, the aim of this study was to determine the most suitable SPD generation method for oncology trials, specifically focusing on both OS and PFS, which are set as the primary evaluation end points in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods: classification and regression trees (CART) [28], random forest (RF) [29], Bayesian network (BN) [30], and the conditional tabular generative adversarial network (CTGAN) approach [31], and the performance of each method was evaluated.

Methods

Overview

To generate the SPD, subject-level clinical trial data were obtained from Project Data Sphere for the following 4 clinical trials (Table 1): (1) each had a different cancer type, (2) included control arm data, (3) contained both OS and PFS data, and (4) had a ready data format for analysis.

Table . List of selected oncology clinical trials in this study.

ClinicalTrials.gov ID	Titles	Phase	Cancer type	Intervention for the control arm	Subjects in the control arm, n
NCT00119613	A Randomized, Double-Blind, Placebo-Controlled Study of Subjects With Previously Untreated Extensive-Stage Small-Cell Lung Cancer (SCLC) Treated With Platinum Plus Etoposide Chemotherapy With or Without Darbepoetin Alfa.	III	Small cell lung cancer	Placebo	232
NCT00339183	A Randomized, Multicenter Phase 3 Study to Compare the Efficacy of Panitumumab in Combination With Chemotherapy to the Efficacy of Chemotherapy Alone in Patients With Previously Treated Metastatic Colorectal Cancer.	III	Metastatic colorectal cancer	FOLFIRI[a] Alone	476
NCT00339183	A Phase 3 Randomized Trial of Chemotherapy With or Without Panitumumab in Patients With Metastatic and/or Recurrent Squamous Cell Carcinoma of the Head and Neck (SCCHN).	III	Recurrent or metastatic (or both) head and neck cancer	Cisplatin and 5-fluorouracil	260
NCT00703326	A Multicenter, Multinational, Randomized, Double-Blind, Phase III Study of IMC-1121B Plus Docetaxel versus Placebo Plus Docetaxel in Previously Untreated Patients With HER2-Negative, Unresectable, Locally-Recurrent or Metastatic Breast Cancer.	III	Breast cancer	Placebo and docetaxel	382

[a]FOLFIRI: panitumumab plus fluorouracil, leucovorin, and irinotecan.

Preparation of the Training Data Set

The patient data for the control arm contained within each trial data set were extracted and used as the actual data for the training data set. The selection of variables in the training data set aimed to include as many variables related to the subjects' background as possible, excluding variables concerning tests and evaluations conducted during the trials. Furthermore, variables that had the same value were excluded, even if they were related to the subjects' background (Multimedia Appendices 1-4).

Generation of Synthetic Data

The SPDs in this study were generated using the following 4 methods:

1. CART: the synthpop package (version 1.8) in R (The R Foundation) was used, specifying the cart method for the syn function's method argument.
2. RF: the synthpop package (version 1.8) in R was used, specifying the Ranger method for the syn function's method argument.
3. BN: the bnlearn package (version 4.9) in R was used to conduct structural learning through the score-based algorithm hill-climbing, followed by parameter estimation using the bn.fit function. The default maximum likelihood estimator was used for parameter estimation.
4. CTGAN: the CTGANSynthesizer module included in the Python package sdv (version 1.3) was used.

In all these generation methods, to ensure the absence of conflicting data regarding the relationship between PFS and OS, constraints were set to ensure that the values of PFS and OS were greater than zero and that PFS was less than or equal to OS. Specific individual patient data in the generated SPD, which did not meet these constraints, were excluded, and new individual patient data were regenerated. The SPDs were generated in a manner that equaled the number of subject-level data to the record count in the actual data.

To ensure the reproducibility of SPD generation, 1000 random numbers were generated as seed values using the Mersenne Twister algorithm. The same seed value set was used for all generation methods.

Statistical Analysis

Histogram

Histograms were created to visually inspect the distributions of the MST of the synthetic data (MSTS) for PFS and OS for the 1000 SPD data sets generated by each method. The histograms also included the MST of the actual data (MSTA) as a vertical line and the range of its 95% CI as a rectangular background. For PFS and OS, a higher percentage of MSTS covered by the 95% CI of the MSTA was determined to indicate a greater level of reliability for the generation method.

Evaluation of Similarity

A hazard ratio (HR) of 1 signifies that the 2 survival functions are entirely identical. Thus, the closer the HR is to 1, the more similar the 2 survival functions are. Accordingly, based on the following calculation formula, the HR distance (HRD) for PFS and OS from the SPD and the actual data were computed and evaluated:

$$HRD=1-abs(HR-1)$$

Kaplan-Meier Plot

In the evaluation of similarity, the SPD that showed the highest HRD value was considered the best case, and the SPD with the lowest HRD value was considered the worst case. Three groups of Kaplan-Meier (KM) plots were created, including the actual data, the best case, and the worst case for each SPD generation method. The best case and worst case for each SPD generation method in both PFS and OS were compared to actual survival by using the log rank test. Multiple comparisons were not performed, nor were *P* values adjusted because controlling for the type I error rate does not affect the conclusions of this study.

Since the purpose of this study was to evaluate the method of generating SPD that closely resemble actual survival data, it might be unnecessary to calculate a *P* value that indicates a significant difference from actual survival, but the *P* value was calculated in this study from the viewpoint that if a significant difference is also observed in the best-case, that method should not be adopted.

All analyses and data generation were performed using R (version 4.3.1; The R Foundation) and Python (version 3.10; Python Software Foundation).

Ethical Considerations

Ethical review was not needed for this simulation study for methodology comparison. All actual clinical trial data sets obtained from Project Data Sphere were used in accordance with relevant guidelines and regulations when the clinical trials were conducted.

Results

Figure 1 shows a histogram of the MSTS for PFS in the NCT00703326 trial. Using CART, RF, and BN, most of the generated MSTS values were within the 95% CI of the MSTA. In contrast, when CTGAN was used, SPD generation resulted in a widened variance in the distribution of MSTS. For the MSTS of PFS in the other 3 trials, RF exhibited a shift in the distribution of the MSTS, shortening the survival period, while BN displayed a shift in the distribution and prolonged the survival period. Similar trends to Figure 1 were observed for CART and CTGAN (Multimedia Appendices 5-7).

Figure 2 displays a histogram of the MSTS for OS in the NCT00460265 trial. The divergence from the PFS findings is that the MSTS of RF was more frequently included within the 95% CI of the MSTA, with similar results observed in other trials (Multimedia Appendices 8-10). In other aspects, similar findings were obtained as with the PFS.

Table 2 presents the number and proportion of the generated MSTS values included within the 95% CI of the MSTA for each trial and each method. In the case of CART for PFS, a high percentage ranging from 88.8% to 98.1% was exhibited for all trials. However, the OS ranged from 60.8% to 96.1%, with some trials displaying a lower percentage than the PFS.

Figure 1. Histogram of the median survival time of the synthetic data for progression-free survival in the NCT00703326 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; MST: median survival time; RF: random forest.

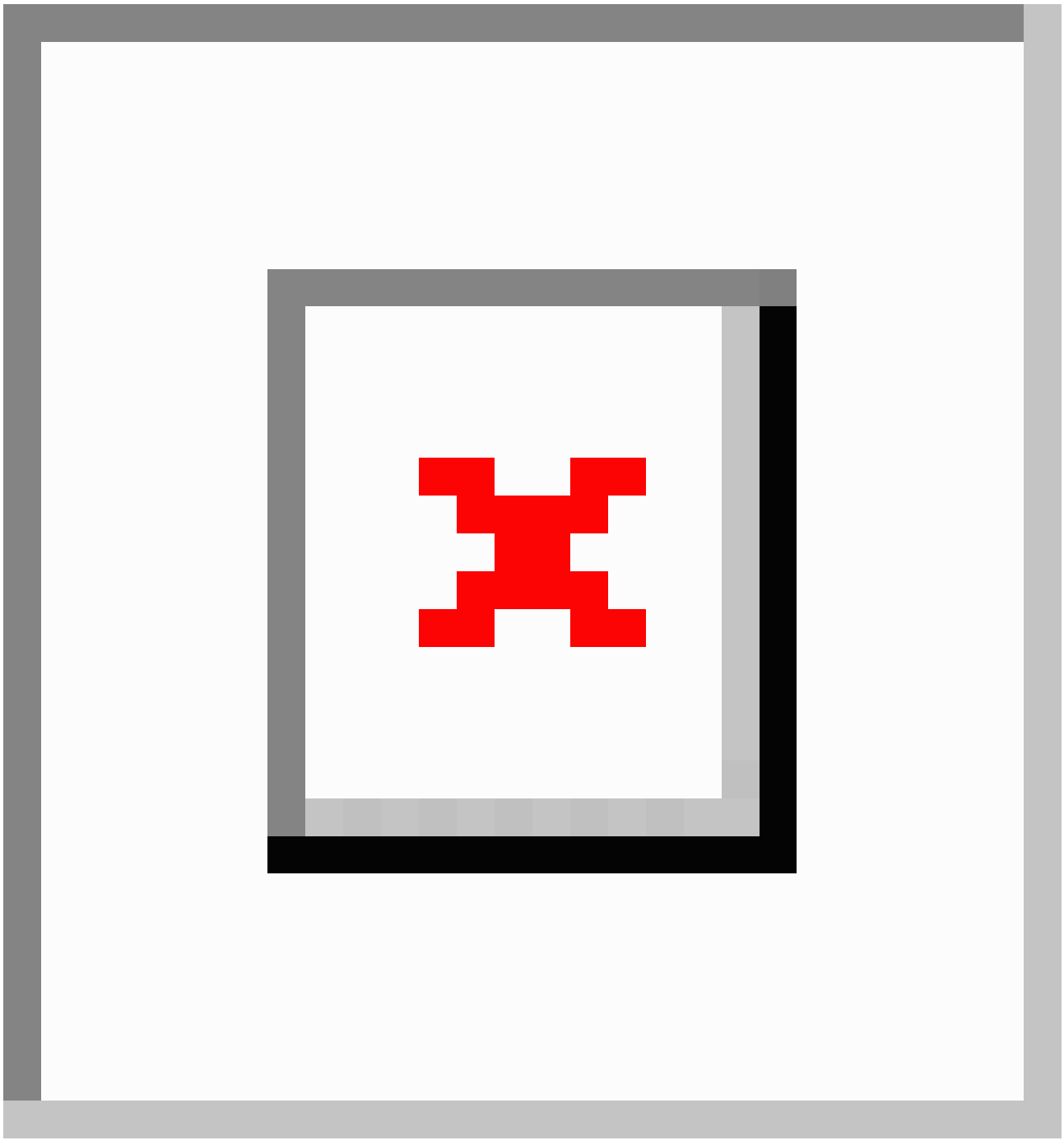


Figure 2. Histogram of the median survival time of the synthetic data of overall survival in the NCT00460265 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; MST: median survival time; RF: random forest.

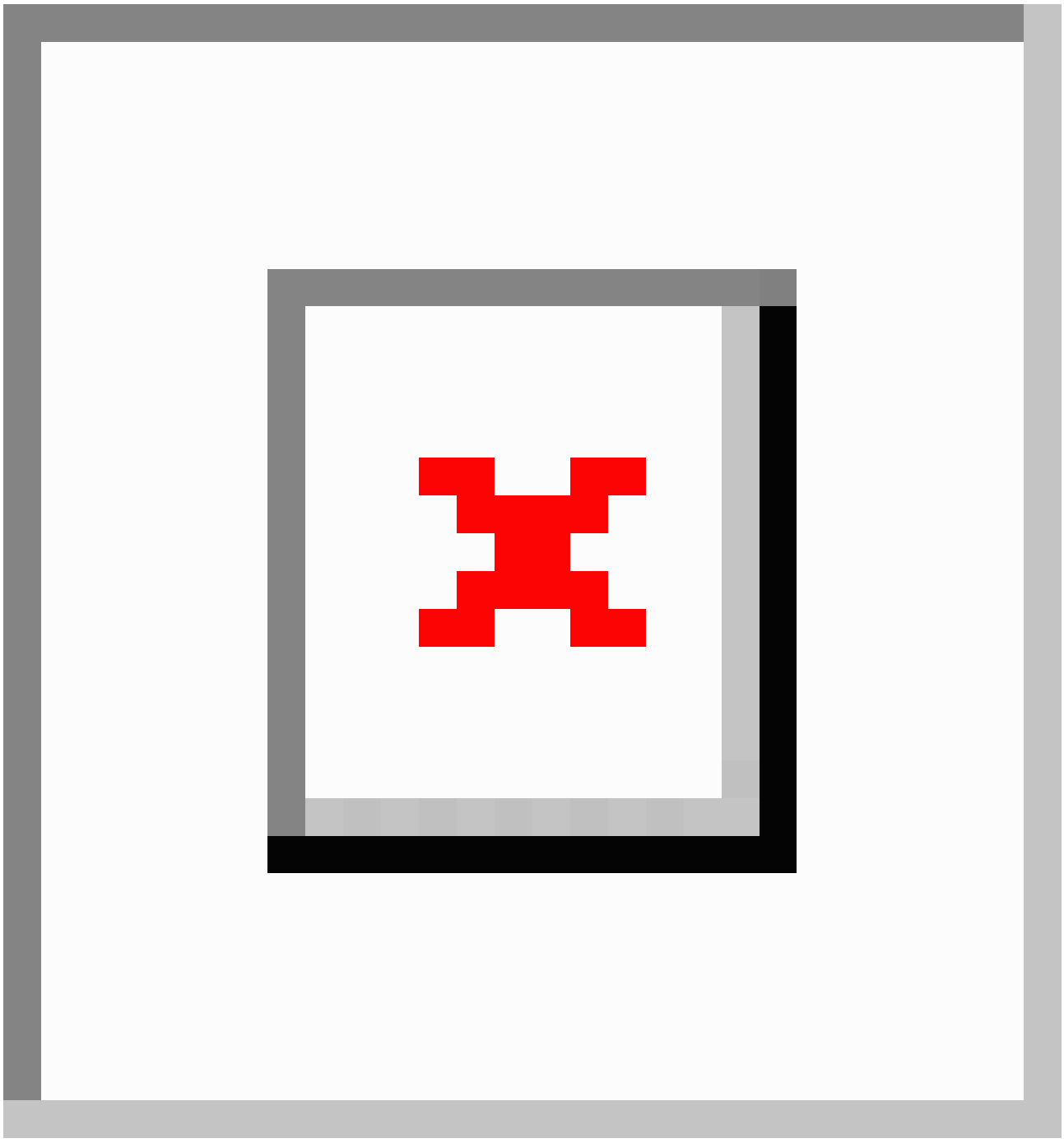


Table . The number and proportion of median survival times of the synthetic data (MSTSs) falling within the 95% CI of the median survival time of the actual data (MSTA).

		ClinicalTrials.gov ID			
		NCT00119613	NCT00339183	NCT00460265	NCT00703326
Progression-free survival					
	MSTA (95% CI)	169 (163-183)	155 (121-168)	133 (121-167)	424 (380-504)
MSTSs, n (%)					
	CART[a] (n=1000)	981 (98.1)	888 (88.8)	955 (95.5)	918 (91.8)
	RF[b] (n=1000)	693 (69.3)	248 (24.8)	426 (42.6)	919 (91.9)
	BN[c] (n=1000)	10 (1.0)	0 (0.0)	37 (3.7)	976 (97.6)
	CTGAN[d] (n=1000)	65 (6.5)	378 (37.8)	322 (32.2)	254 (25.5)
Overall survival					
	MSTA (95% CI)	276 (259-303)	361 (319-393)	286 (255-357)	1452 (1417-1507)
MSTSs, n (%)					
	CART (n=1000)	831 (83.1)	608 (60.8)	719 (71.9)	961 (96.1)
	RF (n=1000)	757 (75.7)	697 (69.7)	980 (98.0)	599 (59.9)
	BN (n=1000)	0 (0.0)	0 (0.0)	0 (0.0)	622 (62.2)
	CTGAN (n=1000)	72 (7.2)	155 (15.5)	197 (19.7)	81 (8.5)

[a]CART: classification and regression tree.

[b]RF: random forest.

[c]BN: Bayesian network.

[d]CTGAN: conditional tabular generative adversarial network.

For RF, a high proportion of 91.9% was observed for PFS in the NCT00703326 trial and 98.0% for OS in the NCT00460265 trial, whereas in other cases, the proportion for RF was not as high as that for CART.

In the case of BN, proportions of 97.6% and 62.2% were observed for PFS and OS, respectively, in the NCT00703326 trial, but in the other 3 trials, BN showed an extremely low percentage ranging from proportion ranging from 0.0% to 3.7%.

CTGAN showed a low proportion ranging from 6.5% to 37.8% for both PFS and OS in all trials.

Figure 3 shows the KM plot for PFS in the NCT00703326 trial. The best-case curves of CART and RF were similar to the actual data curve. In contrast, for BN and CTGAN, even the best-case curves deviated from the actual data curve. In other trials, some SPD did not show a similar trend. However, at least for the best-case scenarios of CART and RF, the generated synthetic survival curves closely resembled the actual survival curve (Multimedia Appendices 11-13).

Figure 4 displays the KM plot for OS in the NCT00460265 trial. Similar to the KM plots for PFS, the best-case curves of CART and RF resembled the actual data curve, whereas those of BN and CTGAN deviated from the actual data curve. These trends were also observed in other trials (Multimedia Appendices 14-16).

Figures 5 and 6 present box plots of the HRD. When using CART, the HRD values for both PFS and OS in all trials were concentrated at approximately 0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS.

Figure 3. Kaplan-Meier plots for progression-free survival in the NCT00703326 trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

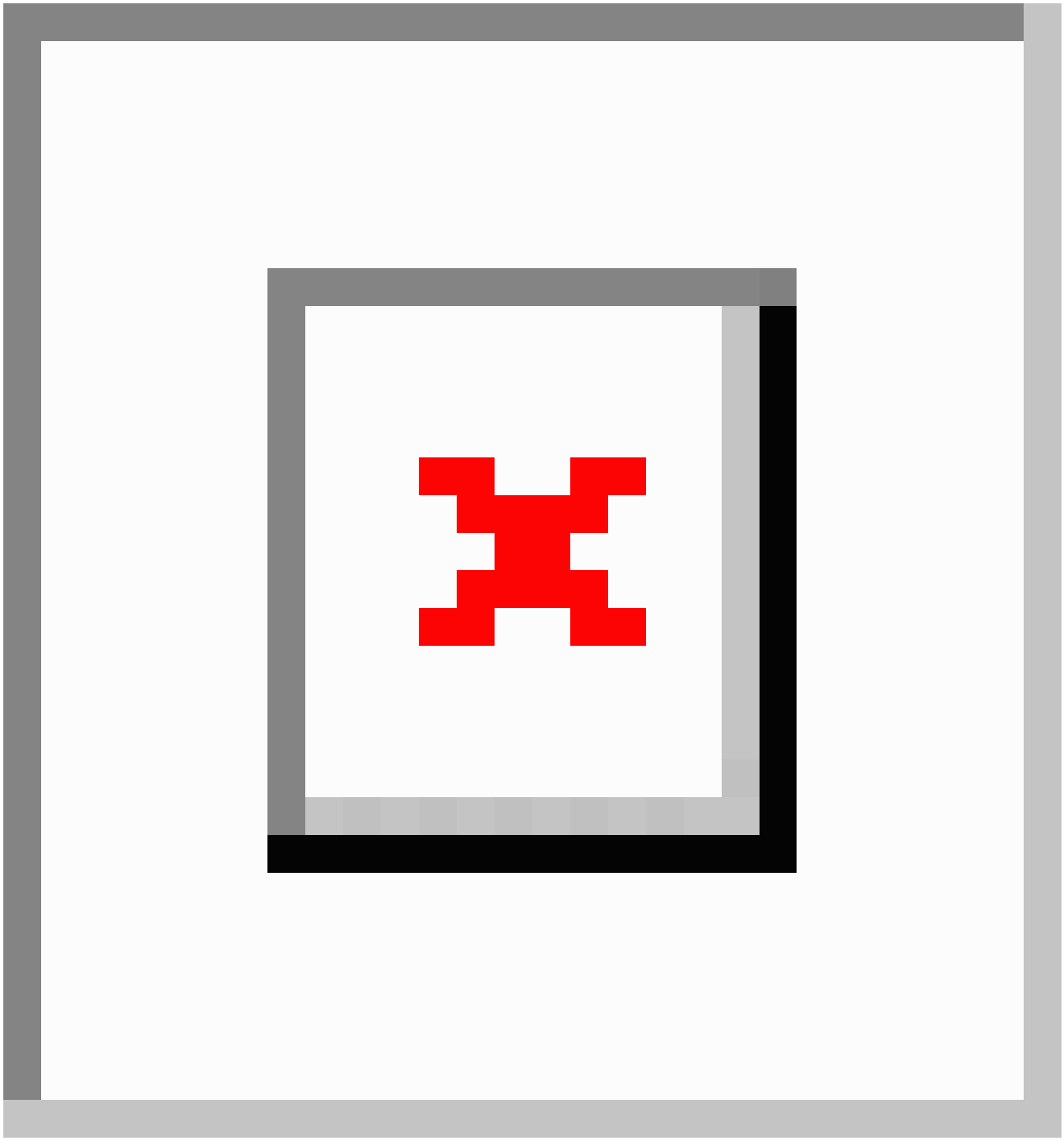


Figure 4. Kaplan-Meier plots for overall survival in the NCT00460265 trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

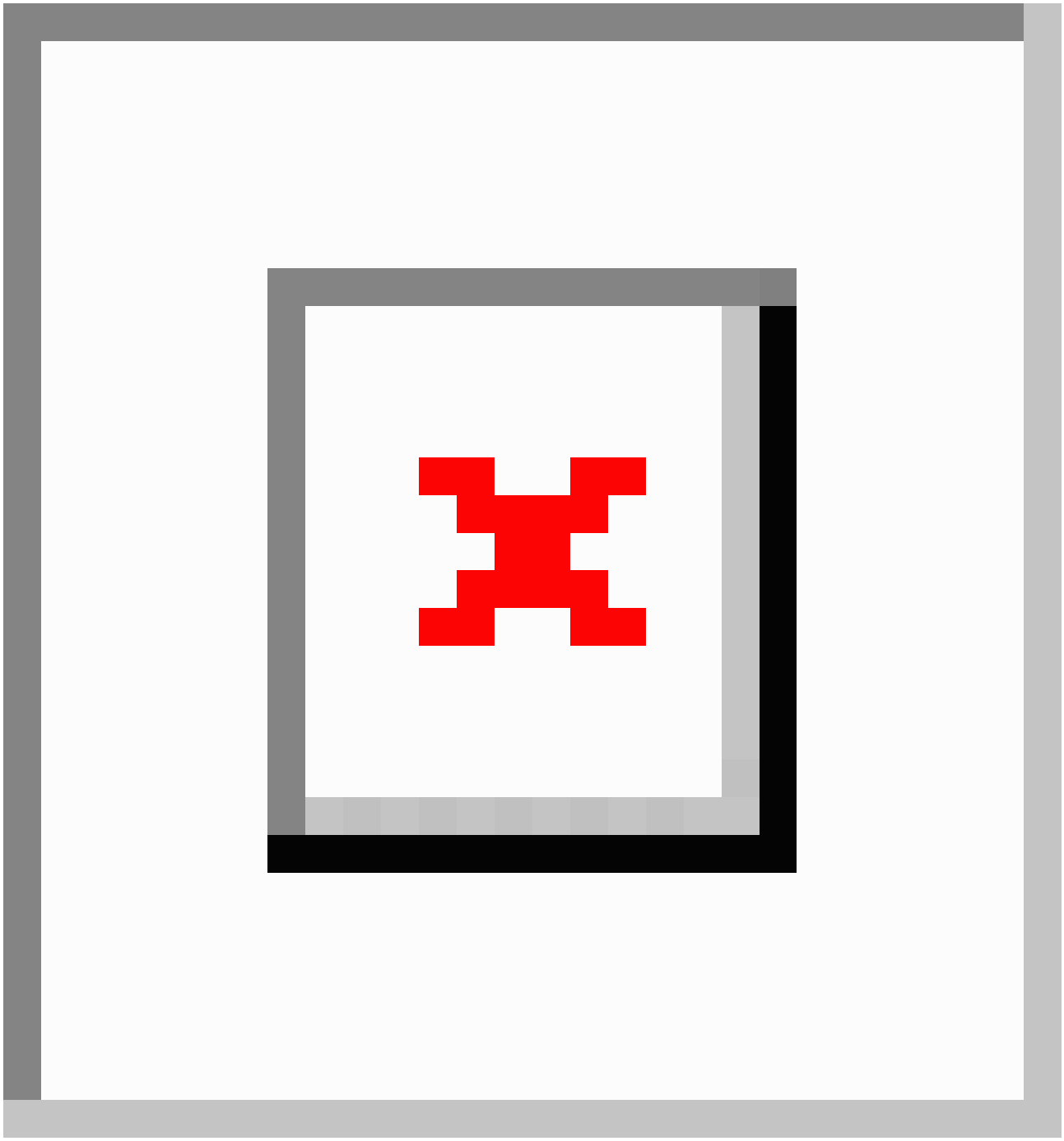


Figure 5. Box plot of progression-free survival hazard ratio distance (HRD) for each method and clinical trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

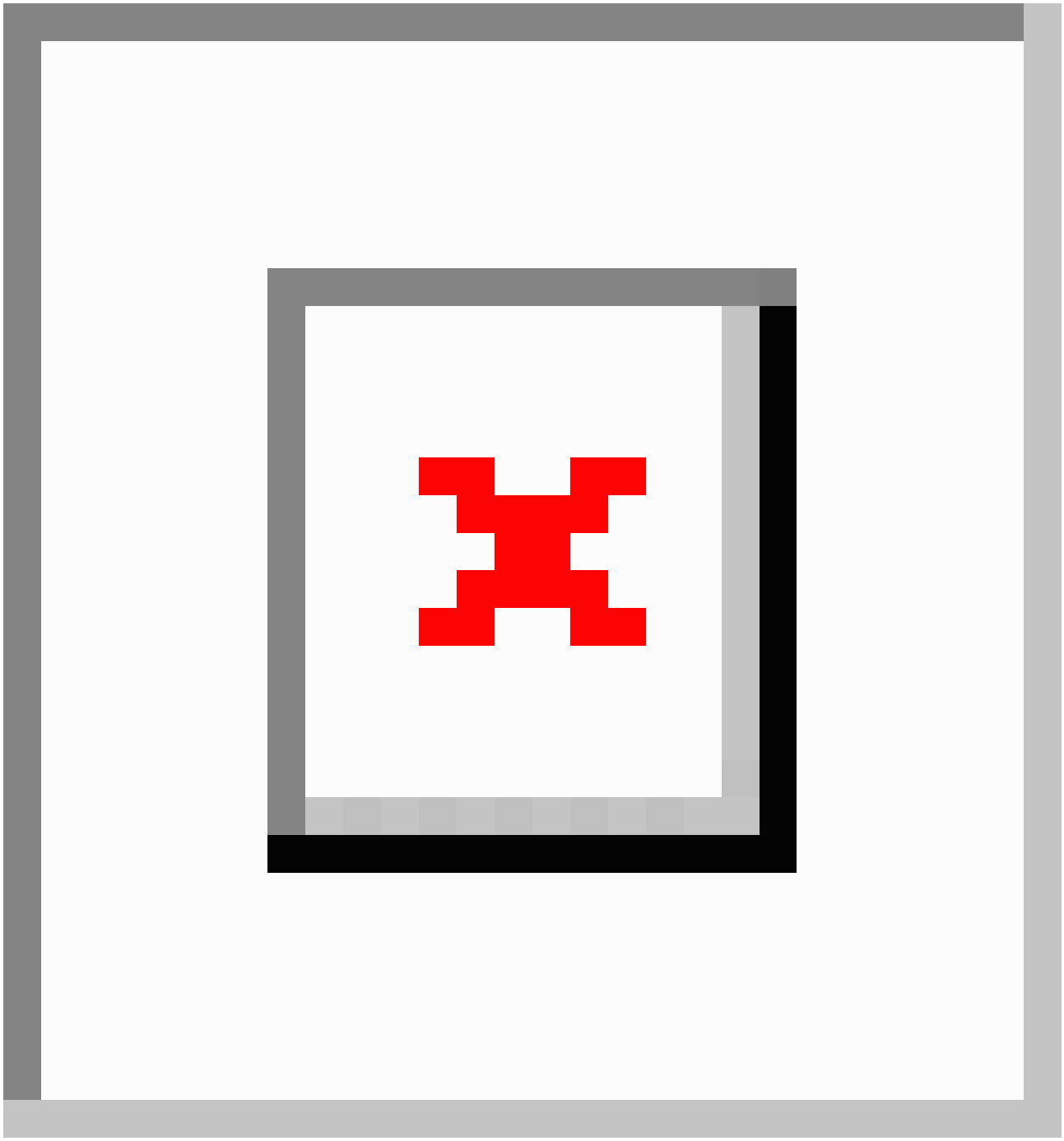
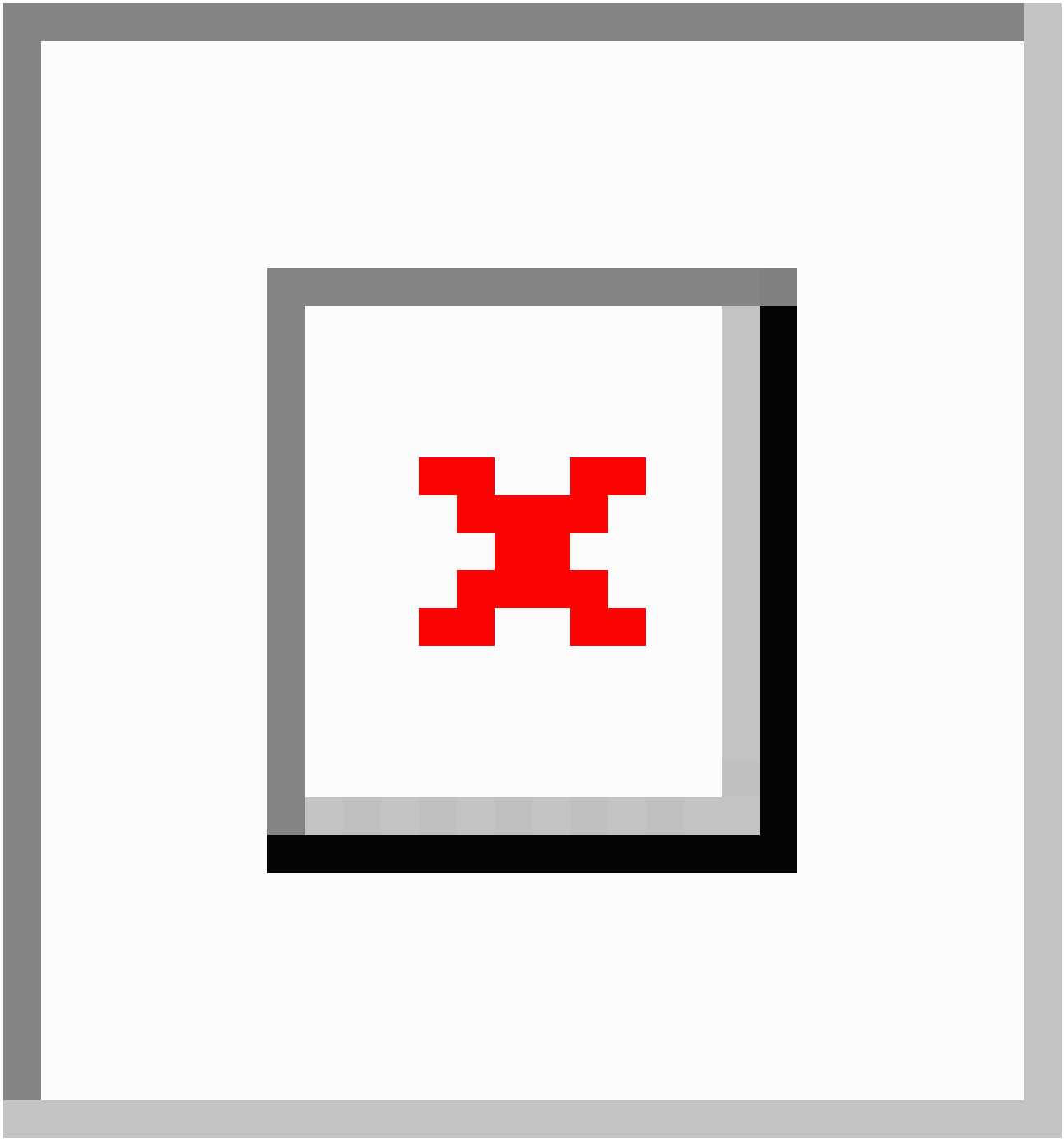


Figure 6. Box plot of overall survival hazard ratio distance (HRD) for each method and clinical trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.



Discussion

Regarding the survival SPD, CART often yielded better results than the other methods in evaluations using MST, HRD, and visual analysis via KM plots. Given the crucial importance of the hazard ratio and MST as end points in oncology trials [26], demonstrating the utility of both of these evaluation metrics is essential. Therefore, using CART for generating survival SPD was suggested as a beneficial approach.

While both CART and RF generally yielded preferable results across all trials, they share the common characteristic of using tree models. RF, with its use of the bootstrap method for resampling and constructing tree models for ensemble learning, is known to prevent overfitting. In general, in terms of constructing machine learning models with high generalization performance, RF performs better than CART. However, CART is prone to overfitting as the layers of the tree become deeper [32]. Although RF is considered a superior method for constructing high–generalization-performance machine learning models, the results from Table 2 and the KM plots in this study suggest that CART is a better approach than RF. This discrepancy might be due to differing views on what is a higher performance between the machine learning prediction model and SPD. In the machine learning prediction model, it is important to prevent overfitting and reduce bias; however, SPD is expected to match its statistical properties with actual data.

Thus, in the case of SPD, the overfitting suppression mechanism possessed by RF might have resulted in inferiority to that of CART from the perspective of improving similarity.

In the case of using BN, the percentage of MSTSs falling within the 95% CI of MSTAs was 0% for the PFS of the NCT00339183 trial, and for OS, this phenomenon also occurred in the NCT00119613, NCT00339183, and NCT0046265 trials. This implies that the SPD failed to accurately reflect the statistical properties of the actual data. Conversely, a high value of 97.6% was observed for the PFS in the NCT00703326 trial. The reason for this discrepancy could not be determined on the basis of the results of this study. Tucker et al [24] reported that they could generate data highly similar to actual data when using BN for the generation of SPD, which differs from our findings. One notable difference is that while Tucker et al [24] used a large-scale actual data set of 27.5 million patients for their study, this study used only a few hundred patients for training data. This difference likely had a significant impact on the accuracy of the SPD generation model, resulting in conflicting results. However, the SPD generated by BN were not distributed in the direction of shortening PFS or OS. Thus, this would not be harmful when the SPD generated by BN is used as a more conservative control arm in clinical trials.

Using CTGAN, the percentage of the MSTSs falling within the 95% CI of the actual data was low, indicating low performance associated with the generation of SPD that reflect the statistical properties of the actual data. However, Krenmayr et al [23] reported favorable performance results when using the same generative adversarial network (GAN)–based methods and RWD. The differences between their study and our study were as follows: their study did not include SPD on survival time or generate multiple SPD data sets from the same actual data, and there was a large amount of individual patient data in their study. In particular, focusing on the amount of individual patient data, the number of patients in each trial included in this study was relatively small, with the NCT00119613 trial having 232 patients, the NCT00339183 trial having 476 patients, the NCT0046265 trial having 260 patients, and the NCT00703326 trial having 382 patients, while the trial conducted by Krenmayr et al [23] had 500 or more patients. GAN-based methods using deep neural networks are known to perform poorly with small amounts of data [25,33]. In this study, although the NCT00339183 trial had the largest number of individual patient data, the best case of CTGAN for NCT00339183 produced a KM plot similar to the actual data, suggesting that a larger data set yields better results. Thus, there is no contradiction. Another characteristic of using CTGAN in this study was the larger variance in the estimated MSTSs, as indicated in Figures 1 and 2. Goncalves et al [34] showed that using MC-MedGAN, a GAN-based method, to generate an SPD from small data resulted in a large SD of the data utility metrics, leading to results with larger variance, similar to those of this study. Therefore, it is extremely challenging to generate useful SPD by applying GAN-based methods to small data sets, such as clinical trial data.

When generating SPDs for survival data and using them as a certain arm in a clinical trial, it is important to verify that the statistical properties closely match those of the actual data with the MST and the hazard ratio with the actual data being close to 1. Based on our results, we conclude that CART, which can concentrate the MSTSs within the range of 95% CI of MSTAs and approximately 0.9 for HRD, is an efficient method for generating SPD that meets the abovementioned conditions. However, even when using CART, slight variations were observed in the MSTSs, and some cases fell outside the 95% CI of the MSTAs, as revealed by our results. Therefore, for practical use, it is necessary to verify that the MSTSs are included in the 95% CI of the MSTAs and that both are close in value. It is also necessary to verify whether the HRD of the actual data and the SPD are close to 1 and then decide whether to adopt the generated SPD. Hence, the generation process must be repeated until an acceptable SPD is obtained. There may also be a need to use statistical methods to match characteristics between the SPD and the actual treatment arm in clinical trials.

In this study, even the most useful CART method produced SPDs that did not meet the requirements of MST and HRD. We expect that this issue will be addressed by incorporating feature engineering, such as dimension reduction, imputing missing values, derived variable creation, and other processing. Additionally, in clinical research, as subgroup analyses are frequently conducted, it is necessary to improve the generation method to reflect the statistical properties of the actual data even when the data are divided into subgroups under certain conditions. Moreover, from the perspective of data privacy, it is essential to incorporate approaches to prevent data reidentification into the generation method [35].

In conclusion, as a method for generating SPD for survival data from small data sets, such as clinical trial data, CART is the most effective method for generating SPD that meet the 2 conditions of having an MSTSs close to the MSTAs and an HRD close to 1. However, as SPD might be generated, which do not meet these 2 conditions, it is necessary to incorporate mechanisms to improve a CART-based generation method in future studies. Overcoming these challenges would make it possible to reduce the recruitment period and costs of clinical trial participants to ≥50% in comparative trials of new drug development against existing therapeutic drugs. This approach could accelerate clinical development, similar to the use of RWD.

Acknowledgments

We would like to express our gratitude to Project Data Sphere, the platform that provided the necessary data for this study, and to the clinical trial data providers Amgen and Eli Lilly.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables used to generate synthetic patient data from the NCT00119613 trial.
[DOCX File, 48 KB - medinform_v12i1e55118_app1.docx]

Multimedia Appendix 2
Variables used to generate synthetic patient data from the NCT00339183 trial.
[DOCX File, 48 KB - medinform_v12i1e55118_app2.docx]

Multimedia Appendix 3
Variables used to generate synthetic patient data from the NCT00460265 trial.
[DOCX File, 47 KB - medinform_v12i1e55118_app3.docx]

Multimedia Appendix 4
Variables used for generating synthetic patient data from the NCT00703326 trial.
[DOCX File, 48 KB - medinform_v12i1e55118_app4.docx]

Multimedia Appendix 5
Histogram of the median survival time of the synthetic data for progression-free survival in the NCT00119613 trial. The dashed vertical line represents the median survival time for the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 202 KB - medinform_v12i1e55118_app5.docx]

Multimedia Appendix 6
Histogram of the median survival times for the synthetic data for progression-free survival in the NCT00339183 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 192 KB - medinform_v12i1e55118_app6.docx]

Multimedia Appendix 7
Histogram of the median survival times of the synthetic data for progression-free survival in the NCT00460265 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 187 KB - medinform_v12i1e55118_app7.docx]

Multimedia Appendix 8
Histogram of the median survival times of the synthetic data for overall survival in the NCT00119613 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 186 KB - medinform_v12i1e55118_app8.docx]

Multimedia Appendix 9
Histogram of the median survival times of the synthetic data for overall survival in the NCT00339183 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 195 KB - medinform_v12i1e55118_app9.docx]

Multimedia Appendix 10
Histogram of the median survival times of the synthetic data for overall survival in the NCT00703326 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 188 KB - medinform_v12i1e55118_app10.docx]

Multimedia Appendix 11
Kaplan-Meier plots for progression-free survival in the NCT00119613 trial.
[DOCX File, 215 KB - medinform_v12i1e55118_app11.docx]

Multimedia Appendix 12
Kaplan-Meier plots for progression-free survival in the NCT00339183 trial.

[DOCX File, 229 KB - medinform_v12i1e55118_app12.docx]

Multimedia Appendix 13
Kaplan-Meier plots for progression-free survival in the NCT00460265 trial.
[DOCX File, 218 KB - medinform_v12i1e55118_app13.docx]

Multimedia Appendix 14
Kaplan-Meier plots for overall survival in the NCT00119613 trial.
[DOCX File, 229 KB - medinform_v12i1e55118_app14.docx]

Multimedia Appendix 15
Kaplan-Meier plots for overall survival in the NCT00339183 trial.
[DOCX File, 252 KB - medinform_v12i1e55118_app15.docx]

Multimedia Appendix 16
Kaplan-Meier plots for overall survival in the NCT00703326 trial.
[DOCX File, 265 KB - medinform_v12i1e55118_app16.docx]

References

1. Huang GD, Bull J, Johnston McKee K, et al. Clinical trials recruitment planning: a proposed framework from the clinical trials transformation initiative. Contemp Clin Trials 2018 Mar;66:74-79. [doi: 10.1016/j.cct.2018.01.003] [Medline: 29330082]
2. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018 Sep;11:156-164. [doi: 10.1016/j.conctc.2018.08.001] [Medline: 30112460]
3. Treweek S, Lockhart P, Pitkethly M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open 2013;3(2):e002360. [doi: 10.1136/bmjopen-2012-002360] [Medline: 23396504]
4. Considerations for the design and conduct of externally controlled trials for drug and biological products. Guidance for industry. US Food and Drug Administration. 2023. URL: https://www.fda.gov/media/164960/download [accessed 2024-06-04]
5. Yap TA, Jacobs I, Baumfeld Andre E, Lee LJ, Beaupre D, Azoulay L. Application of real-world data to external control groups in oncology clinical trial drug development. Front Oncol 2021;11:695936. [doi: 10.3389/fonc.2021.695936] [Medline: 35070951]
6. Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of real‐world evidence to drive drug development strategy and inform clinical trial design. Clin Pharmacol Ther 2022 Jan;111(1):77-89. [doi: 10.1002/cpt.2480] [Medline: 34839524]
7. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature 2021 Apr;592(7855):629-633. [doi: 10.1038/s41586-021-03430-5] [Medline: 33828294]
8. Azizi Z, Lindner S, Shiba Y, et al. A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. Sci Rep 2023 Jul 17;13(1):11540. [doi: 10.1038/s41598-023-38457-3] [Medline: 37460705]
9. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071. [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
10. Kaur D, Sobiesk M, Patil S, et al. Application of Bayesian networks to generate synthetic health data. J Am Med Inform Assoc 2021 Mar 18;28(4):801-811. [doi: 10.1093/jamia/ocaa303] [Medline: 33367620]
11. Mavrogenis AF, Scarlat MM. Artificial intelligence publications: synthetic data, patients, and papers. Int Orthop 2023 Jun;47(6):1395-1396. [doi: 10.1007/s00264-023-05830-w] [Medline: 37162553]
12. Meeker D, Kallem C, Heras Y, Garcia S, Thompson C. Case report: evaluation of an open-source synthetic data platform for simulation studies. JAMIA Open 2022 Oct;5(3):ac067. [doi: 10.1093/jamiaopen/ooac067] [Medline: 35958672]
13. Brownstein JS, Chu S, Marathe A, et al. Combining participatory influenza surveillance with modeling and forecasting: three alternative approaches. JMIR Public Health Surveill 2017 Nov 1;3(4):e83. [doi: 10.2196/publichealth.7344] [Medline: 29092812]
14. Guillaudeux M, Rousseau O, Petot J, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. NPJ Digit Med 2023 Mar 10;6(1):37. [doi: 10.1038/s41746-023-00771-5] [Medline: 36899082]
15. El Emam K. Status of synthetic data generation for structured health data. JCO Clin Cancer Inform 2023 Jun;7:e2300071. [doi: 10.1200/CCI.23.00071] [Medline: 37390378]
16. D'Amico S, Dall'Olio D, Sala C, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. JCO Clin Cancer Inform 2023 Jun;7:e2300021. [doi: 10.1200/CCI.23.00021] [Medline: 37390377]
17. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health 2023 Jan;2(1):e0000082. [doi: 10.1371/journal.pdig.0000082] [Medline: 36812604]

18. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 2023 Oct 9;6(1):186. [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]
19. Ursin G, Sen S, Mottu JM, Nygård M. Protecting privacy in large datasets—first we assess the risk; then we fuzzy the data. Cancer Epidemiol Biomarkers Prev 2017 Aug 1;26(8):1219-1224. [doi: 10.1158/1055-9965.EPI-17-0172] [Medline: 28754793]
20. Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of supervised machine learning using synthetic data in health care: model to preserve privacy for data sharing. JMIR Med Inform 2020 Jul 20;8(7):e18910. [doi: 10.2196/18910] [Medline: 32501278]
21. Summers C, Griffiths F, Cave J, Panesar A. Understanding the security and privacy concerns about the use of identifiable health data in the context of the COVID-19 pandemic: survey study of public attitudes toward COVID-19 and data-sharing. JMIR Form Res 2022 Jul 7;6(7):e29337. [doi: 10.2196/29337] [Medline: 35609306]
22. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, GOING-FWD Collaborators. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open 2021 Apr 16;11(4):e043497. [doi: 10.1136/bmjopen-2020-043497] [Medline: 33863713]
23. Krenmayr L, Frank R, Drobig C, et al. GANerAid: realistic synthetic patient data for clinical trials. Inform Med Unlocked 2022;35:101118. [doi: 10.1016/j.imu.2022.101118]
24. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med 2020 Nov 9;3(1):147. [doi: 10.1038/s41746-020-00353-9] [Medline: 33299100]
25. Santos M. How to generate real-world synthetic data with CTGAN. Medium. 2023. URL: https://medium.com/towards-data-science/how-to-generate-real-world-synthetic-data-with-ctgan-af41b4d60fde [accessed 2024-06-04]
26. Ben-Aharon O, Magnezi R, Leshno M, Goldstein DA. Median survival or mean survival: which measure is the most appropriate for patients, physicians, and policymakers? Oncologist 2019 Nov;24(11):1469-1478. [doi: 10.1634/theoncologist.2019-0175] [Medline: 31320502]
27. Smith A, Lambert PC, Rutherford MJ. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. BMC Med Res Methodol 2022 Jun 23;22(1):176. [doi: 10.1186/s12874-022-01654-1] [Medline: 35739465]
28. Breiman L, editor. Classification and Regression Trees: Chapman and Hall; 1998.
29. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
30. Pearl J. Bayesian networks: a model of self-activated memory for evidential reasoning. Presented at: Proceedings of the 7th Conference of the Cognitive Science Society; Aug 15 to 17, 1985; Irvine, CA URL: https://ftp.cs.ucla.edu/tech-report/198_-reports/850017.pdf [accessed 2024-06-04]
31. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. arXiv. Preprint posted online on Jul 1, 2019. [doi: 10.48550/arXiv.1907.00503]
32. Hayes T, Usami S, Jacobucci R, McArdle JJ. Using classification and regression trees (CART) and random forests to analyze attrition: results from two simulations. Psychol Aging 2015 Dec;30(4):911-929. [doi: 10.1037/pag0000046] [Medline: 26389526]
33. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. arXiv. Preprint posted online on Jun 10, 2016 URL: http://arxiv.org/abs/1606.03498 [accessed 2024-06-04] [doi: 10.48550/arXiv.1606.03498]
34. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 2020 May 7;20(1):108. [doi: 10.1186/s12874-020-00977-1] [Medline: 32381039]
35. El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data: O'Reilly Media; 2020.

Abbreviations

BN: Bayesian network
CART: classification and regression trees
CTGAN: conditional tabular generative adversarial network
GAN: generative adversarial network
HR: hazard ratio
HRD: hazard ratio distance
KM: Kaplan Meier
MST: median survival time
MSTA: median survival time of actual data
MSTS: median survival time of synthetic data
OS: overall survival
PFD: progression-free survival
RF: random forest
RWD: real-world data

SPD: synthetic patient data

Edited by C Lovis; submitted 03.12.23; peer-reviewed by D Hu, J Song; revised version received 06.04.24; accepted 08.05.24; published 18.06.24.

Please cite as:
Akiya I, Ishihara T, Yamamoto K
Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study
JMIR Med Inform 2024;12:e55118
URL: https://medinform.jmir.org/2024/1/e55118
doi:10.2196/55118

The Effect of an Electronic Medical Record–Based Clinical Decision Support System on Adherence to Clinical Protocols in Inflammatory Bowel Disease Care: Interrupted Time Series Study

Reed Taylor Sutton[1], MSc; Kaitlyn Delaney Chappell[1], MSc; David Pincock[2], MBA, MScIB; Daniel Sadowski[1], MD; Daniel C Baumgart[1], MBA, MD, PhD; Karen Ivy Kroeker[1], MSc, MD

1

2

Corresponding Author:
Karen Ivy Kroeker, MSc, MD

Abstract

Background: Clinical decision support systems (CDSSs) embedded in electronic medical records (EMRs), also called electronic health records, have the potential to improve the adoption of clinical guidelines. The University of Alberta Inflammatory Bowel Disease (IBD) Group developed a CDSS for patients with IBD who might be experiencing disease flare and deployed it within a clinical information system in 2 continuous time periods.

Objective: This study aims to evaluate the impact of the IBD CDSS on the adherence of health care providers (ie, physicians and nurses) to institutionally agreed clinical management protocols.

Methods: A 2-period interrupted time series (ITS) design, comparing adherence to a clinical flare management protocol during outpatient visits before and after the CDSS implementation, was used. Each interruption was initiated with user training and a memo with instructions for use. A group of 7 physicians, 1 nurse practitioner, and 4 nurses were invited to use the CDSS. In total, 31,726 flare encounters were extracted from the clinical information system database, and 9217 of them were manually screened for inclusion. Each data point in the ITS analysis corresponded to 1 month of individual patient encounters, with a total of 18 months of data (9 before and 9 after interruption) for each period. The study was designed in accordance with the Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) guidelines for health informatics evaluations.

Results: Following manual screening, 623 flare encounters were confirmed and designated for ITS analysis. The CDSS was activated in 198 of 623 encounters, most commonly in cases where the primary visit reason was a suspected IBD flare. In Implementation Period 1, before-and-after analysis demonstrates an increase in documentation of clinical scores from 3.5% to 24.1% (P<.001), with a statistically significant level change in ITS analysis (P=.03). In Implementation Period 2, the before-and-after analysis showed further increases in the ordering of acute disease flare lab tests (47.6% to 65.8%; P<.001), including the biomarker fecal calprotectin (27.9% to 37.3%; P=.03) and stool culture testing (54.6% to 66.9%; P=.005); the latter is a test used to distinguish a flare from an infectious disease. There were no significant slope or level changes in ITS analyses in Implementation Period 2. The overall provider adoption rate was moderate at approximately 25%, with greater adoption by nurse providers (used in 30.5% of flare encounters) compared to physicians (used in 6.7% of flare encounters).

Conclusions: This is one of the first studies to investigate the implementation of a CDSS for IBD, designed with a leading EMR software (Epic Systems), providing initial evidence of an improvement over routine care. Several areas for future research were identified, notably the effect of CDSSs on outcomes and how to design a CDSS with greater utility for physicians. CDSSs for IBD should also be evaluated on a larger scale; this can be facilitated by regional and national centralized EMR systems.

(JMIR Med Inform 2024;12:e55314) doi:10.2196/55314

KEYWORDS

decision support system; clinical; electronic medical records; electronic health records; health record; medical record; EHR; EHRs; EMR; EMRs; decision support; CDSS; internal medicine; gastroenterology; gastrointestinal; implementation science; implementation; time series; interrupted time series analysis; inflammatory bowel disease; IBD; bowel; adherence; flare; flares; steroid; steroids; standardized care; nurse; clinical practice guidelines; chart; electronic chart; electronic medical chart

Introduction

Limited or delayed adoption of professional society–developed clinical care guidelines into practice is a common problem in medicine [1,2]. In 2007, researchers estimated that it took 17 years on average for only 14% of published evidence in guidelines to be translated into clinical practice [3,4]. One purported reason is that clinical guidelines by themselves are not actionable, as they largely describe what to do but not how to do it [5,6].

Clinical decision support systems (CDSSs) are tools that can be used to support provider decision-making. A CDSS uses clinical, patient, and other health information to supply providers with recommendations to assist in a variety of aspects of care, including diagnosis, treatment, and management [7,8]. Recent systematic reviews suggest that the use of CDSSs in clinical settings can improve practitioner performance in relation to adherence to best practice guidelines [7,9].

There are several demonstrated gaps in the adoption of professional society clinical care guidelines and best practices for inflammatory bowel disease (IBD). These include practices in medication management, preventative care, and bone health [10,11]. The University of Alberta IBD outpatient clinic (Edmonton, Alberta, Canada) has previously developed and implemented several clinical care pathways to consolidate best practices for IBD [10,12]. To further increase adoption, a clinical decision support (CDS) project was undertaken to integrate the pathways into the local electronic medical record (EMR). There are thousands of CDS projects built and deployed within commercial EMRs [13,14], yet there are few published evaluations of EMR-based CDSSsfor IBD [15,16]. Consequently, the objective of this pilot study was to evaluate the effectiveness and provider acceptance of an EMR-integrated CDSS in the context of IBD.

Methods

Ethical Considerations

This study received approval from the University of Alberta Health Research Ethics Board (Pro00083538). A waiver of informed consent was also approved as part of our study by the Health Research Ethics Board.

Organizational Setting

The study was conducted in the Comprehensive Academic Outpatient Center at the University of Alberta Hospital, which provides care for patients with IBD in the Greater Edmonton region as well as rural and remote communities across Alberta, Canada. It also serves a small number of patients with IBD from Saskatchewan, Northwest Territories, and British Columbia.

System Details and System in Use

The clinic's preexisting system was an enterprise EMR based on the 2014 version of Epic EMR (Epic Systems), which was being used for outpatient medical care in Edmonton, Alberta. This system was customized and branded locally as eCLINICIAN. Medication lists, allergies, and health problems are recorded and shared between users as part of clinical documentation, order entry, and planning. The system was implemented for gastroenterology outpatient care in March 2014.

As Epic is a general-purpose EMR, it includes built-in CDS functionality. For example,this includes generic functionality, such as alerting users when duplicate orders exist. More specialty-specific CDS features are often customized at the request and guidance of end users.

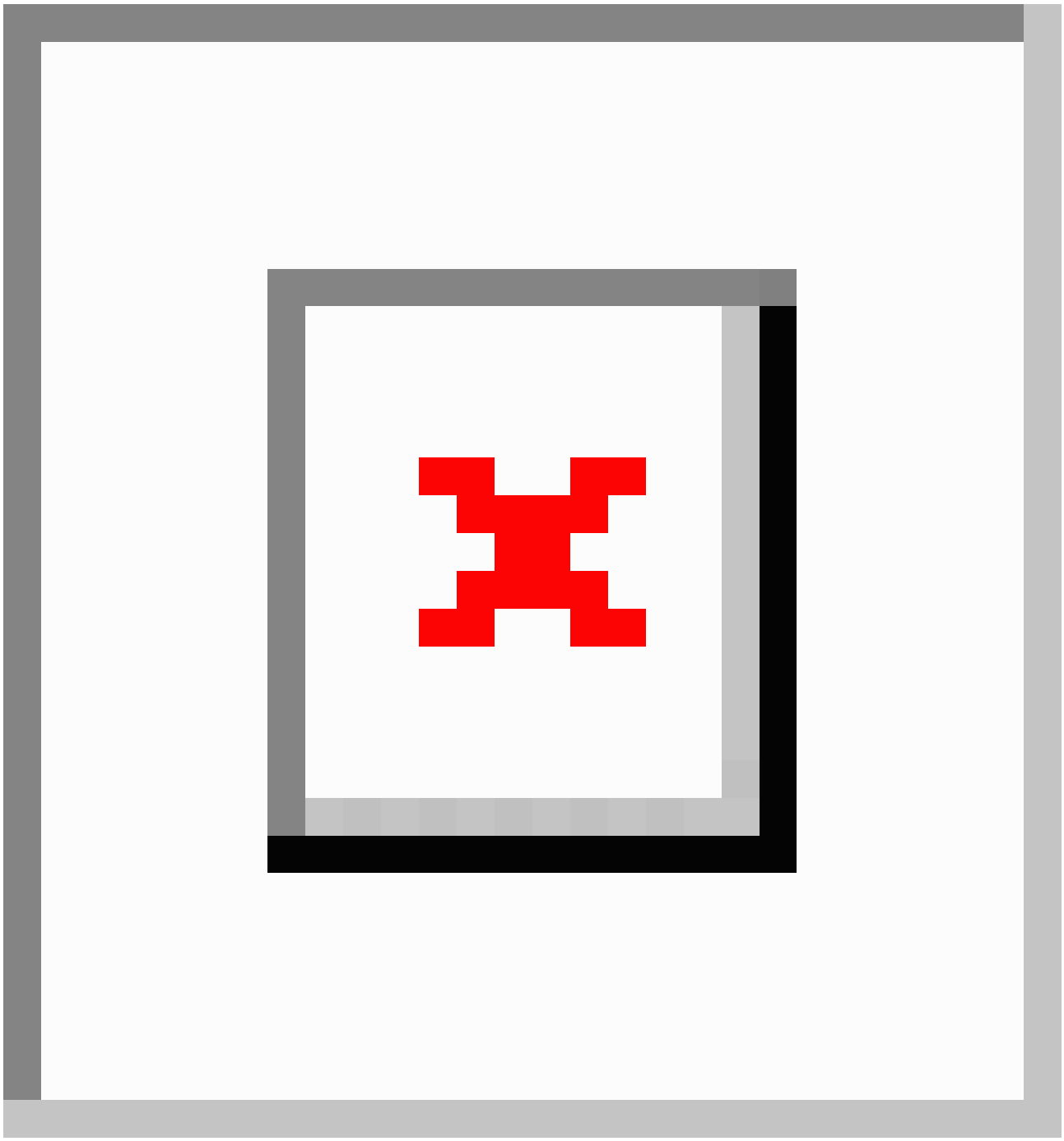
Functionality can be administered through a number of tools, including those referred to by Epic as "Flowsheets" (documentation tables), "Best Practice Advisories (BPAs)" (alerts) [17], and "SmartSets" (ie, grouping of orders and clinical content) [18].

These tools, particularly BPAs and SmartSets, are clinical data and test result driven; they can be triggered by unique combinations of provider characteristics, patient demographics, test results, clinical problems, as well as current and requested medications.

System Interruption and Intervention

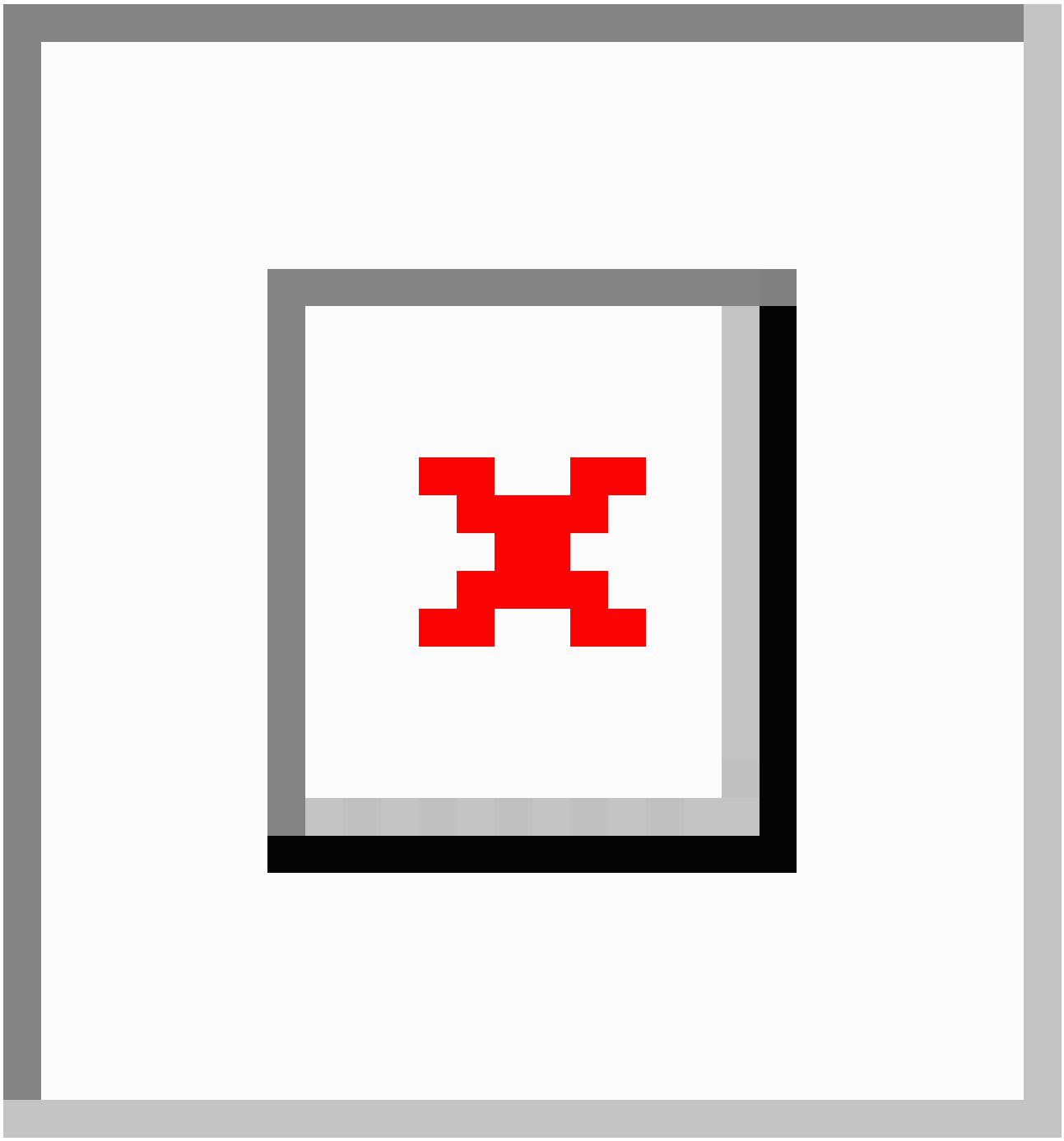
The system interruption and intervention uses BPA appearing in the clinician's navigator workflow. The BPA is triggered by the existence of IBD in the patient problem list or visit diagnosis fields. The BPA (Figure 1) prompts the clinician to complete clinical symptom indices—modified Harvey Bradshaw Index (mHBI) [19] for Crohn disease or partial Mayo (pMayo) score [20] for ulcerative colitis—for the encounter. If the score is indicative of a disease flare, the BPA instructs the user to activate a corresponding SmartSet.

Figure 1. Snapshot of the inflammatory bowel disease (IBD) flare clinical decision support system, showing the initial Best Practice Advisory. Best Practice Advisories act as alerts that present targeted patient-specific guidance to users. They can be active (disruptive pop-ups) or passive (navigation workflow) and can link to actions such as placing orders, order sets, initiating a care plan, or sending a message. This alert appeared passively in the providers' workflow navigation whenever IBD was in the patient problem list.



The SmartSet offers ordering and printing of appropriate lab panels, stool cultures, and other investigations, including imaging, procedures, and medication prescriptions. All recommendations were designed to be consistent with established IBD care guidelines and the flare protocol for the clinic. For example, during a flare encounter, the IBD flare lab panel and fecal calprotectin (FCP) tests are automatically selected for ordering (they can still be deselected by the provider). A snapshot of the SmartSet portion of the CDSS is shown in Figure 2.

Figure 2. Snapshot of the inflammatory bowel disease (IBD) flare clinical decision support system, showing the SmartSet, after activation by Best Practice Advisory. Not all sections of the SmartSet are shown, including sections for medications, imaging investigations, billing, and follow-up appointment booking. ALT: alanine transaminase; AST: aspartate aminotransferase; Cl: chloride; CO2: carbon dioxide; ESR: erythrocyte sedimentation rate; K: potassium; Na: sodium; NO DIFF: no differential.



Study Design

The study used a pre- and postimplementation interrupted time series (ITS) design, the interruption being the enhanced CDSS used within the EMR. Each data point represented 1 month of clinical encounters. For each intervention period, there was a total of 18 data points, 9 before and 9 after the intervention. Multimedia Appendix 1 presents an elaboration on the rationale for using an ITS design.

Physicians at the participating clinic were not guaranteed to have outpatient clinics on a weekly basis due to their service rotation; therefore, it was decided to aggregate the data points by month instead of by week. This avoided the potential week-to-week variation and ensured an adequate number of individual patient encounters (IBD flares) for each data point.

The Quality Criteria for ITS Designs checklist was used in the study design and assessment of appropriateness [21], and the Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) guidelines were used for health informatics evaluations [22,23].

Participants

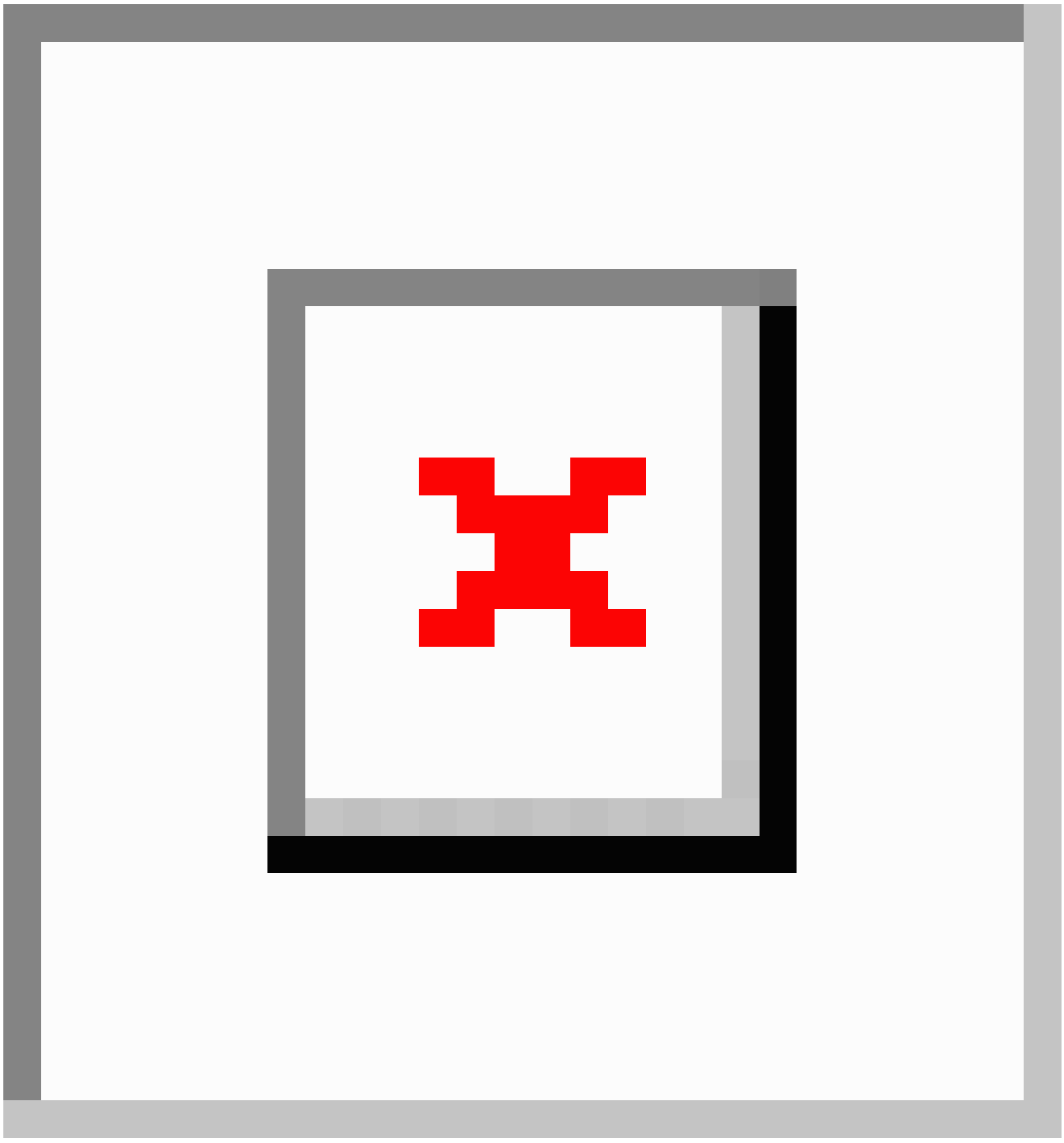
All IBD care providers at the university-based outpatient clinic were included in the study and invited to use the CDSS, including 7 IBD specialist clinicians, 1 IBD nurse practitioner, and 4 IBD specialist nurses. The term "IBD practitioner" will be used to collectively refer to IBD specialists and IBD nurse practitioners.

To be included in the data set, patients had to be under the care of the IBD providers; aged ≥18 years; and diagnosed with either Crohn disease or ulcerative colitis confirmed by imaging, pathology, or endoscopy report. They also had to be experiencing a flare of the disease during the included encounter, as defined by clinical scores (mHBI >5; pMayo >2) or noted symptoms in combination with physician judgment. Only initial encounters in a flare episode spanning multiple encounters were included.

Study Flow

The intervention was implemented and evaluated in 2 continuous periods (Figure 3). First, a pilot version was trialed by IBD nurses (Implementation Period 1), and then, the polished version was implemented across all providers in the division (ie, clinicians, nurse practitioners, and IBD nurses) as Implementation Period 2. The pilot version was trialed beginning in September 2017 and included the following 3 SmartSets available within the BPA, corresponding to different positions along the care path of a patient with flaring IBD: suspected flare, 2 to 4 weeks into the flare, and 16 weeks' postflare assessments. Feedback was gathered informally from providers (Multimedia Appendix 2) to inform further improvement to the CDSS.

Figure 3. Study design diagram of the 2-period interrupted time series design. First, the clinical decision support system (CDSS) was implemented as a limited pilot with inflammatory bowel disease (IBD) nurses (intervention 1), and then, it was fully implemented across all providers (intervention 2). Each data point (abbreviated as D) corresponds to 1 month of clinical encounters by study providers. NP: nurse practitioner.



After collecting feedback from the pilot, further changes were made to the CDSS. Aside from minor modifications to update included lab tests, the most significant change was the consolidation of the 3 separate SmartSets into 1, targeting the "suspected flare," the first step in the care pathway. The activation of the BPA in the initial CDSS was entirely manual and relied on the provider entering a specific visit diagnosis. However, in the full version, the BPA was set to automatically trigger based on the presence of an IBD diagnosis in the patient's problem list. This change was expected to improve the adoption and ease of use of the SmartSet for flare encounters.

The full implementation of the CDSS began on October 10, 2018. An instructional memo with paper-based workflow and educational material was sent to each provider (Multimedia Appendix 3). Over the course of 1 month, each participant was given the opportunity to ask questions about using the system and access to use the system in the sandbox environment. A demonstration of the system was also presented at weekly clinical rounds, with an opportunity to ask questions.

Outcome Measures

Process indicators were used to measure the proportion of adherent IBD practitioner flare encounters. These indicators include completion of clinical scores (mHBI or pMayo);

laboratory testing, such as standard lab panel, FCP, stool cultures, and *Clostridium difficile* toxin (only if diarrhea is present); and of vitamin D or calcium in conjunction with corticosteroid prescription, patient information given and documented, and modification of maintenance therapy. A secondary outcome was the adoption or acceptance of the CDSS measured by application rate (ratio of CDSS uses to CDSS available for activation).

Methods for Data Acquisition and Measurement

Potential encounters in the pre- and postintervention periods were initially identified by querying the eCLINICIAN EMR database for encounters with the included IBD providers, where patients had documentation of IBD in their problem list or diagnosis field (*International Classification of Diseases* coding). A sampling method was used to exclude encounters with specific reasons for visit deemed unlikely to constitute a flare based on exploratory analysis of the data set. Examples of excluded reasons for the visit included "medication refill," "medical insurance coverage," and "review results" (a more detailed description of the sampling method is available in a previous publication [10]). Encounters were then screened manually for inclusion and exclusion eligibility by one of the authors (RTS) and a research assistant.

Data for primary outcome measures were also queried and extracted from the EMR database, in collaboration with the eCLINICIAN reporting team in Alberta Health Services (AHS). The various database codes and IDs as well as the final SQL queries used to extract data are included in the Multimedia Appendix 4.

Methods for Data Analysis

Descriptive statistics were calculated to determine patient characteristics, with data presented as counts and proportions for categorical variables, mean (SD) values for normally distributed continuous variables, and median (IQR) values for nonnormally distributed continuous variables. Proportions were compared by using the Pearson χ^2 test [24].

A segmented regression analysis was performed for each primary outcome variable to determine the level and slope in the preintervention period as well as the change in level and slope in the postintervention period regarding the mean percentage of adherent encounters [25]. Autocorrelation in the residuals was tested using the Durbin-Watson test.

Data analysis was performed using IBM SPSS Statistics (version 23; IBM Corp) and R 3.5.1 (RStudio Inc) [26]. A 95% CI was used in all analyses unless otherwise specified.

Sample Size Determination

The sample size was first calculated for pre- and postimplementation cohorts based on logistic regression (Multimedia Appendix 5). With a power of 0.80 and a type I error set to 5%, the sample size required was approximately 634 for small effects and 145 for medium effects [27]. This assumes equal sample sizes (N) in the comparison groups and an initial proportion of adherence to each guideline component of approximately 70%, chosen based on a recent study by Jackson et al [11]. The sample size was calculated using G*Power 3.2.9.2 [28].

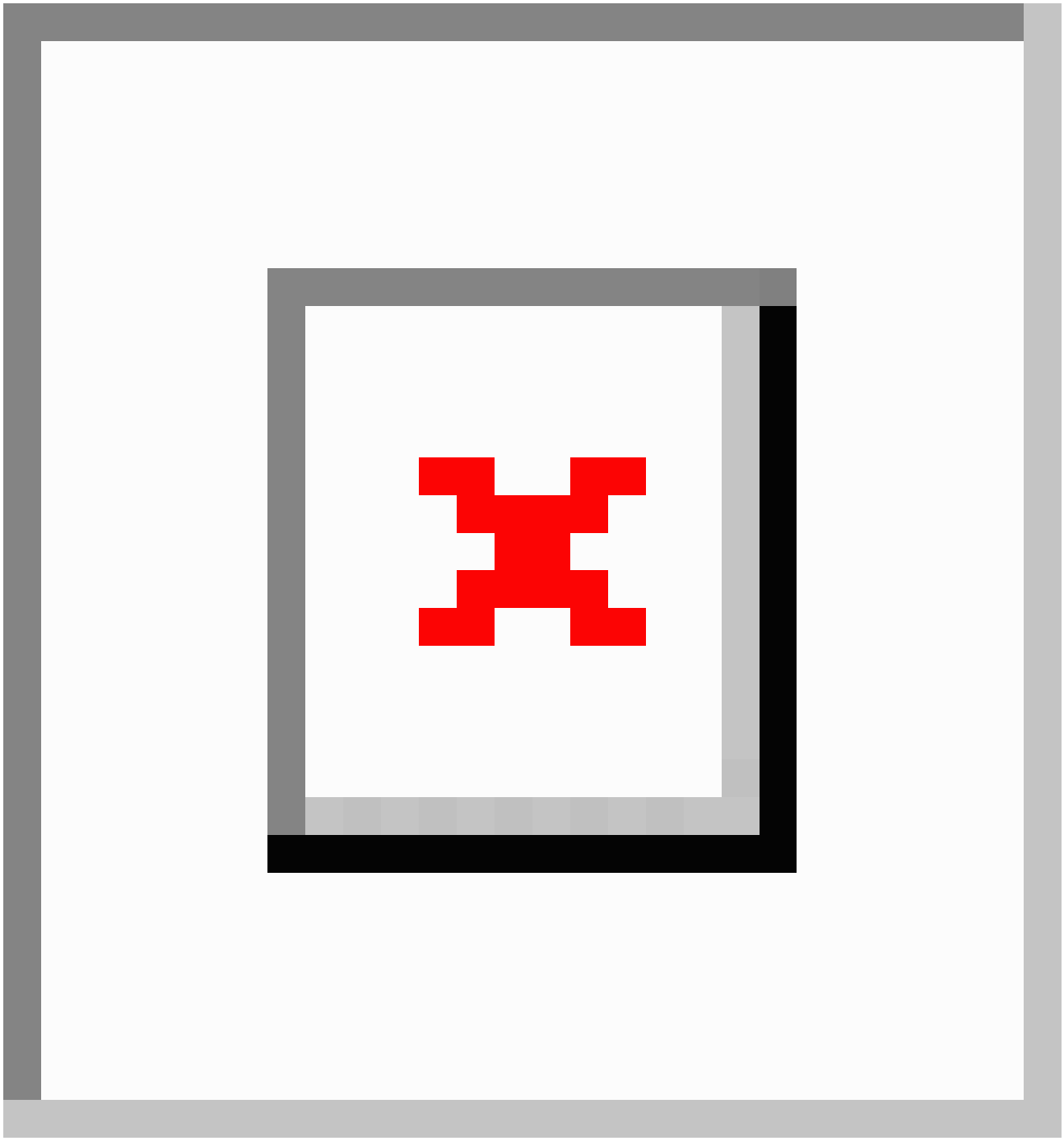
There is no standard method for determining power in time series analyses. However, a simulation-based power calculation displayed that with N=16 (8 data points in the preintervention period and 8 data points in the postintervention period), there is a 70% chance to detect an effect size of 0.5 or more, and over 90% chance to detect an effect size of 1 or more, at an alpha level of .05 [29]. It is also generally recommended in the literature to have over 100 observations per data point [25,30].

Results

Initial Data Set and Preprocessing

Figure 4 shows the study's flow diagram. The complete, extracted data set includes 31,726 encounters from January 1, 2017, to June 30, 2019. When considering only clinic visits (7655), orders (16,485), and telephone (5220) encounter types, the data set totals 29,360 (92.5%) encounters. There was an average of 998 encounters per month, with a minimum of 735 (December 2018) and a maximum of 1202 (May 2017) encounters. Of note, there is an overlap between both implementation periods (Figure 3), and thereby, a number of flare encounters appear in both analyses.

Figure 4. Flow data diagram for data extraction, screening, and analyses. CDSS: clinical decision support system.



Demographics of CDSS-Enabled Encounters

From September 2017 to June 2019, the CDSS was activated a total of 214 times across 214 encounters with 207 patients. Of these, 16 encounters were excluded from analysis due to, upon review, not being used appropriately for a flare or suspected flare encounter with a patient with IBD. This left 198 encounters, which are detailed in Table 1. More detailed demographics of providers using the system are included in Multimedia Appendix 6.

Table . Demographics of users and encounters invoking the inflammatory bowel disease (IBD) flare clinical decision support system.

Demographic variables	Study population (n=198)
Provider characteristics	
Provider type, n (%)	
IBD nurse	172 (86.9)
IBD practitioner	26 (13.1)
Patient characteristics	
Sex, n (%)	
Female	113 (57.1)
Male	85 (42.9)
Age (years), median (IQR)	37.5 (29-49)
Current IBD therapy, n (%)	
None	37 (18.7)
5-aminosalicylic acid only	53 (26.8)
Immunomodulator	18 (9.1)
Biologic monotherapy	59 (29.8)
Biologic combination therapy	31 (15.7)
Encounter characteristics	
Encounter type, n (%)	
Telephone	139 (70.2)
Orders only	32 (16.2)
Clinic visit	27 (13.6)
First encounter diagnosis, n (%)	
None	172 (86.9)
Crohn disease	11 (5.6)
Ulcerative colitis	10 (5.1)
Bloody diarrhea	2 (1.0)
IBD	1 (0.5)
Abdominal bloating	1 (0.5)
Ankylosing spondylitis	1 (0.5)
Visit reason, n (%)	
Suspected IBD flare	113 (57.1)
IBD	39 (19.7)
Disease flare-up	15 (7.6)
None	9 (4.5)
Referral	9 (4.5)
Follow-up	7 (3.5)
Diarrhea	3 (1.5)
Medication change	1 (0.5)
Medication problem	1 (0.5)

Study Findings and Outcome Data

Exploratory Analysis of Adherence to Clinical Protocols

Symptom Documentation

Of 192 patients with clinical scores (mHBI or pMayo) that were applicable (excluding those without pouch or short bowel or those newly diagnosed), 133 (69.3%) had a clinical score completed and documented in their chart at the index dispensation. Of all 198 encounters, 196 (99.0%) had symptoms (ie, pain, number and characteristics of stool, and the presence of blood) documented in the chart by the provider.

Laboratory Investigations

Full flare lab panels, including complete blood count, ferritin, electrolytes, creatinine, albumin, alkaline phosphatase, alanine transaminase, aspartate transaminase, and C-reactive protein (CRP), were ordered for 109/198 (55.1%) patients exactly at the encounter. Including orders up to 1 month prior, full panels were ordered for 183/198 (92.4%) patients. However, 113/198 (57.1%) had at least a partial lab panel, including complete blood count and CRP, ordered at the encounter, and 193/198 (97.5%) had partial lab panels, including complete blood count and CRP ordered up to 1 month prior to the encounter.

FCP was ordered at the encounter for 147/198 (74.2%) patients and within 1 month of the encounter for a further 36/198 (18.2%). This leaves only 15 (7.6%) who had no evaluation of FCP at all. Furthermore, testing for *Clostridium difficile* infection was done in 164/198 (82.8%) patients and for stool cultures in 160/198 (80.8) patients. In 138 patients with liquid stool or diarrhea mentioned in the progress note, 127 (92%) had *Clostridium difficile* testing ordered and 123 (89.1%) had stool cultures ordered.

Provision of Steroid-Sparing Therapy and Osteoprotective Therapy

In this data set, only 12 (6.1%) patients were prescribed steroids at their encounter. Of these, 6 (50%) had maintenance IBD therapy adjusted or added. In contrast, 37 (20%) of the 185 patients who were not prescribed steroids had maintenance therapy adjusted (P=.02 for χ^2).

Vitamin D or calcium supplementation was recommended for 8/12 (67%) patients prescribed steroids and 8/10 (80%) when excluding patients with vitamin D or calcium supplementation already documented in their medication list.

Implementation Period 1: Pilot CDSS Version With IBD Nurses

Implementation Period 1 included data from January 2017 to June 2018 (18 months), where September 2017 and beyond were labelled as the active intervention months (postintervention). Of the total 623 confirmed flare encounters, 502 occurred during Implementation Period 1 (Figure 3). Table 2 compares outcome measures before and after the intervention using chi-square tests. Notably, there was a substantial increase in the proportion of flare encounters with completed clinical scores from 3.5% (8/228) to 24.1% (66/274) post intervention. There was also an increase in the proportion of flare encounters with FCP ordered, from 16.7% (38/228) to 27% (74/274).

Table . Before-and-after analysis of process measures from Implementation Period 1.

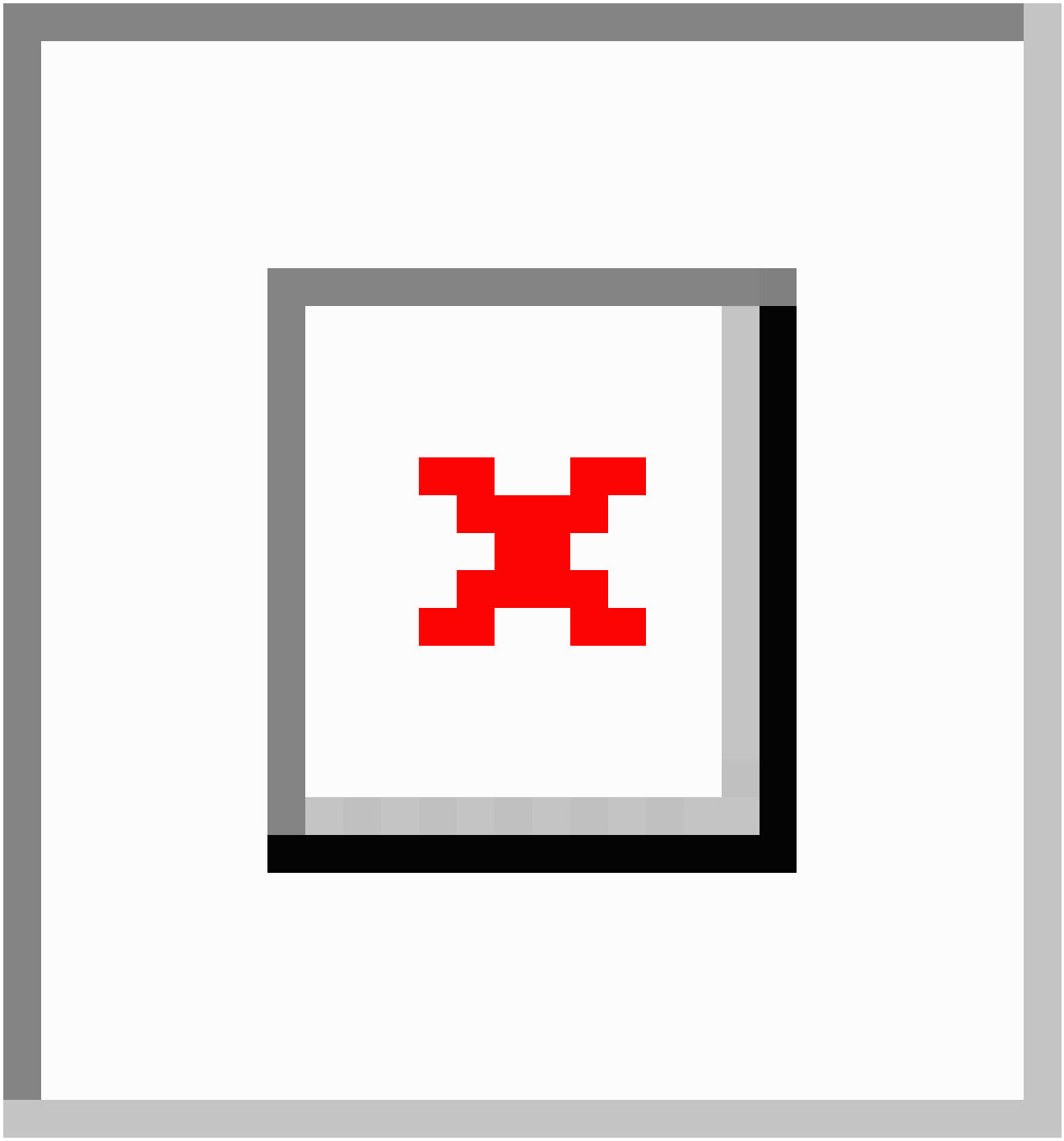
Parameter	Preintervention (n=228), n (%)	Postintervention (n=274), n (%)	*P* value[a]
CDSS[b] activated	0 (0)	66 (24.1)	<.001
Clinical score completed	8 (3.5)	66 (24.1)	<.001
Flare labs ordered	124 (54.4)	132 (48.2)	.33
C-reactive protein ordered	156 (68.4)	178 (65.0)	.56
Fecal calprotectin ordered	38 (16.7)	74 (27.0)	.048
Stool cultures ordered	128 (56.1)	162 (59.1)	.63
Clostridium difficile test ordered	128 (56.1)	172 (62.8)	.29

[a]*P* value of the Pearson chi-square test comparing proportions.

[b]CDSS: clinical decision support system.

ITS analysis was done for outcomes that were significant in the before-and-after analyses (Figure 5). For clinical score completion rates, there was no slope change (estimated β −1.22, 95% CI −4.44 to 2.01; P=.43), but there was a level increase (estimated β 19.0, 95% CI 2.39-35.60; P=.03). For calprotectin testing, there was no slope change (estimated β −2.45, 95% CI −6.21 to 1.32; P=.19) or level change (estimated β 14.77, 95% CI −4.63 to 34.17; P=.13).

Figure 5. Segmented regression for Implementation Period 1 (pilot) of the inflammatory bowel disease flare clinical decision support system on rates of (A) clinical score completion and (B) calprotectin testing.



Implementation Period 2: Full CDSS Implementation With All Providers

Implementation Period 2 included data from January 2018 to June 2019 (18 months), where October 2018 and beyond were postintervention months. Of the total 623 confirmed flare encounters, 492 occurred during Implementation Period 2 (Figure 3). Table 3 compares outcome measures before and after the intervention using chi-square tests. There were increases in the proportion of flare encounters with completed flare labs (109/229, 47.6% to 173/263, 65.8%), CRP ordered (147/229, 64.2% to 207/263, 78.7%), calprotectin ordered (64/229, 27.9% to 98/263, 37.3%), and stool cultures ordered (125/229, 54.6% to 176/263, 66.9%).

Table . Before-and-after analysis of process measures from Implementation Period 2.

Parameter	Preintervention (n=229), n (%)	Postintervention (n=263), n (%)	*P* value[a]
Application of SmartSets	52 (22.7)	72 (27.4)	.23
Clinical score completed	58 (25.3)	75 (28.5)	.43
Flare labs ordered	109 (47.6)	173 (65.8)	<.001
C-reactive protein ordered	147 (64.2)	207 (78.7)	<.001
Fecal calprotectin ordered	64 (27.9)	98 (37.3)	.03
Stool cultures ordered	125 (54.6)	176 (66.9)	.005
Clostridium testing ordered	136 (59.4)	177 (67.3)	.70

[a]*P* value of the Pearson chi-square test comparing proportions.

The ITS analysis for significant outcomes is shown in Figure 6, and accompanying β values for slope change and level change with 95% CIs are shown in Table 4. For Period 2, there were no slope or level increases that reached significance at *P*=.05, although CRP testing and stool culture testing would be significant for a level increase at *P*=.10.

Figure 6. Segmented regression for Implementation Period 2 of the inflammatory bowel disease (IBD) flare clinical decision support system on rates of (A) clinical score completion, (B) flare lab testing, (C) C-reactive protein testing, (D) calprotectin testing, (E) stool culture testing, and (F) *Clostridium difficile* testing.

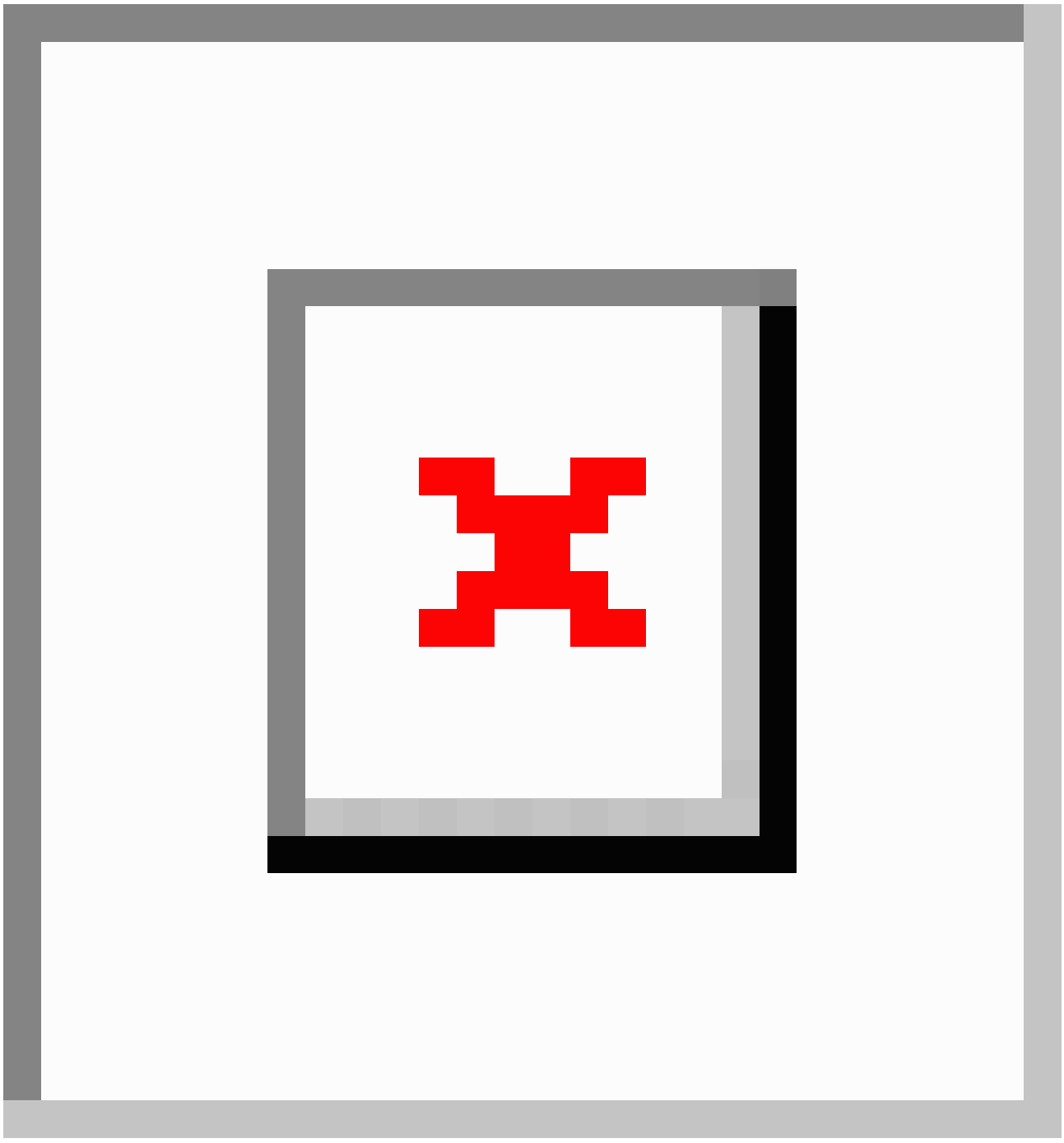


Table . Parameters for segmented logistic regression analysis of the inflammatory bowel disease (IBD) clinical decision support system (CDSS) in Implementation Period 2.

Parameter		ß	95% CI	*P* value
Application rate				
	Preintervention slope (secular trend, per month)	0.151	−3.757 to 4.059	.94
	Change in slope (gradual effect, per month)	2.019	−3.508 to 7.546	.45
	Change in intercept (immediate effect)	−5.048	−33.86 to 23.76	.71
Clinical scores completed and documented				
	Preintervention slope (secular trend, per month)	1.648	−1.596 to 4.893	.29
	Change in slope (gradual effect, per month)	−2.463	−7.051 to 2.125	.27
	Change in intercept (immediate effect)	−0.992	−24.91 to 22.92	.93
IBD flare lab tests ordered				
	Preintervention slope (secular trend, per month)	−0.016	−2.693 to 2.662	.99
	Change in slope (gradual effect, per month)	1.929	−1.858 to 5.715	.29
	Change in intercept (immediate effect)	12.60	−7.137 to 32.34	.19
C-reactive protein ordered				
	Preintervention slope (secular trend, per month)	−0.742	−3.121 to 1.637	.52
	Change in slope (gradual effect, per month)	1.253	−2.111 to 4.618	.44
	Change in intercept (immediate effect)	14.89	−2.645 to 32.43	.09
Fecal calprotectin ordered				
	Preintervention slope (secular trend, per month)	1.298	−2.209 to 4.806	.44
	Change in slope (gradual effect, per month)	0.183	−4.778 to 5.143	.94
	Change in intercept (immediate effect)	−1.034	−26.89 to 24.82	.93
Stool cultures ordered				
	Preintervention slope (secular trend, per month)	−1.060	−3.650 to 1.529	.40
	Change in slope (gradual effect, per month)	1.714	−1.948 to 5.376	.33
	Change in intercept (immediate effect)	15.37	−3.715 to 34.46	.11
	Clostridium difficile ordered			
	Preintervention slope (secular trend, per month)	−0.228	−2.613 to 2.158	.84
	Change in slope (gradual effect, per month)	1.825	−1.549 to 5.198	.27
	Change in intercept (immediate effect)	3.258	−14.33 to 20.84	.70

Discussion

Answering the Study Question

In this study, we evaluated the effectiveness of a CDSS that aimed to standardize protocols for patients with IBD experiencing an acute disease flare. An increase in several practices was demonstrated following the CDSS implementation, including increased FCP use. Completion of clinical scores also increased during Implementation Period 1 (before-and-after analysis and ITS analysis) and remained increased throughout Implementation Period 2.

We did not reach significance in slope changes or level changes in any ITS analysis in Period 2. This could be due to the sample size, which may also account for the large variance seen in some data points. There were, however, some encouraging upward trends in flare lab testing, particularly CRP (P<.10 in the ITS analysis) and stool cultures.

In characterizing the adoption of this CDSS by the application rate, an interesting finding was that the CDSS was used more by IBD nurses compared to nurse practitioners. This could represent the nurses' increased experience with the CDSS from the pilot phase and our CDSS focus on decisions related to patients experiencing a disease flare. In the University of Alberta clinic, patients are instructed to call the IBD nurse flare line if they experience changes in symptoms, and so nurses are often the first point of contact in the flare clinical pathway. This is supported by our data showing flare encounters are primarily telephone encounters. Other research has shown that flares are unlikely to coincide with scheduled clinic appointments, which aligns with the current uptake in remote monitoring and rapid access clinics [31-33].

Our observed CDSS use by specialized IBD nurses is in contrast to several other studies that have demonstrated that nurses are less likely to use CDSSs when making decisions about care they are experienced and confident in delivering, especially in the case of telephone triage decisions [34-36]. Our results could be a product of the integration of the nurses' feedback after the pilot phase, a strategy that may have increased the utility of the CDSS for nurses. This highlights recommendations from other research that emphasize the importance of engaging all stakeholders but especially end users in the CDSS design [37,38].

Limitations of the Study

There are several limitations to this research. Although the ITS design allows for better characterization of temporal changes compared to before-and-after analyses, it is still possible that other changes, such as clinic structure and release or dissemination of guidelines, could have led to the changes observed. However, apart from the intervention activation and the released memo and instructions for use that were disseminated, to our knowledge, there were no other educational campaigns, institutional changes, or major publications promoting the specific care guidelines investigated by the study. There were subtle changes in staff, for example, the joining of a new IBD physician and the leaving of another. However, there were no changes in IBD nurse staff, who were the primary users of the CDSS.

In contrast to the advantage of our 2-phased design regarding the opportunity for feedback from nurses, the design may have hindered our ability to demonstrate change. As we used the same group of IBD nurses in the pilot (Phase 1) and implementation (Phase 2) periods, our baseline use prior to the beginning of Phase 2 had already started. This may have accelerated the observed uptake speed of the CDSS by practitioners and could have also led to an underestimation of the changes before and after Implementation Phase 2.

Sample size is another limitation. In an ITS analysis, it is recommended to have a minimum of 16 data points and 100 observations per data point [25,29,30]. Although we met the data point requirement, the number of flares per month was consistently under 50. Future studies should aim to include more data points, which may require multisite participation. Unfortunately, at the time of this study, the EMR software was only deployed at a single site.

We only captured data from orders that were tied to the encounter. If a decision was made to not order labs for any reason (eg, they were recently completed), they would not be captured by our extraction. As a consequence, estimates of protocol adherence could be deflated.

Finally, it is important to note that for process measures that depend on manual data entry, such as clinical score completion, this research method can only determine whether a process was documented as completed but not necessarily whether it was actually completed. This may have resulted in underestimates of protocol adherence.

Future Directions

The currently available CDSS in this study was limited in its ability to support complex multiprovider pathways and tie together multiple visits along a pathway. More advanced CDSS workflows should be investigated in future versions of the CDSS software and evaluated for effectiveness.

Triggering logic for CDSSs should also be precisely targeted. For example, a CDSS should determine whether a patient has had a test done within a certain time span, and if not, prompt the user to order it. The reverse should also be possible; if a test has been recently ordered (eg, *Clostridium difficile*, which can only be tested once every 2 weeks), the CDSS could automatically deselect or prompt the user to remove this order to save downstream resources. This was not possible with the resources available in our CDSS environment.

In extracting data for analysis, a significant challenge was identifying flare encounters based on EMR data. The problem stems from a lack of discrete data identifying patients with active diseases (clinical scores were not regularly documented as discrete data). Future research should seek to develop a case definition for disease flare through administrative provincial data sets. This could include quantitative metrics, such as CRP and FCP, that predict the likelihood of flare, but it could also include the integration of a case-finding algorithm that uses natural language processing to parse clinical notes. This strategy

has been explored in several other diseases and has been shown to significantly improve case detection [39]. Some work has been done in IBD to identify phenotypic information from clinic notes using natural language processing [40].

The methodology used in this research should be expanded to investigate the effects of improved versions of CDSS for IBD on other community clinics and nonacademic practices throughout Alberta. Cluster-randomized designs or stepped-wedge designs could be explored since multiple clinics could be available for randomization.

This study did not investigate the impact on patient outcomes, which would require a longer follow-up period (ideally 2 or more years). Nonetheless, long-term patient outcomes for the CDSS are of great importance [9] and should be explored in the future.

Conclusions

Through our study, we designed and implemented, in 2 phases, a CDSS for IBD disease flare embedded in existing EMR software and evaluated the impact of the CDSS on provider adoption of clinical guidelines and local best practices. We have shown moderate adoption and acceptance of this system by providers, particularly by IBD nurses, as measured by the system application rate. Findings from the first phase support the hypothesis that the CDSS improved the use of FCP and the documentation of clinical scores. Findings from the second phase support further improvement in ordering flare lab panels, CRP, and stool cultures, as shown in before-and-after analysis and multivariate analysis. In addition, potential improvements in workflow integration were identified through qualitative questionnaires and feedback forms; areas for future research have also been established.

Acknowledgments

The authors acknowledge the faculty and staff of the inflammatory bowel disease (IBD) Unit and Division of Gastroenterology at the University of Alberta Hospital, who helped with the design and implementation of the IBD clinical care pathway (CCP). We also acknowledge the staff of Alberta Health Services (AHS) for their assistance with supplying the data. We thank Mr Darryl Wilson, the reporting systems analyst for the AHS Information Systems, for his assistance with the natural language queries (SQL) data acquisition from eCLINICIAN, and Mr Nathan Stern for helping with the chart review. Finally, we would like to thank the late Dr Richard Fedorak for his contribution to this work.

All results and inferences reported in this manuscript are independent of the funding and support sources.

Authors' Contributions

RTS contributed to study design, data collection, data analysis, and manuscript drafting. KDC contributed to drafting and revision of the manuscript. DP, DCS, and DCB contributed to the critical revision of the manuscript. KIK contributed to the study design as well as the analysis and critical revision of the manuscript. All authors approved the final version. KIK is the guarantor of the paper.

Conflicts of Interest

This study was supported by the Crohn’s and Colitis Canada via the Promoting Access and Care through Centres of Excellence (PACE) initiative. RTS was also supported by studentships from Alberta Innovates, the Faculty of Medicine and Dentistry, University of Alberta, and the Canadian Institutes of Health Research (CIHR). All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
The rationale for using an interrupted time series design.
[DOCX File, 19 KB - medinform_v12i1e55314_app1.docx]

Multimedia Appendix 2
Provider feedback.
[DOCX File, 16 KB - medinform_v12i1e55314_app2.docx]

Multimedia Appendix 3
Materials distributed to providers.
[DOCX File, 781 KB - medinform_v12i1e55314_app3.docx]

Multimedia Appendix 4
eCLINICIAN query information.
[DOCX File, 68 KB - medinform_v12i1e55314_app4.docx]

Multimedia Appendix 5
Sample size calculation.
[DOCX File, 13 KB - medinform_v12i1e55314_app5.docx]

Multimedia Appendix 6
Demographics of users (inflammatory bowel disease nurses and practitioners).
[DOCX File, 13 KB - medinform_v12i1e55314_app6.docx]

References

1. Davis DA, Taylor-Vaisey A. Translating guidelines into practice. A systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines. CMAJ 1997 Aug 15;157(4):408-416. [Medline: 9275952]
2. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH. Why don't physicians follow a framework for improvement. J Am Med Assoc 1999 Oct;282:1458-1465. [doi: 10.1001/jama.282.15.1458] [Medline: 10535437]
3. Westfall JM, Mold J, Fagnan L. "Practice-based research - "blue highways" on the NIH roadmap". JAMA 2007 Jan 24;297(4):403-406. [doi: 10.1001/jama.297.4.403] [Medline: 17244837]
4. Balas EA, Boren SA. Managing clinical knowledge for health care improvement. Yearb Med Inform 2000(1):65-70. [Medline: 27699347]
5. Shortliffe T. Medical thinking: what should we do? Presented at: Conference on Medical Thinking; Jun 23, 2006; London, UK URL: https://slideplayer.com/slide/10838966/ [accessed 2024-03-05]
6. Vander Schaaf EB, Seashore CJ, Randolph GD. Translating clinical guidelines into practice: challenges and opportunities in a dynamic health care environment. N C Med J 2015 Sep;76:230-234. [doi: 10.18043/ncm.76.4.230] [Medline: 26509513]
7. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
8. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
9. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 1;18(3):327-334. [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
10. Sutton RT. Adherence to clinical care protocols for inflammatory bowel disease and evaluation of a clinical decision support system to improve adherence [Master's thesis].: University of Alberta; 2019 URL: https://era.library.ualberta.ca/items/e4651c09-d19f-4d58-b7bc-eeb9368a974e [accessed 2024-03-05]
11. Jackson BD, Con D, Liew D, De Cruz P. Clinicians' adherence to international guidelines in the clinical care of adults with inflammatory bowel disease. Scand J Gastroenterol 2017 May;52(5):536-542. [doi: 10.1080/00365521.2017.1278785] [Medline: 28128675]
12. Lytvyak E, Sutton RT, Dieleman LA, Peerani F, Fedorak RN, Kroeker KI. Management of inflammatory bowel disease patients with clinical care pathways reduces emergency department utilization. Crohns Colitis 360 2020 Oct 13;2(4):taa080. [doi: 10.1093/crocol/otaa080] [Medline: 36777757]
13. Epic UserWeb. 2018. URL: https://comlib.epic.com [accessed 2018-05-08]
14. Pauwen NY, Louis E, Siegel C, Colombel JF, Macq J. Integrated care for Crohn's disease: a plea for the development of clinical decision support systems. J Crohns Colitis 2018 Nov 28;12(12):1499-1504. [doi: 10.1093/ecco-jcc/jjy128] [Medline: 30496446]
15. Breton J, Witmer CM, Zhang Y, et al. Utilization of an electronic medical record-integrated dashboard improves identification and treatment of anemia and iron deficiency in pediatric inflammatory bowel disease. Inflamm Bowel Dis 2021 Aug 19;27(9):1409-1417. [doi: 10.1093/ibd/izaa288] [Medline: 33165613]
16. Jackson B, Begun J, Gray K, et al. Clinical decision support improves quality of care in patients with ulcerative colitis. Aliment Pharmacol Ther 2019 Apr;49(8):1040-1051. [doi: 10.1111/apt.15209] [Medline: 30847962]
17. Epic Userweb. Best Practice Advisories Setup and Support Guide: Epic Userweb Software; 2018.
18. Epic Userweb. Decision Support Strategy Handbook: Epic Userweb Software; 2018.
19. Harvey RF, Bradshaw JM. A simple index of Crohn's disease activity. Lancet 1980 Mar 8;1(8167):514. [doi: 10.1016/s0140-6736(80)92767-1] [Medline: 6102236]
20. Lewis JD, Chuai S, Nessel L, Lichtenstein GR, Aberra FN, Ellenberg JH. Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. Inflamm Bowel Dis 2008 Dec;14(12):1660-1666. [doi: 10.1002/ibd.20520] [Medline: 18623174]

21. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care 2003;19(4):613-623. [doi: 10.1017/s0266462303000576] [Medline: 15095767]
22. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI--Statement on reporting of evaluation studies in Health Informatics. Int J Med Inform 2009 Jan;78(1):1-9. [doi: 10.1016/j.ijmedinf.2008.09.002] [Medline: 18930696]
23. Brender J, Talmon J, de Keizer N, Nykänen P, Rigby M, Ammenwerth E. Statement on reporting of evaluation studies in Health Informatics. Appl Clin Inform 2013;04(3):331-358. [doi: 10.4338/ACI-2013-04-RA-0024]
24. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London Edinburgh Dublin Phil Mag J Sci 1900 Jul;50(302):157-175. [doi: 10.1080/14786440009463897]
25. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther 2002 Aug;27(4):299-309. [doi: 10.1046/j.1365-2710.2002.00430.x] [Medline: 12174032]
26. Muggeo VMR. Segmented: an R package to fit regression models with broken-line relationships. R News 2008;7:1609-3631 [FREE Full text]
27. Chen H, Cohen P, Chen S. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. Commun Stat - Simul C 2010 Mar 31;39(4):860-864. [doi: 10.1080/03610911003650383]
28. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and BIOMEDICAL sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
29. Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. J Clin Epidemiol 2011 Nov;64(11):1252-1261. [doi: 10.1016/j.jclinepi.2011.02.007] [Medline: 21640554]
30. Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. J Clin Epidemiol 2015 Aug;68(8):950-956. [doi: 10.1016/j.jclinepi.2014.12.018] [Medline: 25890805]
31. Kemp K, Griffiths J, Campbell S, Lovell K. An exploration of the follow-up up needs of patients with inflammatory bowel disease. J Crohns Colitis 2013 Oct;7(9):e386-e395. [doi: 10.1016/j.crohns.2013.03.001] [Medline: 23541150]
32. Nene S, Gonczi L, Kurti Z, et al. Benefits of implementing a rapid access clinic in a high-volume inflammatory bowel disease center: access, resource utilization and outcomes. World J Gastroenterol 2020 Feb 21;26(7):759-769. [doi: 10.3748/wjg.v26.i7.759] [Medline: 32116423]
33. Pure N, Mize C. P193 the development of an IBD specialty clinic within a gastroenterology practice. Gastroenterology 2018 Jan;154(1):S107-S108. [doi: 10.1053/j.gastro.2017.11.253]
34. Dowding D, Mitchell N, Randell R, Foster R, Lattimer V, Thompson C. Nurses' use of computerised clinical decision support systems: a case site analysis. J Clin Nurs 2009 Apr;18(8):1159-1167. [doi: 10.1111/j.1365-2702.2008.02607.x] [Medline: 19320785]
35. O'Cathain A, Sampson FC, Munro JF, Thomas KJ, Nicholl JP. Nurses' views of using computerized decision support software in NHS direct. J Adv Nurs 2004 Feb;45(3):280-286. [doi: 10.1046/j.1365-2648.2003.02894.x] [Medline: 14720245]
36. O'Cathain A, Nicholl J, Sampson F, Walters S, McDonnell A, Munro J. Do different types of nurses give different triage decisions in NHS direct? A mixed methods study. J Health Serv Res Policy 2004 Oct;9(4):226-233. [doi: 10.1258/1355819042250221] [Medline: 15509408]
37. Rocque G, Miller-Sonnet E, Balch A, et al. Engaging multidisciplinary stakeholders to drive shared decision-making in oncology. J Palliat Care 2019 Jan;34(1):29-31. [doi: 10.1177/0825859718810723] [Medline: 30382006]
38. Daudelin DH, Ruthazer R, Kwong M, et al. Stakeholder engagement in methodological research: development of a clinical decision support tool. J Clin Transl Sci 2020 Apr;4(2):133-140. [doi: 10.1017/cts.2019.443] [Medline: 32313703]
39. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep;23(5):1007-1015. [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
40. South BR, Shen S, Jones M, et al. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics 2009 Sep 17;10 Suppl 9(Suppl 9):S12. [doi: 10.1186/1471-2105-10-S9-S12] [Medline: 19761566]

Abbreviations

AHS: Alberta Health Services
BPA: Best Practice Advisory
CDS: clinical decision support
CDSS: clinical decision support system
CRP: C-reactive protein

EMR: electronic medical record
FCP: fecal calprotectin
IBD: inflammatory bowel disease
ITS: interrupted time series
mHBI: modified Harvey Bradshaw Index
pMayo: partial Mayo
STARE-HI: Statement on Reporting of Evaluation Studies in Health Informatics

Edited by C Lovis; submitted 13.12.23; peer-reviewed by T Xenodemetropoulos; accepted 02.02.24; published 22.03.24.

Please cite as:
Sutton RT, Chappell KD, Pincock D, Sadowski D, Baumgart DC, Kroeker KI
The Effect of an Electronic Medical Record–Based Clinical Decision Support System on Adherence to Clinical Protocols in Inflammatory Bowel Disease Care: Interrupted Time Series Study
JMIR Med Inform 2024;12:e55314
URL: https://medinform.jmir.org/2024/1/e55314
doi:10.2196/55314

Implementable Prediction of Pressure Injuries in Hospitalized Adults: Model Development and Validation

Thomas J Reese[1], PharmD, PhD; Henry J Domenico[2], MS; Antonio Hernandez[3], MD, MSCI; Daniel W Byrne[1,2], MS; Ryan P Moore[2], MS; Jessica B Williams[4], RN, BSN, CCRN, APRN; Brian J Douthit[1], RN-BC, PhD; Elise Russo[1], MPH, PMP; Allison B McCoy[1], PhD; Catherine H Ivory[1], PhD, RN; Bryan D Steitz[1], PhD; Adam Wright[1], PhD

1

2

3

4

Corresponding Author:
Thomas J Reese, PharmD, PhD

Abstract

Background: Numerous pressure injury prediction models have been developed using electronic health record data, yet hospital-acquired pressure injuries (HAPIs) are increasing, which demonstrates the critical challenge of implementing these models in routine care.

Objective: To help bridge the gap between development and implementation, we sought to create a model that was feasible, broadly applicable, dynamic, actionable, and rigorously validated and then compare its performance to usual care (ie, the Braden scale).

Methods: We extracted electronic health record data from 197,991 adult hospital admissions with 51 candidate features. For risk prediction and feature selection, we used logistic regression with a least absolute shrinkage and selection operator (LASSO) approach. To compare the model with usual care, we used the area under the receiver operating curve (AUC), Brier score, slope, intercept, and integrated calibration index. The model was validated using a temporally staggered cohort.

Results: A total of 5458 HAPIs were identified between January 2018 and July 2022. We determined 22 features were necessary to achieve a parsimonious and highly accurate model. The top 5 features included tracheostomy, edema, central line, first albumin measure, and age. Our model achieved higher discrimination than the Braden scale (AUC 0.897, 95% CI 0.893-0.901 vs AUC 0.798, 95% CI 0.791-0.803).

Conclusions: We developed and validated an accurate prediction model for HAPIs that surpassed the standard-of-care risk assessment and fulfilled necessary elements for implementation. Future work includes a pragmatic randomized trial to assess whether our model improves patient outcomes.

(JMIR Med Inform 2024;12:e51842) doi:10.2196/51842

KEYWORDS

patient safety; electronic health record; EHR; implementation; predictive analytics; prediction; injury; pressure injury; hospitalization; adult; development; routine care; prediction model; pressure sore

Introduction

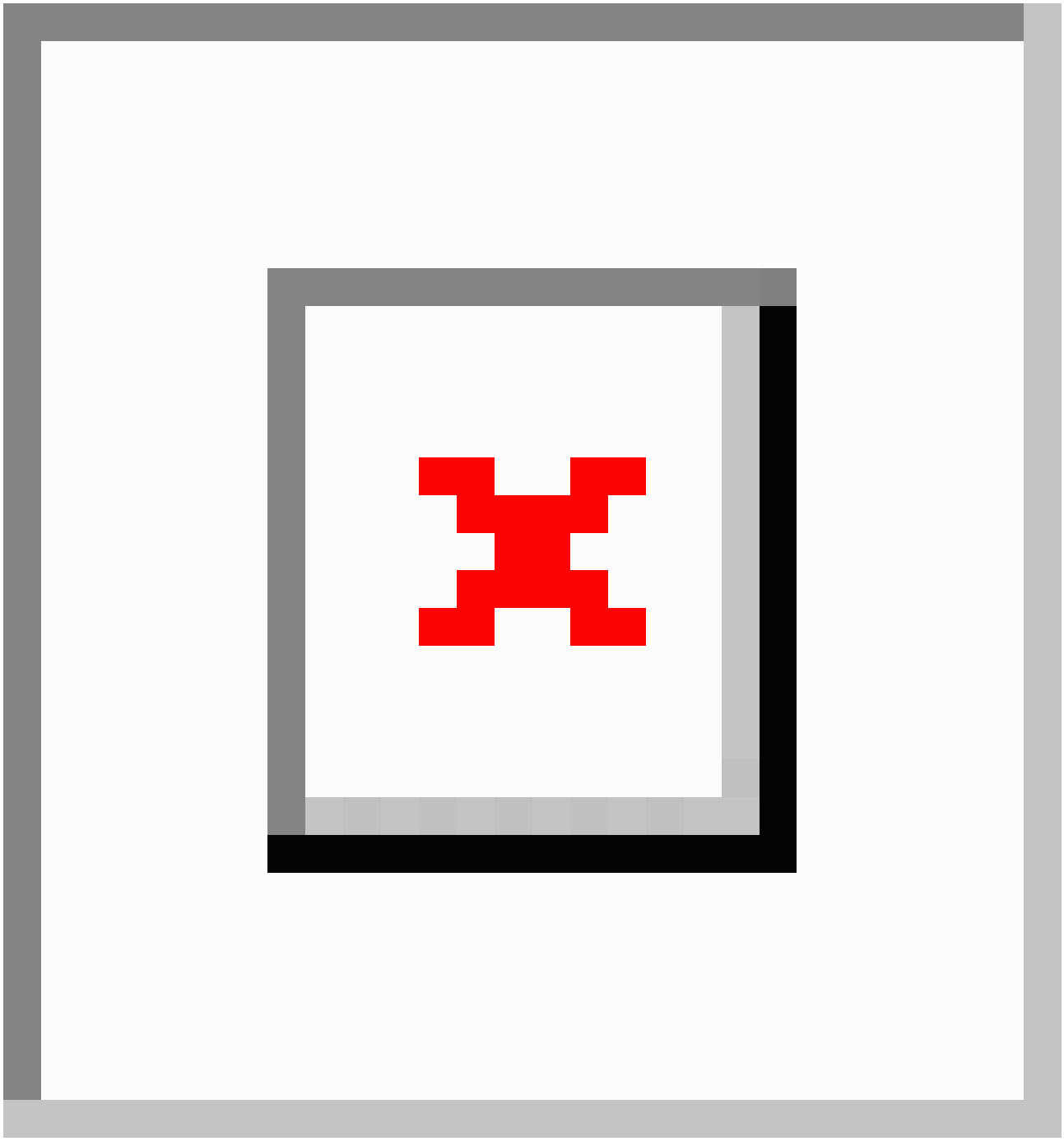
Pressure injuries comprise damage to skin and underlying tissue that usually occurs over a bony prominence but can be related to placement of medical devices [1]. The injury occurs because of intense or prolonged pressure that is combined with shear forces. Pressure injuries are a widespread and costly problem. A recent study found the prevalence of pressure injuries may be close to 30% for patients in intensive care units, which is 10% higher than previous estimates [2,3]. Patients with pressure injuries experience pain and the potential for infection and debilitation, which prolongs hospital stays and impacts recovery. Furthermore, increasing evidence supports the association between severity of pressure injuries and patient mortality [2]. In the United States, health care systems absorb on average US $10,000 per hospital-acquired pressure injury (HAPI), which contributes to a cost burden that will soon exceed US $30 billion [4,5].

Prevention of pressure injuries requires an accurate risk assessment and an interdisciplinary approach with routine repositioning, maintaining dry skin, and padding pressure points to reduce injury [6-8]. Currently, health care systems are striving to accurately measure and prevent HAPIs, since they can be common and negatively impact patient care [9]. Patient factors such as age, vasopressor support, mechanical ventilation, low

albumin, and renal failure can increase the risk for pressure injuries [10,11]. Multiple standardized risk assessment tools have been developed to systematically assess patient factors and assist clinicians in identifying at-risk patients [12,13]. Of these tools, the Braden scale has remained the standard of care across health systems for decades. The Braden scale incorporates components of sensory perception, activity, mobility, and nutrition, as well as skin moisture, friction, and shear force, to produce a score that indicates the risk of developing a pressure injury [14]. Although use of the Braden scale is widespread, its accuracy and reliability in diverse settings and patients is in question; thus, researchers have turned to more advanced risk prediction models that incorporate additional patient factors [12,13,15,16].

Recent literature reviews of advanced risk prediction models have highlighted excellent performance in predicting pressure injuries [17-21]. Zhou and colleagues [20] found that 74% of studies achieved an area under the receiver operating curve (AUC) between 0.68 and 0.99. Although these models were exceptionally accurate at predicting pressure injuries, no studies to our knowledge have implemented such models to reduce the number of pressure injuries. Numerous prediction models have been developed across clinical domains, but few have improved patient outcomes, leading researchers to identify a variety of required elements that may be necessary to implement prediction models in practice [22-24]. For instance, Randall Moorman [23] proposed properties, such as change of risk over time (eg, dynamic risk), for predictive analytics in neonatal intensive care units. Keim-Malpass and colleagues [24] found that potential users want prediction tools to be integrated with the electronic health record (EHR; eg, feasibility). We reviewed and agreed upon 5 elements that applied to HAPI prediction (ie, it should be feasible, broadly applicable, include dynamic risk and actionable criteria, and be rigorously validated) and then applied these elements to 22 recent models from 2020 to 2022 (Figure 1) [17,20,21]. We found no models fulfilled all the necessary elements to impact patient care. To help bridge the gap from model development to implementation, the objective of this study was, therefore, to develop and validate a model that fulfilled these elements and then compare its performance to usual care (ie, the Braden scale).

Figure 1. Comparison of current pressure injury prediction models according to elements of implementable models [25-45].



Methods

Study Population

We used retrospective data from the EHR at Vanderbilt University Medical Center between January 1, 2018, and July 1, 2022. All hospital admissions were included if the length of stay was longer than 24 hours and patient age was greater than 18 years on admission. HAPIs were identified using nurse flowsheet documentation. Nurses use flowsheets to document a variety of assessments, with our institution using a dedicated section for pressure injuries. The presence or absence of a pressure injury is assessed on admission and daily for each patient in the hospital. If a pressure injury is identified, the nurse documents whether it was present on admission and additional characteristics of the pressure injury, including the stage and location. We considered pressure injuries documented with a "no" in the column "present on admission" as HAPIs. For patients who had more than one HAPI, we used the first documented. The cohort included 197,911 hospitalizations, 129,100 patients, and 5458 HAPIs.

Feature Selection and Cohort Development

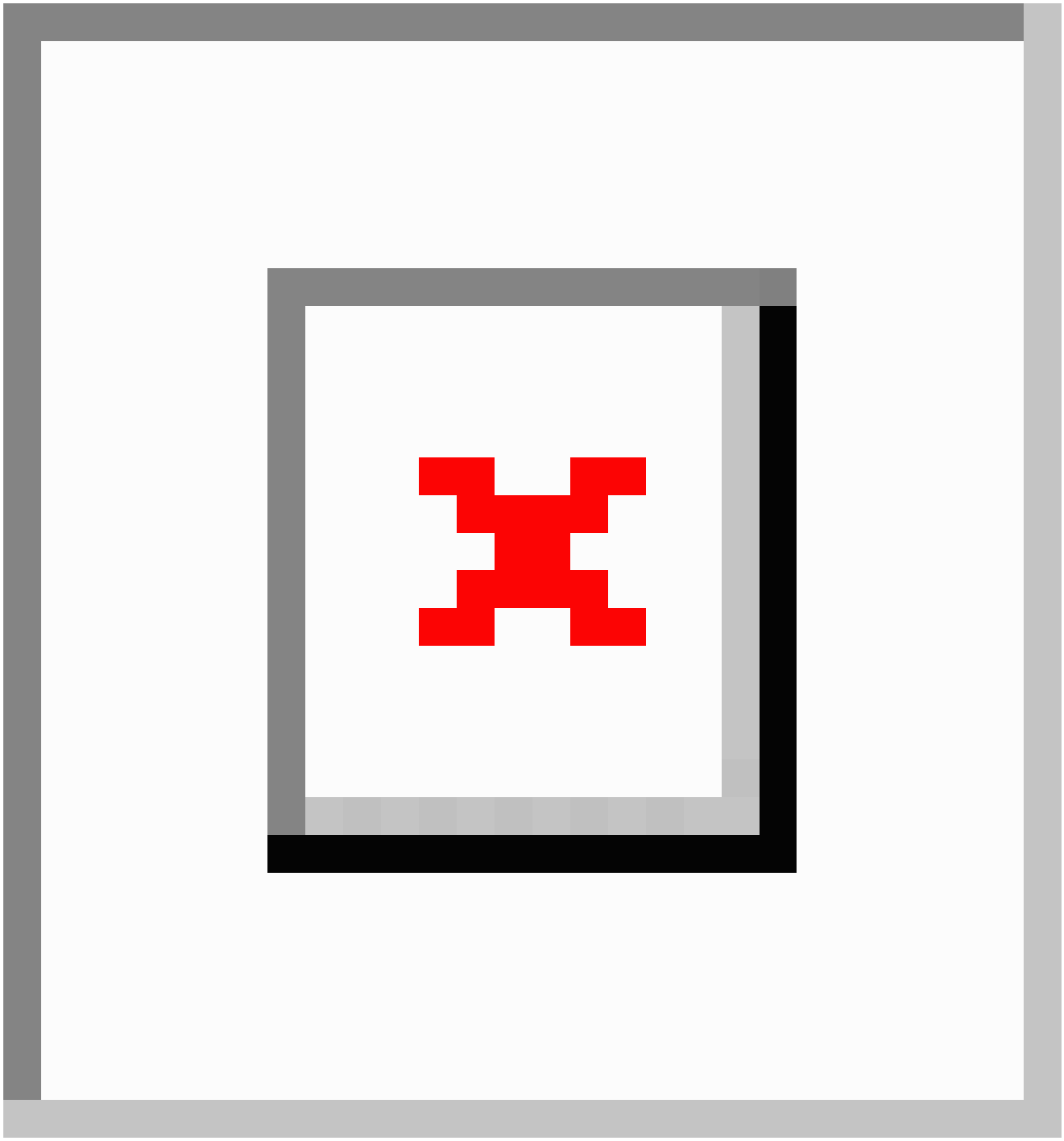
We first identified relevant features associated with pressure injuries from the literature. The list of relevant features was supplemented and pruned by clinical domain experts and informaticians at Vanderbilt University Medical Center. In total, 51 features were extracted as candidate features for predicting HAPIs. Importantly, features were only extracted if they were

available at the time of hospitalization and could be used to update the risk prediction during the encounter (ie, no claims data were used). Table 1 provides a summary of the extracted features. Missing values were imputed with the cohort median [46,47]. Multimedia Appendix 1 provides the full cohort characteristics, including missing values and a full list of measures. We split the full cohort temporally into model development and validation cohorts based on the number of events, with the development and validation cohorts including 80% and 20% of HAPIs, respectively. The development cohort included 161,816 hospitalizations and 4362 HAPIs from January 1, 2018, to August 26, 2021, and the validation cohort included 36,095 hospitalizations and 1096 HAPIs from August 27, 2021, to June 29, 2022 (Figure 2). Outcomes and features were identified and extracted in the same manner for the development and validation cohorts.

Table . Overview of extracted features.

Source	Feature
Patient demographics and social history	Age; gender; race; ethnicity; smoking status
Administration	Hospital admission through emergency department; intensive care unit admission; length of stay
Flowsheets	Hospital-acquired pressure injury (primary outcome); temperature; respiratory rate; heart rate; BMI; oxygen saturation; blood pressure; Braden scale (items and composite score); consciousness; gait transfer; Glasgow Coma Scale; malnutrition score; spinal cord injury; dialysis during hospitalization; tracheostomy; gastric tube; central line; chest tube; ostomy; drain; extracorporeal membrane oxygenation
Laboratory results	Hemoglobin; hemoglobin A_{1C}; hematocrit; mean corpuscular hemoglobin concentration; red cell distribution width; platelet count; chloride; blood urea nitrogen; creatinine; lactate; albumin; glucose

Figure 2. Model development and validation cohorts.



Model Development

We developed 3 models for comparison using logistic regression. The present model (Vanderbilt) used a broad set of candidate features (Table 1). The second model used the sum of the individual item measures from the Braden scale (ie, continuous Braden) [14]. Finally, since the Braden scale is typically operationalized using a single composite score (ie, less than 18=high risk; greater than or equal to 18=low risk), we included the dichotomous Braden for comparison as well. Logistic regression is the most frequently used model in clinical care [20,48]. The primary advantages of using logistic regression are that feature importance is easily interpretable and that the mathematical equation used to extract features and calculate a risk prediction is readily available in most commercial EHRs. Currently, the output from many machine learning models is not operationalizable for patient care in the EHR. To account for nonlinearity of the numeric features, we tested 3 knot-restricted cubic splines but found the discrimination failed to improve by using the nonlinear model [49]. Since the purpose was to develop a model that could be easily implemented in the EHR and compare it to standard care, we focused on use of logistic regression for the Vanderbilt and continuous Braden models.

We first included all 51 candidate features in the present (Vanderbilt) model to examine complexity versus accuracy as measured by cross-validation AUC. Again, included features were derived from the literature and refined by clinical domain

experts and informaticians. We tested for multicollinearity by examining the proportion of variance in each candidate feature that could be explained by other candidate features and removed hemoglobin. Included features had to be structured and readily available for automated processing in the EHR without additional input by the user. Using the conservative 15:1 rule, we were able to include 290.8 degrees of freedom in the model. To ensure the model was broadly applicable across settings and patients, we used a least absolute shrinkage and selection operator (LASSO) approach to identify important candidate features. Candidate features were standardized (scaled and centered) prior to running the LASSO regression. LASSO introduces a penalty term to the standard regression model, which forces some of the regression coefficients to shrink toward zero, effectively performing feature selection [50]. Variables with nonzero coefficients were included in the final model. The model was designed to calculate a risk prediction on admission and daily while the patient was in the hospital. Missing numeric measures were to be imputed with the cohort median until measures became available.

Model Evaluation

The final model was assessed in an external cohort that was temporally separated from the model development cohort. We evaluated the model using traditional and novel performance measures, which included the AUC, Brier score, slope, intercept, integrated calibration index, and calibration curve. AUC is a performance measure for the discrimination of HAPI versus no HAPI. It combines the true and false positive rates, with an AUC of 0.5 indicating no meaningful discrimination. The Brier score accounts for the predicted HAPI outcome as well as the estimate and is calculated by the squared difference between the prediction (0 to 1) and outcome (0=no HAPI and 1=HAPI) [51]. For example, if a patient had a 90% probability of developing a HAPI and did develop a HAPI during that encounter, the Brier score would be 0.01. A Brier score of 0 indicates perfect accuracy and a score of 1 indicates perfect inaccuracy. The integrated calibration index is a numeric summary of model calibration across the predicted probabilities [52]. It is the weighted average of the absolute difference between the observed and predicted probabilities; therefore, a lower integrated calibration index indicates better calibration. A slope equal to 1 indicates agreement between the observed response and the predicted probability, while a slope greater than 1 indicates potential underfitting, and a slope lower than 1 indicates potential overfitting [52]. Similarly, an intercept of zero is ideal. As with prior models, no adjustments were made for multiple comparisons [47,53,54]. We used the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) reporting guidelines (Checklist 1) and performed all analyses in R (version 4.2.3; R Foundation for Statistical Computing) with relevant extension packages [55].

Ethical Considerations

This study was approved by the Vanderbilt University Medical Center Institutional Review Board (220644), and data were deidentified.

Results

Cohort Characteristics

The full cohort of patient encounters was split temporally, based on the number of HAPIs, into model development and validation cohorts. The characteristics for each cohort are provided in Table 2. Among the model development cohort, those who developed HAPIs were older and male. Table 3 provides the model development cohort characteristics divided by whether a HAPI occurred.

Table . Characteristics for model development and validation cohorts. Measures were first taken during the hospital stay unless specified otherwise. Race and ethnicity were not included as candidate features.

		Development cohort (n=161,816 encounters)	Validation cohort (n=36,095 encounters)
Age (years), median (IQR)		56 (37-69)	56 (37-69)
Female, n (%)		84,727 (52.4)	17,060 (47.3)
Race, n (%)			
	White	125,322 (77.4)	27,649 (76.6)
	African American	26,299 (16.3)	5659 (15.7)
	Asian	2325 (1.4)	480 (1.3)
	American Indian or Alaska Native	289 (0.2)	66 (0.2)
	Pacific Islander	104 (0.1)	25 (0.1)
	Multiple	1185 (0.7)	231 (0.6)
Hispanic, n (%)		7406 (4.6)	2074 (5.7)
Physiological and clinical features			
	Temperature (°C), median (IQR)	36.7 (36.5-36.9)	36.7 (36.5-36.9)
	Respiratory rate (breaths per minute), median (IQR)	18 (16-19)	18 (16-20)
	Heart rate (beats per minute), median (IQR)	87 (74-100)	87 (75-101)
	BMI (kg/m^2), median (IQR)	28.2 (24.1-33.4)	28.3 (24.2-33.5)
	Oxygen saturation (%), median (IQR)	98 (96-99)	98 (96-99)
	Systolic blood pressure (mm Hg), median (IQR)	131 (117-147)	130 (117-146)
	Diastolic blood pressure (mm Hg), median (IQR)	77 (68-88)	77 (67-87)
	Emergency department admissions, n (%)	91,363 (56.5)	20,972 (58.1)
	Intensive care admissions, n (%)	34,190 (21.1)	7831 (21.7)
	Length of stay (days), median (IQR)	4 (2-6)	4 (2-7)
	Smokers, n (%)	56,750 (35.1)	12,561 (34.8)
	Edema, n (%)	86,582 (53.5)	19,846 (55)
	Spinal cord injury, n (%)	5908 (3.7)	1428 (4)
	Dialysis, n (%)	92 (0.1)	74 (0.2)
	Tracheostomy, n (%)	2122 (1.3)	520 (1.4)
	Gastric tube, n (%)	35 (0)	5 (0)
	Central line, n (%)	20,648 (12.8)	4803 (13.3)
	Chest tube, n (%)	5186 (3.2)	1278 (3.5)
	Ostomy, n (%)	2059 (1.3)	459 (1.3)
	Drain, n (%)	17,800 (11)	4005 (11.1)
	ECMO[a], n (%)	414 (0.3)	71 (0.2)
Laboratory results, median (IQR)			
	Hemoglobin A_{1C} (%)	6.1 (5.5-7.5)	6.1 (5.6-7.5)
	Hemoglobin (g/dL)	12.0 (10.3-13.6)	11.9 (10.2-13.5)
	Hematocrit (%)	36.0 (32.0-41.0)	36.0 (32.0-40.0)
	MCHC[b] (g/dL)	33.0 (32.0-34.0)	32.9 (31.9-33.9)
	Red cell distribution width (%)	13.9 (13.0-15.5)	14.0 (13.0-15.6)
	Platelet count ($\times 10^9$/L)	228 (174-291)	234 (179-298)
	Chloride (mEq/L)	105 (101-108)	104 (101-107)

		Development cohort (n=161,816 encounters)	Validation cohort (n=36,095 encounters)
	Lactate (mmol/L)	1.1 (0.8-1.9)	1.2 (0.8-2.0)
	Albumin (g/dL)	3.6 (3.1-4.0)	3.5 (3.0-3.9)
	Urine blood urea nitrogen	412 (260-603)	415 (275-609)
	Creatinine (mg/dL)	0.9 (0.9-1.3)	0.9 (0.8-1.3)
	Glucose (mmol/L)	114 (96-146)	114 (96-145)
Nursing assessment features			
	Braden scale score, median (IQR)	20 (18-22)	20 (17-21)
	Level of consciousness=2, n (%)	21,357 (13.2)	5043 (14)
	Gait transfer=20, n (%)	10,673 (6.6)	2190 (6.1)
	Glasgow Coma Scale=3, n (%)	4872 (3)	961 (2.7)
	Malnutrition score=5, n (%)	1241 (0.8)	345 (1)
Outcomes, n (%)			
	Any pressure injury	9259 (5.7)	2143 (5.9)
	Hospital-acquired pressure injury	4362 (2.7)	1096 (3)

[a]ECMO: extracorporeal membrane oxygenation.

[b]MCHC: mean corpuscular hemoglobin concentration.

Table . Model development cohort characteristics with and without hospital acquired pressure injury. Measures were the first taken during the hospital stay unless specified otherwise. Race and ethnicity were not included as candidate features.

		No hospital-acquired pressure injury (n=157,454 encounters)	Hospital-acquired pressure injury (n=4362 encounters)
Age (years), median (IQR)		56 (37-68)	64 (52-74)
Female, n (%)		82,999 (52.7)	1728 (39.6)
Race, n (%)			
	White	121,786 (77.3)	3536 (81.1)
	African American	25,654 (16.3)	645 (14.8)
	Asian	2290 (1.5)	35 (0.8)
	American Indian or Alaska Native	285 (0.2)	4 (0.1)
	Pacific Islander	102 (0.1)	2 (0)
	Multiple	1154 (0.7)	31 (0.7)
Hispanic, n (%)		7306 (4.6)	100 (2.3)
Physiologic and clinical features			
	Temperature (°C), median (IQR)	36.7 (36.5-36.9)	36.7 (36.4-37.0)
	Respiratory rate (breaths per minute), median (IQR)	18.0 (16.0-19.0)	18.0 (16.0-22.0)
	Heart rate (beats per minute), median (IQR)	87.0 (74.0-100.0)	91.0 (77.0-106.0)
	BMI (kg/m^2), median (IQR)	28.2 (24.1-33.5)	26.8 (22.6-32.2)
	Oxygen saturation (%), median (IQR)	98.0 (96.0-99.0)	97.0 (95.0-99.0)
	Systolic blood pressure (mm Hg), median (IQR)	131.0 (117.0-147.0)	124.0 (107.0-142.0)
	Diastolic blood pressure (mm Hg), median (IQR)	78.0 (68.0-88.0)	72.0 (61.0-84.0)
	Emergency department admissions, n (%)	88,552 (56.2)	2811 (64.4)
	Intensive care admissions, n (%)	31,795 (20.2)	2395 (54.9)
	Length of stay (days), median (IQR)	3 (2-6)	15 (8-26)
	Smokers, n (%)	55,278 (35.1)	1472 (33.7)
	Edema, n (%)	82,640 (52.5)	3942 (90.4)
	Spinal cord injury, n (%)	5398 (3.4)	510 (11.7)
	Dialysis, n (%)	81 (0.1)	11 (0.3)
	Tracheostomy, n (%)	1491 (0.9)	631 (14.5)
	Gastric tube, n (%)	24 (0)	11 (0.3)
	Central line, n (%)	18,350 (11.7)	2298 (52.7)
	Chest tube, n (%)	4598 (2.9)	588 (13.5)
	Ostomy, n (%)	1881 (1.2)	178 (4.1)
	Drain, n (%)	16,888 (10.7)	912 (20.9)
	ECMO[a], n (%)	242 (0.2)	172 (3.9)
Laboratory results, median (IQR)			
	Hemoglobin A_{1C} (%)	6.1 (5.5-7.5)	6.0 (5.4-7.1)
	Hemoglobin (g/dL)	12.0 (10.3-13.6)	12.0 (10.3-13.6)
	Hematocrit (%)	37.0 (32.0-41.0)	34.0 (29.0-40.0)

		No hospital-acquired pressure injury (n=157,454 encounters)	Hospital-acquired pressure injury (n=4362 encounters)
	MCHC[b] (g/dL)	33.0 (32.0-34.0)	32.6 (31.5-33.7)
	Red cell distribution width (%)	13.9 (13.0-15.5)	14.9 (13.5-16.8)
	Platelet count ($\times 10^9$/L)	228.0 (175.0-291.0)	215.0 (151.0-298.0)
	Chloride (mEq/L)	105.0 (101.0-108.0)	104.0 (99.0-108.0)
	Lactate (mmol/L)	1.1 (0.8-1.3)	1.4 (0.9-2.5)
	Albumin (g/dL)	3.6 (3.1-4.0)	3.1 (2.6-3.5)
	Urine blood urea nitrogen	412.0 (263.0-603.0)	410.0 (244.0-605.5)
	Creatinine (mg/dL)	0.9 (0.8-1.3)	1.2 (0.8-1.9)
	Glucose (mmol/L)	114.0 (96.0-145.0)	125.0 (101.0-168.0)
Nursing assessment features			
	Braden scale score, median (IQR)	20.0 (18.0-22.0)	15.0 (13.0-18.0)
	Level of consciousness=2, n (%)	20,712 (13.2)	645 (14.8)
	Gait transfer=20, n (%)	10,055 (6.4)	618 (14.2)
	Glasgow Coma Scale=3, n (%)	4406 (2.8)	466 (10.7)
	Malnutrition score=5, n (%)	1166 (0.7)	75 (1.7)

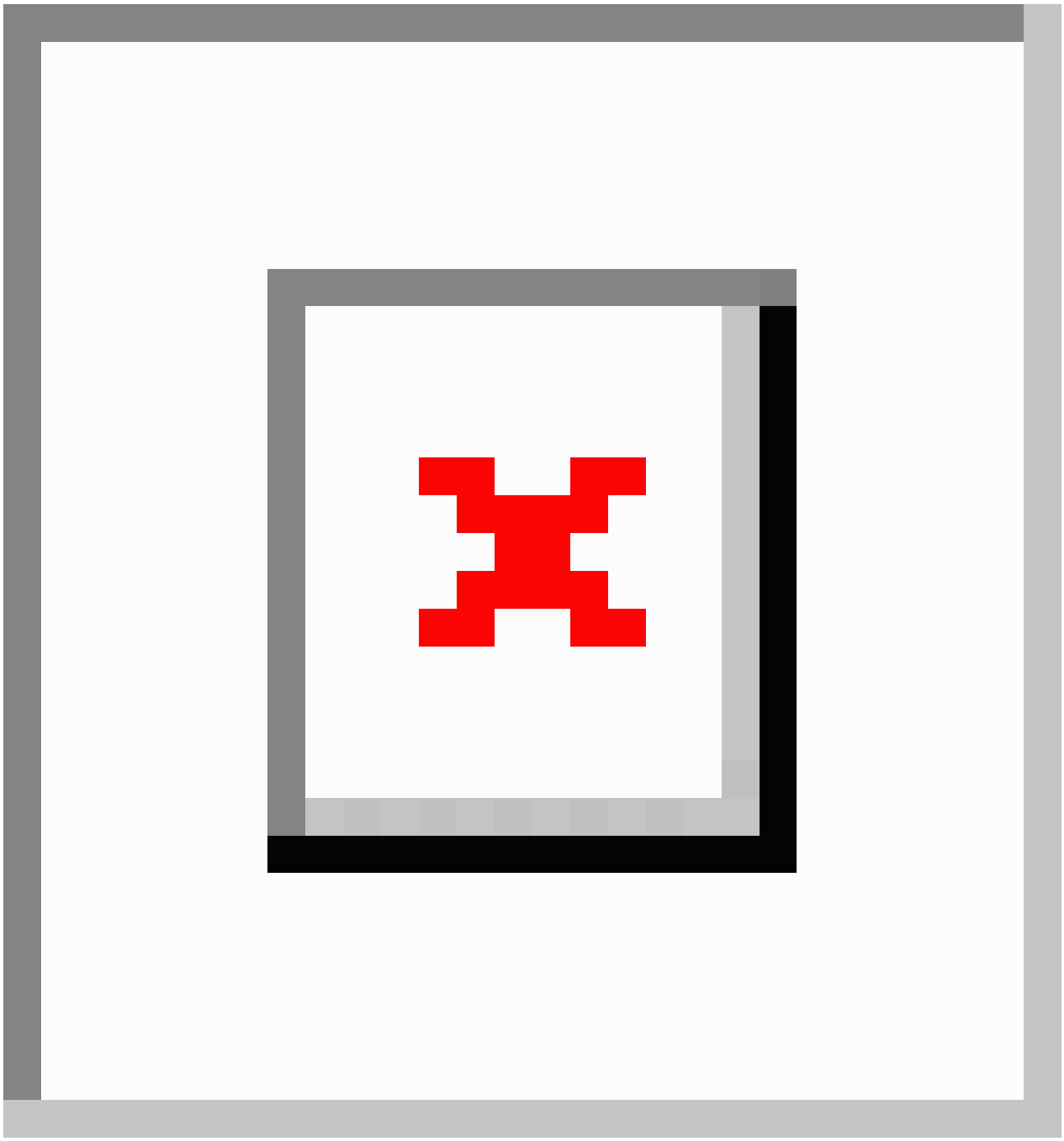
[a]ECMO: extracorporeal membrane oxygenation.

[b]MCHC: mean corpuscular hemoglobin concentration.

Model Description

We determined 22 features were necessary to achieve a parsimonious yet highly accurate model. Again, features were selected using a LASSO approach. We fit the final model with 4362 HAPI encounters and 291 degrees of freedom, which indicated the model was unlikely to overfit the data. Of the 40 features that exhibited association with developing a HAPI, the top 5 features included tracheostomy (odds ratio [OR] 4.5, 95% CI 4.0-5.1), peripheral edema (OR 2.9, 95% CI 2.6-3.2), central line (OR 2.1, 95% CI 1.9-2.3), first albumin measure (OR 0.6, 95% CI 0.6-0.6), and age (OR 1.2, 95% CI 1.2-1.2) (Figure 3). Although the directionality for each feature may vary, the relative importance in Figure 3 was ranked on a single scale. Additional significant features included whether the patient was on sympathomimetic medications, had a spinal cord injury or chest tube, and individual Braden score component measures. The final Vanderbilt model with 22 features provided excellent discriminatory ability with an AUC of 0.897 (95% CI 0.893-0.901). Multimedia Appendix 2 depicts the probability density plot for the development and validation cohorts.

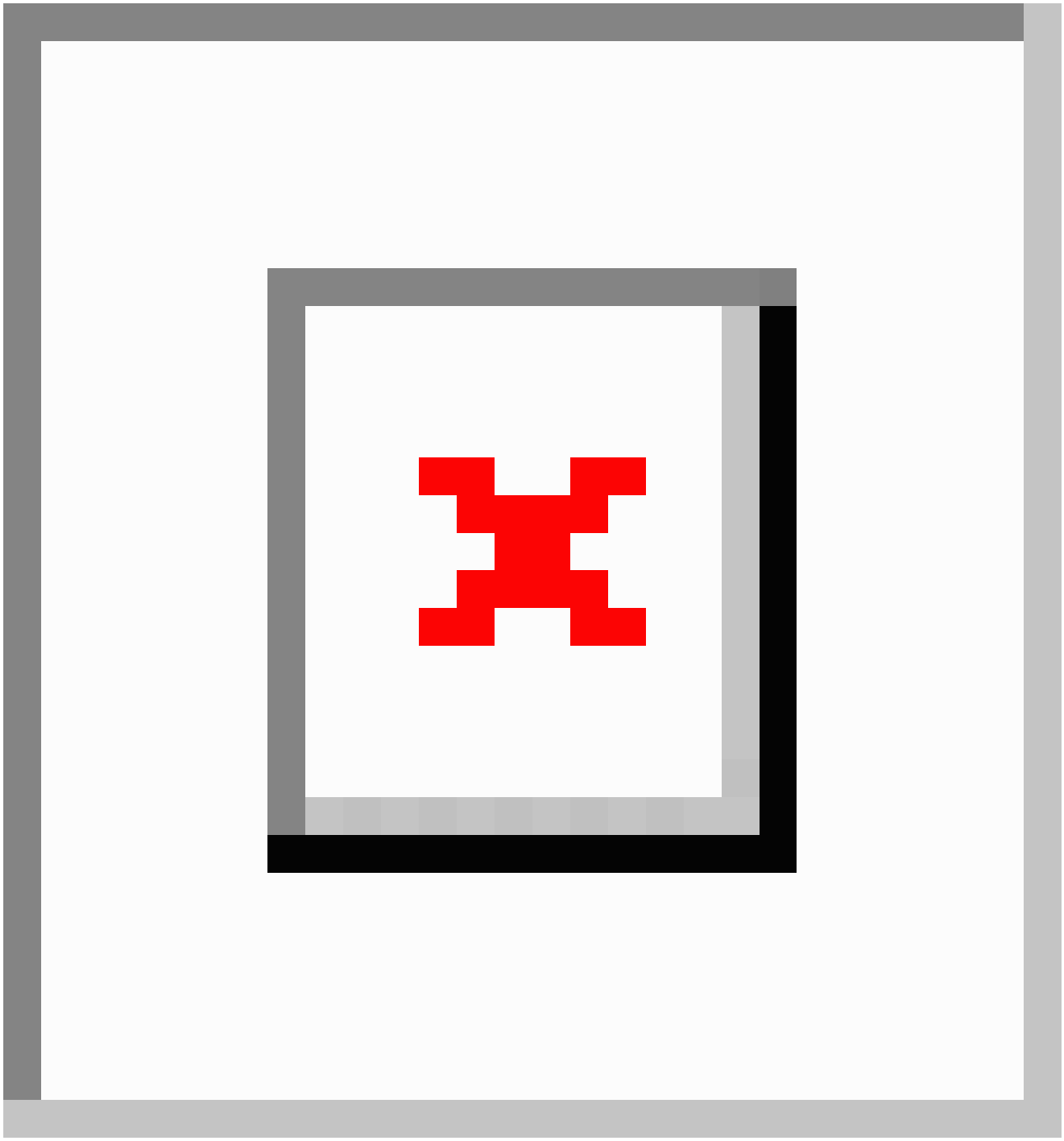
Figure 3. Relative importance of features used in the final Vanderbilt model. Gray subfeatures represent item comparisons used to generate features. *P* values for variable significance were derived using the Wald χ^2 test. BUN: blood urea nitrogen; ECMO: extracorporeal membrane oxygenation; ICU: intensive care unit; RDW: red cell distribution width.



Comparison With the Braden Scale

Using the model development cohort, the Vanderbilt model achieved an AUC of 0.897 (95% CI 0.893-0.901), compared to 0.798 (95% CI 0.791-0.803) and 0.733 (95% CI 0.725-0.740) for the continuous and dichotomous Braden, respectively (Figure 4).

Figure 4. Area under the receiver operating characteristic curve comparing the Vanderbilt (gold), continuous Braden (blue), and dichotomous Braden (gray) models.



Model Validation

The validation cohort consisted of 34,999 hospitalizations without a HAPI and 1096 hospitalizations with at least one HAPI. Model development and validation cohorts were compared to confirm that each had similar characteristics. Overall, characteristics were similar between the 2 cohorts (Table 3). We applied the same model from the development cohort to the validation cohort without adjusting coefficients, which provided a concordance statistic of 0.893 (95% CI 0.885-0.899; Table 4). Model calibration was consistent between the development and validation cohorts. The calibration curve indicated the model most accurately predicted risk for patients in the range of 0%-25% predicted risk (Figure 5); above this, the model could overpredict a HAPI. Since the model was intended to bring nurse attention and interventions to patients who would otherwise be overlooked, we believe the miscalibration at higher percentages was less clinically relevant. There was no evidence of collinearity. We are confident that this model performs well for most patients across the intensive care and general hospital settings, as 98.2% of the cohort had a predicted risk of less than 25%.

Since the model was designed to be used broadly in the general adult hospital, we performed a post hoc analysis among subpopulations for age (older than 65 years), gender, race, ethnicity, intensive care unit admission, and Braden score (greater than 18). The subpopulation analysis revealed only

slight changes in discrimination performance (Multimedia Appendix 3).

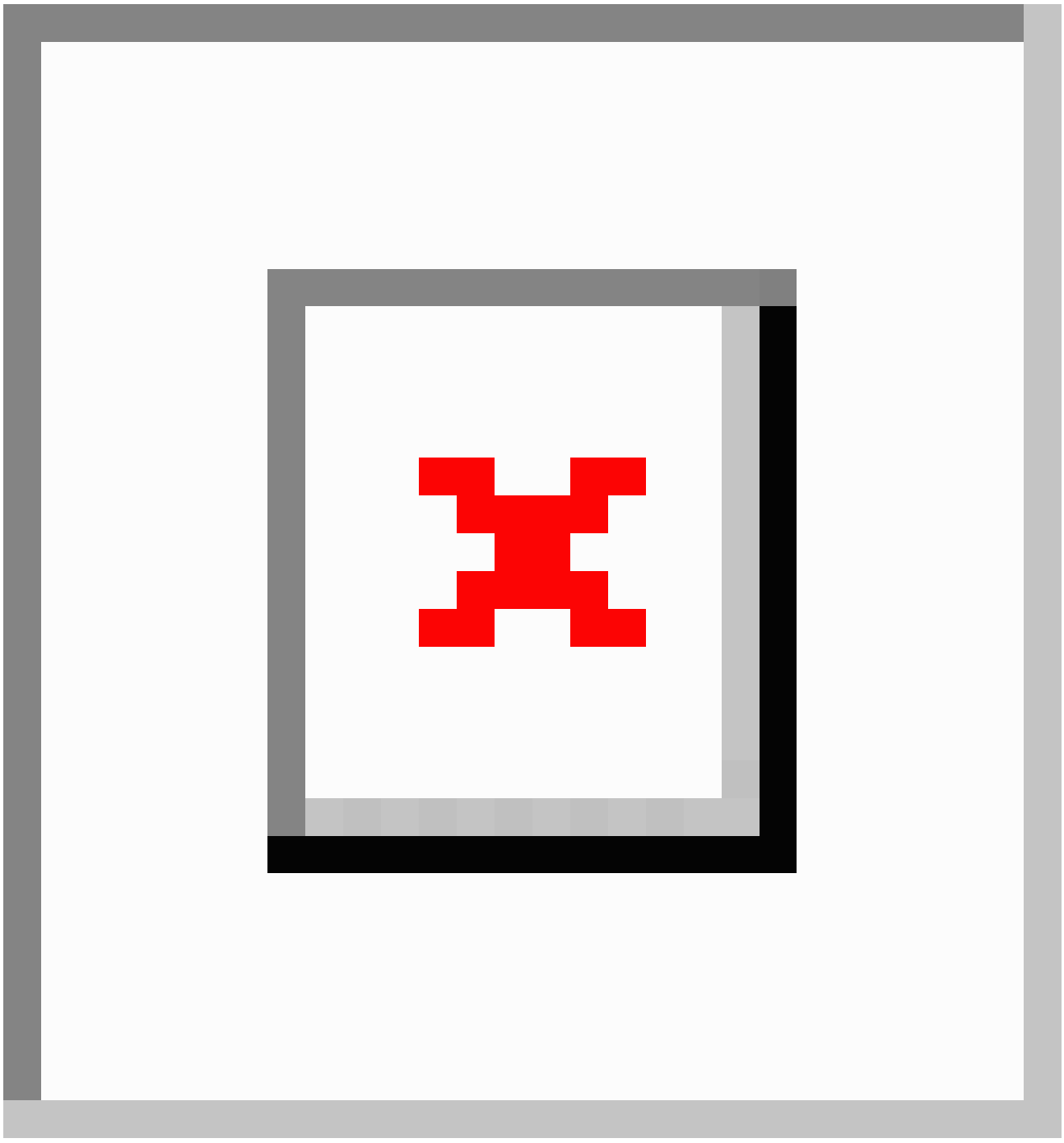
To operationalize the Vanderbilt model in the EHR (Epic), we generated the equation below. The output from the equation is a numeric probability from 0 to 1. Z is the sum of –4.1812002 and the product of the coefficient and measured value (eg, first albumin) for each feature. In Multimedia Appendix 4, we provide the coefficients for the equation. The model has been deployed as a population management tool to generate risk prediction data at Vanderbilt University Medical Center, but the output is only available for the research team until a trial period has been completed and governance has approved it for patient care. Within a report for multiple patients, output from the model is available as a column among other relevant factors to prioritize pressure injury interventions. As part of the implementation plan, we have created an application for potential users to test the model [56].

$$\text{Probability of hospital-acquired pressure injury} = 1/(1+\exp(-Z))$$

Table . Prediction model performance for hospital-acquired pressure injury.

Model	Area under the curve (95% CI)	Brier score	Integrated calibration index	Intercept	Slope
Vanderbilt (logistic regression)	0.893 (0.885-0.899)	0.026	0.006	−0.041	0.977
Continuous Braden (logistic regression)	0.799 (0.789-0.811)	0.028	0.006	0.178	1.034
Dichotomous Braden (score<18)	0.733 (0.725-0.740)	0.025	Too few levels to compute	0.0	1.0

Figure 5. Calibration curves for model development (left) and validation (right). Logistic calibration (solid line) represents parameter-based calibration (logistic regression model fit between predicted and observed values). Nonparametric calibration (dotted line) represents locally estimated scatterplot smoothing trend between predicted and observed values.



Discussion

Principal Findings

We developed and validated a risk prediction model for HAPIs that can be used in the general adult population. The model achieved excellent discrimination and adequate calibration (Table 4). Although several recent models have achieved similar performance, our model may have the greatest likelihood of reducing HAPIs because it was built with the foresight of overcoming known barriers to implementation of risk-prediction clinical decision support (Figure 1). According to the scoring criteria in Figure 1, the present model would have achieved 8 of a possible 10, compared to the current highest score of 6. It lost points for being limited to adults from a single institution (broadly applicable) and partially specified intervention (actionable criteria). Limiting development of the model to a single institution could limit the generalizability due to documentation patterns and data availability. Although we specified how to deploy the model in the EHR, the intervention components and implementation strategies were underspecified for implementation and evaluation. The next step is to test the effectiveness of the model in a pragmatic randomized clinical trial in which the intervention will be fully specified [57].

Although our model achieved similar performance and used the same regression approach as the top 3 models in Figure 1 (Ladios-Martin et al [25], Levy et al [27], and Song et al [26]), many of the most important features among the models varied. Among the most important features in the Ladios-Martin et al [25] model (eg, medical service, days of antidiabetic therapy, ability to eat, number of red blood cell units transfused, and hemoglobin range), only medical service was similar to our model. Relatedly, 2 important features in the Levy et al [27] model overlapped (friction and mobility). However, several important features from the Song et al [26] model (albumin, gait/transferring, activity, blood urea nitrogen, chloride, and spinal cord injury) overlapped with our model. We anticipate the similarity in features between our model and the Song et al [26] model was due to use of the same EHR and the models being developed at academic medical centers in the United States.

Limited implementation of risk prediction models in the EHR presents a critical challenge in health care today; the barrier is now less about the performance of risk prediction models and more the sociotechnical obstacles to uptake in patient care [58-60]. Despite the growing availability and sophistication of these models, their integration into routine clinical practice remains inadequate. Of the 22 models identified, we were unable to find one that decreased HAPIs. Even when prespecified elements for an implementable model are fulfilled, concerted efforts are needed from various stakeholders. Collaboration between health care organizations, technology developers, and regulatory bodies is essential to establish standards and guidelines for incorporating risk prediction models into EHR systems [61]. Enhancing data infrastructure, promoting data standardization, and developing robust privacy and security frameworks are crucial steps toward facilitating the implementation of these models [62]. Additionally, targeted education and training initiatives can help build trust and confidence among health care providers, encouraging their acceptance and use of risk prediction models in clinical practice, along with actionable steps to take for patients at highest risk [63,64]. Furthermore, there are significant socio-organizational barriers that impede the implementation of risk prediction models in EHRs. Resistance to change, lack of awareness or understanding among health care providers, and concerns regarding liability and accountability are common challenges faced by health care institutions. Clinicians may be skeptical of relying on risk prediction models, fearing that their judgment and decision-making autonomy may be compromised. The integration of risk prediction models also requires extensive training and education for health care providers, which may be resource-intensive and time-consuming [65,66]. Only when these barriers are addressed in a pragmatic manner can risk-prediction clinical decision support models improve patient outcomes.

Pragmatic trials are crucial in testing the real-world effectiveness and utility of interventions in health care settings [57,67,68]. These trials provide valuable insights into how interventions perform when integrated into routine clinical practice, considering factors such as patient outcomes, workflow integration, and usability. Institutions are beginning to develop the infrastructure and stakeholder engagement to support pragmatic trials. At our institution, Semler and colleagues [69] tested the effectiveness of balanced crystalloids and saline for fluids in critically ill adults. This pragmatic trial was cluster-randomized with 5 intensive care units. The authors found that use of balanced crystalloids resulted in a lower rate of death. A key aspect that makes pragmatic trials feasible is the use of existing infrastructure and real-world practice, which typically includes an inclusive patient population, minimal staff training, flexible protocols, minimally disruptive interventions, and outcomes captured as part of care. For pressure injuries specifically, the intervention infrastructure and guidance already exist as part of routine care; however, risk prediction will help identify and prioritize the most at-risk patients for targeted intervention. Preliminarily, we envision a clinician will use a list of patients ranked highest to lowest risk for HAPI.

Strengths and Limitations

Pressure injury prediction models have shown promise in identifying individuals at risk of developing pressure injuries. However, there are several limitations with these models, including ours, that should be considered. First, documentation of pressure injuries varies by institution and can lead to misclassification. We found that documentation of some pressure injuries carried over from previous encounters. On further testing, we found that missing measures (eg, albumin) can lead to inaccurate prediction. Thus, we chose to use a replicable imputation method with the median. Although our prediction model was developed and validated using incident HAPIs, documentation errors should be carefully considered. To increase the generalizability of our model, we chose not to include text from notes, despite evidence that use of clinical notes may have predictive power. Although we had a relatively large sample size that was sufficient to include all important features, the patient cohort was from a single institution and may not generalize to institutions in different geographical areas or using different EHRs. Finally, we chose to use an interpretable model that could be operationalized in current EHRs; however, other models may provide slightly higher performance. We anticipate certain EHR vendors will continue to develop capabilities for implementing complex machine learning models for more complicated prediction tasks. In anticipation of this, we performed a preliminary analysis of random forest, generalized additive model, and XGBoost. Of these models, we found that XGBoost had higher discrimination than ours in the model development cohort (AUC 0.960, 95% CI 0.957-0.962 vs AUC 0.893, 95% CI 0.885-0.899). In the model validation cohort, however, performance was not superior to logistic regression (AUC 0.869, 95% CI 0.861-0.877 vs AUC 0.893, 95% CI 0.885-0.899). Future work is needed to fully optimize the machine learning models and explore the tradeoff between interpretability and performance.

Conclusion

Despite numerous models developed to predict pressure injuries, studies demonstrating improved patient outcomes are missing. This is because implementing risk prediction models for routine patient care is complex and requires model developers, clinicians, and researchers to address challenges early in the

process. Therefore, we developed and validated an accurate prediction model for HAPIs that fulfilled necessary elements for implementation. The next step is to overcome socio-organizational barriers to rigorously evaluate the model through a pragmatic randomized clinical trial that includes targeted intervention for patients at highest risk. Our approach to developing an implementable risk prediction model, with feasible plans to evaluate its effectiveness, is generalizable to risk prediction and may be necessary to unlock the potential of this technology and improve decision-making.

Acknowledgments

We would like to thank Donald Sengstack for helping with the data extraction and Lance Mailloux for guidance on pressure injury quality improvement. This study was funded by the National Institutes of Health (R01 AG062499) and the Advanced Vanderbilt Artificial Intelligence Laboratory.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full cohort characteristics.
[DOCX File, 33 KB - medinform_v12i1e51842_app1.docx]

Multimedia Appendix 2
Density and probability plots for model development (A) and validation (B).
[PNG File, 51 KB - medinform_v12i1e51842_app2.png]

Multimedia Appendix 3
Subpopulation analysis of adult general hospital patients.
[DOCX File, 18 KB - medinform_v12i1e51842_app3.docx]

Multimedia Appendix 4
Features and coefficients of model and equation.
[DOCX File, 16 KB - medinform_v12i1e51842_app4.docx]

Checklist 1
Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist.
[PDF File, 115 KB - medinform_v12i1e51842_app5.pdf]

References

1. Edsberg LE, Black JM, Goldberg M, McNichol L, Moore L, Sieggreen M. Revised National Pressure Ulcer Advisory Panel pressure injury staging system. J Wound Ostomy Continence Nurs 2016;43(6):585-597. [doi: 10.1097/WON.0000000000000281] [Medline: 27749790]
2. Labeau SO, Afonso E, Benbenishty J, et al. Prevalence, associated factors and outcomes of pressure injuries in adult intensive care unit patients: the DecubICUs study. Intensive Care Med 2021 Feb;47(2):160-169. [doi: 10.1007/s00134-020-06234-9] [Medline: 33034686]
3. Mervis JS, Phillips TJ. Pressure ulcers: pathophysiology, epidemiology, risk factors, and presentation. J Am Acad Dermatol 2019 Oct;81(4):881-890. [doi: 10.1016/j.jaad.2018.12.069] [Medline: 30664905]
4. Padula WV, Mishra MK, Makic MBF, Sullivan PW. Improving the quality of pressure ulcer care with prevention: a cost-effectiveness analysis. . 2011 Apr(4) p. 385-392. [doi: 10.1097/MLR.0b013e31820292b3] [Medline: 21368685]
5. Padula WV, Delarmente BA. The national cost of hospital-acquired pressure injuries in the United States. Int Wound J 2019 Jun;16(3):634-640. [doi: 10.1111/iwj.13071] [Medline: 30693644]
6. Wound, Ostomy and Continence Nurses Society-Wound Guidelines Task Force. WOCN 2016 Guideline for Prevention and Management of Pressure Injuries (Ulcers): an executive summary. J Wound Ostomy Continence Nurs 2017;44(3):241-246. [doi: 10.1097/WON.0000000000000321] [Medline: 28472816]
7. Reddy M, Gill SS, Kalkar SR, Wu W, Anderson PJ, Rochon PA. Treatment of pressure ulcers: a systematic review. JAMA 2008 Dec 10;300(22):2647-2662. [doi: 10.1001/jama.2008.778] [Medline: 19066385]
8. Reddy M, Gill SS, Rochon PA. Preventing pressure ulcers: a systematic review. JAMA 2006 Aug 23;296(8):974-984. [doi: 10.1001/jama.296.8.974] [Medline: 16926357]

9. Kavanagh KT, Dykes PC. Hospital pressure injury metrics, an unfulfilled need of paramount importance. J Patient Saf 2021 Apr 1;17(3):189-191. [doi: 10.1097/PTS.0000000000000694] [Medline: 32805091]
10. Alderden J, Rondinelli J, Pepper G, Cummins M, Whitney J. Risk factors for pressure injuries among critical care patients: a systematic review. Int J Nurs Stud 2017 Jun;71:97-114. [doi: 10.1016/j.ijnurstu.2017.03.012] [Medline: 28384533]
11. Serrano ML, Méndez MIG, Cebollero FMC, Rodríguez JSL. Risk factors for pressure ulcer development in intensive care units: a systematic review. Med Intensiva 2017 Aug;41(6):339-346. [doi: 10.1016/j.medine.2017.04.006]
12. Liao Y, Gao G, Mo L. Predictive accuracy of the Braden Q scale in risk assessment for paediatric pressure ulcer: a meta-analysis. Int J Nurs Sci 2018 Oct 10;5(4):419-426. [doi: 10.1016/j.ijnss.2018.08.003] [Medline: 31406858]
13. Wei M, Wu L, Chen Y, Fu Q, Chen W, Yang D. Predictive validity of the Braden scale for pressure ulcer risk in critical care: a meta-analysis. Nurs Crit Care 2020 May;25(3):165-170. [doi: 10.1111/nicc.12500] [Medline: 31985893]
14. Papanikolaou P, Lyne P, Anthony D. Risk assessment scales for pressure ulcers: a methodological review. Int J Nurs Stud 2007 Feb;44(2):285-296. [doi: 10.1016/j.ijnurstu.2006.01.015] [Medline: 17141782]
15. Huang C, Ma Y, Wang C, et al. Predictive validity of the Braden scale for pressure injury risk assessment in adults: a systematic review and meta-analysis. Nurs Open 2021 Sep;8(5):2194-2207. [doi: 10.1002/nop2.792] [Medline: 33630407]
16. Hyun S, Vermillion B, Newton C, et al. Predictive validity of the Braden scale for patients in intensive care units. Am J Crit Care 2013 Nov;22(6):514-520. [doi: 10.4037/ajcc2013991] [Medline: 24186823]
17. Dweekat OY, Lam SS, McGrath L. Machine learning techniques, applications, and potential future opportunities in pressure injuries (bedsores) management: a systematic review. Int J Environ Res Public Health 2023 Jan 1;20(1):796. [doi: 10.3390/ijerph20010796] [Medline: 36613118]
18. Jiang M, Ma Y, Guo S, et al. Using machine learning technologies in pressure injury management: systematic review. JMIR Med Inform 2021 Mar 10;9(3):e25704. [doi: 10.2196/25704] [Medline: 33688846]
19. Ribeiro F, Fidalgo F, Silva A, Metrôlho J, Santos O, Dionisio R. Literature review of machine-learning algorithms for pressure ulcer prevention: challenges and opportunities. Informatics 2021 Dec 1;8(4):76. [doi: 10.3390/informatics8040076]
20. Zhou Y, Yang X, Ma S, Yuan Y, Yan M. A systematic review of predictive models for hospital-acquired pressure injury using machine learning. Nurs Open 2023 Mar;10(3):1234-1246. [doi: 10.1002/nop2.1429] [Medline: 36310417]
21. Qu C, Luo W, Zeng Z, et al. The predictive effect of different machine learning algorithms for pressure injuries in hospitalized patients: a network meta-analyses. Heliyon 2022 Nov;8(11):e11361. [doi: 10.1016/j.heliyon.2022.e11361] [Medline: 36387440]
22. Kitzmiller RR, Vaughan A, Skeeles-Worley A, et al. Diffusing an innovation: clinician perceptions of continuous predictive analytics monitoring in intensive care. Appl Clin Inform 2019 Mar;10(2):295-306. [doi: 10.1055/s-0039-1688478] [Medline: 31042807]
23. Randall Moorman J. The principles of whole-hospital predictive analytics monitoring for clinical medicine originated in the neonatal ICU. NPJ Digit Med 2022 Mar 31;5(1):41. [doi: 10.1038/s41746-022-00584-y] [Medline: 35361861]
24. Keim-Malpass J, Kitzmiller RR, Skeeles-Worley A, et al. Advancing continuous predictive analytics monitoring: moving from implementation to clinical action in a learning health system. Crit Care Nurs Clin North Am 2018 Jun;30(2):273-287. [doi: 10.1016/j.cnc.2018.02.009] [Medline: 29724445]
25. Ladios-Martin M, Fernández-de-Maya J, Ballesta-López FJ, Belso-Garzas A, Mas-Asencio M, Cabañero-Martínez MJ. Predictive modeling of pressure injury risk in patients admitted to an intensive care unit. Am J Crit Care 2020 Jul 1;29(4):e70-e80. [doi: 10.4037/ajcc2020237] [Medline: 32607572]
26. Song W, Kang MJ, Zhang L, et al. Predicting pressure injury using nursing assessment phenotypes and machine learning methods. J Am Med Inform Assoc 2021 Mar 18;28(4):759-765. [doi: 10.1093/jamia/ocaa336] [Medline: 33517452]
27. Levy JJ, Lima JF, Miller MW, Freed GL, O'Malley AJ, Emeny RT. Machine learning approaches for hospital acquired pressure injuries: a retrospective study of electronic medical records. Front Med Technol 2022 Jun;4:926667. [doi: 10.3389/fmedt.2022.926667] [Medline: 35782577]
28. Ossai CI, O'Connor L, Wickramasinghe N. Real-time inpatients risk profiling in acute care: a comparative study of falls and pressure injuries vulnerabilities. In: Pucihar A, Kljajic Borstnar M, Bons R, editors. 33rd BLED eConference: Enabling Technology for a Sustainable Society: University of Maribor Press; 2021:35-50. [doi: 10.18690/978-961-286-362-3.3]
29. Cai JY, Zha ML, Song YP, Chen HL. Predicting the development of surgery-related pressure injury using a machine learning algorithm model. J Nurs Res 2020 Dec 21;29(1):e135. [doi: 10.1097/JNR.0000000000000411] [Medline: 33351552]
30. Xu J, Chen D, Deng X, et al. Development and validation of a machine learning algorithm–based risk prediction model of pressure injury in the intensive care unit. Int Wound J 2022 Nov;19(7):1637-1649. [doi: 10.1111/iwj.13764] [Medline: 35077000]
31. Šín P, Hokynková A, Marie N, Andrea P, Krč R, Podroužek J. Machine learning-based pressure ulcer prediction in modular critical care data. Diagnostics (Basel) 2022 Mar 30;12(4):850. [doi: 10.3390/diagnostics12040850] [Medline: 35453898]
32. Do Q, Lipatov K, Ramar K, Rasmusson J, Pickering BW, Herasevich V. Pressure injury prediction model using advanced analytics for at-risk hospitalized patients. J Patient Saf 2022 Oct 1;18(7):e1083-e1089. [doi: 10.1097/PTS.0000000000001013] [Medline: 35588068]

33. Walther F, Heinrich L, Schmitt J, Eberlein-Gonska M, Roessler M. Prediction of inpatient pressure ulcers based on routine healthcare data using machine learning methodology. Sci Rep 2022 Mar 23;12(1):5044. [doi: 10.1038/s41598-022-09050-x] [Medline: 35322109]
34. Anderson C, Bekele Z, Qiu Y, Tschannen D, Dinov ID. Modeling and prediction of pressure injury in hospitalized patients using artificial intelligence. BMC Med Inform Decis Mak 2021 Aug 30;21(1):253. [doi: 10.1186/s12911-021-01608-5] [Medline: 34461876]
35. Cheng FM, Jin YJ, Chien CW, Chuang YC, Tung TH. The application of Braden scale and rough set theory for pressure injury risk in elderly male population. J Mens Health 2021 Sep;17(4):156-165. [doi: 10.31083/jomh.2021.022]
36. Song J, Gao Y, Yin P, et al. The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms. Risk Manag Healthc Policy 2021 Mar;14:1175-1187. [doi: 10.2147/RMHP.S297838] [Medline: 33776495]
37. Nakagami G, Yokota S, Kitamura A, et al. Supervised machine learning-based prediction for in-hospital pressure injury development using electronic health records: a retrospective observational cohort study in a university hospital in Japan. Int J Nurs Stud 2021 Jul;119:103932. [doi: 10.1016/j.ijnurstu.2021.103932] [Medline: 33975074]
38. Delparte JJ, Flett HM, Scovil CY, Burns AS. Development of the spinal cord injury pressure sore onset risk screening (SCI-Presors) instrument: a pressure injury risk decision tree for spinal cord injury rehabilitation. Spinal Cord 2021 Feb;59(2):123-131. [doi: 10.1038/s41393-020-0510-y] [Medline: 32694750]
39. Vyas K, Samadani A, Milosevic M, Ostadabbas S, Parvaneh S. Additional value of augmenting current subscales in Braden scale with advanced machine learning technique for pressure injury risk assessment. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE; 2020:2993-2995. [doi: 10.1109/BIBM49941.2020.9313401]
40. Sotoodeh M, Gero ZH, Zhang W, Hertzberg VS, Ho JC. Pressure ulcer injury in unstructured clinical notes: detection and interpretation. AMIA Annu Symp Proc 2021 Jan 25;2020:1160-1169. [Medline: 33936492]
41. Alderden J, Drake KP, Wilson A, Dimas J, Cummins MR, Yap TL. Hospital acquired pressure injury prediction in surgical critical care patients. BMC Med Inform Decis Mak 2021 Jan 6;21(1):12. [doi: 10.1186/s12911-020-01371-z] [Medline: 33407439]
42. Choi BK, Kim MS, Kim SH. Risk prediction models for the development of oral-mucosal pressure injuries in intubated patients in intensive care units: a prospective observational study. J Tissue Viability 2020 Nov;29(4):252-257. [doi: 10.1016/j.jtv.2020.06.002] [Medline: 32800513]
43. Hu YH, Lee YL, Kang MF, Lee PJ. Constructing inpatient pressure injury prediction models using machine learning techniques. Comput Inform Nurs 2020 Aug;38(8):415-423. [doi: 10.1097/CIN.0000000000000604] [Medline: 32205474]
44. Goodwin TR, Demner-Fushman D. A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision. J Am Med Inform Assoc 2020 Apr 1;27(4):567-576. [doi: 10.1093/jamia/ocaa004] [Medline: 32065628]
45. Ahmad MA, Larson B, Overman S, et al. Machine learning approaches for pressure injury prediction. In: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI): IEEE; 2021:427-431. [doi: 10.1109/ICHI52183.2021.00069]
46. Walker SC, French B, Moore RP, et al. Model-guided decision-making for thromboprophylaxis and hospital-acquired thromboembolic events among hospitalized children and adolescents: the CLOT randomized clinical trial. JAMA Netw Open 2023 Oct 2;6(10):e2337789. [doi: 10.1001/jamanetworkopen.2023.37789] [Medline: 37831448]
47. Walker SC, Creech CB, Domenico HJ, French B, Byrne DW, Wheeler AP. A real-time risk-prediction model for pediatric venous thromboembolic events. Pediatrics 2021 Jun;147(6):e2020042325. [doi: 10.1542/peds.2020-042325] [Medline: 34011634]
48. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
49. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989 May;8(5):551-561. [doi: 10.1002/sim.4780080504] [Medline: 2657958]
50. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat 1996 Jan;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
51. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagn Progn Res 2017 Dec;1:19. [doi: 10.1186/s41512-017-0020-3] [Medline: 31093548]
52. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019 Sep 20;38(21):4051-4065. [doi: 10.1002/sim.8281] [Medline: 31270850]
53. Agarwal R, Domenico HJ, Balla SR, et al. Palliative care exposure relative to predicted risk of six-month mortality in hospitalized adults. J Pain Symptom Manage 2022 May;63(5):645-653. [doi: 10.1016/j.jpainsymman.2022.01.013] [Medline: 35081441]
54. Freundlich RE, Li G, Domenico HJ, Moore RP, Pandharipande PP, Byrne DW. A predictive model of reintubation after cardiac surgery using the electronic health record. Ann Thorac Surg 2022 Jun;113(6):2027-2035. [doi: 10.1016/j.athoracsur.2021.06.060] [Medline: 34329600]

55. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Eur Urol 2015 Jun;67(6):1142-1151. [doi: 10.1016/j.eururo.2014.11.025] [Medline: 25572824]
56. Domenico H, Reese T, Moore R, Byrne D, Hernandez T. Predicted risk of hospital acquired pressure injury calculator. Vanderbilt University Medical Center. 2023. URL: https://cqs.app.vumc.org/shiny/PressureInjuryPrediction/ [accessed 2023-08-13]
57. Byrne DW. Artificial Intelligence for Improved Patient Outcomes: Principles for Moving Forward with Rigorous Science: Lippincott Williams & Wilkins; 2022.
58. Reese TJ, Liu S, Steitz B, et al. Conceptualizing clinical decision support as complex interventions: a meta-analysis of comparative effectiveness trials. J Am Med Inform Assoc 2022 Sep 12;29(10):1744-1756. [doi: 10.1093/jamia/ocac089] [Medline: 35652167]
59. Weaver CGW, McAlister FA. Machine learning, predictive analytics, and the emperor's new clothes: why artificial intelligence has not yet replaced conventional approaches. Can J Cardiol 2021 Aug;37(8):1156-1158. [doi: 10.1016/j.cjca.2021.03.003] [Medline: 33711476]
60. Reese TJ, Mixon AS, Matheny ME, et al. Using intervention mapping to design and implement a multicomponent intervention to improve antibiotic and NSAID prescribing. Transl Behav Med 2023 Dec 15;13(12):928-943. [doi: 10.1093/tbm/ibad063] [Medline: 37857368]
61. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Affairs 2014 Jul;33(7):1148-1154. [doi: 10.1377/hlthaff.2014.0352] [Medline: 25006140]
62. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar;28(3):231-237. [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
63. Gomez Lumbreras A, Reese TJ, Del Fiol G, et al. Shared decision-making for drug-drug interactions: formative evaluation of an anticoagulant drug interaction. JMIR Form Res 2022 Oct 19;6(10):e40018. [doi: 10.2196/40018] [Medline: 36260377]
64. Reese TJ, Del Fiol G, Morgan K, et al. A shared decision-making tool for drug interactions between warfarin and nonsteroidal anti-inflammatory drugs: design and usability study. JMIR Hum Factors 2021 Oct 26;8(4):e28618. [doi: 10.2196/28618] [Medline: 34698649]
65. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
66. Reese TJ, Schlechter CR, Kramer H, et al. Implementing lung cancer screening in primary care: needs assessment and implementation strategy design. Transl Behav Med 2022 Feb 16;12(2):187-197. [doi: 10.1093/tbm/ibab115] [Medline: 34424342]
67. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
68. Cabitza F, Zeitoun JD. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Ann Transl Med 2019 Apr;7(8):161. [doi: 10.21037/atm.2019.04.07] [Medline: 31168442]
69. Semler MW, Self WH, Wanderer JP, et al. Balanced crystalloids versus saline in critically ill adults. N Engl J Med 2018 Mar 1;378(9):829-839. [doi: 10.1056/NEJMoa1711584] [Medline: 29485925]

Abbreviations

AUC: area under the receiver operating curve
EHR: electronic health record
HAPI: hospital-acquired pressure injury
LASSO: least absolute shrinkage and selection operator
OR: odds ratio
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 14.08.23; peer-reviewed by J Walsh, L Zhang, N Fareed, P Okoro, P Kukhareva, W Wei; revised version received 08.03.24; accepted 10.03.24; published 08.05.24.

Please cite as:
Reese TJ, Domenico HJ, Hernandez A, Byrne DW, Moore RP, Williams JB, Douthit BJ, Russo E, McCoy AB, Ivory CH, Steitz BD, Wright A
Implementable Prediction of Pressure Injuries in Hospitalized Adults: Model Development and Validation
JMIR Med Inform 2024;12:e51842
URL: https://medinform.jmir.org/2024/1/e51842
doi:10.2196/51842

A Machine Learning Application to Classify Patients at Differing Levels of Risk of Opioid Use Disorder: Clinician-Based Validation Study

Tewodros Eguale[1,2,*], MD, PhD; François Bastardot[3,4,*], MSACI, MD; Wenyu Song[2,5], PhD; Daniel Motta-Calderon[6], MD; Yasmin Elsobky[2,7], BPharmSci, MSc; Angela Rui[2], MA; Marlika Marceau[8], BA; Clark Davis[2], BA; Sandya Ganesan[2], BA; Ava Alsubai[2], MA; Michele Matthews[1,9], PharmD; Lynn A Volk[8], MHS; David W Bates[2,5,8,10], MSc, MD; Ronen Rozenblum[2,5], MPH, PhD

1

2

3

4

5

6

7

8

9

10

[*]these authors contributed equally

Corresponding Author:
Ronen Rozenblum, MPH, PhD

Abstract

Background: Despite restrictive opioid management guidelines, opioid use disorder (OUD) remains a major public health concern. Machine learning (ML) offers a promising avenue for identifying and alerting clinicians about OUD, thus supporting better clinical decision-making regarding treatment.

Objective: This study aimed to assess the clinical validity of an ML application designed to identify and alert clinicians of different levels of OUD risk by comparing it to a structured review of medical records by clinicians.

Methods: The ML application generated OUD risk alerts on outpatient data for 649,504 patients from 2 medical centers between 2010 and 2013. A random sample of 60 patients was selected from 3 OUD risk level categories (n=180). An OUD risk classification scheme and standardized data extraction tool were developed to evaluate the validity of the alerts. Clinicians independently conducted a systematic and structured review of medical records and reached a consensus on a patient's OUD risk level, which was then compared to the ML application's risk assignments.

Results: A total of 78,587 patients without cancer with at least 1 opioid prescription were identified as follows: not high risk (n=50,405, 64.1%), high risk (n=16,636, 21.2%), and suspected OUD or OUD (n=11,546, 14.7%). The sample of 180 patients was representative of the total population in terms of age, sex, and race. The interrater reliability between the ML application and clinicians had a weighted kappa coefficient of 0.62 (95% CI 0.53-0.71), indicating good agreement. Combining the high risk and suspected OUD or OUD categories and using the review of medical records as a gold standard, the ML application had a corrected sensitivity of 56.6% (95% CI 48.7%-64.5%) and a corrected specificity of 94.2% (95% CI 90.3%-98.1%). The positive and negative predictive values were 93.3% (95% CI 88.2%-96.3%) and 60.0% (95% CI 50.4%-68.9%), respectively. Key themes for disagreements between the ML application and clinician reviews were identified.

Conclusions: A systematic comparison was conducted between an ML application and clinicians for identifying OUD risk. The ML application generated clinically valid and useful alerts about patients' different OUD risk levels. ML applications hold promise for identifying patients at differing levels of OUD risk and will likely complement traditional rule-based approaches to generating alerts about opioid safety issues.

(JMIR Med Inform 2024;12:e53625) doi:10.2196/53625

KEYWORDS

opioid-related disorders; opioid use disorder; machine learning; artificial intelligence; electronic health record; clinical decision support; model validation; patient medication safety; medication safety; clinical decision; decision making; decision support; patient safety; opioid use; drug use; opioid safety; medication; OUD; EHR; AI

Introduction

In the past few decades, the "opioid epidemic" has become a public health crisis. According to a 2020 US survey, 2.7 million people aged 12 years or older had an opioid use disorder (OUD), and only 1 in 9 (11.2%) received medication-assisted therapy [1]. OUD is a frequently underdiagnosed condition, and it is estimated that for every patient with an OUD diagnosis, there are at least 2 who remain undiagnosed [2]. In 2021, nearly 92,000 drug overdose deaths were reported in the United States [3]. Furthermore, 54% and 46% of the US $1.02 trillion aggregate annual societal costs in 2020 in the United States were attributed to overdose deaths and OUD, respectively [4].

There is an immediate urgency to identify patients at high risk of OUD and those with OUD. Clinicians have reported major barriers to adequately assessing patients' risk, including time pressure, incomplete or restricted medical records, and a lack of robust clinical decision support systems (CDSSs) [5,6]. The current rule-based approaches, such as Medicare Part D's Overutilization Monitoring System or statewide Prescription Drug Monitoring Programs, fail to incorporate clinical data and are often underused [7]. Moreover, unless CDSSs use individual patient-specific clinical data in generating alerts, many false positive alerts may be presented to clinicians contributing to alert fatigue [8].

Artificial intelligence and machine learning (ML) algorithms have recently demonstrated their usefulness in CDSSs; however, compared with conventional statistical methods, their black-box nature and a lack of studies assessing the clinical validity of these interventions have created uneasiness in the medical community [9-12]. MedAware is a commercial software application that uses various statistical and ML methods to identify and prevent medication safety issues, including the risk of OUD [13]. It uses an iterative development process and has conducted pilot testing to optimize its OUD risk prediction algorithm to increase its accuracy in patient risk identification.

The goals of this study were to assess the clinical validity of the ML application by (1) determining the agreement between the ML algorithm's output and the outcomes of structured clinicians' review of medical records in classifying patients into distinct categories of OUD risk, including not high risk, high risk, or suspected OUD or OUD; (2) determining the potential utility of using the ML application as an alerting tool by evaluating its test characteristics against the gold standard; and (3) identifying major factors contributing to discrepancies between the ML application and clinician risk assignments to provide a knowledge base for future system improvement.

Methods

Ethical Considerations

This study was approved by the Mass General Brigham Institutional Review Boards (#2014P002167) that granted a patient waiver of consent for this study. Patients did not receive any compensation.

Evaluation of the ML Application

MedAware (Ra'anana, Israel) has developed an ML software application to identify prescription errors and adverse drug events [13]. This application identifies medication issues based on ML methods including random forest algorithms—a widely used ML method in medical applications [14]. Multiple studies using ML models for disease prediction have achieved robust performance [15,16].

Based on clinical data in the electronic health record (EHR), the ML application's algorithms generate patient-specific alerts on medication orders that deviate from predominant prescribing patterns in similar patient situations. Previously, it was found that the ML application generates medication error alerts that might otherwise be missed with existing applications with a high degree of alert usefulness, and it has the potential to reduce costs [17,18].

The ML application has been enhanced to generate alerts in real time to identify patients at risk of OUD and overdose based on clinical, psychosocial, and medication data. The input features used in the model were age, gender, opioid and nonopioid medication history (for each prescription: drug name, route of administration, duration, and dosage), and diagnosis history found in *ICD-9* (*International Classification of Diseases*) diagnoses codes and problem lists. The application can also produce aggregate alert data about the risk of OUD or overdose, which may be used for population health management.

The model outcome was defined by MedAware by combining OUD diagnosis codes, medication use, and experts' annotation. The test cohort was independent from the training set to avoid overfitting. Random data splitting was conducted to separate training (50%) and test (50%) sets. MedAware used a scikit-learn (1.2.0) implementation of the random forest algorithm. It was used in a cross-fold manner and some of its hyperparameters (mainly: n_estimators, max_depth, class_weight) were tuned for optimization while leaving others at their default values. Additional details of the ML algorithm were not available to the research team because of intellectual property protections and were not the focus of this study; our study aimed to clinically validate OUD alerts generated by the algorithm against clinician judgement.

Study Setting and Patient Population

The patient population of this study comprised patients who had at least 1 outpatient encounter between January 1, 2012,

and December 31, 2013, and were prescribed at least 1 opioid medication between January 1, 2010, and December 31, 2013, in an outpatient setting at 2 large academic medical centers in the United States. Patients diagnosed with cancer and those with incomplete data were excluded. Once a patient had a documented OUD diagnosis or started receiving opioid rehabilitation drugs (eg, suboxone, naltrexone, methadone, and bezitramide), any subsequent patient data were excluded from the analysis as the patient's status was known.

The evaluated application classified patients into 3 levels of OUD risk: not high risk, high risk, and suspected OUD or OUD. Alerts to clinicians are generated for only the high risk and suspected OUD or OUD categories. The risk alerts are generated when a clinician initiates an opioid medication prescription. A short textual description is created by the application for each alert generated to explain why the alert fired, for example, *the patient has a long opioid sequence, concurrent benzodiazepines use*. This explanation enables clinicians to understand the general reasoning underlying the alert. To improve study efficiency, the validation study comprised a random sample of 60 patients from each risk category for a total of 180 cases for which a retrospective review was performed by clinicians [19].

Data Collection and Transfer

Clinical and encounter data on the patient population from 2010 to 2013 were extracted and sent to MedAware, including demographics, diagnoses, problem lists, outpatient and inpatient encounters, encounter clinicians, clinician specialties, procedures, medications, allergies, vital signs, and selected blood test outcomes. Patient and clinician names and medical record numbers were removed from the data set, and a random study ID was assigned to each patient and clinician before the limited data set was sent through a secure transfer application (password-protected and encrypted) for analysis.

Development of a Risk Classification Scheme and Pilot Testing

Evaluation criteria for risk assignment by clinicians using the clinical data were developed with an extensive review of established guidelines, such as those of the Centers for Disease Control and Prevention and *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*), and risk factors for OUD through an iterative process in consultation with experts in the field of pain and opioid management [20-24]. The research clinicians and team reviewed the Centers for Disease Control and Prevention's and *DSM-5* guidelines and created draft criteria based on these guidelines to reflect 3 levels of risk, and then these criteria were reviewed by 2 pain management experts (a physician and a pharmacist). After modifications, this risk classification scheme was piloted to evaluate its effectiveness and compatibility with the ML application. We conducted the pilot review of medical records with 25 randomly selected medical records. One research assistant (CD) extracted data from the medical records using a standardized data collection tool as described below and 2 physician reviewers (FB and TE) individually reviewed the data. The reviewers reached a consensus on their risk determinations, and revisions were made to criteria, as needed, to standardize assessments and support a more transparent, generalizable validation process. MedAware sent a list of those patients for whom a risk assessment was conducted to be used for selecting the random sample for review of medical records.

Structured Clinicians' Review of Medical Records Using a Standardized Data Collection Tool

In total, 180 patients with a history of opioid use were randomly selected from those patients classified by the ML application into 3 risk categories (60 in each group), and structured reviews of medical records were conducted to evaluate patients' OUD risk. Clinicians were blinded to the patients' risk assignment by the application. A data abstraction tool was developed to organize relevant patients' clinical data from an EHR and facilitate the process for the review of medical records (Figure 1). This tool contains important demographic, patient, and family medical history including psychiatric and psychosocial information, patient complaints as documented in relevant clinical notes, relevant laboratory findings and drug history with graphical representation of opioid drug start and stop dates (ie, medication timeline; Figure 2), clinical events relevant to pain management such as surgeries or dates of major accidents, admission and emergency room visits, and curated clinical notes related to relevant clinical events. Collected data included both structured and free-text data that were extracted by research staff and organized into the abstraction tool. Data collection was focused on relevant information during the 2010-2013 time period; however, as the complete medical record was available for review, relevant information available prior to 2010 may have been considered. After training, 5 research assistants (CD, AA, SG, AR, and MM) individually extracted clinical data. Information from medical records was reviewed by extractors and clinicians up to the ML application's first alert date (index date). For patients determined to be not high risk by the ML application, a random date was assigned up to which medical record data were extracted and reviewed.

Figure 1. Tool used to extract data from patients' medical records. Template used to organize patient information extracted during the review of electronic health records (EHRs). A patient's demographics and relevant past medical, psychosocial, family, and medication histories were captured. Provider notes and encounters relevant to opioid use and pain management were described and recorded by date. Any patient's laboratory findings relevant to opioid use or other medications of interest were also recorded. Clinician reviewers recorded their risk categorization and rationalization after reviewing the information captured on the data extraction tool and reviewing the EHR, as needed. OUD: opioid use disorder; PRN: pro re nata; SUD: substance use disorder.

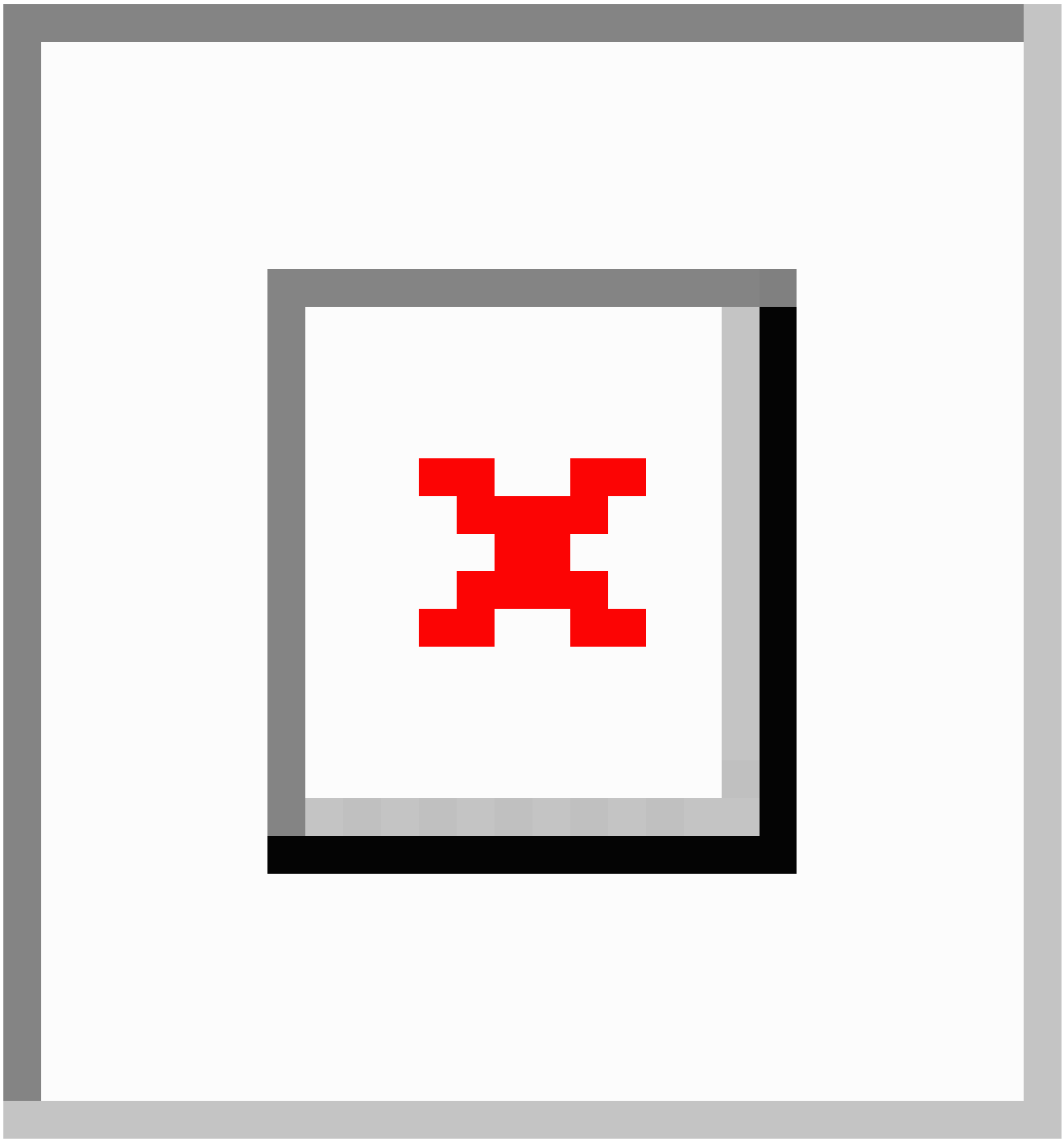
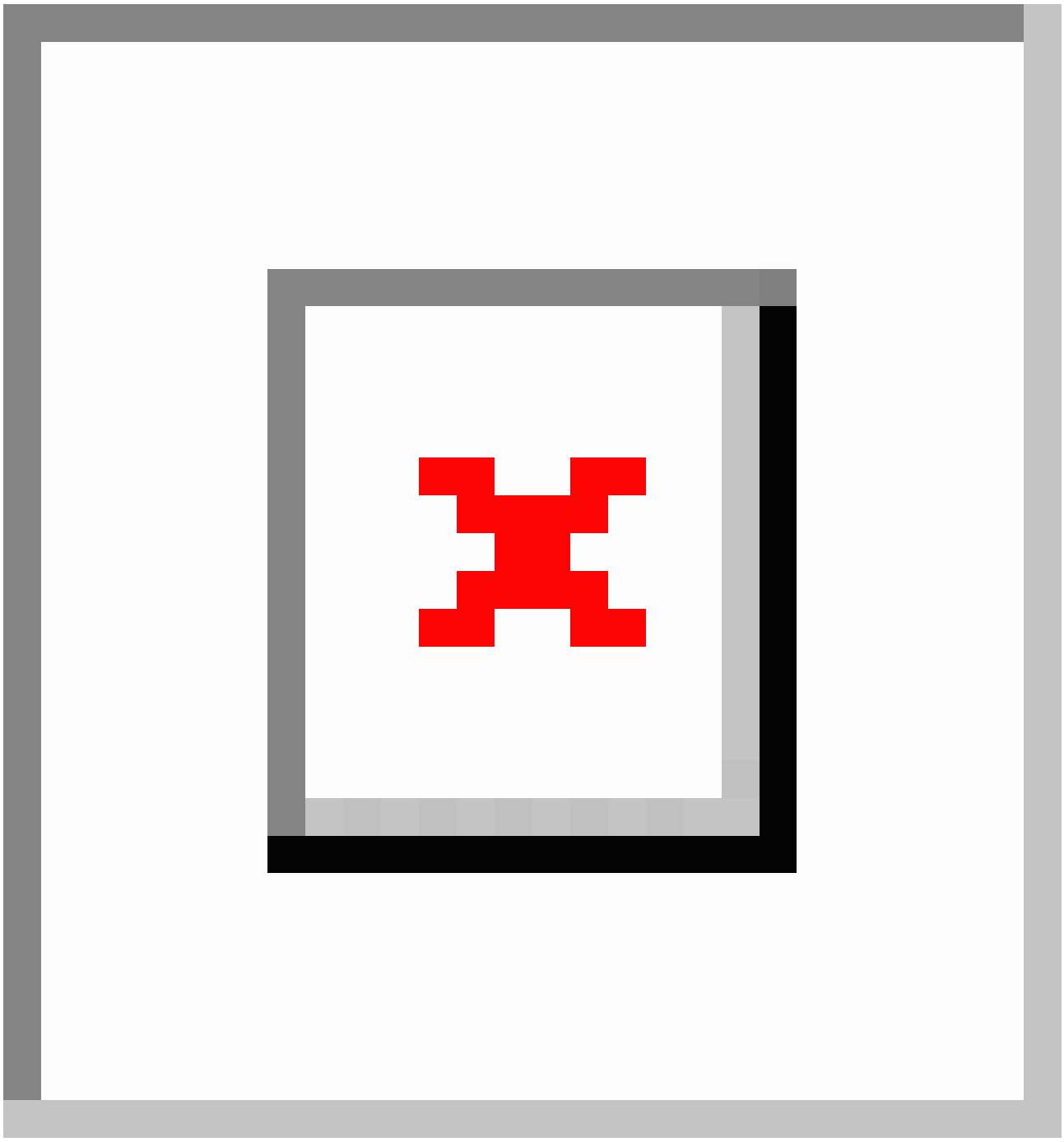


Figure 2. Example medication timeline from the tool used to extract data from the review of medical records. This timeline was created after medication history was recorded and morphine milligram equivalent (MME) conversion for opioid medications was included. Relevant encounters were recorded by date to provide context for the medication timeline (eg, surgery). Medications of interest included opioids, benzodiazepines, antidepressants, anticonvulsants, and other nonopioid medications contributing to risk. BID: twice a day; Q6H: every 6 hours; Q8H: every 8 hours; QD: every day, daily; SR: sustained release.



Four clinician reviewers (FB, TE, DM, and YE) individually examined data extracted by the research assistants and reviewed the EHRs directly, as needed, to holistically understand the clinical context of opiate prescription for the patients. The reviewers comprised 2 general internal medicine physicians, a hospitalist with extensive daily opioid prescribing experience, one with a PhD focusing on pharmacoepidemiology and drug safety, a recent medical student, and a pharmacist. All medical records were reviewed by 2 independent clinicians. After the primary review of medical records, a second reviewer blinded to the risk assignment of the first reviewer determined the risk level for OUD. The 2 clinicians discussed the case to reach consensus when their risk assignments differed. This consensus determination was then compared to the ML application's alert. Statistical analyses were conducted to evaluate the level of agreement between the clinician reviewers and the ML application's risk classifications.

Evaluation of Reasons for Disagreement Between Risk Assignments

To evaluate and identify the main reasons for disagreement between the clinician reviewers and the ML application's risk

classifications, a qualitative analysis was also conducted. For cases where there was disagreement, additional information contributing to the system risk assessment was requested from MedAware. Using a thematic analysis approach, 3 members of the research team (AR, LAV, and MM) independently conducted a qualitative analysis of the alert information. They reviewed the ML application's reasoning for assigning a particular risk category, information from the data extraction sheet, and information from the clinician reviewer's final risk assignment consensus. Then, this information was systematically coded to identify, categorize, and sort key concepts for the disagreements. Codes were then grouped into emergent themes and relationships after iterative review and discussion. In cases where there was disagreement, all 3 researchers reviewed and discussed the case together to reach consensus.

Statistical Analysis

We used descriptive statistics to summarize demographic characteristics of the study population, patients in each of the 3 risk categories identified by the ML application, and the 180 patients sampled for the validation study. We assessed the validity of the application by comparing them to the structured clinicians' review of medical records. The agreements between the 2 methods were evaluated with the following parameters:

1. Overall percent agreements were calculated, including percent agreements for the 3 risk categories. Disagreements were reported for the overall validated sample and the 3 opioid risk categories.
2. Weighted kappa and 95% CIs were reported because of the ordered nature of the risk categories to measure the agreement between the 2 methods.
3. Naïve sensitivity and naïve specificity were calculated along with positive and negative predictive values for the ML application using the structured clinicians' review of medical records as a gold standard and combining the 2 opioid risk categories, namely high risk and suspected OUD or OUD.
4. Corrected sensitivity and corrected specificity were calculated to account for verification bias, that is, overestimation of sensitivity and underestimation of specificity [19,25,26]. Verification bias occurs when disease status (eg, the presence or absence of OUD) is not ascertained in all participants by the gold-standard method (review of medical records) and proportionately more high risk and suspected OUD or OUD patients identified by the test methodology (eg, the ML algorithm) were selected for verification. This verification-biased sampling increases sensitivity and decreases specificity, and these parameters are mathematically corrected to adjust for the biased sampling method.
5. Descriptive statistics were calculated for evaluating risk assignments to determine the most frequently occurring themes for disagreement between the 2 methods.

Results

Patient Risk Categories and Demographics

Of the 649,504 eligible patients with at least 1 prescription in the source data, 78,587 (12.1%) were classified by the ML application into the 3 risk categories after excluding patients with no opioid prescription, patients without sufficient data to evaluate opioid risk, or patients with a diagnosis of cancer (Figure 3). Patients were excluded due to insufficient data if they did not have 1 day before and 1 year of data after their first opioid prescription, or if they were identified as having OUD (based on a diagnosis or rehabilitation drug) and did not have a first opioid prescription before identification of OUD. Patients with opioids prescribed within 2 years of a cancer diagnosis based on *ICD-9* (*International Classification of Diseases, Ninth Revision*) codes were excluded. Accordingly, 50,405 (64.1%) patients were classified by the ML application as being in the not high risk category, 16,636 (21.2%) as being in the high risk category, and 11,546 (14.7%) as being in the suspected OUD or OUD category. We excluded patients who do not have 1 day before and 1 year of data after the first opioid Rx or, if identified as having OUD (based on diagnosis or rehabilitation drug) and do not have a first opioid Rx before identification.

Table 1 details the distribution of eligible patients by demographic characteristics across the different ML application risk assignment categories and sampled patients. Female sex and age 30-64 years were overrepresented in the groups with opioid prescriptions and validation samples for medical records review compared to the eligible patient pool. The sample randomly selected for validation with the structured review of medical records was representative of the patients on opioid treatment with regard to age, sex, and race.

Figure 3. Patient flow diagram with the final verification sample. Patients were excluded from the overall population if they did not have any opioid prescriptions since 2010, were diagnosed with cancer, or had insufficient data to predict opioid risk. The remaining patients were evaluated for opioid risk and stratified by risk classification category. A total of 60 patients were randomly sampled from each risk classification category to be used for the review of medical records and clinician evaluation.

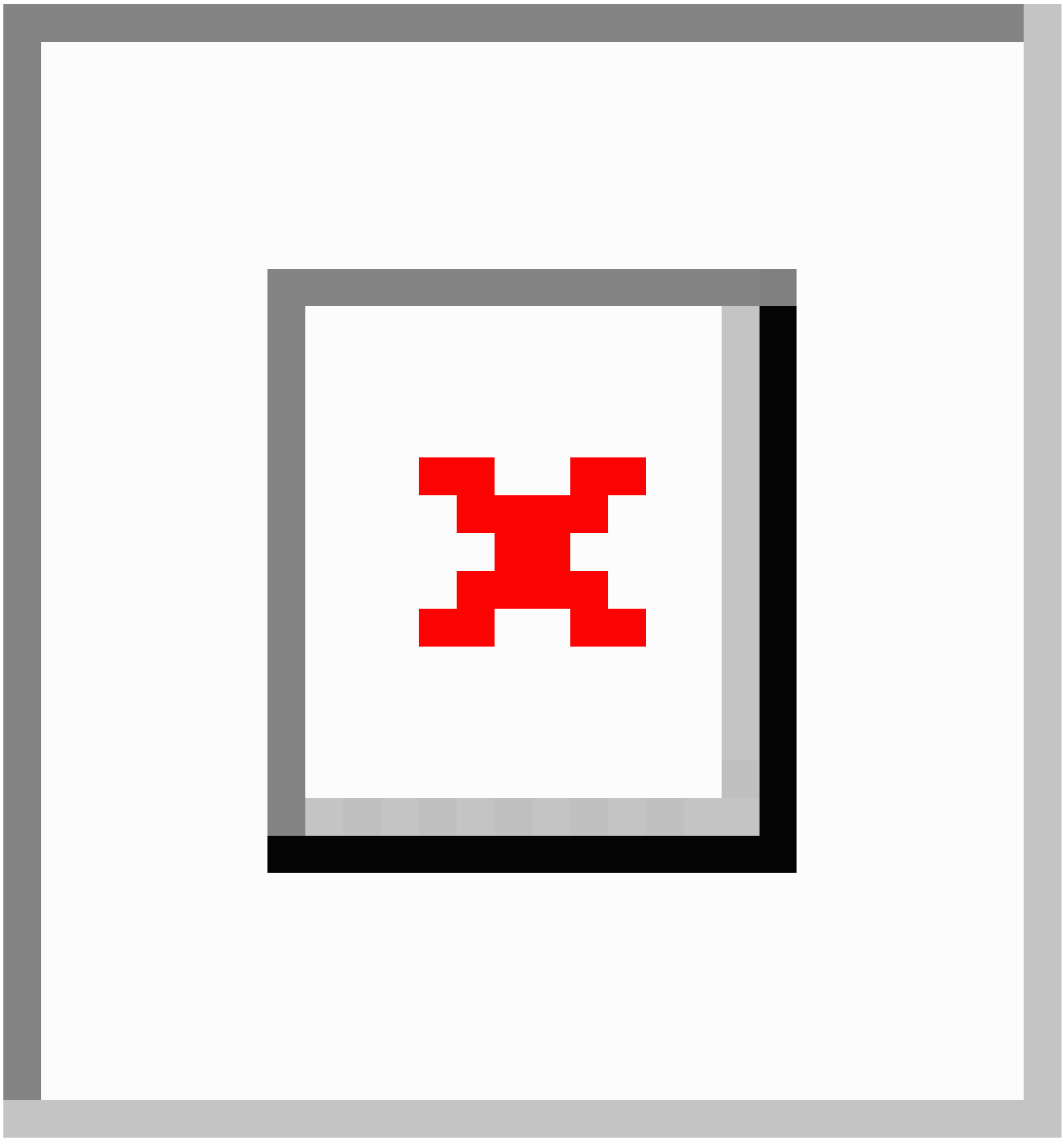


Table . Demographic characteristics of the study populations, including the overall patient population, patients who met the criteria for opioid risk evaluation, patients stratified by the machine learning (ML) application's risk categories "not high risk," "high risk," or "suspected OUD or OUD," and for the validation sample for total review of medical records.

Patient characteristics		Patients with at least 1 prescription (n=649,504), n (%)	Patients meeting criteria for opioid risk evaluation (n=78,587), n (%)	Not high risk (n=50,405), n (%)	High risk (n=16,636), n (%)	Suspected OUD or OUD (n=11,546), n (%)	Sampled patients (n=180), n (%)
Sex							
	Female	385,959 (59.4)	50,064 (63.7)	31,184 (61.9)	11,860 (71.3)	7020 (60.8)	116 (64.4)
	Male	263,535 (40.6)	28,521 (36.3)	19,221 (38.1)	4775 (28.7)	4525 (39.2)	64 (35.6)
	Unknown	10 (0.0)	2 (0.0)	0 (0.0)	1 (0.0)	1 (0.0)	0 (0.0)
Age (years)							
	0 - 17	76,024 (11.7)	737 (0.9)	635 (1.3)	77 (0.5)	25 (0.2)	4 (2.2)
	18 - 29	83,216 (12.8)	8467 (10.8)	5953 (11.8)	1645 (9.9)	869 (7.5)	22 (12.2)
	30 - 49	180,603 (27.8)	27,666 (35.2)	17,777 (35.3)	5834 (35.1)	4055 (35.1)	64 (35.6)
	50 - 64	164,188 (25.3)	24,329 (31.0)	14,442 (28.7)	5564 (33.4)	4323 (37.4)	51 (28.3)
	≥65	145,473 (22.4)	17,388 (22.1)	11,598 (23.0)	3516 (21.1)	2274 (19.7)	39 (21.7)
Race [a]							
	American Indian or Native Alaskan	719 (0.1)	89 (0.1)	60 (0.1)	16 (0.1)	13 (0.1)	0 (0)
	Asian	28,328 (4.4)	2211 (2.8)	1763 (3.5)	297 (1.8)	151 (1.3)	3 (1.7)
	Black or African American	41,794 (6.4)	6189 (7.9)	4119 (8.2)	1245 (7.5)	825 (7.1)	11 (6.1)
	Native Hawaiian or Pacific Islander	286 (0.0)	26 (0.0)	20 (0.0)	5 (0.0)	1 (0.0)	0 (0)
	White	475,939 (73.3)	58,617 (74.6)	36,972 (73.3)	12,373 (74.4)	9272 (80.3)	135 (75.0)
	Other (unknown, declined, or bi- or multiracial)	102,438 (15.8)	11,455 (14.6)	7471 (14.8)	2700 (16.2)	1284 (11.1)	31 (17.2)
Ethnicity [a]							
	Hispanic or Latino	43,119 (6.6)	1874 (2.4)	1279 (2.5)	420 (2.5)	175 (1.5)	17 (9.4)
	Other[b]	606,385 (93.4)	76,713 (97.6)	49,126 (97.5)	16,216 (97.5)	11,371 (98.5)	163 (90.6)

[a]Race and ethnicity data are based on coded fields in the electronic health record.

[b]Other refers to non-Hispanic or non-Latino, declined to respond, and unknown.

Percent Agreement and Kappa Statistics

Prior to conducting final consensus assessments, the independent clinician reviewers' assessment of the levels of risk matched exactly for 70% (126/180) of the patients. When comparing assessments of the not high risk group and those of the high risk and suspected OUD or OUD groups, the clinician reviewer assessments matched 88% of the time.

The overall percent agreement between the ML application and clinician reviewers in stratifying patients into 3 risk categories was 70% (126/180 patients; Table 2). Of the 30% disagreements, 22.8% (n=41) and 7.2% (n=13) indicated underestimation and overestimation of risk by the ML application, respectively, compared to the clinicians' structured review of medical records. Among different risk categories, percent agreement was the highest (90%) for the suspected OUD or OUD category than for the not high risk and high risk categories (60% each). Of the patients classified to the suspected OUD or OUD category by the ML application, 8.3% and 1.7% of them were classified to the high risk and not high risk categories, respectively, by the clinicians' review of medical records. Of the patients classified to the not high risk category by the ML application, clinician reviews classified 40% of patients to the 2 higher risk categories: 30% of patients to the high risk category and 10% of patients to the suspected OUD or OUD category.

Table . Distribution in opioid risk assignment between the machine learning (ML) application and clinicians' structured review of medical records of 180 randomly sampled patients (percent agreement 70%, 95% CI 63.3%-76.7%; weighted kappa coefficient 0.62, 95% CI 0.52-0.71).

ML system risk assignment	Clinician reviewer risk assignment, n			
	Not high risk	High risk	Suspected OUD/OUD	Total
Not high risk	36	18	6	60
High risk	7	36	17	60
Suspected OUD[a] or OUD	1	5	54	60

[a]OUD: opioid use disorder.

The interrater reliability, as expressed using the weighted kappa coefficient for the 2 methods, was 0.62 (95% CI 0.53-0.71), indicating good or substantial agreement [27].

Corrected Sensitivity, Corrected Specificity, and Positive and Negative Predictive Values

Table 3 presents a revised version of Table 2, where the 2 higher-level opioid risk categories (high risk and suspected OUD or OUD) were combined to investigate the potential utility of the ML application in generating signals or alerts to prescribing clinicians, that is, how complete and accurate the ML application is in identifying patients who are at the risk of developing or who may already have OUD. The naïve sensitivity of the ML application was 82.4% (95% CI 75.9%-88.9%), and its naïve specificity was 81.8% (95% CI 70.2%-93.4%). After accounting for verification-biased sampling, the corrected sensitivity of the ML application was 56.6% (95% CI 48.7%-64.5%) and its corrected specificity was 94.2% (95% CI 90.3%-98.1%). The positive and negative predictive values of the ML application were 93.3% (95% CI 88.2%-96.3%) and 60.0% (95% CI 50.4%-68.9%), respectively.

Table . Distribution in opioid use disorder (OUD) risk assignment between the machine learning (ML) application and clinicians' structured review of medical records when the 2 higher-risk categories were combined to investigate the utility of an OUD risk alert at the time of prescribing.

ML system risk assignment	Clinician reviewer risk assignment, n		
	High risk and suspected OUD or OUD	Not high risk	Total
High risk and suspected OUD or OUD	112	8	120
Not high risk	24	36	60
Total	136	44	180

Key Reasons for Disagreements in OUD Risk Categories Between the ML Application and Clinician Reviewers

Table 4 contains the 6 themes that emerged as reasons for disagreements between the ML application and the clinicians' structured review of medical records after conducting a qualitative analysis. Disagreement between the 2 methods was noted for 54 patients, among whom the ML application underestimated the OUD risk in 41 patients and overestimated it in 13 patients. Two or more themes were identified as reasons for most of the disagreements (74.9%). Of the 6 themes, the theme "differences in risk assessment of medication information," accounted for most of the disagreements (72%), followed by the theme "information in clinical notes not available to the ML application" (55.6%).

Table . Key reasons for disagreements in opioid use disorder (OUD) risk assignments between the machine learning (ML) application and clinician reviewers. The reasons for discrepancies were categorized into 6 major themes. More than 1 reason might be identified for a given patient. Results are displayed by whether the assigned risk category was underestimated or overestimated by the ML application in comparison with the clinician reviewers.

Themes of reasons for disagreements in OUD risk assignment	Description of the themes	Patients with at least 1 reason coded in a given theme category, n (%)		
		Cases underestimated by MedAware[a] (n=41)	Cases overestimated by MedAware[b] (n=13)	Total discrepant cases (n=54)
I. Differences in risk assessment of medication information	Medication information available to both the clinician reviewers and the MedAware system contributed to differing risk assessments (eg, medication duration, dose, indication, and gaps in medication timelines).	30 (73.2)	9 (69.2)	39 (72.2)
II. Information in clinical notes not available to MedAware system	Information in patients' clinical notes was available to the clinician reviewers but not to the MedAware system (eg, psychosocial information, experience with opioids and other medications, patient participation in pain management and substance abuse services, and medication information not on the medication list).	27 (65.9)	3 (23.1)	30 (55.6)
III. Differences in risk assessment of psychosocial issues	Psychosocial or psychiatric information available to both the clinician reviewers and the MedAware system contributed to differing risk assessments (eg, patient history of substance abuse, family members with a history of psychosocial or psychiatric issues, and the presence of patients' individual psychiatric conditions contributed to differing risk assessments).	17 (41.5)	2 (15.4)	19 (35.2)
IV. Differences in risk assessment of nonopioid medications	Information on nonopioid medications available to both research reviewers and the MedAware system, which reflects an increased complexity of the patient's medical situation (eg, pain level) or a higher risk when combined with opioids, contributed to differences in risk assessments (eg, zolpidem and gabapentinoids).	10 (24.4)	2 (15.4)	12 (22.2)
V. Bugs identified in the MedAware system	Bugs in the MedAware system included inaccurate mapping of data elements (eg, dosage units and incorrect medication), missing medication in drug class, and incorrectly constructed alert messages.	5 (12.2)	5 (38.5)	10 (18.5)

Themes of reasons for disagreements in OUD risk assignment	Description of the themes	Patients with at least 1 reason coded in a given theme category, n (%)		
		Cases underestimated by MedAware[a] (n=41)	Cases overestimated by MedAware[b] (n=13)	Total discrepant cases (n=54)
VI. Presence of other clinical information not considered by the MedAware system or the clinician reviewers	Clinical information that may indicate the risk of OUD not considered by the clinician reviewers or the MedAware system, but not both, such as hepatitis C diagnosis, urine toxicity tests, and MedAware system access to *ICD-9*[c] diagnostic information that clinician reviewers did not see.	6 (14.6)	0 (0.0)	6 (11.1)

[a]The ML application's risk assignment was lower in severity compared to the clinician reviewers' risk assignment.

[b]The ML application's risk assignment was higher in severity compared to the clinician reviewers' risk assignment.

[c]*ICD-9*: *International Classification of Diseases, Ninth Revision*.

Discussion

Principal Results

ML algorithms can leverage large-scale EHR and medical claims data and potentially identify patients at risk of OUD [28-32]. However, very few studies have assessed the clinical validity and potential utility of ML algorithms designed to differentiate among levels of patients' OUD risk. In this study, we examined the agreement between an ML application and clinicians' structured review of medical records in classifying patients on opioid drug treatment into 3 distinct categories of OUD risk (ie, not high risk, high risk, or suspected OUD or OUD). We also assessed the application's utility in identifying clinically valid alerts and identified and quantified reasons that could lead to disagreements between clinicians' judgment and outputs of ML applications. The ML application was validated in an outpatient database, and it appeared to have value.

There was substantial agreement between the application and the clinician reviewers' structured review of medical records. The agreement between the 2 methods was the highest for the suspected OUD or OUD category. The ML application correctly identified this most vulnerable group of patients to increase clinician awareness and responsiveness to improve patient management, including modifications to their medication regimen or referral to a specialized treatment service to mitigate the complications of opioid use. Moreover, if the ML application is used to generate alerts on patients at high risk of OUD or those who already have OUD, it will identify approximately 60% of these patients with a 93.3% precision (positive predictive value). Thus, the results of this study show that this ML application was able to generate clinically valid and useful alerts to screen for patients at risk of OUD. It is important to recognize that alerting clinicians regarding patients at risk of OUD should be coupled with clinician education on appropriate treatment guidelines and practices to avoid undertreatment of pain and patient stigma [33,34].

Comparison With Prior Work

Previous studies have shown that artificial intelligence tools using ML algorithms can improve treatment, enhance quality of care and patient safety, reduce burden on providers, and generally increase the efficiency with which resources are used, resulting in potential cost savings or health gains [7,32,35-38]. In addition, our findings align with those of previous studies that highlight the potential of ML applications to predict individual patients' risk of specific medical conditions and associated complications to offer specialized care programs to high-risk patients [39,40]. Our study also confirms and extends the findings of a few studies that examined other ML applications and highlighted the potential to identify patients at risk for substance misuse and abuse, including OUD and opioid overdose [31,38,41]. Nevertheless, these comparable ML applications were plagued with very low positive predictive values due, in part, to low OUD prevalence as a result of suboptimal definitions of OUD by relying solely on *ICD* (*International Classification of Diseases*) codes [42]. A few previous studies identified additional limitations and challenges related to comparable ML applications. For example, Afshar et al [43] described the use of an algorithm to identify patients at risk for any substance misuse at the time of admission, based on clinical notes from the first 24 hours after hospital admission. In this study, we found that the positive predictive value of this tool was 61%-72%, which was lower than that of the ML application. The tool that Afshar et al [43] studied does not identify patients outside of the hospital setting and depends on physicians' notes. As a result, this tool is not suited for more general screening using structured clinical EHR data and medical claims data. Another recent study by Lo-Ciganic et al [41] described an algorithm to predict the occurrence of overdose episodes, but does not identify patients who are most at risk of OUD in the future.

We believe that the substantial agreement, high specificity, and high positive predictive value of the ML application was achieved because we pilot-tested the ML models in comparison with clinician assessments and then used an iterative process

with continuous calibration of model parameters to optimize the accurate identification of OUD risk categories. In addition, we used a composite definition of OUD not restricted to *ICD* codes resulting in a higher prevalence of OUD identified in the patient population. The ML application classified 1 in 7 and about one-fifth of the eligible population with prescribed opioids in the suspected OUD or OUD and high risk categories, respectively, compared to other studies that reported a prevalence of OUD in the range of 1%-5% [44,45]. Furthermore, the full accessibility of the EHR at the time of case evaluation, coupled with standardized data extraction and a medication timeline visualization tool, allowed seamless analysis of cases contributing to the high accuracy rates.

Our study also identified the main reasons for disagreements between the clinician reviewers and the ML application's risk assignments. These reasons included information available in the clinical notes not being accessible to the ML application (eg, psychosocial issues and patients' participation in substance abuse services), and different interpretation of available information such as differences in the impact of antidepressant treatments. Clinicians considered stable and sufficiently treated depression as not being a risk factor for OUD [46]. In analyzing the reasons for discrepancies, we observed factors related to model training processes, data quality, and outcome definitions. The knowledge gained through our analytic process could be useful to further optimize their ML algorithm development pipeline. As of today, it is critical to standardize the ML development process and make it more understandable to clinical end users. However, to our knowledge, few efforts have been made to systematically analyze each component of the model development process from the clinician's point of view and further evaluate its impact on the model's clinical implementation. We believe that our work can facilitate a better bridging of the gap between ML model builders and clinicians.

Limitations

Our study has some limitations. We used retrospective data to evaluate an algorithm primarily designed to be used in real time. Although many of the findings from our retrospective analysis should be applicable to real-time alerting, it is difficult to predict whether some alerts would perform differently or how clinicians would respond to real-time alerts. Second, although our clinician reviewers were carefully trained and a coding manual was developed with clear operational definitions, each risk assessment required a degree of judgment on the part of the reviewers; human factors could impact the final risk assignment. Finally, our study was limited to outpatients at 2 large academic medical centers in the United States, which limits the generalizability of our results. Additional biases may have been introduced into the ML application in ways that the research team were not able to assess [7,47]. Although the total population of patients receiving outpatient care within an academic medical center was included, there may have been biases in patients who were able to access care, those receiving opioid prescriptions, and in the clinical documentation of concerns regarding opioid use and substance abuse. Validation across different sites and populations (eg, veterans' facilities) may reveal site-specific differences and may require unique models or warrant the identification and capture of new descriptive features.

Conclusions

We tested an ML application that assessed OUD risk in an extensive outpatient EHR database and found that it appeared to classify patients into differing levels of OUD risk, and that there was substantial agreement with clinicians' review of medical records. We identified key themes for disagreements between the commercial application and clinician review, which can be used to further enhance ML applications. ML algorithms applied to available EHR clinical data hold promise for identifying patients at differing levels of OUD risk and supporting better clinical decision-making regarding treatment. Such tools will likely complement traditional, rule-based approaches to provide alerts about potential opioid prescribing safety issues.

Acknowledgments

This work was supported in part by MedAware, Ltd. DWB reports grants and personal fees from EarlySense, personal fees from Center for Digital Innovation Negev, equity from ValeraHealth, equity from Clew, equity from MDClone, personal fees and equity from AI-Enhanced Safety of Prescription, and grants from Merative, outside the submitted work. RR reports having equity from Hospitech Respiration, equity from Tri.O Medical Device, equity from AEYE Health, equity from RxE2, equity from OtheReality; equity from Co-Patient Support, and equity from Medyx.ai, all of which are unrelated to this work. He is also receiving research funding from Telem, Calosense Health, Breath of Health, and BriefCam.

MedAware's contributions to this study were limited to running the patient data through their ML application, providing risk assessment results from their system, and additional information on selected patients to clarify their risk assessment results. They were not involved in any data analysis or interpretation. They had no influence on the results, and they were not part of the manuscript writing process.

Data Availability

The data sets generated in this study are not publicly available due to hospital institutional review board (IRB) regulations and patient privacy policies, but deidentified data sets are available from the corresponding author upon reasonable request. These deidentified data sets would include the model training set to facilitate independent model replications, patient demographics, and risk level assessments generated by the MedAware system and clinician review for the study cohort of 180 patients. Detailed

electronic medical record data extracted on this cohort of patients in support of clinician risk assessments will not be available due to IRB and institutional policy restricting the use of clinical notes and sharing of patient-sensitive data. The MedAware system algorithm will not be available for sharing as this is a proprietary commercial product.

Conflicts of Interest

None declared.

References

1. 2020 National Survey on Drug Use and Health (NSDUH): methodological summary and definitions. Substance Abuse and Mental Health Services Administration. 2021. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt35330/2020NSDUHMethodSummDefs091721.pdf [accessed 2023-07-25]
2. Kirson NY, Shei A, Rice JB, et al. The burden of undiagnosed opioid abuse among commercially insured individuals. Pain Med 2015 Jul;16(7):1325-1332. [doi: 10.1111/pme.12768] [Medline: 25929289]
3. Drug overdose death rates. National Institute on Drug Abuse. 2023. URL: https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates [accessed 2023-07-25]
4. Kuehn BM. Massive costs of the US opioid epidemic in lives and dollars. JAMA 2021 May 25;325(20):2040. [doi: 10.1001/jama.2021.7464]
5. Satterwhite S, Knight KR, Miaskowski C, et al. Sources and impact of time pressure on opioid management in the safety-net. J Am Board Fam Med 2019;32(3):375-382. [doi: 10.3122/jabfm.2019.03.180306] [Medline: 31068401]
6. Harle CA, Bauer SE, Hoang HQ, Cook RL, Hurley RW, Fillingim RB. Decision support for chronic pain care: how do primary care physicians decide when to prescribe opioids? A qualitative study. BMC Fam Pract 2015 Apr 14;16:48. [doi: 10.1186/s12875-015-0264-3] [Medline: 25884340]
7. Artificial intelligence in health care: benefits and challenges of technologies to augment patient care. United States Government Accountability Office. 2022. URL: https://www.gao.gov/assets/720/711471.pdf [accessed 2023-07-25]
8. McCoy AB, Thomas EJ, Krousel-Wood M, Sittig DF. Clinical decision support alert appropriateness: a review and proposal for improvement. Ochsner J 2014;14(2):195-202. [Medline: 24940129]
9. Petersen C, Smith J, Freimuth RR, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. J Am Med Inform Assoc 2021 Mar 18;28(4):677-684. [doi: 10.1093/jamia/ocaa319] [Medline: 33447854]
10. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750. [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
11. Chen D, Liu S, Kingsbury P, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med 2019;2:43. [doi: 10.1038/s41746-019-0122-0] [Medline: 31304389]
12. Ghassemi M, Mohamed S. Machine learning and health need better values. NPJ Digit Med 2022 Apr 22;5(1):51. [doi: 10.1038/s41746-022-00595-9] [Medline: 35459793]
13. Technology: your safety layer within. MedAware. 2023. URL: https://www.medaware.com/technology/ [accessed 2022-11-04]
14. Syrowatka A, Song W, Amato MG, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: a scoping review. Lancet Digit Health 2022 Feb;4(2):e137-e148. [doi: 10.1016/S2589-7500(21)00229-6] [Medline: 34836823]
15. Hanko M, Grendár M, Snopko P, et al. Random forest-based prediction of outcome and mortality in patients with traumatic brain injury undergoing primary decompressive craniectomy. World Neurosurg 2021 Apr;148:e450-e458. [doi: 10.1016/j.wneu.2021.01.002] [Medline: 33444843]
16. Song W, Kang MJ, Zhang L, et al. Predicting pressure injury using nursing assessment phenotypes and machine learning methods. J Am Med Inform Assoc 2021 Mar 18;28(4):759-765. [doi: 10.1093/jamia/ocaa336] [Medline: 33517452]
17. Rozenblum R, Rodriguez-Monguio R, Volk LA, et al. Using a machine learning system to identify and prevent medication prescribing errors: a clinical and cost analysis evaluation. Jt Comm J Qual Patient Saf 2020 Jan;46(1):3-10. [doi: 10.1016/j.jcjq.2019.09.008] [Medline: 31786147]
18. Schiff GD, Volk LA, Volodarskaya M, et al. Screening for medication errors using an outlier detection system. J Am Med Inform Assoc 2017 Mar 1;24(2):281-287. [doi: 10.1093/jamia/ocw171] [Medline: 28104826]
19. Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. Am J Epidemiol 1994 Oct 15;140(8):759-769. [doi: 10.1093/oxfordjournals.aje.a117323] [Medline: 7942777]
20. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain—United States, 2016. MMWR Recomm Rep 2016 Mar 18;65(1):1-49. [doi: 10.15585/mmwr.rr6501e1] [Medline: 26987082]
21. American Psychiatric Association. Opioid use disorder. In: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition: American Psychiatric Association; 2013:541-546 URL: https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596 [accessed 2024-05-08] [doi: 10.1176/appi.books.9780890425596]
22. Webster LR. Risk factors for opioid-use disorder and overdose. Anesth Analg 2017 Nov;125(5):1741-1748. [doi: 10.1213/ANE.0000000000002496] [Medline: 29049118]

23. Burcher KM, Suprun A, Smith A. Risk factors for opioid use disorders in adult postsurgical patients. Cureus 2018 May 11;10(5):e2611. [doi: 10.7759/cureus.2611] [Medline: 30018867]
24. Zhao S, Chen F, Feng A, Han W, Zhang Y. Risk factors and prevention strategies for postoperative opioid abuse. Pain Res Manag 2019;2019:7490801. [doi: 10.1155/2019/7490801] [Medline: 31360271]
25. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983 Mar;39(1):207-215. [Medline: 6871349]
26. Punglia RS, D'Amico AV, Catalona WJ, Roehl KA, Kuntz KM. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. N Engl J Med 2003 Jul 24;349(4):335-342. [doi: 10.1056/NEJMoa021659] [Medline: 12878740]
27. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363. [Medline: 15883903]
28. Burgess-Hull AJ, Brooks C, Epstein DH, Gandhi D, Oviedo E. Using machine learning to predict treatment adherence in patients on medication for opioid use disorder. J Addict Med 2023;17(1):28-34. [doi: 10.1097/ADM.0000000000001019] [Medline: 35914118]
29. Lo-Ciganic WH, Donohue JM, Yang Q, et al. Developing and validating a machine-learning algorithm to predict opioid overdose among medicaid beneficiaries in two US States: a prognostic modeling study. Lancet Digit Health 2022 Jun;4(6):e455-e465. [doi: 10.1016/S2589-7500(22)00062-0] [Medline: 35623798]
30. Afshar M, Sharma B, Dligach D, et al. Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study. Lancet Digit Health 2022 Jun;4(6):e426-e435. [doi: 10.1016/S2589-7500(22)00041-3] [Medline: 35623797]
31. Heo KN, Lee JY, Ah YM. Development and validation of a risk-score model for opioid overdose using a national claims database. Sci Rep 2022 Mar 23;12(1):4974. [doi: 10.1038/s41598-022-09095-y] [Medline: 35322156]
32. Dong X, Deng J, Rashidian S, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. J Am Med Inform Assoc 2021 Jul 30;28(8):1683-1693. [doi: 10.1093/jamia/ocab043] [Medline: 33930132]
33. Keister LA, Stecher C, Aronson B, McConnell W, Hustedt J, Moody JW. Provider bias in prescribing opioid analgesics: a study of electronic medical records at a hospital emergency department. BMC Public Health 2021 Aug 6;21(1):1518. [doi: 10.1186/s12889-021-11551-9] [Medline: 34362330]
34. Pergolizzi JV, Lequang JA, Passik S, Coluzzi F. Using opioid therapy for pain in clinically challenging situations: questions for clinicians. Minerva Anestesiol 2019 Aug;85(8):899-908. [doi: 10.23736/S0375-9393.19.13321-4] [Medline: 30871302]
35. Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. Cell 2020 Apr 2;181(1):92-101. [doi: 10.1016/j.cell.2020.03.022] [Medline: 32243801]
36. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health 2018 Dec;8(2):020303. [doi: 10.7189/jogh.08.020303] [Medline: 30405904]
37. Li Q, Wright J, Hales R, Voong R, McNutt T. A digital physician peer to automatically detect erroneous prescriptions in radiotherapy. NPJ Digit Med 2022 Oct 21;5(1):158. [doi: 10.1038/s41746-022-00703-9] [Medline: 36271138]
38. Canan C, Polinski JM, Alexander GC, Kowal MK, Brennan TA, Shrank WH. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. J Am Med Inform Assoc 2017 Nov 1;24(6):1204-1210. [doi: 10.1093/jamia/ocx066] [Medline: 29016967]
39. Chen JH, Asch SM. Machine learning and prediction in medicine: beyond the peak of inflated expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509. [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
40. Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med 2021 Mar 30;4(1):62. [doi: 10.1038/s41746-021-00426-3] [Medline: 33785839]
41. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. JAMA Netw Open 2019 Mar 1;2(3):e190968. [doi: 10.1001/jamanetworkopen.2019.0968] [Medline: 30901048]
42. Lagisetty P, Garpestad C, Larkin A, et al. Identifying individuals with opioid use disorder: validity of International Classification of Diseases diagnostic codes for opioid use, dependence and abuse. Drug Alcohol Depend 2021 Apr 1;221:108583. [doi: 10.1016/j.drugalcdep.2021.108583] [Medline: 33662670]
43. Afshar M, Sharma B, Bhalla S, et al. External validation of an opioid misuse machine learning classifier in hospitalized adult patients. Addict Sci Clin Pract 2021 Mar 17;16(1):19. [doi: 10.1186/s13722-021-00229-7] [Medline: 33731210]
44. Barocas JA, White LF, Wang J, et al. Estimated prevalence of opioid use disorder in Massachusetts, 2011-2015: a capture-recapture analysis. Am J Public Health 2018 Dec;108(12):1675-1681. [doi: 10.2105/AJPH.2018.304673] [Medline: 30359112]
45. Han B, Compton WM, Blanco C, Crane E, Lee J, Jones CM. Prescription opioid use, misuse, and use disorders in U.S. adults: 2015 National Survey on Drug Use and Health. Ann Intern Med 2017 Sep 5;167(5):293-301. [doi: 10.7326/M17-0865] [Medline: 28761945]
46. Brooner RK, King VL, Kidorf M, Schmidt CW, Bigelow GE. Psychiatric and substance use comorbidity among treatment-seeking opioid abusers. Arch Gen Psychiatry 1997 Jan;54(1):71-80. [doi: 10.1001/archpsyc.1997.01830130077015] [Medline: 9006403]

47. Boyd AD, Gonzalez-Guarda R, Lawrence K, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. J Am Med Inform Assoc 2023 Aug 18;30(9):1561-1566. [doi: 10.1093/jamia/ocad115] [Medline: 37364017]

Abbreviations

CDSS: clinical decision support system
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
EHR: electronic health record
***ICD*:** *International Classification of Diseases*
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
ML: machine learning
OUD: opioid use disorder

Edited by A Benis; submitted 20.10.23; peer-reviewed by L Zhang, X Dong; revised version received 15.03.24; accepted 20.04.24; published 04.06.24.

Please cite as:
Eguale T, Bastardot F, Song W, Motta-Calderon D, Elsobky Y, Rui A, Marceau M, Davis C, Ganesan S, Alsubai A, Matthews M, Volk LA, Bates DW, Rozenblum R
A Machine Learning Application to Classify Patients at Differing Levels of Risk of Opioid Use Disorder: Clinician-Based Validation Study
JMIR Med Inform 2024;12:e53625
URL: https://medinform.jmir.org/2024/1/e53625
doi:10.2196/53625

Impact of a Nationwide Medication History Sharing Program on the Care Process and End-User Experience in a Tertiary Teaching Hospital: Cohort Study and Cross-Sectional Study

Jungwon Cho[1,2], PhD; Sooyoung Yoo[3], PhD; Eunkyung Euni Lee[1,2,*], PharmD, PhD; Ho-Young Lee[3,4,*], MD, PhD

[1]

[2]

[3]

[4]

[*]these authors contributed equally

Corresponding Author:
Ho-Young Lee, MD, PhD

Abstract

Background: Timely and comprehensive collection of a patient's medication history in the emergency department (ED) is crucial for optimizing health care delivery. The implementation of a medication history sharing program, titled "Patient's In-home Medications at a Glance," in a tertiary teaching hospital aimed to efficiently collect and display nationwide medication histories for patients' initial hospital visits.

Objective: As an evaluation was necessary to provide a balanced picture of the program, we aimed to evaluate both care process outcomes and humanistic outcomes encompassing end-user experience of physicians and pharmacists.

Methods: We conducted a cohort study and a cross-sectional study to evaluate both outcomes. To evaluate the care process, we measured the time from the first ED assessment to urgent percutaneous coronary intervention (PCI) initiation from electronic health records. To assess end-user experience, we developed a 22-item questionnaire using a 5-point Likert scale, including 5 domains: information quality, system quality, service quality, user satisfaction, and intention to reuse. This questionnaire was validated and distributed to physicians and pharmacists. The Mann-Whiteny *U* test was used to analyze the PCI initiation time, and structural equation modeling was used to assess factors affecting end-user experience.

Results: The time from the first ED assessment to urgent PCI initiation at the ED was significantly decreased using the patient medication history program (mean rank 42.14 min vs 28.72 min; Mann-Whitney U=346; P=.03). A total of 112 physicians and pharmacists participated in the survey. Among the 5 domains, "intention to reuse" received the highest score (mean 4.77, SD 0.37), followed by "user satisfaction" (mean 4.56, SD 0.49), while "service quality" received the lowest score (mean 3.87, SD 0.79). "User satisfaction" was significantly associated with "information quality" and "intention to reuse."

Conclusions: Timely and complete retrieval using a medication history-sharing program led to an improved care process by expediting critical decision-making in the ED, thereby contributing to value-based health care delivery in a real-world setting. The experiences of end users, including physicians and pharmacists, indicated satisfaction with the program regarding information quality and their intention to reuse.

(JMIR Med Inform 2024;12:e53079) doi:10.2196/53079

KEYWORDS

health information system; HIS; medication history; history; histories; patients' own medication; satisfaction; DeLone and McLean Model of information systems success; value-based health care; emergency department; information system; information systems; emergency; urgent; drug; drugs; pharmacy; pharmacies; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications; sharing; user experience; survey; surveys; intention; intent; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; acceptance; adoption

Introduction

Health information systems (HISs) play a vital role in the delivery of health care services, as they provide access to the patient's medical records, help track treatment progress, and support health care providers in making care decisions [1-3]. Although the development of HISs has revolutionized the provision of patient care and handling of patients' health information, in the transitional period toward the era of the fourth industrial revolution, studies that evaluate humanistic outcomes as well as clinical or economic outcomes caused by

HIS, are needed [4]. As leaders in health care settings have made various investments, such as time, money, and manpower, in managing HIS [5], the multifaceted evaluation of whether end users can use the HIS skillfully and achieve satisfaction in functionality and usability would be increasingly important in the future [4,6].

Health care organizations can ensure effective HIS use and improve the quality of patient care by conducting evaluations of HISs. These evaluations could allow health care organizations to proactively address issues related to system performance, integration, and data accuracy. However, evaluating the diversity and complexity of HISs in real-world clinical settings is a significant challenge [5,7]. Hospitals use different HISs depending on their work process, and the program related to direct patient care, including documentation and retrieval of medical records, or clinical decision support systems varies [8-10]. In addition, health care environments are constantly evolving with the emergence of innovative technologies [11]. Newly developed information systems or programs tend to be integrated into homegrown HISs after establishing a fully electronic medical record system. Thus, although HIS evaluations reporting economic, clinical, and humanistic outcomes could provide a balanced picture of the comprehensive impact of the health care interventions implemented, comprehensive evaluations of HISs are rarely conducted [12].

Acquisition of patients' complete medication use history could greatly enhance medication management and support physicians in making informed decisions. Accurate and efficient compilation of information can be more important when time-sensitive clinical decisions and subsequent interventions are made [13], especially in the emergency department (ED). However, previous studies have demonstrated that accurate and timely collection of patients' medication histories is challenging especially in the ED for various reasons, including patients with altered mental status due to confusion or intoxication, patients taking multiple outpatient prescriptions, and first-time patients to the hospital [14-16]. Since the treatment plan would change depending on the medication history, the prompt and complete evaluation of the medication history is vital. The process of collecting medication history was also described as a labor-intensive process, often requiring manual retrieval of information from outside the hospital [17,18]. Thus, a medication history sharing program called "Patient's In-home Medications at a Glance" was developed and successfully launched within a homegrown HIS known as BESTCare in Seoul National University Bundang Hospital (SNUBH) on January 11, 2021. The program enabled health professionals to access the patients' nationwide medication history swiftly and accurately from the Healthcare Insurance Review and Assessment Service database in South Korea with added features about the patient instructions and the identification guide for each medication. The rate of identification of patients' medication history within 24 hours was significantly improved at the ED after the implementation of the program [19]. However, comprehensive evaluations of querying patient medication history were necessary to provide a balanced picture of the medication history program, as an HIS intervention could have had an impact not only on the care process but also on humanistic outcomes, such as end-user experience about its functionality and usability, which may evolve over time.

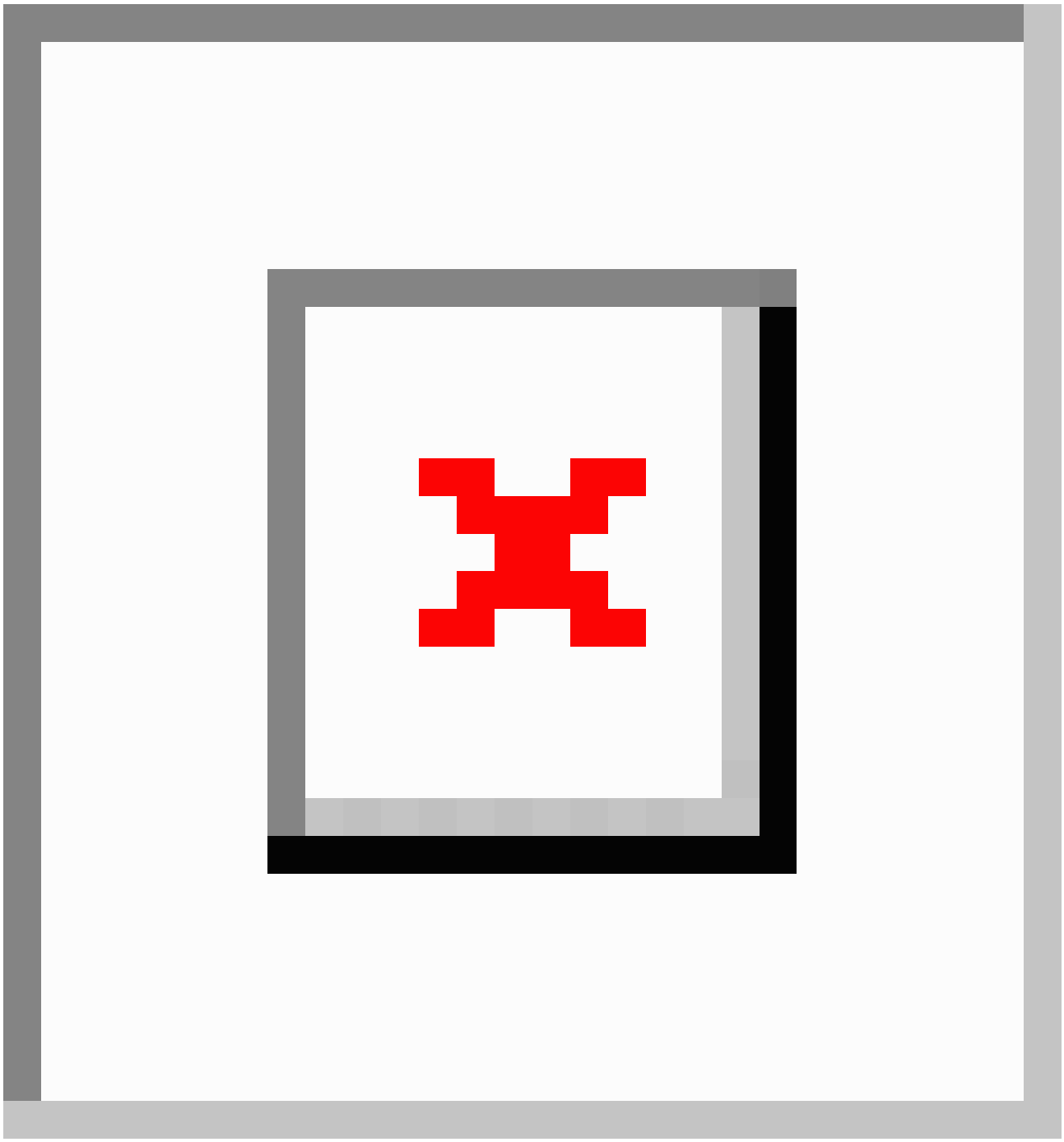
Therefore, this study aimed to evaluate the impact of an HIS intervention on health care delivery, namely medication history retrieval, using the "Patient's In-home Medications at a Glance" program. Specifically, we evaluated the care process outcome, that is, the time from the first ED assessment to urgent percutaneous coronary intervention (PCI) initiation, and the humanistic outcome, that is, the end-user experience among physicians and pharmacists.

Methods

Study Design

We conducted a cohort study and a cross-sectional study to evaluate both outcomes. We evaluated the impact of medication history retrieval using the "Patient's In-home Medications at a Glance" program on two aspects: (1) the care process outcome and (2) the end-user experience among physicians and pharmacists. Figure 1 shows the ED process and medication history check to describe the 2 outcomes of this study.

Figure 1. Emergency department (ED) process and medication history check depicting two outcomes: (1) time from the first ED assessment to urgent percutaneous coronary intervention (PCI) initiation as the care process outcome and (2) the end-user experience among physicians and pharmacists using the program as a humanistic outcome. Delayed medication history checks could increase the time of PCI initiation at the ED, especially in urgent clinical situations. The "Patient's In-home Medications at a Glance" program linking to the nationwide personal medication records provides more rapid and complete collections of medication history compared to manual retrievals that often require interviews with patients or caregivers at the ED (icons are made by Freepik).



First, we analyzed the care process to determine whether physicians' use of the program could expedite the time from the first ED assessment to urgent PCI initiation. Second, to assess end-user experience, we developed a questionnaire consisting of 22 survey items that were validated. We then conducted a website-based survey among physicians and pharmacists who served as end users of the program.

Care Process Outcome

Data Collection

For the care process, patients who were admitted to the ED for the first time from January 1, 2021, to December 31, 2022, were included to estimate the impact of the program on the collection of patients' drug therapy. The outcome was defined as the time of initiating urgent PCI after the first assessment by ED physicians from January 1, 2021, to December 31, 2022. Urgent PCI was defined as PCI performed within an hour of admission

to the ED. As the identification of the patient's medication use history was required to further improve the care plan, the time from the first ED assessment to urgent PCI initiation was analyzed.

Data Analysis

To analyze the impact of the program on the care process, data were extracted from the SNUBH electronic database. We performed a Mann-Whitney *U* test to evaluate the difference in the time from the first ED assessment to urgent PCI initiation between patients who were queried about their medication use history by physicians via the program and those who were not.

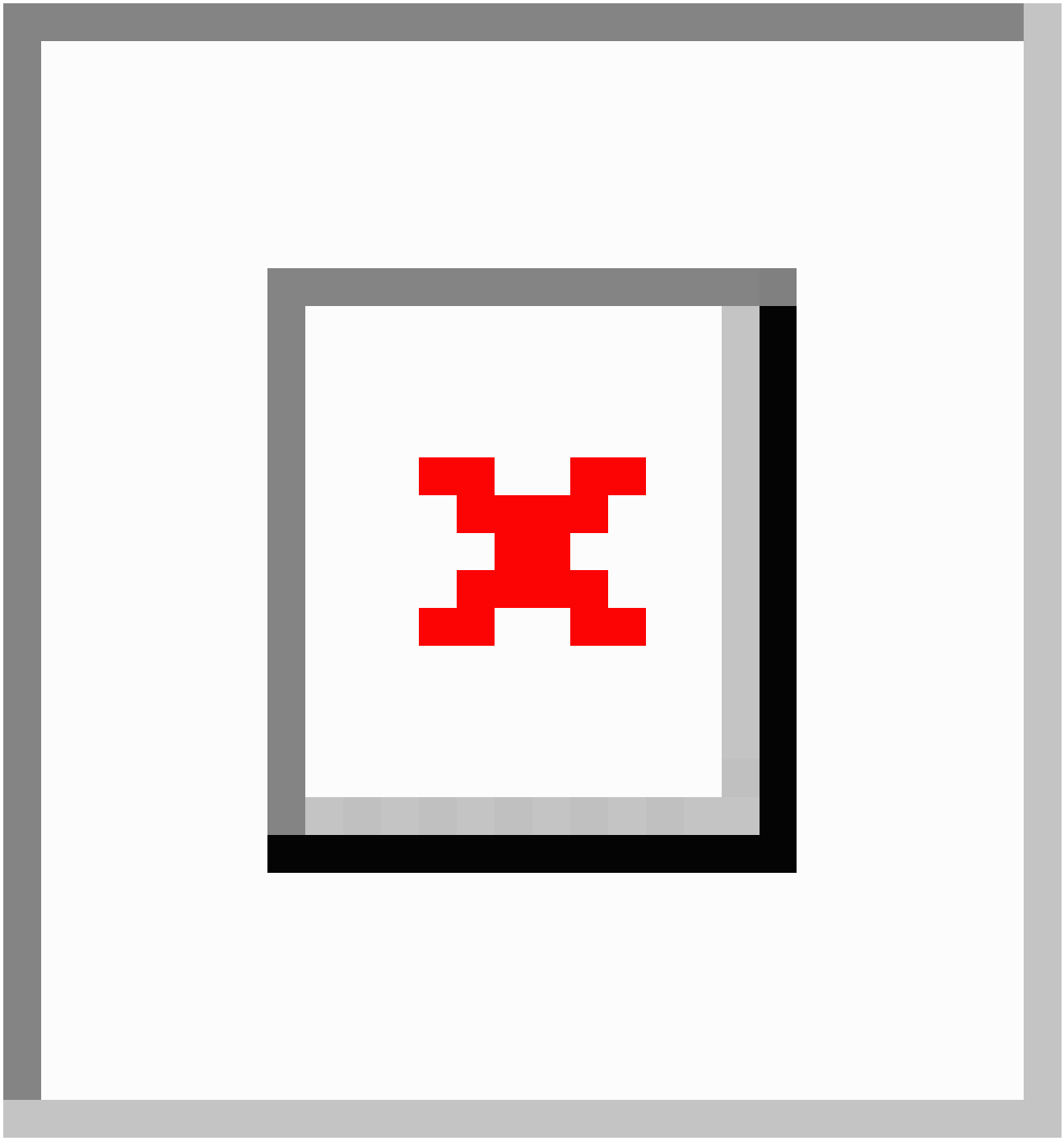
All analyses were performed using IBM SPSS Statistics (version 22.0; IBM Corp) and R (version 4.0.2; R Foundation for Statistical Computing).

Survey and Assessing Factors Affecting End-User Experience on the Program

Survey Development With a Conceptual Framework

To assess end-user experience and whether end users are satisfied with HIS and their intention to reuse it, we adopted the updated DeLone and McLean Model of Information Systems Success (DMISM) [20] for survey development. The updated DMISM provides a conceptual framework to suggest the factors necessary for the provision of use and benefits from the HIS. Based on the updated DMISM, we proposed that the quality of the information system consists of 3 quality domains: information quality, system quality, and service quality. These domains are necessary for user satisfaction and are instrumental in driving users' intentions to reuse the system. In this study, we narrowed the scope to physicians and pharmacists who were already using the program. Therefore, we adjusted the factor of "intention to use" and "use" in the updated DMISM to "intention to reuse." Due to the nature of the HIS, "intention to reuse" of the program by end users is considered the ultimate and crucial goal. By setting it as the final outcome variable, "intention to reuse" is influenced by preceding user satisfaction. Therefore, we established the research model with the relationship that "user satisfaction" affects "intention to reuse." These domains were used to develop the survey (Figure 2).

Figure 2. A research model for the survey development. The updated DeLone and McLean Model of Information Systems Success [20] provides domain variables consisting of 3 quality domains (ie, information, system, and service quality) as well as outcome domains.



We collected 32 survey questionnaires that assessed each quality domain regarding previous studies [3,21-24]. Through face validation with 6 pharmacists, a physician, and a medical informatics professor every 2 weeks for 3 months, the survey questionnaires were classified according to each domain. The questionnaires were eliminated or revised to reflect the contextual significance of the program. The draft survey finally consisted of 22 questionnaires, and a pilot study was conducted with 10 pharmacists and 2 physicians at SNUBH.

The survey was conducted from December 15, 2022, to December 28, 2022, at SNUBH. We used a web-based survey to collect data on the end-user experience efficiently and rapidly. The survey link was distributed to all physicians and pharmacists at the hospital via email. Survey completion was expected to take approximately 5 minutes. The items in the survey were rated on a 5-point Likert scale (1=not at all; 5=very much). Only those who provided consent after receiving an explanation of the background and purpose of the survey were included.

Data Analysis

An exploratory factor analysis of the results was then performed to determine how the items were classified into components. We used the Kaiser-Meyer-Olkin measure to assess sampling adequacy and obtained a specific value of 0.858, surpassing the recommended threshold of 0.5. The suitability of the data for factor analysis was further confirmed through the Bartlett test

of sphericity, yielding a statistically significant result (χ^2_{105}=723.6; P<.001). The analysis of communality, indicating the explanatory power between measurement variables and extracted factors, was performed. Considering the general criterion that variables with communality below 0.4 are deemed low and should be excluded from factor analysis, 8 questions were excluded. Consequently, 14 questionnaires were retained (Table S1 in Multimedia Appendix 1).

Subsequently, we conducted a reliability analysis of the survey items and calculated Cronbach α. We analyzed the convergent and discriminant validity of the constructs. We used SPSS to conduct statistical analyses, including factor and reliability analyses. Finally, structural equation modeling (SEM) was used to evaluate the structural correlations among the domains using the AMOS 25 software (version 25.0; IBM Corp). SEM was chosen to provide a comprehensive understanding of the relationships among survey variables and to help validate the theoretical models with a visual representation.

Ethical Considerations

This study was approved by the Institutional Review Board of SNUBH (B-2203-746-001; April 21, 2022), and the requirement of obtaining written consent was waived, as this study did not contain sensitive personally identifiable information.

Results

Care Process Outcome

Of the 162 patients who were admitted to the ED and visited the hospital for the first time over a 2-year period, 77 who underwent urgent PCIs within an hour from the first ED assessment to urgent PCI initiation were included. Patients who were regularly visiting hospitals with chronic diseases were excluded. Table 1 describes the demographic characteristics of patients, including gender, age, department, tests, and diagnosis, between the patient group (n=59), for which the doctor did not use the program, and the patient group (n=18), whose medications were accessed through the program.

Table . Demographics of patients receiving urgent percutaneous coronary intervention by use of the medication history program during the study period at an emergency department (n=77[a]).

Characteristics		No (without the program; n=59)	Yes (with the program; n=18)
Sex (male), n (%)		50 (84.7)	11 (61.1)
Age (years), mean (SD)		64.3 (12.1)	68.9 (12.4)
Department at discharge, n (%)			
	Cardiology	54 (91.5)	16 (88.9)
	Others	5 (8.5)	2 (11.1)
Had CT[b] scan, n (%)		12 (20.3)	3 (16.7)
Diagnosis, n (%)			
	ST elevation myocardial infarction	53 (89.8)	16 (88.9)
	Others	10 (16.9)	4 (22.2)

[a]Patients receiving percutaneous coronary intervention within an hour at an emergency department from January 12, 2021, to December 31, 2022.

[b]CT: computed tomography.

Changes in time from the first ED assessment to urgent PCI initiation significantly decreased in patients who used the program (n=18; mean rank 28.72 min) versus patients who did not use the program (n=59; mean rank 42.14 min; Mann-Whitney U=346; P=.03).

Survey and Assessing Factors Affecting End-User Experience on the Program

Survey Participants' Characteristics

During the 2-week survey period, we received survey responses from 112 participants in the hospital. Among them, we removed the responses of 10 participants who never used the "Patient's In-home Medication at a Glance" based on their answers to the first question. In addition, the responses of 5 participants who gave the same rating to the negative and positive questions were removed, as they were considered either not meaningful or not sincere to the survey, leaving 97 responses for analysis. Table 2 presents the characteristics. Participants included 62 (63.9%) physicians and 35 (36.1%) pharmacists, and the mean use count during the week was approximately 10.8 (SD 13.9).

Table . Participants' characteristics (N=97).

Characteristics				Values, n (%)
Occupation				
	Physician (n=62, 63.9%)			
		Position		
			Professor	35 (56.5)
			Resident	27 (43.5)
		Department		
			Internal medicine	50 (80.6)
			Surgery	12 (19.3)
		Workplace		
			Ambulatory clinic	24 (38.7)
			General ward	21 (33.9)
			Emergency room	11 (17.7)
			Intensive care unit	6 (9.7)
	Pharmacist			35 (36.1)
EHR [a] experience (years)				
	1			5 (5.2)
	3			20 (20.6)
	5			22 (22.7)
	10			20 (20.6)
	>10			30 (30.9)
Sex				
	Male			32 (33.0)
	Female			65 (67.0)
Age (years)				
	≤30			14 (14.4)
	31-40			58 (59.8)
	41-50			20 (20.6)
	>50			5 (5.2)
Weekly frequency of using the program				
	Mean (SD)			10.7 (13.9)
	Median (IQR)			6 (4-10)

[a]EHR: electronic health record.

Evaluation of the Survey Results

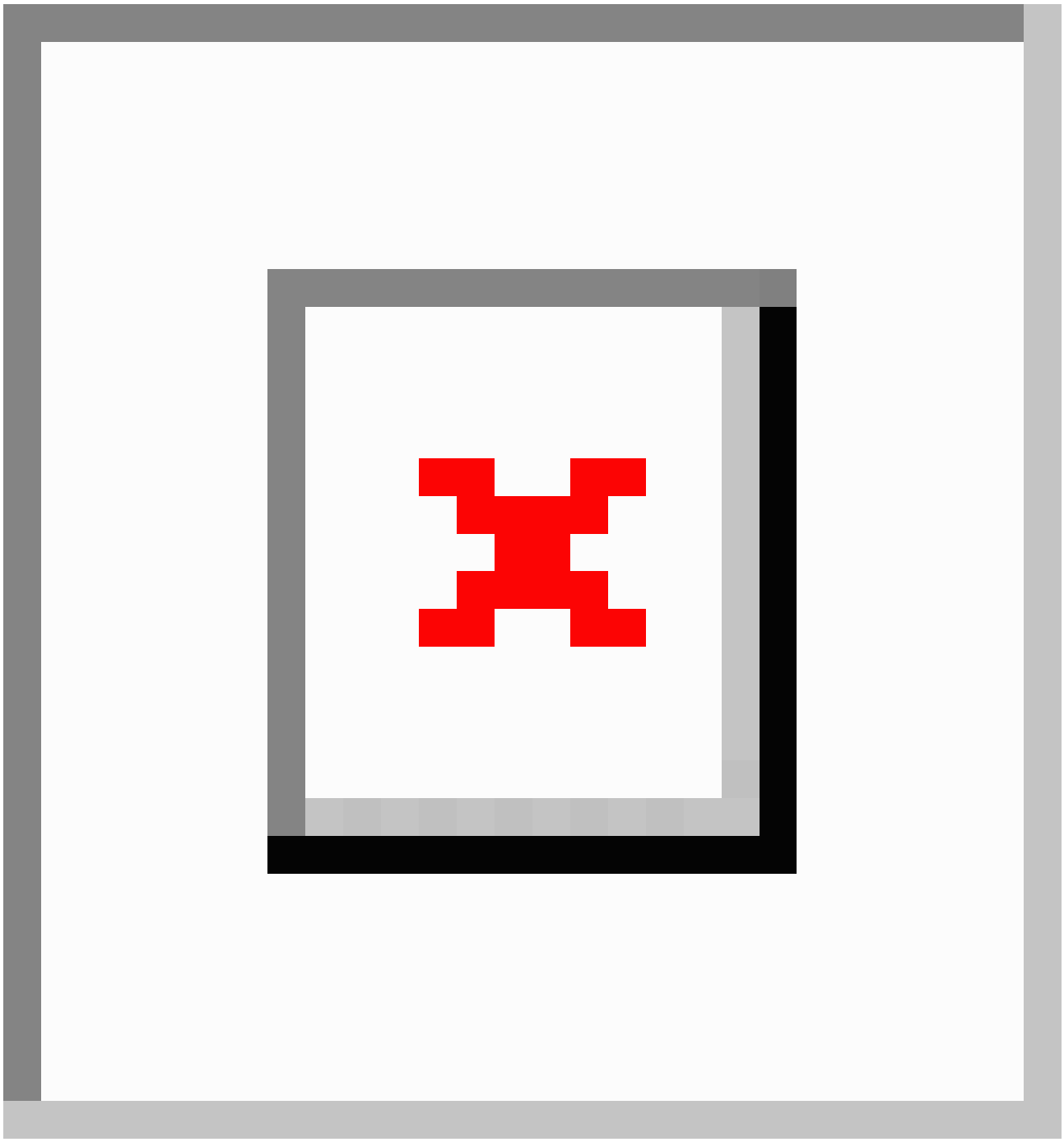
Of the 22 survey questions, the updated DMISM comprised 14 questions in 5 domains. After performing exploratory factor analysis, we calculated the mean score of each domain and Cronbach α to confirm the consistency of the items. This reliability analysis revealed that Cronbach α for all variables exceeded 0.80 (information quality: 0.808; system quality: 0.834; and service quality: 0.800), except for user satisfaction (Cronbach α=0.788) and intention to use (Cronbach α=0.795).

On a 5-point scale, the mean scores values for the information, system, and service quality of the program were 4.11 (SD 0.76), 4.24 (SD 0.75), and 3.87 (SD 0.79), respectively. User satisfaction (4.56, SD 0.49) and intention to reuse (4.77, SD 0.37) were measured. Among the 5 domains of the survey questionnaire, intention to reuse obtained the highest score. The estimates and weights of all 5 domains were analyzed, and no issues were observed in the convergent validity of the constructs (Table S2 in Multimedia Appendix 1). In addition, the subsequent analysis revealed the absence of discriminant validity (Table S3 in Multimedia Appendix 1).

Structural Correlations Between Domains

The SEM images are shown in Figure 3. The model fit indices were calculated as follows: χ^2_{70}=103.413 (P<.001); goodness-of-fit index=0.868 (recommended: 0-1.0); root mean square residual=0.039 (recommended: 0-0.05); and root mean square error of approximation=0.071 (recommended: 0.05-0.08). The comparative fit index and the Tucker-Lewis index for the model exceeded 0.9. The normed fit index and adjusted goodness-of-fit index values were lower than the recommended values of 0.859 and 0.802, respectively. Thus, this model was confirmed to be appropriate for assessing the factors affecting the “intention to reuse” program as an end-user experience.

Figure 3. Results of the research model using the structural equation modeling analysis. Significant paths are indicated with solid lines, while nonsignificant paths are shown with dotted lines. Two significant paths are shown: information quality toward user satisfaction and user satisfaction toward intention to reuse. The standardized beta values are presented. *P=.01; ***P<.001. CFI: Comparative Fit Index; RMSEA: root mean square error of approximation; TLI: Tucker-Lewis Index.



The associations between the latent variables were positive, supporting our hypotheses. Among the 3 quality domains, “information quality” had a significantly positive influence on “user satisfaction.” Consequently, the influence of “information quality” in “user satisfaction” and the influence of “user satisfaction” in “intent to reuse” were significantly associated.

Discussion

Principal Results

This study aimed to evaluate the impact of medication history retrieval using the "Patient's In-home Medications at a Glance" program in homegrown HISs during the 2-year maintenance phase after program implementation. The significance of our findings was twofold. First, we conducted a comprehensive evaluation of the impact of the nationwide medication history-sharing program, consisting of care process outcomes and end-user experiences as humanistic outcomes. We elaborately planned both the care process and humanistic outcomes of 2-year use, which allowed the program to stabilize, after its implementation in the HIS [23]. The care process, focusing on the time required for urgent PCI initiation, was improved in the patient group, whose physicians used the program and experienced expedited urgent PCI initiation. Thus, the use of the program could help identify whether patients are taking an antiplatelet or anticoagulant agent when they are unconscious or are unable to identify their medications. Regarding humanistic outcomes, the survey showed high scores overall, especially for "user satisfaction" and "intention to reuse." The increasing trend in the use of the "Patient's In-home Medications at a Glance" program by physicians and pharmacists indicates the successful integration of the newly developed program into the HIS, as evidenced by a positive end-user experience.

Second, we assessed factors affecting end-user experience using SEM; "information quality" significantly influenced "user satisfaction," and "user satisfaction," in turn, positively enhanced "intention to reuse." Since the survey was developed with the updated DMISM, which is a conceptual framework to suggest factors necessary for the "intention to reuse" the program, we could examine whether and how the 3 quality domains, including information, system, and service, affect "user satisfaction" and how "user satisfaction" affects "intention to reuse." These findings highlight the potential of the HIS in supporting clinical decision-making and contributing to value-based health care through the provision of a comprehensive medication use history.

Implications

Value-based health care is an approach to health care delivery in which providers are paid based on the patient's health outcomes [25], while reducing costs [26]. The benefits of a value-based health care system include reduced treatment costs, increased care efficiency, and reduced risks [27]. Measuring a patient's clinical outcomes is a major aim of value-based health care. In our study, we measured both care process outcomes and end-user experiences, which help present humanistic outcomes. Hence, a comprehensive evaluation was conducted by selecting both outcomes to determine the impact of the interventions using the HIS. Health service providers should provide patient-centered team care, share patients' medical information, and measure the care process using the HIS. The physicians were able to collect the patients' complete medication use histories in a friendly manner, even if the patients were unable to identify the exact medications they were taking. As access to a complete medication use history could help physicians make clinical decisions and collaborate care within the hospital [28], the HIS could help improve the patient's outcomes. Thus, HISs can play a vital role in value-based health care by delivering comprehensive and up-to-date information, including medication use history, laboratory results, and other medical records.

In terms of the association between the survey domains, the updated DMISM was applied to identify the quality factors that contribute to "user satisfaction," which affects end users' "intention to reuse." According to Alzahrani et al [29], 3 quality domains are significantly related to "user satisfaction" and "intention to reuse" and consequently affect actual usage. By conducting an SEM analysis of the survey results, our model revealed a significant effect of "information quality" on "user satisfaction," as well as "user satisfaction" on "intention to reuse." These results indicate that providing complete, accurate, and regent information is important for "user satisfaction," ultimately driving the "intention to reuse." A previous study stated that studies assessing the acceptance of HISs have been conducted from the physicians' perspective, not the clinical pharmacists' [30]. Since the program has been used by physicians and pharmacists, we could assess the factors affecting end-user experience in both professional groups. If the quality of information in an HIS is not guaranteed, health care professionals will not use specific programs in the HIS.

Limitations

This study had some limitations. First, we developed and implemented the "Patient's In-home Medications at a Glance" program in a single hospital. Thus, outcomes, such as care processes or factors affecting end-user experience, cannot be generalized to other hospitals in South Korea. However, as the Healthcare Insurance Review and Assessment Service has established guidelines for program development, further studies that use similar HISs could be conducted in other hospitals. Second, the pretest and posttest studies had the inherent limitations of nonrandomized, uncontrolled study designs. Although we showed the impact of the program on the time to PCI as the care process, we could not capture the long-term effects on clinical outcomes, such as survival rates or extended hospital stays. Nevertheless, our findings regarding the care process, specifically the reduction in time from the first ED assessment to urgent PCI initiation, could be meaningful not only in expediting clinical decisions but also in the evaluation of HISs in a real-world health care setting. Third, a notable limitation of our study is the imbalanced distribution of participants between the patient groups with or without the program (18 vs 59 participants) and the small number of patients in the group using the program. This uneven and small sample size raises concerns about the statistical robustness of our findings. Future research endeavors should prioritize achieving a more equitable number and distribution of patients to enhance the reliability and generalizability of our conclusions. Although our study offers valuable insights, the limitation of uneven and small sample sizes underscores the importance of cautious interpretation and highlights a potential area for improvement in subsequent research. Fourth, in the results of the SEM analysis, "information quality" was a standalone significant

factor among 3 quality domains influencing "user satisfaction." It is possible that the developed survey item may not adequately address the measurement of the quality domain. Lastly, our focus in this study was on system acceptability rather than the direct improvement in the health of the patients. We plan to focus more on the clinical outcome of the program, which includes not only medication information but also ensuring comprehensive disease management. This approach should be followed up for future measurements in subsequent studies.

Conclusions

Our findings highlight the impact of the rapid and complete medication history retrieval using the "Patient's In-home Medications at a Glance" program on the care process and end-user experience. A significantly positive effect was found on the care process by expediting urgent PCI initiation time at the ED, thereby contributing to value-based healthcare delivery in a real-world setting. Moreover, the HIS intervention provided high-quality information to physicians and pharmacists, resulting in high satisfaction. Long-term assessments can provide valuable insights into the sustained impact of the program, further optimizing patient outcomes.

Acknowledgments

We would like to thank the project team and program developer of the Medical Informatics Team for developing and implementing the program and assisting with data retrieval. We are grateful to all the pharmacists and physicians who participated in patient care. We would like to acknowledge the contributions of EEL as a cocorresponding author. We would also like to thank the Brain Korea (BK) 21 Plus Project of the National Research Foundation of Korea.

Authors' Contributions

JC, SY, HYL, and EEL contributed to the conception and design of the research, the acquisition and analysis of the data, as well as the interpretation of the data. They also drafted the manuscript. All authors critically revised the manuscript, agreed to be fully accountable for ensuring the integrity and accuracy of the work, and read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey items, convergent validity, and discriminant validity.
[DOCX File, 20 KB - medinform_v12i1e53079_app1.docx]

References

1. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 2009 Jan;78(1):22-31. [doi: 10.1016/j.ijmedinf.2008.06.005] [Medline: 18644745]
2. Aggelidis VP, Chatzoglou PD. Hospital information systems: measuring end user computing satisfaction (EUCS). J Biomed Inform 2012 Jun;45(3):566-579. [doi: 10.1016/j.jbi.2012.02.009] [Medline: 22426283]
3. Ojo AI. Validation of the DeLone and McLean information systems success model. Healthc Inform Res 2017 Jan;23(1):60-66. [doi: 10.4258/hir.2017.23.1.60] [Medline: 28261532]
4. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform 2014 Nov;83(11):779-796. [doi: 10.1016/j.ijmedinf.2014.06.011] [Medline: 25085286]
5. Laukka E, Huhtakangas M, Heponiemi T, Kanste O. Identifying the roles of healthcare leaders in HIT implementation: a scoping review of the quantitative and qualitative evidence. Int J Environ Res Public Health 2020 Apr 21;17(8):2865. [doi: 10.3390/ijerph17082865] [Medline: 32326300]
6. Jones SS, Rudin RS, Perry T, Shekelle PG. Health information technology: an updated systematic review with a focus on meaningful use. Ann Intern Med 2014 Jan 7;160(1):48-54. [doi: 10.7326/M13-1531] [Medline: 24573664]
7. Joseph AL, Stringer E, Borycki EM, Kushniruk AW. Evaluative frameworks and models for health information systems (HIS) and health information technologies (HIT). Stud Health Technol Inform 2022 Jan 14;289:280-285. [doi: 10.3233/SHTI210914] [Medline: 35062147]
8. Bates DW, Gawande AA. Improving safety with information technology. N Engl J Med 2003 Jun 19;348(25):2526-2534. [doi: 10.1056/NEJMsa020847] [Medline: 12815139]
9. Mogharbel A, Dowding D, Ainsworth J. Physicians' use of the computerized physician order entry system for medication prescribing: systematic review. JMIR Med Inform 2021 Mar 4;9(3):e22923. [doi: 10.2196/22923] [Medline: 33661126]

10. Neame MT, Sefton G, Roberts M, Harkness D, Sinha IP, Hawcutt DB. Evaluating health information technologies: a systematic review of framework recommendations. Int J Med Inform 2020 Oct;142:104247. [doi: 10.1016/j.ijmedinf.2020.104247] [Medline: 32871491]
11. Zeadally S, Siddiqui F, Baig Z, Ibrahim A. Smart healthcare: challenges and potential solutions using Internet of Things (IOT) and big data analytics. PSU Res Rev 2019 Feb;4:149-168. [doi: 10.1108/PRR-08-2019-0027]
12. Gunter MJ. The role of the ECHO model in outcomes research and clinical practice improvement. Am J Manag Care 1999 Apr;5(4 Suppl):S217-S224. [Medline: 10387542]
13. Marshall J, Hayes BD, Koehl J, et al. Effects of a pharmacy-driven medication history program on patient outcomes. Am J Health Syst Pharm 2022 Sep 22;79(19):1652-1662. [doi: 10.1093/ajhp/zxac143] [Medline: 35596269]
14. Cadwallader J, Spry K, Morea J, Russ AL, Duke J, Weiner M. Design of a medication reconciliation application: facilitating clinician-focused decision making with data from multiple sources. Appl Clin Inform 2013 Mar 13;4(1):110-125. [doi: 10.4338/ACI-2012-12-RA-0057] [Medline: 23650492]
15. Cornish PL, Knowles SR, Marchesano R, et al. Unintended medication discrepancies at the time of hospital admission. Arch Intern Med 2005 Feb 28;165(4):424-429. [doi: 10.1001/archinte.165.4.424] [Medline: 15738372]
16. Hart C, Price C, Graziose G, Grey J. A program using pharmacy technicians to collect medication histories in the emergency department. P T 2015 Jan;40(1):56-61. [Medline: 25628508]
17. Lau HS, Florax C, Porsius AJ, De Boer A. The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. Br J Clin Pharmacol 2000 Jun;49(6):597-603. [doi: 10.1046/j.1365-2125.2000.00204.x] [Medline: 10848724]
18. Tamblyn R, Huang AR, Meguerditchian AN, et al. Using novel Canadian resources to improve medication reconciliation at discharge: study protocol for a randomized controlled trial. Trials 2012 Aug 27;13:150. [doi: 10.1186/1745-6215-13-150] [Medline: 22920446]
19. Cho J, Lee E, Lee K, Lee HY, Lee E. Continuity of care with a one-click medication history program: patient's in-home medications at a glance. Int J Med Inform 2022 Apr;160:104710. [doi: 10.1016/j.ijmedinf.2022.104710] [Medline: 35183048]
20. William HD, Ephraim RM. The DeLone and McLean model of information systems success: a ten-year update. J Manag Info Syst 2003 Apr;19(4):9-30. [doi: 10.1080/07421222.2003.11045748]
21. Cho HH. Study on influence of perceived quality factor of smartphone on satisfaction & continued use intention - from the standpoint of updated DeLone & McLean's information system success model -. Entrue J Inf Technol 2012 Aug;11(2):167-180 [FREE Full text]
22. Shim M, Jo HS. What quality factors matter in enhancing the perceived benefits of online health information sites? application of the updated DeLone and McLean information systems success model. Int J Med Inform 2020 May;137:104093. [doi: 10.1016/j.ijmedinf.2020.104093] [Medline: 32078918]
23. Bossen C, Jensen LG, Udsen FW. Evaluation of a comprehensive EHR based on the DeLone and McLean model for IS success: approach, results, and success factors. Int J Med Inform 2013 Oct;82(10):940-953. [doi: 10.1016/j.ijmedinf.2013.05.010] [Medline: 23827768]
24. Song T, Deng N, Cui T, et al. Measuring success of patients' continuous use of mobile health services for self-management of chronic conditions: model development and validation. J Med Internet Res 2021 Jul 13;23(7):e26670. [doi: 10.2196/26670] [Medline: 34255685]
25. NEJM Catalyst. What is value-based healthcare? NEJM Catalyst. 2017 Jan 1. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.17.0558 [accessed 2023-03-01]
26. Ibanez-Sanchez G, Fernandez-Llatas C, Martinez-Millana A, et al. Toward value-based healthcare through interactive process mining in emergency rooms: the stroke case. Int J Environ Res Public Health 2019 May 20;16(10):1783. [doi: 10.3390/ijerph16101783] [Medline: 31137557]
27. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
28. Vargas V, Blakeslee WW, Banas CA, Teter C, Dupuis-Dobson K, Aboud C. Use of complete medication history to identify and correct transitions-of-care medication errors at psychiatric hospital admission. PLoS One 2023;18(1):e0279903. [doi: 10.1371/journal.pone.0279903] [Medline: 36696376]
29. Alzahrani AI, Mahmud I, Ramayah T, Alfarraj O, Alalwan N. Modelling digital library success using the DeLone and McLean information system success model. J Librariansh Inf Sci 2019 Jun;51(2):291-306. [doi: 10.1177/0961000617726123]
30. English D, Ankem K, English K. Acceptance of clinical decision support surveillance technology in the clinical pharmacy. Inform Health Soc Care 2017 Mar;42(2):135-152. [doi: 10.3109/17538157.2015.1113415] [Medline: 26890621]

Abbreviations

DMISM: DeLone and McLean Model of Information Systems Success
ED: emergency department
HIS: health information system
PCI: percutaneous coronary intervention

SEM: structural equation modeling
SNUBH: Seoul National University Bundang Hospital

Edited by C Lovis; submitted 25.09.23; peer-reviewed by D Carvalho, G Vergeire-Dalmacion; revised version received 16.01.24; accepted 04.02.24; published 20.03.24.

Please cite as:
Cho J, Yoo S, Lee EE, Lee HY
Impact of a Nationwide Medication History Sharing Program on the Care Process and End-User Experience in a Tertiary Teaching Hospital: Cohort Study and Cross-Sectional Study
JMIR Med Inform 2024;12:e53079
URL: https://medinform.jmir.org/2024/1/e53079
doi:10.2196/53079

A Patient Similarity Network (CHDmap) to Predict Outcomes After Congenital Heart Surgery: Development and Validation Study

Haomin Li[1,*], PhD; Mengying Zhou[1,2,*], MSc; Yuhan Sun[1,2,*], BSc; Jian Yang[1,2], BSc; Xian Zeng[1,2], PhD; Yunxiang Qiu[3], MD; Yuanyuan Xia[3], MD; Zhijie Zheng[3], MD; Jin Yu[4], MD; Yuqing Feng[1], MSc; Zhuo Shi[5], MD; Ting Huang[5], MD; Linhua Tan[3], MD; Ru Lin[5], MD; Jianhua Li[5], MD; Xiangming Fan[5], MD; Jingjing Ye[4], MD; Huilong Duan[2], PhD; Shanshan Shi[3,*], MD; Qiang Shu[5,*], MD

1

2

3

4

5

[*]these authors contributed equally

Corresponding Author:
Haomin Li, PhD

Abstract

Background: Although evidence-based medicine proposes personalized care that considers the best evidence, it still fails to address personal treatment in many real clinical scenarios where the complexity of the situation makes none of the available evidence applicable. “Medicine-based evidence” (MBE), in which big data and machine learning techniques are embraced to derive treatment responses from appropriately matched patients in real-world clinical practice, was proposed. However, many challenges remain in translating this conceptual framework into practice.

Objective: This study aimed to technically translate the MBE conceptual framework into practice and evaluate its performance in providing general decision support services for outcomes after congenital heart disease (CHD) surgery.

Methods: Data from 4774 CHD surgeries were collected. A total of 66 indicators and all diagnoses were extracted from each echocardiographic report using natural language processing technology. Combined with some basic clinical and surgical information, the distances between each patient were measured by a series of calculation formulas. Inspired by structure-mapping theory, the fusion of distances between different dimensions can be modulated by clinical experts. In addition to supporting direct analogical reasoning, a machine learning model can be constructed based on similar patients to provide personalized prediction. A user-operable patient similarity network (PSN) of CHD called CHDmap was proposed and developed to provide general decision support services based on the MBE approach.

Results: Using 256 CHD cases, CHDmap was evaluated on 2 different types of postoperative prognostic prediction tasks: a binary classification task to predict postoperative complications and a multiple classification task to predict mechanical ventilation duration. A simple poll of the *k*-most similar patients provided by the PSN can achieve better prediction results than the average performance of 3 clinicians. Constructing logistic regression models for prediction using similar patients obtained from the PSN can further improve the performance of the 2 tasks (best area under the receiver operating characteristic curve=0.810 and 0.926, respectively). With the support of CHDmap, clinicians substantially improved their predictive capabilities.

Conclusions: Without individual optimization, CHDmap demonstrates competitive performance compared to clinical experts. In addition, CHDmap has the advantage of enabling clinicians to use their superior cognitive abilities in conjunction with it to make decisions that are sometimes even superior to those made using artificial intelligence models. The MBE approach can be embraced in clinical practice, and its full potential can be realized.

(JMIR Med Inform 2024;12:e49138) doi:10.2196/49138

KEYWORDS

medicine-based evidence; general prediction model; patient similarity; congenital heart disease; echocardiography; postoperative complication; similarity network; heart; cardiology; NLP; natural language processing; predict; predictive; prediction; complications; complication; surgery; surgical; postoperative

Introduction

Congenital heart disease (CHD) is the most common type of birth defect, with birth prevalence reported to be 1% of live births worldwide [1]. Despite remarkable success in the surgical and medical management that has increased the survival of children with CHD [2], the quality of treatment and prognosis after congenital heart surgery remains unsatisfactory and varies across centers [3,4]. The reason for this is that the complexity of the disease, clinical heterogeneity within lesions, and small number of patients with specific forms of CHD severely degrade the precision and value of estimates of average treatment effects provided by randomized controlled trials on the average patient. Some visionary researchers have proposed a new paradigm called "medicine-based evidence" (MBE), in which big data and machine learning techniques are embraced to interrogate treatment responses among appropriately matched patients in real-world clinical practice [5,6].

Postoperative complications in congenital heart surgery have been inconsistently reported but have important contributions to mortality, hospital stay, cost, and quality of life [7-9]. Heart centers with the best outcomes might not report fewer complications but rather have systems in place to recognize and correct complications before deleterious outcomes ensue [8]. The early detection of deterioration after congenital heart surgery enables prompt initiation of therapy, which may result in reduced impairment and earlier rehabilitation. Several risk scoring systems, such as the Risk Adjustment for Congenital Heart Surgery 1 (RACHS-1) method, Aristotle score, and Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery (STS-EACTS) score, have been developed and used to adjust the risk of in-hospital morbidity and mortality [10-13]. However, most of these consensus-based risk models only focus on the procedures themselves and ignore the differences between centers and patients. Specific patient characteristics, such as lower weight [14] and longer cardiopulmonary bypass time [15], especially the quantitative echocardiographic indicators used by clinicians to understand CHD conditions, were not incorporated into these models nor can they be adjusted for. Based on the increasing number of CHD databases being built, some machine learning–based predictive models have recently been used to identify independent risk factors and predict complications after congenital heart surgery [16-18]. These predictive models achieved outstanding performance compared to traditional risk scores, but these models are usually only capable of performing a single task. In addition, such models often contain hundreds of features, so for clinicians, understanding how to interpret the prediction from a complicated machine learning model is still a challenge [19]. Based on our previous studies [16-18], as the model becomes more complex and more variables are included, the results are better, but it is more difficult to understand and accept clinically. Although some explainable artificial intelligence (AI) techniques continue to evolve [20,21], machine learning prediction models are still a black box for clinicians. Due to the lack of understanding and manipulation of the model, clinicians often lack confidence in the predicted outcomes, which severely hampers the entry of these machine learning models into routine care.

Patient similarity networks (PSNs) are an emerging paradigm for precision medicine, in which patients are clustered or classified based on their similarities in various features [22,23]. PSNs address many challenges in data analytics and is naturally interpretable. In a PSN, each node is an individual patient, and the distance (or edge) between 2 nodes corresponds to pairwise patient similarity for given features. PSNs naturally handle heterogeneous data, as any data type can be converted into a similarity network by defining similarity measures [24,25]. A PSN generated based on a large cohort of patients will show several subgroups of patients who are tightly connected. If a new patient is located on the PSN, neighbors that have similar features with known risk or prognosis will inform clinicians of the potential risk and prognosis of the patient. This mimics the clinical reasoning of many experienced clinical experts, who often relate a patient to similar patients they have seen. Moreover, representing patients by similarity is conceptually intuitive and explainable because it can convert the data into network views, where the decision boundary can be visually evident [26]. PSNs can also provide a feasible engineering solution for the MBE framework, which, based on a library of "approximate matches" consisting of a group of patients who share the greatest similarity with the index case, can be examined to estimate the effects of various treatments within the context of the individual patient's specific characteristics [6].

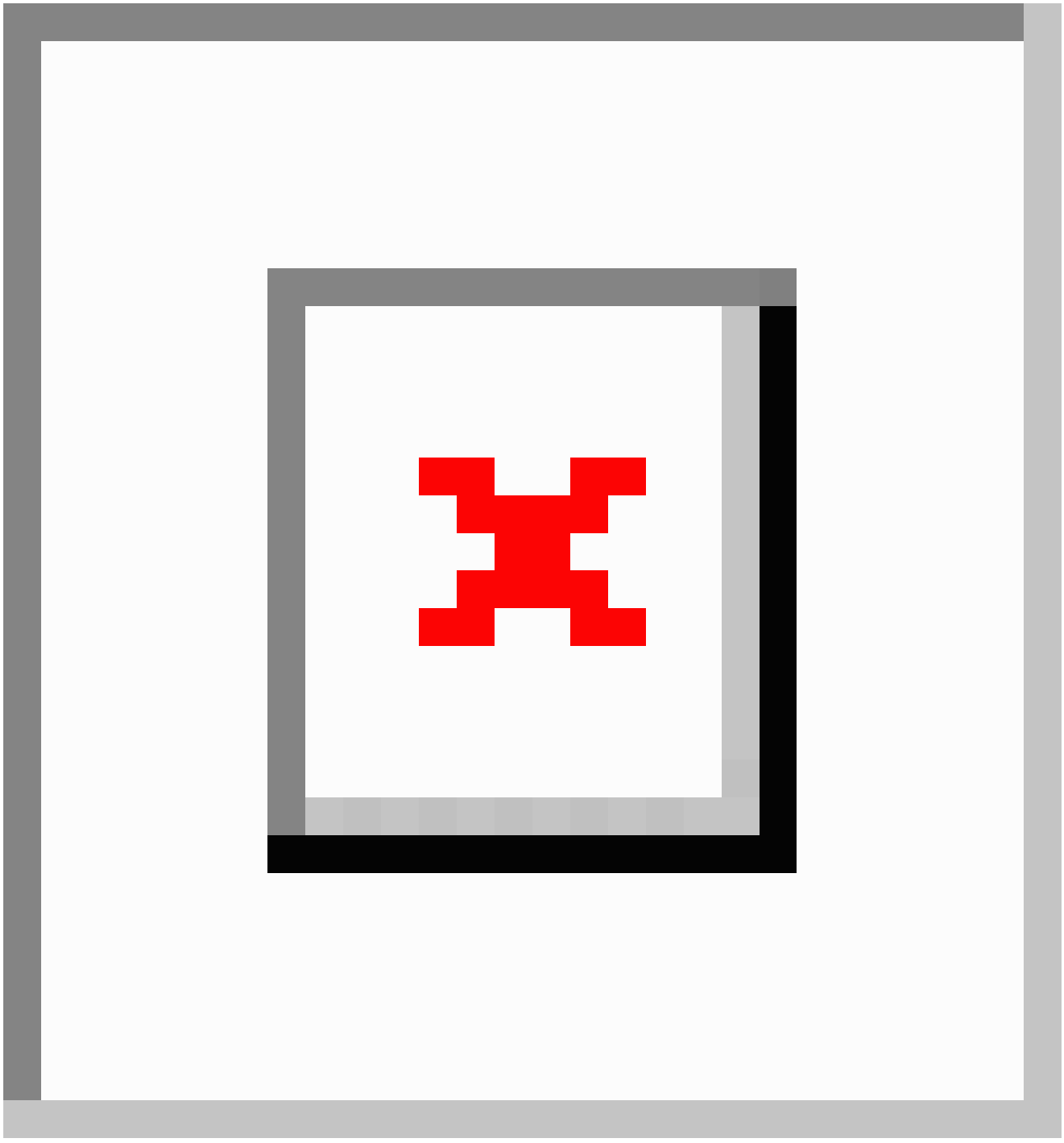
PSNs have been reported in many studies. Although early PSN studies have focused on using omics data in precision medicine [27-29], with the development of electronic health record (EHR) systems, abundant, complex, high-dimensional, and heterogeneous data are being captured during daily care, and some EHR-based patient similarity frameworks have been proposed for diagnosis [30], subgroup patients [31,32], outcome prediction [33], drug recommendation [34,35], and disease screening [36]. However, studies of PSNs that predict the outcome after CHD surgery have not been reported. A perspective article proposed an MBE conceptual framework for CHD [6], in which similarity analysis is used to generate a library of "approximate matches." However, they did not provide any technical solution for this framework. The challenge in applying PSNs in a real clinical setting is, first of all, to assess the distance between patients with complex conditions such as CHD in a computable way. However, mimicking clinical analogy reasoning is not a simple math formula based on various patients' attributes. The structure-mapping theory in cognitive science argues that advanced cognitive functions are involved in the analysis of relationship similarity above attribute similarity [37]. Analogy inference requires advanced cognitive activity, which current AI technology lacks but clinical experts are good at. However, all established models ignore this important feature of patient similarity analysis, in that it should not only measure patients' distance but also put clinicians back behind the wheel to generate MBE for clinical decision-making. In this study, we aimed to develop and evaluate a clinician-operable PSN of CHD to try to mitigate the above problems.

Methods

Study Design and Population

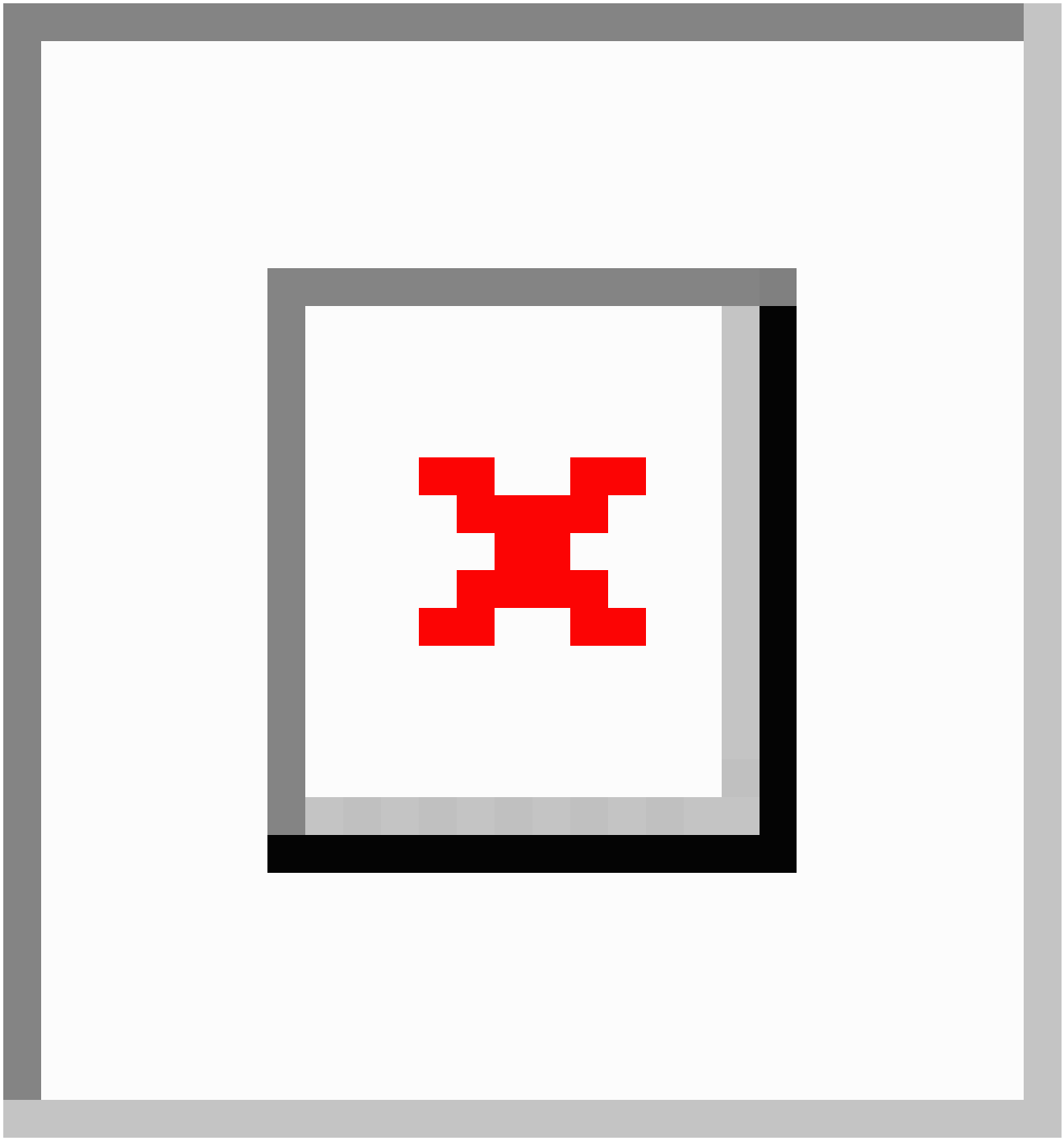
As shown in Figure 1, using data available at different stages, 4 PSNs were generated and named as screening map, echo map, patient map, and surgery map. These data were obtained from the ultrasound reporting system and EHR system of the Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China.

Figure 1. CHDmap contains 4 patient similarity networks generated from 4 different clinical phases, with different data obtained at each phase. CHD: congenital heart disease; ICU: intensive care unit; LOS: length of stay.



A schematic of the data processing and workflow for the construction of the PSN is shown in Figure 2 and described below.

Figure 2. Schematic of data processing and workflow of the construction of the congenital heart disease (CHD) patient similarity network. NLP: natural language processing; t-SNE: t-distributed stochastic neighbor embedding.



Ethical Considerations

This retrospective study was performed according to relevant guidelines and approved by the institutional review board of the Children's Hospital of Zhejiang University School of Medicine with a waiver of informed consent (2018_IRB_078). All cases included in this study were anonymized. Intensive care unit (ICU) clinicians who participated in the trial received cash compensation (RMB ¥100 [US $14.06] per day), which complied with local regulatory requirements for scientific labor.

Data Collection and Preprocessing

In addition to preoperative echocardiography reports that described the CHD conditions, the following patient and surgical characteristics were also collected: age, sex, height, weight, preoperative oxygen saturation of the right-upper limb, surgery time, cardiopulmonary bypass time, aortic cross-clamping time, mechanical ventilation time, duration of postoperative hospital stay, duration of ICU stay, and postoperative complications (the detailed definitions of postoperative complications are shown in Table S1 in Multimedia Appendix 1 [38-40]).

The most challenging part of patient similarity analysis was defining all the semantic concepts in the domain. An ontology of CHD was developed based on reviewing a large number of clinical guidelines for CHD to cover 436 CHD conditions and 87 related echocardiographic indicators. The OWL format ontology file is available on the CHDmap website [41]. The

ontology was used to normalize all concepts and measure semantic similarity among them. It was also used to identify quantitative indicators from the unstructured text of echocardiography reports. In addition to recording some routine cardiac structure indicators, the echocardiography report also provided quantitative indicators regarding various malformations, such as the size of various defects, shunt flow velocity, and pressure difference at the defect, depending on the specific CHD structural malformation. Natural language processing (NLP) technology [38] was used to extract 66 commonly used quantitative indicators. A range of processing and computational methods were used to assess similarity between patients (details information are shown in the supplemental methods and Tables S2 Table S3 in Multimedia Appendix 1). The various automatically extracted measurement values were subject to quality control, and any abnormal data (outside the reasonable range of the corresponding values) were modified or removed after manual verification. The diagnosis in the report was also extracted and mapped to the normalized terms defined in the CHD ontology.

Measuring Patient Similarity

In this study, the similarity of patients with CHD was measured using 4 groups of features: the quantitative echocardiographic indicators, the specific CHD diagnosis, preoperative clinical features, and surgical features. Different distance measurement methods were adopted for different groups of features, as described in the supplemental methods in Multimedia Appendix 1. We provided 3 types of methods to handle the echocardiographic indicators: the origin value, the *z* score, and the indicator combination ratio. The similarity between 2 diagnoses was calculated using the depth of the corresponding nodes in the CHD ontology, which organizes hundreds of CHD diagnoses in a hierarchical structure. Two approaches were used to measure the distance between diagnosis lists: one treats all diagnoses equally, referred to in the result section as “ungrade,” whereas the other distinguishes between basic and other diagnoses, referred to as “grade.” Finally, the patient distance was measured as the weighted sum of the 4 distances as shown in equation (1), and the final distances were also normalized to [0,1].

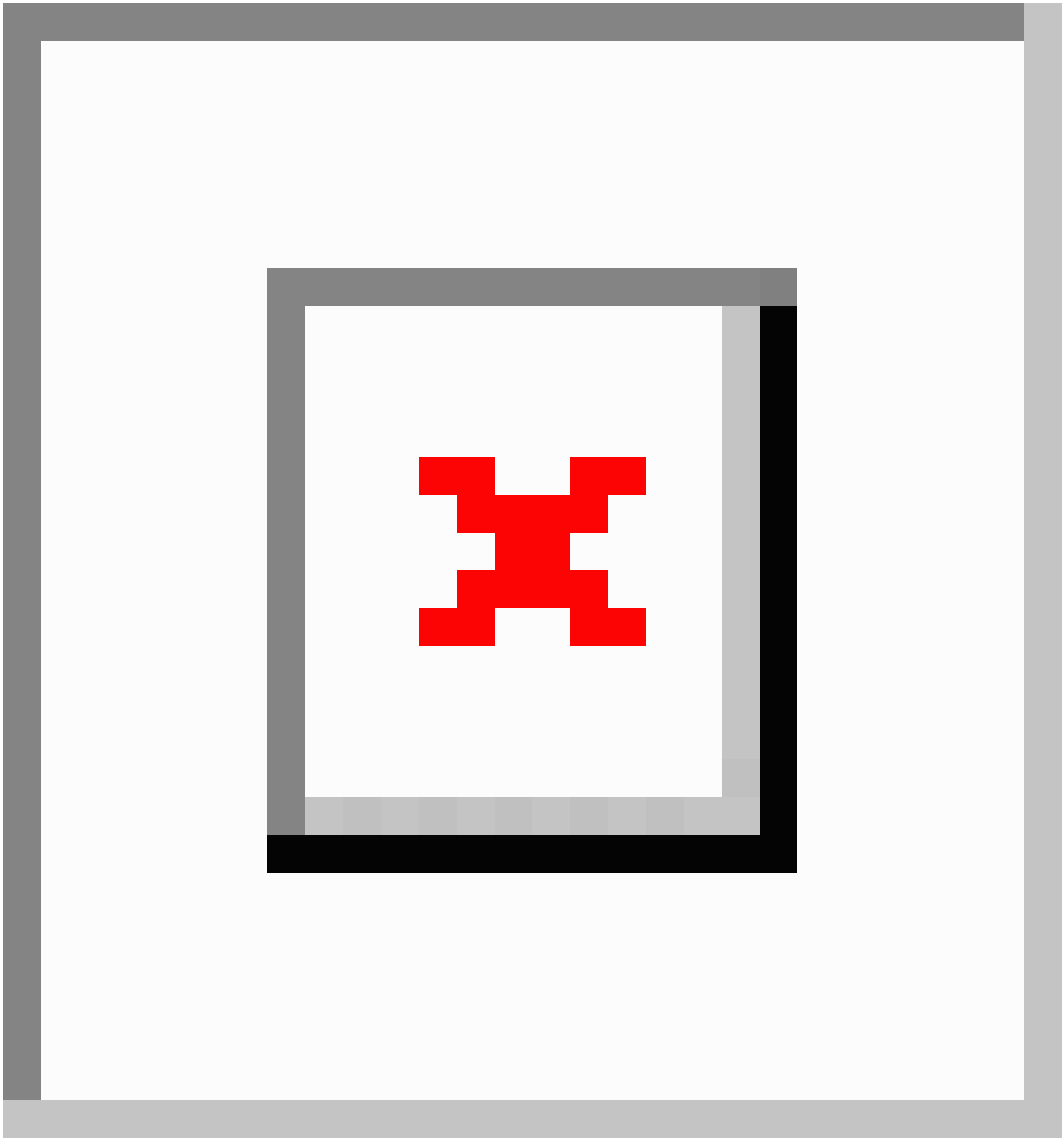
(1)

The weights in equation (1) and the different methods used to measure distance can also be modified by users depending on their experience in different tasks to fully exploit the advanced cognitive ability of clinical professionals. The distance matrix among historical patients can be calculated based on the aforementioned methods. We used t-distributed stochastic neighbor embedding [42] to convert the distance matrix into 2D points, which can be visualized as a map. The user-operable CHDmap was developed based on ECharts [43] using React (Meta) and Node.js (OpenJS Foundation). The patient similarity analysis engine, which measures the distances between a new patient and patients in CHDmap, was developed using Python (Python Software Foundation).

CHDmap

A user-operable CHD PSN called CHDmap was developed and published on the web [44]. The introduction video of this tool is also available in Multimedia Appendix 2. Based on the different available data for each clinical phase, as shown in Figure 1, CHDmap provides 4 different PSNs: the screening map, echo map, patient map, and surgery map. The workspace of CHDmap comprises 3 major modules: (1) map view, (2) cockpit view, and (3) outcome view (as shown in Figure 3).

Figure 3. Screenshot of CHDmap. The map view, cockpit view, and outcome view of the workspace are marked separately. CHDmap was published on the web [44]. CHD: congenital heart disease.



The map view presents the PSN as a zoomable electronic map, in which a node presents a patient and the distance between nodes shows their similarity. The map can be enhanced by using different colors to show the diagnostic labels as well as relevant prognostic indicators (eg, length of stay and complications). Different methods to handle the echocardiographic indicators, such as normal, z score, or combination ratio value, can be selected on the web. The similar patient group is also highlighted on the map view during similarity analysis.

The cockpit view provides a navigation function that helps clinicians locate cases based on specified query conditions, such as age, gender, and CHD subtypes. In practice, clinicians were allowed to create a new case, in which an NLP-based information extraction tool will assist users in filling in most of the echocardiographic indicators based on Chinese echocardiography reports. The top *k* value, or threshold of patient similarity, is used to customize the similar group. For advanced users, a customized map can be generated by adjusting the weights for the patient similarity measurement defined in the *Methods* section.

The outcome view provides an overview of outcomes, including the length of hospital stay, mechanical ventilation time, length of ICU stay, complications, and hospital survival of the selected similar patient group. Multiple charts are used to show the difference between the selected patient group and others. The Mann-Whitney *U* test and the χ^2 test are used to determine the

significance of differences between groups. When there are significant differences between the selected patient group and other patients, the color of the check box at the top of the outcome view will turn red; otherwise, it will stay gray. Checking the box will show detailed charts and tables of the outcome. This real-time feedback will help clinicians adjust the parameters in the cockpit view based on the requirements of the scenario for clinical decision-making. Based on a selected group of similar patients, CHDmap provides machine learning models to personalize the prediction of relevant outcome metrics for the current patient. Therefore, for each case, different parameters can be applied and compared to ultimately assess the credibility of the relevant decision support information.

Evaluation Method

The closer 2 patients are located on the CHDmap, the more similar their conditions and postoperative outcomes are considered to be. When a new patient is admitted to the hospital, historical patients can be divided into similar and nonsimilar groups based on some criteria. There are 2 criteria to define patient similarity groups: one is to use the most similar k patients, also known as k-nearest neighbor (KNN), to form a patient similarity group, and the other is to define a threshold above which patients form a similarity group. The statistical characteristics or regression value of postoperative outcomes in the similarity group are used to predict the outcomes of the current patient.

In this paper, we evaluated the performance of the surgery map of CHDmap on 2 tasks: predicting postoperative complications as a binary classification task, in which more than 50% of patients in the similarity group with complications were assigned ‘True” for the target patient, and predicting mechanical ventilation duration as a multiple-label classification task (I: 0-12 h, II: 12-24 h, III: 24-48 h, and IV: >48 h), in which the category with the highest proportion in the similarity group was assigned to the target patient.

As the optimum k of KNN to form a similarity group for a specific case is always different, the unified population-level optimized k on the training data set was used to evaluate CHDmap on the test data set without individual customization. Different data preprocessing methods (original, z score, and combination ratio) and whether to distinguish primary diagnoses (grade and ungrade) were tested and compared.

Making decisions may not be straightforward if the outcome of a similar patient group is extremely heterogeneous, whereby a machine learning model based on a similar patient population can provide a more personalized prediction of the relevant prognostic indicators. Although there are numerous machine learning models to choose from, the focus of this study was to demonstrate the advantages of basing the model on similar patient populations, so we chose to use the most conventional and easily understood logistic regression (LR) model. Clinical users obtained a population of similar patients after various parameter adjustments and threshold settings on CHDmap, and the data from this population were used to train an LR model (KNN+LR), which can be accomplished on the web in real time because this population of similar patients is usually not very large. To demonstrate the effect of similar patient populations, we trained another LR model (k-Random+LR) based on randomly collected cases of the same size in parallel in the evaluation. We evaluated such approaches and compared the LR models based on k similar patients and k random patients.

The accuracy, recall, F_1-score, and area under the receiver operating characteristic curve (AUC), which are defined below, were adopted to evaluate the performance of the classification. Accuracy is defined as the total correctly classified example including true positive (TP) and true negative (TN) divided by the total number of classified examples. Recall quantifies the number of correct positive predictions made out of all positive predictions that could have been made. F_1-score is a weighted average of precision and recall. As we know, in precision and recall, there are false positive (FP) and false negative (FN), so F_1-score also considers both of them. AUC provides an aggregate measure of the performance across all possible classification thresholds. The higher the accuracy, recall, F_1-score, and AUC, the better the model’s performance is at distinguishing between the positive and negative classes.

(2)

(3)

(4)

(5)

The performance was evaluated on an independent test set, which included 256 patients with CHD. These test cases were also available on CHDmap when users created a new case. Three clinicians working in the cardiac ICU with extensive experience were also asked to make relevant judgments for these test cases based on their clinical experience. After half a year following the initial trial, we conducted an experiment where the 3 clinicians were asked to make further predictions based on the output of CHDmap, and this prediction was compared with the previous results based on clinical experience alone to validate the benefits of CHDmap in supporting clinical decision-making.

Results

Population Characteristics

A total of 4774 patients who underwent congenital heart surgery between June 2016 and June 2021 at the Children’s Hospital of Zhejiang University School of Medicine were used to generate the CHD PSN. The performance of the PSN in predicting complications and mechanical ventilation duration was evaluated on an independent test data set, which included 256 pediatric patients who underwent congenital heart surgery between July 2021 and November 2021 at the Children’s Hospital of Zhejiang University School of Medicine. The characteristics of patients used to generate the PSN and for evaluation are described in Table 1. Since the test data and the

data used by the PSN were generated and collected in different time periods, as shown in Table 1, they are somewhat statistically different. The test data were older; therefore, the patients were significantly larger in terms of height and weight (P<.001), and there were also relatively large differences in the distribution of outcomes, lower complication rates, and shorter duration of mechanical ventilation. It should be noted that the diagnostic label is not the complete diagnostic information; we just use a few of the most common CHD subtypes to facilitate statistics and visualization, and this cohort contains a complete range of epidemiological characteristics as well as a variety of complex CHD subtypes such as transposition of the great arteries, tetralogy of Fallot, etc, which may appear in various diagnostic labels that they are combined with. When the case has 2 common CHD subtypes, such as ventricular septal defect and patent ductus arteriosus, only the more common subtype, ventricular septal defect, is labeled.

Table . Characteristics of patients with CHD[a] used to generate CHDmap and in the test data set.

Characteristic		Patients of CHDmap (n=4774)	Patients of the test data set (n=256)	*P* value
Gender (male), n (%)		2336 (48.9)	111 (43.4)	.09
Age (mo), median (IQR)		12.0 (4.0-32.0)	22.1 (7.8-50.9)	<.001
Height (cm), median (IQR)		75.0 (63.0-94.0)	85.5 (67.0-106.3)	<.001
Weight (kg), median (IQR)		9.2 (6.0-13.4)	10.8 (6.8-16.5)	<.001
Preoperative oxygen saturation (%), median (IQR)		98.0 (97.0-99.0)	98.0 (97.0-99.0)	.007
Surgery time (min), median (IQR)		119.0 (96.0-147.0)	120.0 (100.0-147.0)	.25
Cardiopulmonary bypass time (min), median (IQR)		60.0 (48.0-82.0)	61.5 (49.3-80.0)	.55
Aortic cross-clamping time (min), median (IQR)		40.0 (28.0-54.0)	38.5 (27.0-52.0)	.55
Duration of hospital stay (d), median (IQR)		9.0 (7.0-13.0)	7.0 (6.0-11.0)	.003
Duration of ICU[b] stay (d), median (IQR)		3.0 (1.0-4.0)	3.0 (1.0-4.0)	.49
Diagnostic label, n (%)				.46
	ASD[c] and VSD[d]	1659 (34.8)	78 (30.5)	
	VSD	1522 (31.9)	94 (36.7)	
	ASD	1228 (25.7)	65 (25.4)	
	PFO[e]	134 (2.8)	5 (2)	
	PDA[f]	123 (2.6)	9 (3.5)	
	Others	108 (2.3)	5 (2)	
Mechanical ventilation time (%), n (%)				.001
	I (<12 h)	3009 (63.0)	180 (70.3)	
	II (12-24 h)	918 (19.2)	54 (21.1)	
	III (24-48 h)	433 (9.1)	7 (2.7)	
	IV (≥48 h)	414 (8.7)	15 (5.9)	
Complication, n (%)		1229 (25.7)	48 (18.8)	.02

[a]CHD: congenital heart disease.

[b]ICU: intensive care unit.

[c]ASD: atrial septal defect.

[d]VSD: ventricular septal defect.

[e]PFO: patent foramen ovale.

[f]PDA: patent ductus arteriosus.

Performance of CHDmap

Three methods for preprocessing the echocardiographic indicators (origin, *z* score, combination) and 2 distinguishing primary diagnoses (grade and ungrade) were used to compare their effect on CHDmap performance. The performance of the CHDmap and 3 clinicians is shown in Table 2 and Figure 4.

Table . Evaluation results in the 2 tasks.

Methods		Prediction of postoperative complications				Prediction of mechanical ventilation duration			
		Accuracy	Recall	F_1-score	AUC[a]	Accuracy	Recall	F_1-score	AUC
KNN[b]									
	Origin+ungrade	0.832	0.438	0.494	0.757	0.813	0.444	0.459	0.862
	Origin+grade	0.836	0.417	0.489	0.773	0.797	0.437	0.467	0.860
	z score+ungrade	0.828	0.458	0.500	0.738	0.836	0.554	0.574	0.902
	z score+grade	0.848	0.458	0.530	0.747	0.855	0.564	0.573	0.895
	Combination+ungrade	0.836	0.500	0.533	0.767	0.828	0.468	0.488	0.900
	Combination+grade	0.859	0.458	0.550	0.768	0.855	0.521	0.545	0.873
KNN+LR[c]									
	Origin+ungrade	0.813	0.604	0.547	*0.810*[d]	0.848	0.558	0.602	0.921
	Origin+grade	0.813	*0.667*	0.571	0.799	0.863	0.589	0.632	0.920
	z score+ungrade	0.809	0.604	0.542	0.809	0.840	0.537	0.561	0.888
	z score+grade	0.813	0.646	0.564	0.805	0.855	0.549	0.562	0.886
	Combination+ungrade	0.805	0.583	0.528	0.801	0.840	0.537	0.555	0.900
	Combination+grade	0.805	0.604	0.537	0.798	0.824	0.500	0.522	*0.926*
k-Random+LR		0.809	0.500	0.495	0.774	0.809	0.484	0.488	0.895
Clinicians[e]									
	C1	0.875	0.396	0.543	N/A[f]	0.844	*0.614*	0.618	N/A
	C2	0.758	0.646	0.500	N/A	0.734	0.535	0.496	N/A
	C3	0.840	0.208	0.328	N/A	0.797	0.498	0.536	N/A
	Clinician average	0.824	0.417	0.457	N/A	0.792	0.549	0.550	N/A
	C1+CHDmap	*0.883*	0.426	*0.580*	N/A	*0.943*	0.612	*0.647*	N/A
	C2+CHDmap	0.816	0.5625	0.534	N/A	0.874	0.587	0.542	N/A
	C3+CHDmap	0.852	0.313	0.441	N/A	0.916	0.511	0.546	N/A
	Clinician+CHDmap average	0.850	0.434	0.518	N/A	0.911	0.570	0.578	N/A

[a]AUC: area under the receiver operating characteristic curve.

[b]KNN: *k*-nearest neighbor.

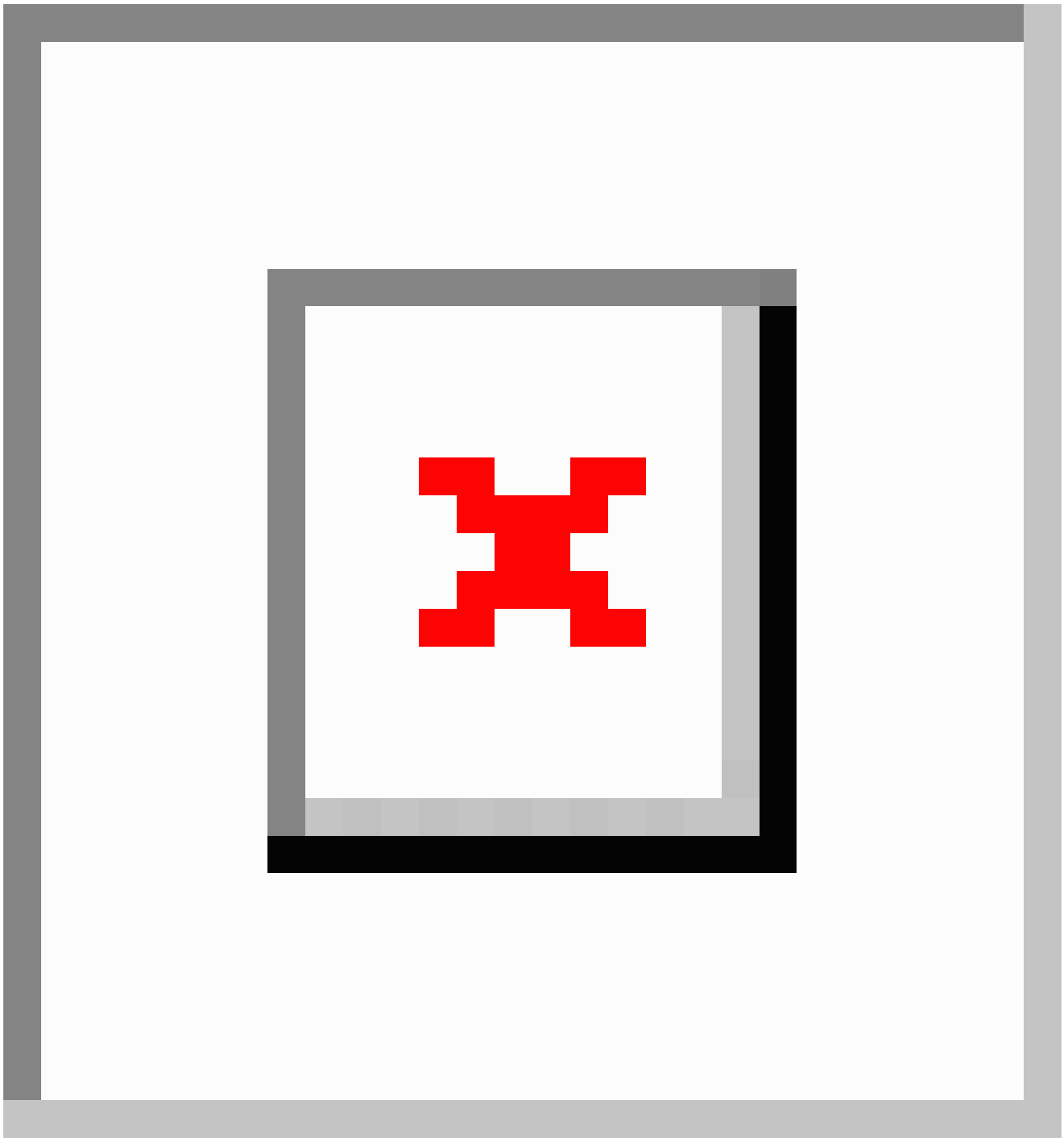
[c]LR: logistic regression.

[d]In each column, the maximum value is italicized.

[e]The performance of the 3 clinicians are labeled as C1, C2, and C3.

[f]N/A: not applicable.

Figure 4. Evaluation result based on receiver operating characteristic curves. (A) Binary postoperative complication prediction using KNN; (B) to (E) multilabel mechanical ventilation duration prediction (I: 0-12 h, II: 12-24 h, III: 24-48 h, and IV: >48 h) using KNN, respectively; (F) binary postoperative complication prediction using KNN+LR; (G) to (J) multilabel mechanical ventilation duration prediction (I: 0-12 h, II: 12-24 h, III: 24-48 h, and IV: >48 h) using KNN+LR, respectively. The performance of 3 clinicians are labeled as black stars in different tasks as C1, C2, and C3. The performance of 3 clinicians enhanced by CHDmap are labeled as red stars. CHD: congenital heart disease; KNN: *k*-nearest neighbor; LR: logistic regression.



In the postoperative complication prediction task, the F_1-score of methods using KNN exceeded the average of the 3 clinicians, although 1 clinician achieved the best accuracy when dropping a high recall value. In all 6 KNN methods, introducing the indicator combination ratio and distinguishing the primary diagnosis in the similarity measurement can truly improve the overall performance of the F_1-score. LR models constructed using the KNN-obtained patient groups were able to generally achieve better predictions compared to simple voting of similar patients and the LR model based on k random patients. Interestingly, both the model with the best F_1-score performance and the model with the best AUC used the original values. This may be because original values are more reflective of individualized patient differences in a similar patient population. The main improvement of CHDmap on this task is reflected in the general improvement in recall values, with the best recall method being 0.250 higher than the clinician average.

In another multiclassification task that predicts mechanical ventilation duration, the differences among these different KNN methods in overall performance were not consistent. The

KNN+LR approaches also achieved better composite performance (F_1-score and AUC), although 1 of the human experts got the best recall value.

From the test result, clinicians do not have the same performance for such predictive judgments. Some raise the standard and thus miss some events; on the other hand, some lower the judgment threshold, and thus the accuracy of the judgment decreases. At the same time, the performance of clinical experts on different tasks is inconsistent. A simple poll of the *k*-most similar patients provided by the CHDmap can achieve better results than the clinician average. When 3 clinicians were allowed to use the results of CHDmap (KNN+LR) as a reference to give predictions again, all 3 clinicians achieved a substantial improvement in their prediction ability. The averages of accuracy, recall, and F_1-score in the first task improved by 0.026, 0.017, and 0.061, respectively. The averages of accuracy, recall, and F_1-score in the second task improved by 0.119, 0.021, and 0.028, respectively. One of the enhanced clinicians also surpassed the KNN+LR CHDmap.

It is important to note that the evaluation is performed with population-optimized parameters, whereas in practice, clinicians can adjust the relevant parameters such as *k* or similarity threshold for each case in a personalized manner, which theoretically leads to better results. The use of the obtained similar patient population to construct modern deep learning models for prediction can further improve the performance of each prediction task. Especially important is that the experience and cognitive ability of the clinical expert combined with CHDmap can further enhance the accuracy of the prediction.

Discussion

Principal Findings

Medicine remains both an art and a science, which are congruent to the extent that the individual patient resembles the average subject in randomized controlled trials. Although the evidence-based medicine approach proposes personalized care, it still fails to address the physician's most important question—"How to treat the unique patient in front of me?"—in many real clinical scenarios where the complexity of the situation makes none of the available evidence applicable [45]. The proposal of MBE represents a fundamental change in clinical decision-making [5,6]. Although how to construct an MBE clinical decision support tool still faces many challenges, the CHDmap seems to be a very promising first step in realizing what has been coined MBE.

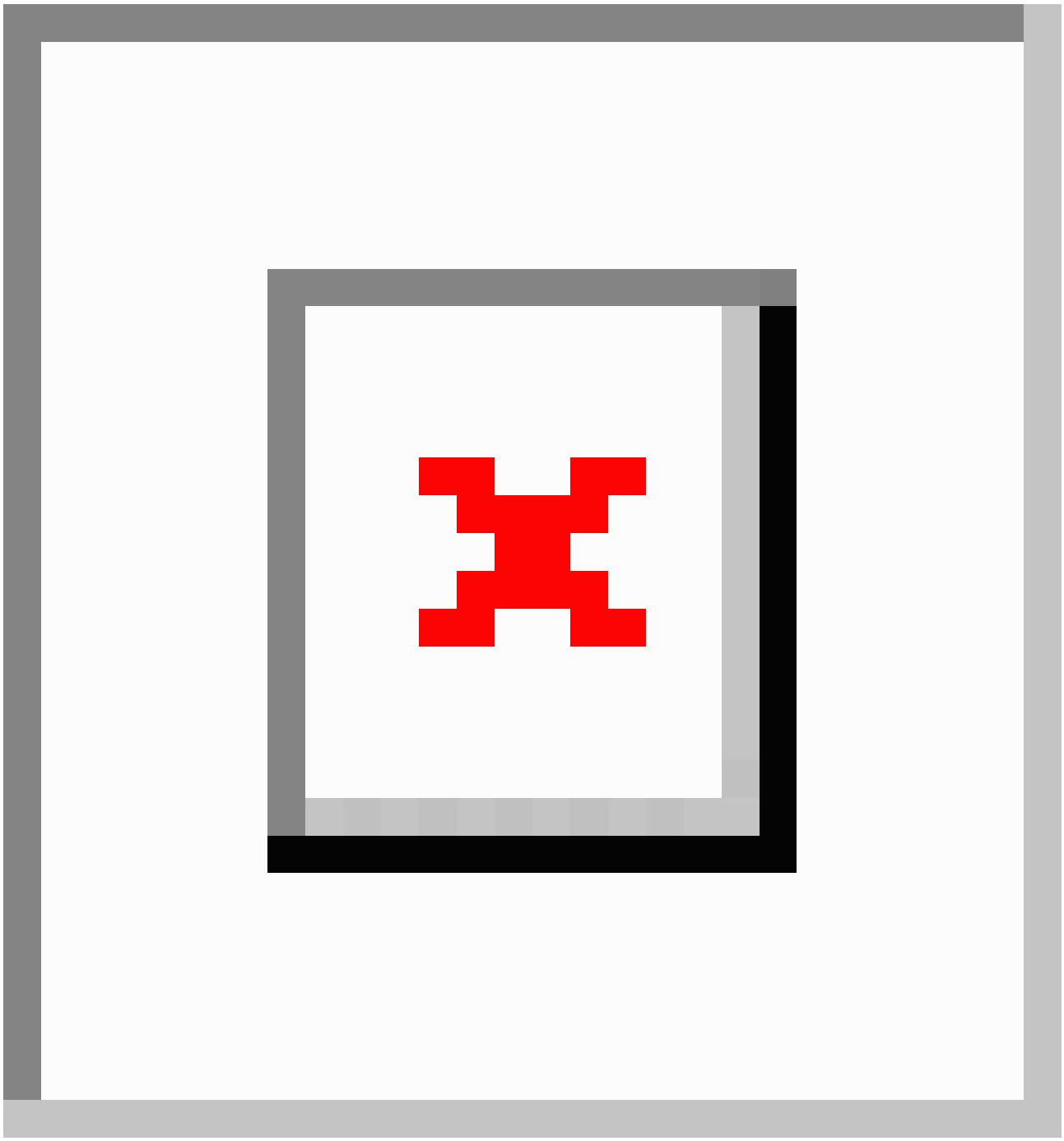
AI is poised to reshape health care. Many AI applications, especially modern deep learning models, have been developed in recent years to improve clinical prediction abilities. In addition to supervised and unsupervised machine learning, PSNs, another form of data-driven AI, have shown many unique properties in the clinical field, especially in complex clinical settings such as surgery for CHD. Moreover, their potential to construct a "library of clinical experience" will gradually be recognized, discovered, and used in the context of the continuous accumulation of medical big data.

In many other popular AI paradigms, such as supervised or unsupervised machine learning, models are usually trained toward a specific task, and thus, the models are only capable of performing that single task. This, coupled with the black-box nature of many machine learning models, especially deep learning models, makes it difficult to widely apply these techniques in practice. In contrast, patient similarity analysis exhibits many natural advantages. First, PSNs usually do not serve a single task; all characteristics exhibited by the patient similarity group, such as disease risk, various prognostic outcomes, and cost of care, can be used as MBE for decision support. Second, instead of a model that simply gives black-box predictions, CHDmap allows users to see how the patient similarity group is segmented and bounded across the patient population and then adjust the size of the patient similarity group or set custom quantitative thresholds based on their knowledge and experience. On CHDmap, the results after parameter adjustments during user manipulation are reflected in the visualized map in real time, and the statistical characteristics of multiple predictors that distinguish the current patient's similar group from other patients are also highlighted by the color of the title of the outcome view. The process of continuously adjusting and optimizing parameters through visualized feedback combines the computational advantages of computers and the advanced cognitive abilities of the human brain and truly puts the clinician, who is responsible for the decision, in control of the decision-making. Third, many machine learning models tend to require that the test and training data have consistent statistical distribution characteristics, but as shown in this evaluation, similarity analyses are still very compatible with test data with different characteristics. Finally, this PSN framework does not exclude any type of machine learning models, and all models constructed based on similar patient populations are expected to be more adaptable to individualized decision-making needs than models trained on heterogeneous populations.

Because the goal of patient similarity analysis is to be able to mimic clinical analogy reasoning, the major challenge is constructing computational patient similarity measurements that are consistent with sophisticated clinical reasoning. This is especially true when faced with complex scenarios containing a large number of dynamic features with different dimensions. Some deep learning models have been introduced to address this challenge [46-49], but they do not exhibit the interpretability and tractability of PSNs. Another way to address this challenge is to open up the computational process to clinicians, allowing them to determine and adjust the weights of different dimensions and thresholds for the similarity group themselves, thus better simulating their clinical reasoning process, as shown in Figure 5. We believe that clinical users will be able to learn how to better optimize these parameters as they continue to gain experience and understanding of this "large history data set" in the process of using CHDmap. Using a data-driven approach on how to customize the parameters of PSNs to be able to self-optimize and adapt to different tasks is also a good research direction for the future. In this study, CHDmap serves as a personalized decision aid for clinicians, using the computer's power in data storage and processing while giving clinicians more control over the decision-making process. We believe

CHDmap can perform better with the full involvement of clinicians.

Figure 5. Collaborative decision-making based on the congenital heart disease patient similarity network (PSN). The right half shows the storage and computational capacity of the PSN for a large number of cases; the left half shows the role of the clinical user who, by receiving a variety of feedback and his or her own experience, can autonomously adjust the parameters of the similarity group and reconstruct the similarity network so that the strengths of both can be used to make collaborative decisions. ASD: atrial septal defect; PDA: patent ductus arteriosus; PFO: patent foramen ovale; VSD: ventricular septal defect.



CHDmap can be used in several scenarios: for the intensivists in cardiac ICUs, CHDmap can be used to predict postoperative complications after cardiac surgery, as evaluated in this paper; for surgeons, CHDmap can also be used to assess the prognosis of surgical procedures; and for departmental managers, CHDmap can be used to assess the lengths of stay and costs. By far, CHDmap is still in the early stages of a research project. Transforming this tool into routine care is dependent on the availability of funding and the willingness of users to change their existing working patterns. The publication of this paper will also facilitate the advancement of our subsequent translational work.

It is important to note that associations between treatments and outcomes obtained by observation in similar patient populations may not be causal. The real causal effects often rely on a matching process to control for the bias introduced by the treatment itself in the selection of patients [50]. An initial demo feature is available on CHDmap to estimate treatment outcome

effects based on matched patient groups. CHDmap can match 1 or *k* patients for each patient receiving the treatment using a PSN and then allow for a more visual and unbiased assessment of treatment outcomes by showing the difference in prognosis between these 2 groups of patients. It is important to note that this causal assessment assumes that there are no other factors outside the variables covered by the patient's similarity analysis that may influence treatment choice or prognosis. Thus, the reliability of this real world–generated evidence usually relies on clinical experts to judge it as well. In future versions, we hope to incorporate more modern frameworks for causal inference (such as DoWhy [51]) to automatically quantitatively assess causal effects as well as their reliability.

There are several limitations to this study. First, limited clinical features were used to measure the similarity of patients with CHD. In addition to the information presented by the echocardiography, there is a wealth of other clinical information that can be used to assess the patient's status. Second, the use of NLP to automatically extract measurement information can also be subject to errors or mismatches, and although manual quality control is carried out, it is still not possible to ensure that all of the measurements are 100% accurate. Third, just as clinicians gain clinical experience by continuously treating different patients, PSNs need to expand their ability to dynamically accumulate cases. A PSN with a web-based automatic update mechanism will be the next key research step. Fourth, data from only a single center were used to evaluate this tool, and the introduction of data from multiple centers during PSN construction may pose unknown risks that require attention in future studies. Finally, different clinicians may have different decision-making philosophies, and different weights can be assigned to different indicators for different tasks. CHDmap offers only a limited number of customizations that may be difficult to adapt to all scenarios. A way to attribute weights to each of the indicators and dimensions by AI for specific tasks may potentially improve the performance of CHDmap in the future.

Conclusions

A clinician-operable PSN for CHD was proposed and developed to help clinicians make decisions based on thousands of previous surgery cases. Without individual optimization, CHDmap can obtain competitive performance compared to clinical experts. Statistical analysis of data based on patient similarity groups is intuitive and clear to clinicians, whereas the operable, visual user interface puts clinicians in real control of decision-making. Clinicians supported by CHDmap can make better decisions than both pure experience-based decisions and AI model output results. Such a PSN-based framework can become a routine method of CHD case management and use. The MBE can be embraced in clinical practice, and its full potential can be realized.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81871456).

Authors' Contributions

HL, SS, and QS contributed equally to the paper as cocorresponding authors. SS can be contacted at Sicu1@zju.edu.cn, and QS can be contacted at shuqiang@zju.edu.cn.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental methods, definitions of postoperative complications, features used to measure patient similarity, and echocardiographic indicators used in different calculations.
[DOCX File, 1306 KB - medinform_v12i1e49138_app1.docx]

Multimedia Appendix 2
Video introduction for CHDmap.
[MP4 File, 102353 KB - medinform_v12i1e49138_app2.mp4]

References

1. Bernier PL, Stefanescu A, Samoukovic G, Tchervenkov CI. The challenge of congenital heart disease worldwide: epidemiologic and demographic facts. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu 2010;13(1):26-34. [doi: 10.1053/j.pcsu.2010.02.005] [Medline: 20307858]
2. van der Linde D, Konings EEM, Slager MA, et al. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol 2011 Nov 15;58(21):2241-2247. [doi: 10.1016/j.jacc.2011.08.025] [Medline: 22078432]

3. Jacobs JP, Mayer JEJ, Mavroudis C, et al. The Society of Thoracic Surgeons Congenital Heart Surgery Database: 2016 update on outcomes and quality. Ann Thorac Surg 2016 Mar;101(3):850-862. [doi: 10.1016/j.athoracsur.2016.01.057] [Medline: 26897186]
4. Triedman JK, Newburger JW. Trends in congenital heart disease. Circulation 2016 Jun 21;133(25):2716-2733. [doi: 10.1161/CIRCULATIONAHA.116.023544] [Medline: 27324366]
5. Horwitz RI, Hayes-Conroy A, Caricchio R, Singer BH. From evidence based medicine to medicine based evidence. Am J Med 2017 Nov;130(11):1246-1250. [doi: 10.1016/j.amjmed.2017.06.012] [Medline: 28711551]
6. van den Eynde J, Manlhiot C, van de Bruaene A, et al. Medicine-based evidence in congenital heart disease: how artificial intelligence can guide treatment decisions for individual patients. Front Cardiovasc Med 2021 Dec;8:798215. [doi: 10.3389/fcvm.2021.798215] [Medline: 34926630]
7. Benavidez OJ, Gauvreau K, del Nido P, Bacha E, Jenkins KJ. Complications and risk factors for mortality during congenital heart surgery admissions. Ann Thorac Surg 2007 Jul;84(1):147-155. [doi: 10.1016/j.athoracsur.2007.02.048] [Medline: 17588402]
8. Pasquali SK, He X, Jacobs JP, Jacobs ML, O'Brien SM, Gaynor JW. Evaluation of failure to rescue as a quality metric in pediatric heart surgery: an analysis of the STS Congenital Heart Surgery Database. Ann Thorac Surg 2012 Aug;94(2):573-580. [doi: 10.1016/j.athoracsur.2012.03.065] [Medline: 22633496]
9. Kansy A, Tobota Z, Maruszewski P, Maruszewski B. Analysis of 14,843 neonatal congenital heart surgical procedures in the European Association for Cardiothoracic Surgery Congenital Database. Ann Thorac Surg 2010 Apr;89(4):1255-1259. [doi: 10.1016/j.athoracsur.2010.01.003] [Medline: 20338347]
10. Jenkins KJ, Gauvreau K, Newburger JW, Spray TL, Moller JH, Iezzoni LI. Consensus-based method for risk adjustment for surgery for congenital heart disease. J Thorac Cardiovasc Surg 2002 Jan;123(1):110-118. [doi: 10.1067/mtc.2002.119064] [Medline: 11782764]
11. Lacour-Gayet F, Clarke D, Jacobs J, et al. The Aristotle score: a complexity-adjusted method to evaluate surgical results1. Eur J Cardiothorac Surg 2004 Jun;25(6):911-924. [doi: 10.1016/j.ejcts.2004.03.027] [Medline: 15144988]
12. O'Brien SM, Clarke DR, Jacobs JP, et al. An empirically based tool for analyzing mortality associated with congenital heart surgery. J Thorac Cardiovasc Surg 2009 Nov;138(5):1139-1153. [doi: 10.1016/j.jtcvs.2009.03.071] [Medline: 19837218]
13. Jacobs ML, O'Brien SM, Jacobs JP, et al. An empirically based tool for analyzing morbidity associated with operations for congenital heart disease. J Thorac Cardiovasc Surg 2013 Apr;145(4):1046-1057.E1. [doi: 10.1016/j.jtcvs.2012.06.029] [Medline: 22835225]
14. Kalfa D, Krishnamurthy G, Duchon J, et al. Outcomes of cardiac surgery in patients weighing <2.5 kg: affect of patient-dependent and -independent variables. J Thorac Cardiovasc Surg 2014 Dec;148(6):2499-2506.E1. [doi: 10.1016/j.jtcvs.2014.07.031] [Medline: 25156464]
15. Agarwal HS, Wolfram KB, Saville BR, Donahue BS, Bichell DP. Postoperative complications and association with outcomes in pediatric cardiac surgery. J Thorac Cardiovasc Surg 2014 Aug;148(2):609-616.E1. [doi: 10.1016/j.jtcvs.2013.10.031] [Medline: 24280709]
16. Zeng X, An J, Lin R, et al. Prediction of complications after paediatric cardiac surgery. Eur J Cardiothorac Surg 2020 Feb 1;57(2):350-358. [doi: 10.1093/ejcts/ezz198] [Medline: 31280308]
17. Zeng X, Hu Y, Shu L, et al. Explainable machine-learning predictions for complications after pediatric congenital heart surgery. Sci Rep 2021 Aug 26;11(1):17244. [doi: 10.1038/s41598-021-96721-w] [Medline: 34446783]
18. Zeng X, Shi S, Sun Y, et al. A time-aware attention model for prediction of acute kidney injury after pediatric cardiac surgery. J Am Med Inform Assoc 2022 Dec 13;30(1):94-102. [doi: 10.1093/jamia/ocac202] [Medline: 36287639]
19. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
20. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
21. Hu Y, Gong X, Shu L, et al. Understanding risk factors for postoperative mortality in neonates based on explainable machine learning technology. J Pediatr Surg 2021 Dec;56(12):2165-2171. [doi: 10.1016/j.jpedsurg.2021.03.057] [Medline: 33863558]
22. Pai S, Bader GD. Patient similarity networks for precision medicine. J Mol Biol 2018 Sep 14;430(18 Pt A):2924-2938. [doi: 10.1016/j.jmb.2018.05.037] [Medline: 29860027]
23. Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: a systematic review. J Biomed Inform 2018 Jul;83:87-96. [doi: 10.1016/j.jbi.2018.06.001] [Medline: 29864490]
24. Zeng X, Jia Z, He Z, et al. Measure clinical drug–drug similarity using electronic medical records. Int J Med Inform 2019 Apr;124:97-103. [doi: 10.1016/j.ijmedinf.2019.02.003] [Medline: 30784433]
25. Jia Z, Lu X, Duan H, Li H. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. BMC Med Inform Decis Mak 2019 Apr 25;19(1):91. [doi: 10.1186/s12911-019-0807-y] [Medline: 31023325]
26. Cheng F, Liu D, Du F, et al. VBridge: connecting the dots between features and data to explain healthcare models. IEEE Trans Vis Comput Graph 2022 Jan;28(1):378-388. [doi: 10.1109/TVCG.2021.3114836] [Medline: 34596543]

27. Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol 2019 Mar 14;15(3):e8497. [doi: 10.15252/msb.20188497] [Medline: 30872331]
28. Yang J, Dong C, Duan H, Shu Q, Li H. RDmap: a map for exploring rare diseases. Orphanet J Rare Dis 2021 Feb 25;16(1):101. [doi: 10.1186/s13023-021-01741-4] [Medline: 33632281]
29. Zhang G, Peng Z, Yan C, Wang J, Luo J, Luo H. A novel liver cancer diagnosis method based on patient similarity network and DenseGCN. Sci Rep 2022 Apr 26;12(1):6797. [doi: 10.1038/s41598-022-10441-3] [Medline: 35474072]
30. Jia Z, Zeng X, Duan H, Lu X, Li H. A patient-similarity-based model for diagnostic prediction. Int J Med Inform 2020 Mar;135:104073. [doi: 10.1016/j.ijmedinf.2019.104073] [Medline: 31923816]
31. Li L, Cheng WY, Glicksberg BS, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med 2015 Oct 28;7(311):311ra174. [doi: 10.1126/scitranslmed.aaa9364] [Medline: 26511511]
32. Tokodi M, Shrestha S, Bianco C, et al. Interpatient similarities in cardiac function: a platform for personalized cardiovascular medicine. JACC Cardiovasc Imaging 2020 May;13(5):1119-1132. [doi: 10.1016/j.jcmg.2019.12.018] [Medline: 32199835]
33. Wang N, Wang M, Zhou Y, et al. Sequential data-based patient similarity framework for patient outcome prediction: algorithm development. J Med Internet Res 2022 Jan 6;24(1):e30720. [doi: 10.2196/30720] [Medline: 34989682]
34. Wu J, Dong Y, Gao Z, Gong T, Li C. Dual attention and patient similarity network for drug recommendation. Bioinformatics 2023 Jan 1;39(1):btad003. [doi: 10.1093/bioinformatics/btad003] [Medline: 36617159]
35. Tan WY, Gao Q, Oei RW, Hsu W, Lee ML, Tan NC. Diabetes medication recommendation system using patient similarity analytics. Sci Rep 2022 Dec 3;12(1):20910. [doi: 10.1038/s41598-022-24494-x] [Medline: 36463296]
36. Chen X, Faviez C, Vincent M, et al. Patient-patient similarity-based screening of a clinical data warehouse to support ciliopathy diagnosis. Front Pharmacol 2022 Mar 25;13:786710. [doi: 10.3389/fphar.2022.786710] [Medline: 35401179]
37. Gentner D. Structure‐mapping: a theoretical framework for analogy. Cognitive Science 1983;7(2):155-170 [FREE Full text]
38. Shi Y, Li Z, Jia Z, et al. Automatic knowledge extraction and data mining from echo reports of pediatric heart disease: application on clinical decision support. In: Sun M, Liu Z, Zhang M, Liu Y, editors. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. CCL 2015, NLP-NABD 2015. Lecture Notes in Computer Science, vol 9427: Springer; 2015:417-424. [doi: 10.1007/978-3-319-25816-4_34]
39. Lopez L, Colan S, Stylianou M, et al. Relationship of echocardiographic z scores adjusted for body surface area to age, sex, race, and ethnicity: the Pediatric Heart Network Normal Echocardiogram Database. Circ Cardiovasc Imaging 2017 Nov;10(11):e006979. [doi: 10.1161/CIRCIMAGING.117.006979] [Medline: 29138232]
40. Zhou M, Yu J, Duan H, et al. Study on the correlation between preoperative echocardiography indicators and postoperative prognosis in children with ventricular septal defect. Article in Chinese. Chinese J Ultrason 2022 Sep 25;31(9):767-773. [doi: 10.3760/cma.j.cn131148-20220127-00076]
41. Download. CHDmap. URL: http://chdmap.nbscn.org/Help#download [accessed 2024-01-04]
42. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(86):2579-2605 [FREE Full text]
43. Li D, Mei H, Shen Y, et al. ECharts: a declarative framework for rapid construction of web-based visualization. Vis Inform 2018 Jun;2(2):136-146. [doi: 10.1016/j.visinf.2018.04.011]
44. CHDmap. URL: http://chdmap.nbscn.org/ [accessed 2024-01-04]
45. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ Digit Med 2020 May 11;3:67. [doi: 10.1038/s41746-020-0277-8] [Medline: 32411828]
46. Suo Q, Ma F, Yuan Y, et al. Deep patient similarity learning for personalized healthcare. IEEE Trans Nanobioscience 2018 Jul;17(3):219-227. [doi: 10.1109/TNB.2018.2837622] [Medline: 29994534]
47. Gu Y, Yang X, Tian L, et al. Structure-aware Siamese graph neural networks for encounter-level patient similarity learning. J Biomed Inform 2022 Mar;127:104027. [doi: 10.1016/j.jbi.2022.104027] [Medline: 35181493]
48. Sun Z, Lu X, Duan H, Li H. Deep dynamic patient similarity analysis: model development and validation in ICU. Comput Methods Programs Biomed 2022 Oct;225:107033. [doi: 10.1016/j.cmpb.2022.107033] [Medline: 35905698]
49. Navaz AN, El-Kassabi HT, Serhani MA, Oulhaj A, Khalil K. A novel patient similarity network (PSN) framework based on multi-model deep learning for precision medicine. J Pers Med 2022 May 10;12(5):768. [doi: 10.3390/jpm12050768] [Medline: 35629190]
50. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010 Feb 1;25(1):1-21. [doi: 10.1214/09-STS313] [Medline: 20871802]
51. Sharma A, Syrgkanis V, Zhang C, Kıcıman E. DoWhy: addressing challenges in expressing and validating causal assumptions. arXiv. Preprint posted online on Aug 27, 2021. [doi: 10.48550/arXiv.2108.13518]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CHD: congenital heart disease

EHR: electronic health record
FN: false negative
FP: false positive
ICU: intensive care unit
KNN: *k*-nearest neighbor
LR: logistic regression
MBE: medicine-based evidence
NLP: natural language processing
PSN: patient similarity network
RACHS-1: Risk Adjustment for Congenital Heart Surgery 1
STS-EACTS: Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery
TN: true negative
TP: true positive

Edited by C Lovis; submitted 24.05.23; peer-reviewed by JVD Eynde, Y Kim; revised version received 21.08.23; accepted 16.11.23; published 19.01.24.

Please cite as:
Li H, Zhou M, Sun Y, Yang J, Zeng X, Qiu Y, Xia Y, Zheng Z, Yu J, Feng Y, Shi Z, Huang T, Tan L, Lin R, Li J, Fan X, Ye J, Duan H, Shi S, Shu Q
A Patient Similarity Network (CHDmap) to Predict Outcomes After Congenital Heart Surgery: Development and Validation Study
JMIR Med Inform 2024;12:e49138
URL: https://medinform.jmir.org/2024/1/e49138
doi:10.2196/49138

Original Paper

Prediction of Antibiotic Resistance in Patients With a Urinary Tract Infection: Algorithm Development and Validation

Nevruz İlhanlı[1,2*], MSc; Se Yoon Park[1,3,4*], MD, PhD; Jaewoong Kim[1,3], MSc; Jee An Ryu[1], BA; Ahmet Yardımcı[2], PhD; Dukyong Yoon[1,4,5], MD, PhD

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Yongin, Republic of Korea

[2]Department of Biostatistics and Medical Informatics, Akdeniz University, Antalya, Turkey

[3]Department of Hospital Medicine, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Republic of Korea

[4]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Republic of Korea

[5]Institute for Innovation in Digital Healthcare, Severance Hospital, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Dukyong Yoon, MD, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
363, Dongbaekjukjeon-daero
Yongin, 16995
Republic of Korea
Phone: 82 3151898450
Email: dukyong.yoon@yonsei.ac.kr

Abstract

Background: The early prediction of antibiotic resistance in patients with a urinary tract infection (UTI) is important to guide appropriate antibiotic therapy selection.

Objective: In this study, we aimed to predict antibiotic resistance in patients with a UTI. Additionally, we aimed to interpret the machine learning models we developed.

Methods: The electronic medical records of patients who were admitted to Yongin Severance Hospital, South Korea were used. A total of 71 features extracted from patients' admission, diagnosis, prescription, and microbiology records were used for classification. UTI pathogens were classified as either sensitive or resistant to cephalosporin, piperacillin-tazobactam (TZP), carbapenem, trimethoprim-sulfamethoxazole (TMP-SMX), and fluoroquinolone. To analyze how each variable contributed to the machine learning model's predictions of antibiotic resistance, we used the Shapley Additive Explanations method. Finally, a prototype machine learning–based clinical decision support system was proposed to provide clinicians the resistance probabilities for each antibiotic.

Results: The data set included 3535, 737, 708, 1582, and 1365 samples for cephalosporin, TZP, TMP-SMX, fluoroquinolone, and carbapenem resistance prediction models, respectively. The area under the receiver operating characteristic curve values of the random forest models were 0.777 (95% CI 0.775-0.779), 0.864 (95% CI 0.862-0.867), 0.877 (95% CI 0.874-0.880), 0.881 (95% CI 0.879-0.882), and 0.884 (95% CI 0.884-0.885) in the training set and 0.638 (95% CI 0.635-0.642), 0.630 (95% CI 0.626-0.634), 0.665 (95% CI 0.659-0.671), 0.670 (95% CI 0.666-0.673), and 0.721 (95% CI 0.718-0.724) in the test set for predicting resistance to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone, respectively. The number of previous visits, first culture after admission, chronic lower respiratory diseases, administration of drugs before infection, and exposure time to these drugs were found to be important variables for predicting antibiotic resistance.

Conclusions: The study results demonstrated the potential of machine learning to predict antibiotic resistance in patients with a UTI. Machine learning can assist clinicians in making decisions regarding the selection of appropriate antibiotic therapy in patients with a UTI.

(JMIR Med Inform 2024;12:e51326) doi:10.2196/51326

KEYWORDS

antibiotic resistance; machine learning; urinary tract infections; UTI; decision support

Introduction

Urinary tract infection (UTI) refers to an infection that occurs in any part of the urinary system, including the kidneys, ureters, urinary bladder, urethra, and other auxiliary structures [1,2]. Globally, UTIs are the most prevalent type of infectious disease, with around 150-250 million cases occurring each year [3]. Considerable morbidity and mortality result from these infections [4]. Typically, the most effective treatment for UTIs is the administration of antibiotics [3]. However, inappropriate use of antibiotics can permanently affect the normal microbiota of the urinary tract system and lead to antibiotic resistance [5].

The antibiotic susceptibility test is commonly used to identify antibiotic resistance, but it takes 24-48 hours to obtain test results [6,7]. However, in the clinical workflow, clinicians need to identify antibiotic resistance quickly to provide effective treatment for patients with UTIs. For this reason, early prediction of antibiotic resistance in patients with UTIs is important to guide the selection of appropriate antibiotic therapy. Machine learning can be used to develop prediction models and clinical decision support systems (CDSSs) to identify antibiotic resistance and support the selection of appropriate antibiotic therapy for patients with a UTI.

Several efforts have been made to predict antibiotic resistance in patients with UTIs using data from patients' electronic medical records (EMRs), including demographics, prescriptions, comorbidities, procedures, and laboratory tests. These investigations have yielded promising results. Some of these studies were limited to specific patient groups, including patients with uncomplicated UTIs [8] and patients treated in the emergency department [9]. In other studies, researchers worked with heterogeneous data that were not limited to a specific patient group [10-12]. However, prior studies that analyzed heterogeneous data did not address the interpretation of machine learning models. The black-box nature of machine learning is a limiting factor not only in its use for antibiotic resistance prediction but also in its wider clinical use [13,14]. Thus, interpreting the results obtained by the machine learning model is crucial in increasing users' trust in the machine learning model [15,16]. Furthermore, these studies did not address the development of the CDSS with the prediction models they built.

In this study, we aimed to predict antibiotic resistance in patients with a UTI. Heterogeneous data that were not limited to a specific patient group were used. UTI pathogens were classified as either sensitive or resistant to 5 commonly used antibiotics in UTI treatment: cephalosporin, piperacillin-tazobactam (TZP), carbapenem, trimethoprim-sulfamethoxazole (TMP-SMX), and fluoroquinolone. In addition, our objective was to understand and explain the inner workings of the machine learning models we developed. Eventually, a prototype CDSS was developed to provide clinicians the resistance probabilities for each antibiotic.

Methods

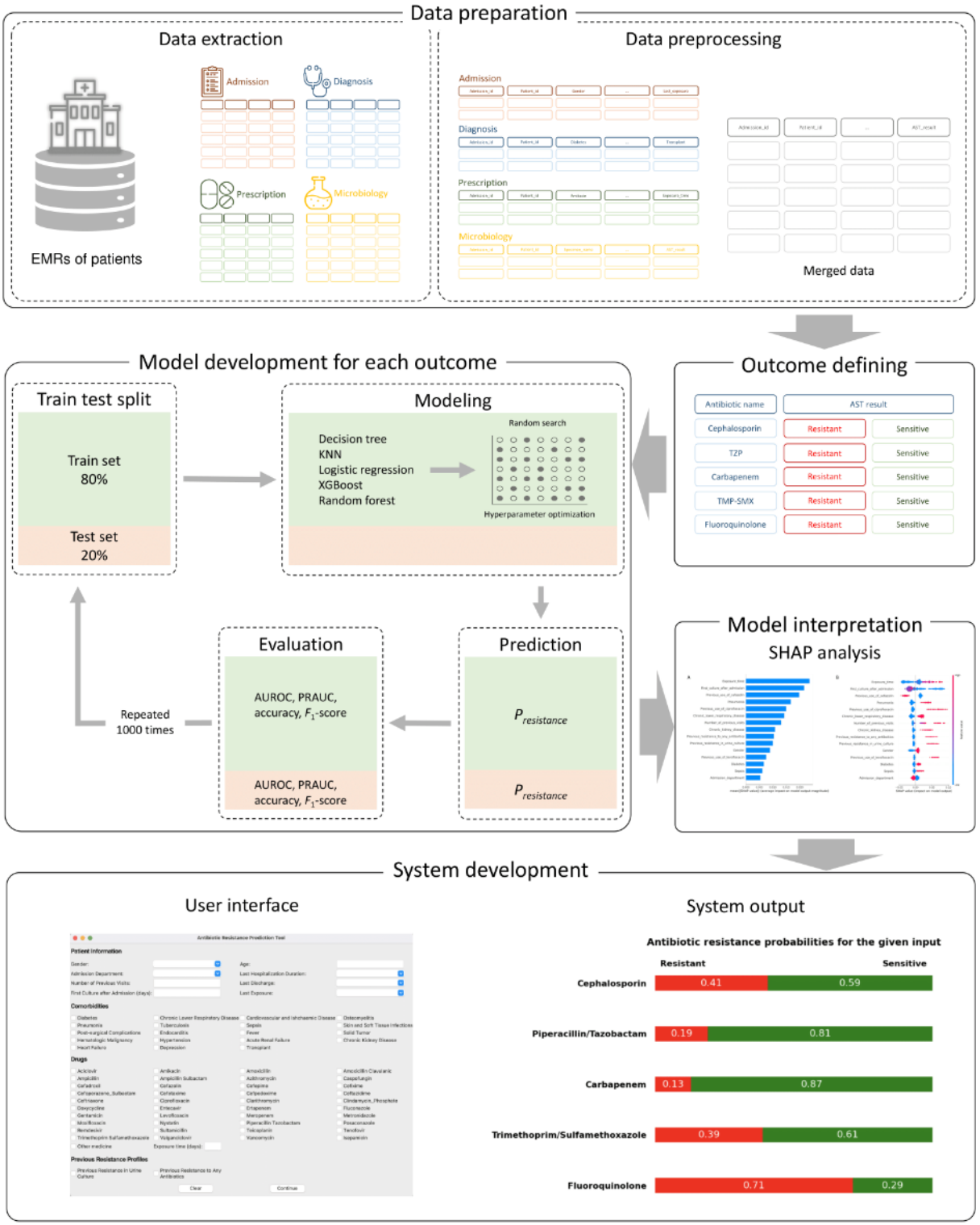
Ethical Considerations

Ethics approval for the study was obtained from the institutional review board of Yonsei University Severance Hospital on June 6, 2022 (approval 9-2023-0095). The informed consent was not required due to the retrospective nature of the study.

Data Set Description and Study Design

In this study, we used the EMRs of patients who were admitted to Yongin Severance Hospital, South Korea, between October 2012 and October 2022. To build the prediction models, admission, diagnosis, prescription, and microbiology records were extracted. The summary of the research process is presented in Figure 1.

Figure 1. Summary of the research process. AST: antibiotic susceptibility test; AUROC: area under the receiver operating characteristic curve; EMR: electronic medical record; KNN: k-nearest neighbor; PRAUC: precision-recall area under the curve; SHAP: Shapley Additive Explanations; TMP-SMX: trimethoprim-sulfamethoxazole; TZP: piperacillin-tazobactam; XGBoost: Extreme Gradient Boosting.



Data Preprocessing

The microbiology table contained 143,114 urine cultures collected from 6011 patients during 7719 admissions. Since positive samples typically indicate the presence of bacteriuria, and urine culture samples were typically collected from patients with UTI symptoms, we considered these to be indicative of a UTI [10]. The resistance profiles were evaluated based on the Clinical and Laboratory Standards Institute guidelines, where intermediate-level resistance was considered sensitive. To assess the resistance of UTI pathogens to antibiotic classes, antibiotics were grouped as cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone. The antibiotics included in each antibiotic class are presented in Multimedia Appendix 1. The patients' demographic information was extracted from the admission

table, their comorbidities were extracted from the diagnosis table, and their drug use information was extracted from the prescription table. For all input variables, the time of the first culture test was considered as the end point, and only data collected before the first culture test were used. After preprocessing and variable extraction from the raw data, the tables were combined using the admission number as the primary key. Missing data were excluded from the study. Patients aged 19 years and older and 100 years and younger at admission were included in the study, and numerical variables were standardized. A total of 71 features were used to classify UTI pathogens as either sensitive or resistant to each antibiotic. The predictors for the prediction models were selected by considering related works and using clinical judgment. Additionally, the threshold values for binarization were selected according to the literature [17] and the expert assessment of a specialist in infectious diseases. Detailed information about the predictors can be found in Multimedia Appendix 2.

Machine Learning Model Development

We used a repeated train test split approach for modeling. The data sets were split into training and test sets using an 80:20 ratio, and the training sets were used for the development of the machine learning models. When splitting the data into training and test sets, data points from the same patient and admission were exclusively included in either the training or test data set to prevent potential data leakage and ensure the models were evaluated on previously unseen data. At each iteration, we created different training and test data sets by changing the random seed. Decision tree, k-nearest neighbor, logistic regression, Extreme Gradient Boosting, and random forest were used for modeling. The hyperparameters of the machine learning models were optimized by using the random search hyperparameter optimization method with 10-fold cross-validation on the training data set. We stored the performance of the prediction models at each iteration, and the mean of performance metrics was calculated. The procedure of splitting the data, optimizing hyperparameters, modeling, and evaluation was iteratively repeated 1000 times to classify UTI pathogens as either sensitive or resistant to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone. The machine learning models were built using Python (version 3.10.4; Python Software Foundation).

Machine Learning Model Interpretation

To analyze the contribution of the variables to the machine learning models in predicting antibiotic resistance, we used the Shapley Additive Explanations (SHAP) method. The SHAP values of the random forest models that showed superior performance compared to other machine learning methods were evaluated. The random forest model with the highest area under the receiver operating characteristic curve (AUROC) on the test set across all iterations for each antibiotic was used for SHAP analysis. Python (version 3.10.4; Python Software Foundation) was used for SHAP analysis.

CDSS Development

To develop the CDSS prototype, the random forest model with the highest AUROC on the test set across all iterations for each antibiotic was used. The CDSS prototype was developed using the *tkinter* package in Python (version 3.10.4; Python Software Foundation).

Evaluation

The performance of the machine learning model for predicting antibiotic resistance was evaluated on the training and test sets using the AUROC with 95% CIs, precision-recall area under the curve (PRAUC), accuracy, and F_1-score performance metrics. Herein, the AUROC value was considered the main evaluation metric. The definitions of the performance metrics we used are provided below.

- AUROC: The AUROC is a widely used metric that represents a classifier's ability to discriminate between positive instances and negative instances [18].
- PRAUC: PRAUC refers to the area under the precision-recall curve that plots precision as a function of recall for all the possible decision thresholds [19].
- Accuracy: Accuracy is the ratio of correctly classified samples to all samples.

- F_1-score: F_1-score is the harmonic mean of precision and recall metrics.

Python (version 3.10.4; Python Software Foundation) was used to evaluate the prediction models.

Results

Data Set Characteristics

The general characteristics of the data set used in this study are presented in Table 1. The data set included 3535, 737, 708, 1582, and 1365 samples for cephalosporin, TZP, TMP-SMX, fluoroquinolone, and carbapenem resistance prediction models, respectively. *Escherichia coli* was the most frequently isolated bacterial specimen across all antibiotics.

Table 1. General characteristics of the data set.

	Cephalosporin	TZP[a]	TMP-SMX[b]	Fluoroquinolone	Carbapenem
Samples, n	3535	737	708	1582	1365
Admissions, n	396	366	374	571	392
Patients, n	390	360	368	557	386
Resistance, n (%)	1492 (42.2)	169 (22.9)	281 (39.7)	1014 (64.1)	142 (10.4)
Age (years), mean (SD)	71.5 (14.4)	71.4 (14.4)	71.4 (14.4)	71.9 (14.4)	71.7 (14.3)
Female, n (%)	2597 (73.5)	523 (71)	507 (71.6)	1013 (64)	994 (72.8)
Most common bacteria (*Escherichia coli*), n (%)	1650 (46.7)	312 (42.3)	331 (46.7)	349 (22)	624 (45.7)
Second-most common bacteria (*Klebsiella pneumoniae*), n (%)	556 (15.7)	109 (14.8)	111 (15.7)	305 (19.3)[c]	220 (16.1)
Third-most common bacteria (*Pseudomonas aeruginosa*), n (%)	168 (4.7)	69 (9.4)	21 (3)[d]	180 (11.4)[e]	83 (6.1)

[a]TZP: piperacillin-tazobactam.

[b]TMP-SMX: trimethoprim-sulfamethoxazole.

[c]The isolated bacterial specimen is *Enterococcus faecium*.

[d]The isolated bacterial specimen is *Citrobacter freundii*.

[e]The isolated bacterial specimen is *Enterococcus faecalis*.

Model Performance

The performance analysis of the random forest models is presented in Table 2. The AUROC values were 0.777 (95% CI 0.775-0.779), 0.864 (95% CI 0.862-0.867), 0.877 (95% CI 0.874-0.880), 0.881 (95% CI 0.879-0.882), and 0.884 (95% CI 0.884-0.885) in the training set and 0.638 (95% CI 0.635-0.642), 0.630 (95% CI 0.626-0.634), 0.665 (95% CI 0.659-0.671), 0.670 (95% CI 0.666-0.673), and 0.721 (95% CI 0.718-0.724) in the test set for predicting resistance to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone, respectively. The performance analysis of the other machine learning models is presented in Multimedia Appendices 3-6.

Table 2. Classification performances of the random forest models.

	Training set				Test set			
	AUROC[a] (95% CI)	PRAUC[b]	Accuracy	F_1-score	AUROC (95% CI)	PRAUC	Accuracy	F_1-score
Cephalosporin	0.777 (0.775-0.779)	0.725	0.715	0.676	0.638 (0.635-0.642)	0.547	0.603	0.556
TZP[c]	0.864 (0.862-0.867)	0.688	0.808	0.652	0.630 (0.626-0.634)	0.332	0.641	0.313
Carbapenem	0.877 (0.874-0.880)	0.539	0.822	0.493	0.665 (0.659-0.671)	0.222	0.725	0.220
TMP-SMX[d]	0.881 (0.879-0.882)	0.829	0.822	0.781	0.670 (0.666-0.673)	0.568	0.638	0.560
Fluoroquinolone	0.884 (0.884-0.885)	0.938	0.802	0.832	0.721 (0.718-0.724)	0.813	0.657	0.706

[a]AUROC: area under the receiver operating characteristic curve.

[b]PRAUC: precision-recall area under the curve.

[c]TZP: piperacillin-tazobactam.

[d]TMP-SMX: trimethoprim-sulfamethoxazole.

Important Features

The SHAP values of the 15 most important features in the random forest models are presented in Figure 2.

The SHAP feature importance bar plot (Figure 3A) and SHAP summary plot (Figure 3B) of the fluoroquinolone resistance prediction model are presented in Figure 3. The SHAP feature importance plot and SHAP summary plot of the other antibiotic prediction models are presented in Multimedia Appendices 7-10.

Figure 2. SHAP values of the 15 most important features in the prediction models. SHAP: Shapley Additive Explanations; TMP-SMX: trimethoprim-sulfamethoxazole; TZP: piperacillin-tazobactam.

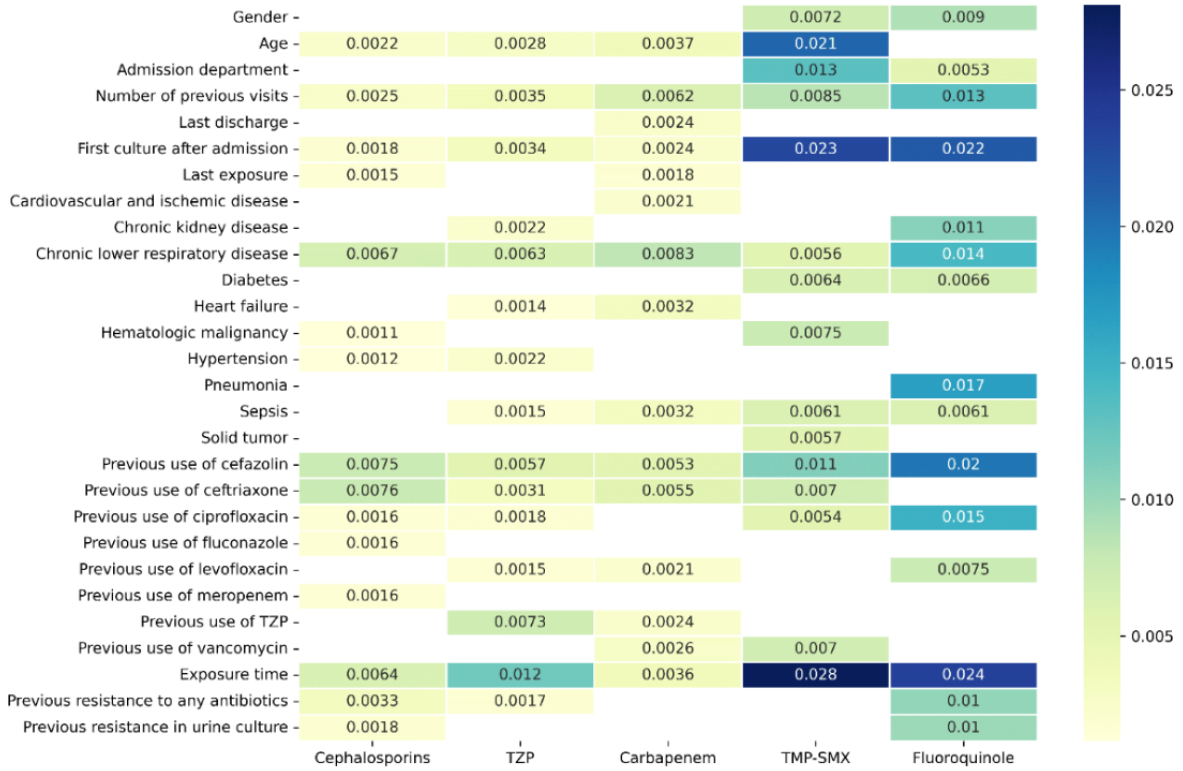
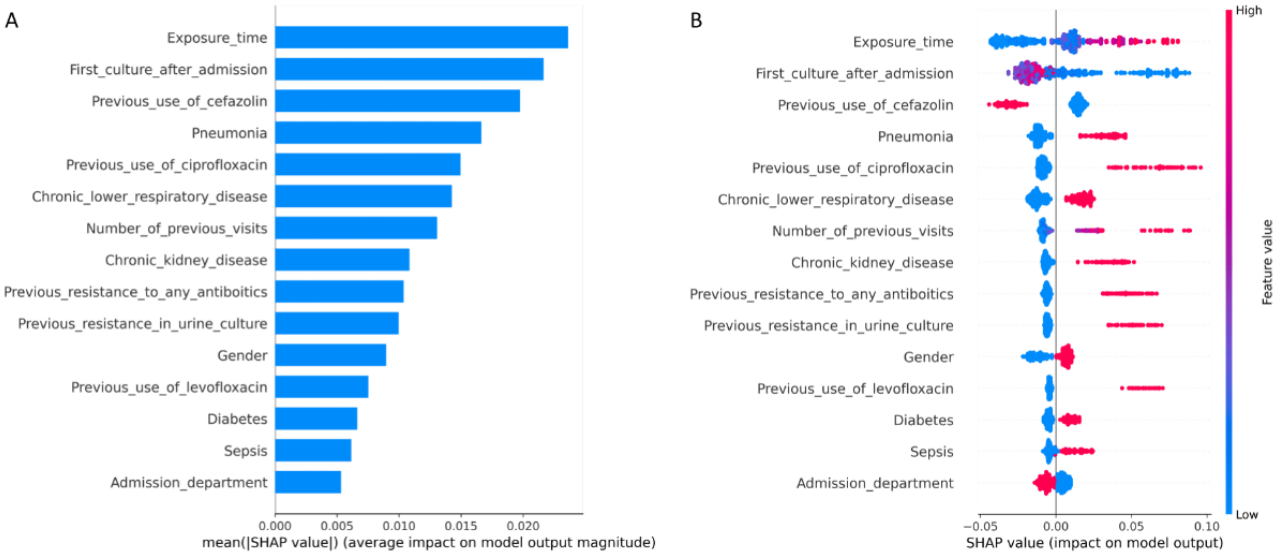


Figure 3. SHAP analysis results of fluoroquinolone resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations.



Clinical Decision Support System

The user interface of the CDSS is shown in Figure 4. The CDSS prototype obtains data from the user and produces antibiotic resistance probabilities for each antibiotic.

We presented the CDSS prototype on a scenario. In this case, a female aged 55 years was admitted to the hospital's outpatient department. The patient previously visited the hospital 3 times and was readmitted to the hospital within 30 days of her last 3-day stay. The duration between the patient's admission to the hospital and the first culture was 1 day. The patient was previously diagnosed with diabetes and chronic lower respiratory disease. Additionally, the patient had a history of cefazolin use in the last 30 days and resistance in urine culture.

The system output for the given scenario is shown in Figure 5. The system produced resistance probabilities for each antibiotic. For the given scenario, the system produced a 71% probability of fluoroquinolone resistance, a 41% probability of cephalosporin resistance, a 39% probability of TMP-SMX resistance, a 19% probability of TZP resistance, and a 13% probability of carbapenem resistance.

Figure 4. The user interface of the clinical decision support system.

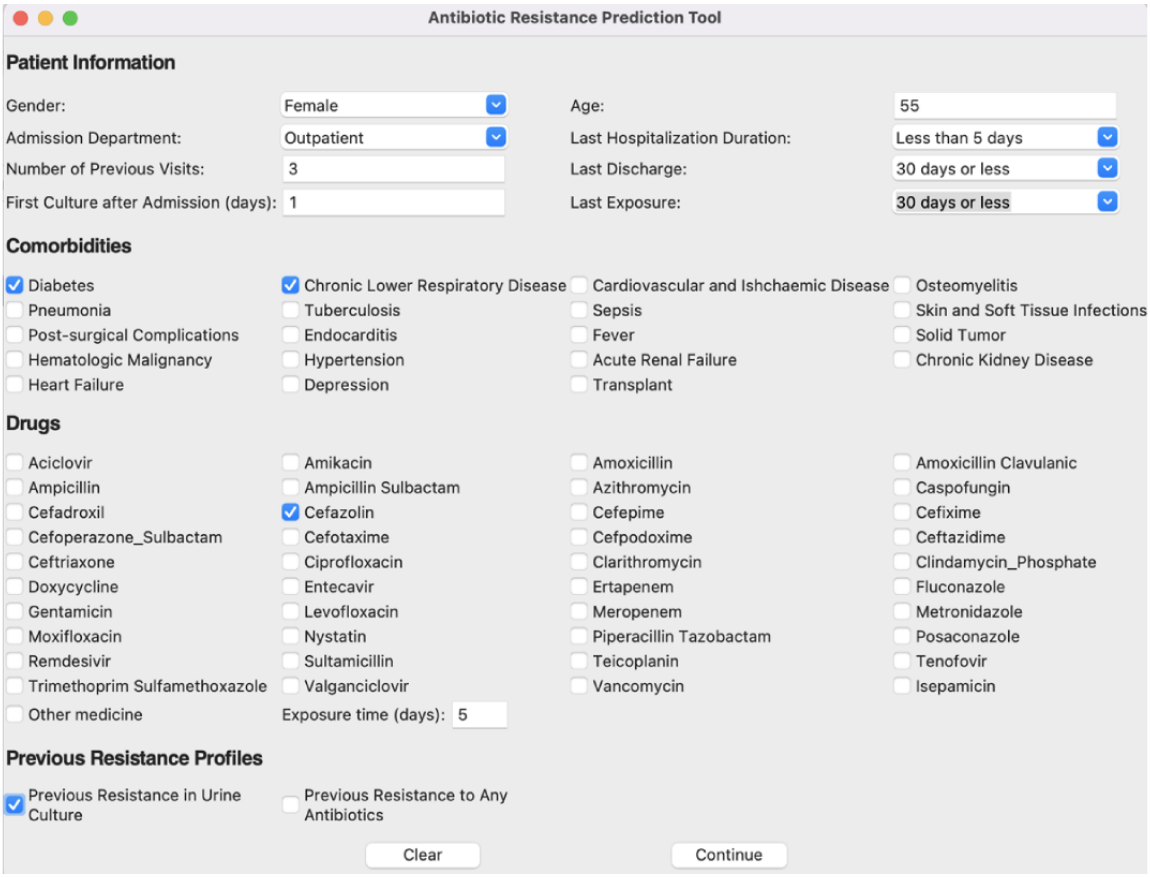
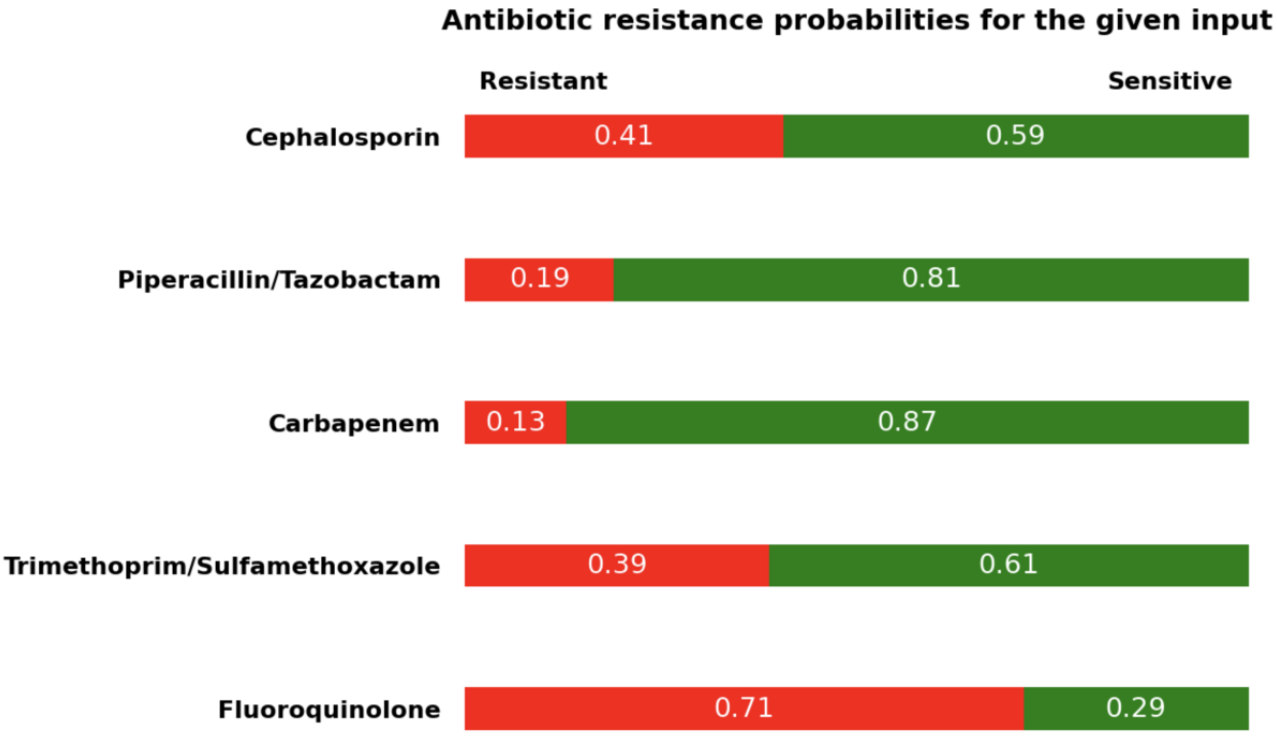


Figure 5. The screenshot of system output for the given data.



Discussion

Principal Findings

In this study, our main objective was to predict cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone resistance in patients with UTI and develop a CDSS with the machine learning models we built. Moreover, we identified the most important features for predicting antibiotic resistance in patients with UTI using SHAP analysis.

Our prediction models achieved AUROCs of 0.777 (95% CI 0.775-0.779), 0.864 (95% CI 0.862-0.867), 0.877 (95% CI 0.874-0.880), 0.881 (95% CI 0.879-0.882), and 0.884 (95% CI 0.884-0.885) in the training set and 0.638 (95% CI 0.635-0.642), 0.630 (95% CI 0.626-0.634), 0.665 (95% CI 0.659-0.671), 0.670 (95% CI 0.666-0.673), and 0.721 (95% CI 0.718-0.724) in the test set for predicting resistance to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone, respectively. The fluoroquinolone resistance prediction model showed superior performance, as confirmed by its high AUROC values in both the training and test sets. On the other hand, the cephalosporin resistance prediction model showed poor performance, as confirmed by the low AUROC values in both training and test sets.

According to SHAP analysis, the contribution of the variables varied for each antibiotic; however, we found that the number of previous visits, first culture after admission, chronic lower respiratory diseases, administration of drugs before infection, and exposure time to these drugs were important predictors across all antibiotics. Factors such as the first culture after admission, exposure time, and the number of previous visits were found to affect resistance, which can be explained by the impact of health care–associated infections. Chronic lower respiratory and kidney diseases are also likely to be associated with frequent visits to health care facilities, although it is difficult to confirm the actual number of visits. However, this suggests that the characteristics of health care–seeking behavior in patients with specific underlying diseases may influence resistance [20]. Interestingly, the use of cefazolin had a negative impact on the development of resistance for all antibiotics. This is because cefazolin is one of the narrow-spectrum antibiotics used in less severe patients. Further research is needed to examine these results.

Comparison to Prior Work

Past efforts to predict antibiotic resistance in patients with UTIs have had promising results, with the lowest AUROC being 0.58 for predicting TMP-SMX resistance [12] and the highest AUROC being 0.83 for predicting ciprofloxacin resistance [9]. In comparison, our prediction models demonstrated comparable performance to these prior works. Some previous studies on predicting antibiotic resistance in patients with UTIs were limited to specific patient groups, including patients with uncomplicated UTIs [8] and patients treated in the emergency department [9]. We analyzed heterogeneous data that were not limited to a specific patient group or bacteria. This approach provides a more comprehensive insight into the prediction of antibiotic resistance in patients with UTIs. Similarly, Lewin-Epstein et al [21] analyzed heterogeneous data and were able to achieve AUROC values ranging from 0.73 to 0.79 for the prediction of ceftazidime, gentamicin, imipenem, ofloxacin, and TMP-SMX resistance. Their data contained multiple culture tests, which provided a more comprehensive approach to predicting antibiotic resistance. Although urine cultures can be used to infer colonized resistance in patients, further research is needed to extend culture results beyond urine.

Limitations

While this study provides insights into predicting antibiotic resistance in patients with UTIs, it has some limitations. First, this study is the lack of multidrug resistance classification. The data set we used in this study did not contain a sufficient amount of multidrug resistance outcomes to build a classification model for the prediction of multidrug resistance. Furthermore, our prediction models were developed using prescription records within the hospital setting. However, patients may have used antibiotics outside of the hospital setting during visits to other hospitals. The lack of information about past drug use could have negatively impacted the performance of our prediction models. To overcome this limitation, we intend to conduct further studies using data from the National Health Insurance Service of South Korea, which contain all past drug use information of the patients. Thus, we will have a more comprehensive data set. By using this approach, we may be able to develop more accurate machine learning models to predict antibiotic resistance and improve our ability to guide appropriate antibiotic therapy selection. Additionally, further development is required to address the limitations of prototype CDSS, including the integration of real-time patient data and validation in larger patient cohorts. Moreover, the prototype CDSS only gives the resistance risk probability to the user. However, a more comprehensive system that can provide decision support on the selection of appropriate therapy, dosage, and duration of treatment can be developed in further studies. Such a system has the potential to reduce the duration of treatment, number of antibiotics used, cost, mortality, and morbidity [22,23].

Conclusions

In conclusion, our study results demonstrated that prediction models to predict antibiotic resistance in patients with UTIs can be constructed using routinely collected EMR data alone, without requiring additional laboratory tests or specialized tests. Machine learning techniques can be used to develop systems that can guide clinicians in selecting appropriate antibiotic therapy. This has the potential to prevent the risk of inappropriate antibiotic administration, thereby reducing patients' risk of developing antibiotic resistance.

Acknowledgments

This study was supported by the National Institute for International Education of the Government of the Republic of Korea, The Scientific and Technological Research Council of Turkey (grant 2214-A), and a faculty research grant of Yonsei University College of Medicine (6-2022-0118).

Authors' Contributions

Nİ, AY, SYP, and DY contributed to the conceptualization of the study and to the funding acquisition. JK, JAR, and SYP were responsible for data curation. Nİ performed the formal analysis of the collected data and wrote the paper. Nİ, SYP, and DY contributed to the development of the study methodology. SYP and DY reviewed and edited the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The antibiotics included in each antibiotic class.
[DOCX File , 17 KB - medinform_v12i1e51326_app1.docx]

Multimedia Appendix 2
Description of input variables.
[DOCX File , 15 KB - medinform_v12i1e51326_app2.docx]

Multimedia Appendix 3
Classification performances of the decision tree models.
[DOCX File , 20 KB - medinform_v12i1e51326_app3.docx]

Multimedia Appendix 4
Classification performances of the k-nearest neighbor models.
[DOCX File , 20 KB - medinform_v12i1e51326_app4.docx]

Multimedia Appendix 5
Classification performances of the logistic regression models.
[DOCX File , 20 KB - medinform_v12i1e51326_app5.docx]

Multimedia Appendix 6
Classification performances of the Extreme Gradient Boosting models.
[DOCX File , 20 KB - medinform_v12i1e51326_app6.docx]

Multimedia Appendix 7
SHAP analysis results of cephalosporin resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations.
[PNG File , 121 KB - medinform_v12i1e51326_app7.png]

Multimedia Appendix 8
SHAP analysis results of TZP resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations; TZP: piperacillin-tazobactam.
[PNG File , 117 KB - medinform_v12i1e51326_app8.png]

Multimedia Appendix 9
SHAP analysis results of carbapenem resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations; TZP: piperacillin-tazobactam.
[PNG File , 122 KB - medinform_v12i1e51326_app9.png]

Multimedia Appendix 10
SHAP analysis results of TMP-SMX resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations; TMP-SMX: trimethoprim-sulfamethoxazole.

[PNG File , 123 KB - medinform_v12i1e51326_app10.png]

References

1. Tan CW, Chlebicki MP. Urinary tract infections in adults. Singapore Med J 2016;57(9):485-490 [FREE Full text] [doi: 10.11622/smedj.2016153] [Medline: 27662890]
2. Belete MA, Saravanan M. A systematic review on drug resistant urinary tract infection among pregnant women in developing countries in Africa and Asia; 2005-2016. Infect Drug Resist 2020;13:1465-1477 [FREE Full text] [doi: 10.2147/IDR.S250654] [Medline: 32547115]
3. Santos M, Mariz M, Tiago I, Martins J, Alarico S, Ferreira P. A review on urinary tract infections diagnostic methods: laboratory-based and point-of-care approaches. J Pharm Biomed Anal 2022;219:114889 [FREE Full text] [doi: 10.1016/j.jpba.2022.114889] [Medline: 35724611]
4. Suskind AM, Saigal CS, Hanley JM, Lai J, Setodji CM, Clemens JQ, Urologic Diseases of America Project. Incidence and management of uncomplicated recurrent urinary tract infections in a national sample of women in the United States. Urology 2016;90:50-55 [FREE Full text] [doi: 10.1016/j.urology.2015.11.051] [Medline: 26825489]
5. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. Nat Rev Microbiol 2015;13(5):269-284 [FREE Full text] [doi: 10.1038/nrmicro3432] [Medline: 25853778]
6. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 2019;20(6):356-370 [FREE Full text] [doi: 10.1038/s41576-019-0108-4] [Medline: 30886350]
7. Benkova M, Soukup O, Marek J. Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice. J Appl Microbiol 2020;129(4):806-822. [doi: 10.1111/jam.14704] [Medline: 32418295]
8. Kanjilal S, Oberst M, Boominathan S, Zhou H, Hooper DC, Sontag D. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. Sci Transl Med 2020;12(568):eaay5067 [FREE Full text] [doi: 10.1126/scitranslmed.aay5067] [Medline: 33148625]
9. Lee HG, Seo Y, Kim JH, Han SB, Im JH, Jung CY, et al. Machine learning model for predicting ciprofloxacin resistance and presence of ESBL in patients with UTI in the ED. Sci Rep 2023;13(1):3282 [FREE Full text] [doi: 10.1038/s41598-023-30290-y] [Medline: 36841917]
10. Yelin I, Snitser O, Novich G, Katz R, Tal O, Parizade M, et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. Nat Med 2019;25(7):1143-1152 [FREE Full text] [doi: 10.1038/s41591-019-0503-6] [Medline: 31273328]
11. Hebert C, Gao Y, Rahman P, Dewart C, Lustberg M, Pancholi P, et al. Prediction of antibiotic susceptibility for urinary tract infection in a hospital setting. Antimicrob Agents Chemother 2020;64(7):e02236-19 [FREE Full text] [doi: 10.1128/AAC.02236-19] [Medline: 32312778]
12. Rich SN, Jun I, Bian J, Boucher C, Cherabuddi K, Morris JG, et al. Development of a prediction model for antibiotic-resistant urinary tract infections using integrated electronic health records from multiple clinics in North-Central Florida. Infect Dis Ther 2022;11(5):1869-1882 [FREE Full text] [doi: 10.1007/s40121-022-00677-x] [Medline: 35908268]
13. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886. [doi: 10.1136/bmj.l886] [Medline: 30862612]
14. Macesic N, Polubriaginof F, Tatonetti NP. Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. Curr Opin Infect Dis 2017;30(6):511-517. [doi: 10.1097/QCO.0000000000000406] [Medline: 28914640]
15. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. J Med Artif Intell 2022;5:4-4 [FREE Full text] [doi: 10.21037/jmai-21-25]
16. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
17. Tacconelli E. New strategies to identify patients harbouring antibiotic-resistant bacteria at hospital admission. Clin Microbiol Infect 2006;12(2):102-109 [FREE Full text] [doi: 10.1111/j.1469-0691.2005.01326.x] [Medline: 16441446]
18. Janssens ACJW, Martens FK. Reflection on modern methods: revisiting the area under the ROC curve. Int J Epidemiol 2020;49(4):1397-1403 [FREE Full text] [doi: 10.1093/ije/dyz274] [Medline: 31967640]
19. Cook J, Ramadas V. When to consult precision-recall curves. Stata J 2020;20(1):131-148 [FREE Full text] [doi: 10.1177/1536867x20909693]
20. Park H, Son MJ, Jung DW, Lee H, Lee JY. National trends in hospitalization for ambulatory care sensitive conditions among Korean adults between 2008 and 2019. Yonsei Med J 2022;63(10):948-955 [FREE Full text] [doi: 10.3349/ymj.2022.0110] [Medline: 36168248]
21. Lewin-Epstein O, Baruch S, Hadany L, Stein GY, Obolski U. Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records. Clin Infect Dis 2021;72(11):e848-e855 [FREE Full text] [doi: 10.1093/cid/ciaa1576] [Medline: 33070171]
22. Curtis CE, Al Bahar F, Marriott JF. The effectiveness of computerised decision support on antibiotic use in hospitals: a systematic review. PLoS One 2017;12(8):e0183062 [FREE Full text] [doi: 10.1371/journal.pone.0183062] [Medline: 28837665]

23. Laka M, Milazzo A, Merlin T. Can evidence-based decision support tools transform antibiotic management? A systematic review and meta-analyses. J Antimicrob Chemother 2020;75(5):1099-1111 [FREE Full text] [doi: 10.1093/jac/dkz543] [Medline: 31960021]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CDSS: clinical decision support system
EMR: electronic medical record
PRAUC: precision-recall area under the curve
SHAP: Shapley Additive Explanations
TMP-SMX: trimethoprim-sulfamethoxazole
TZP: piperacillin-tazobactam
UTI: urinary tract infection

Edited by A Benis; submitted 27.07.23; peer-reviewed by MO Khursheed, YJ Tseng; comments to author 25.08.23; revised version received 17.11.23; accepted 08.01.24; published 29.02.24.

Please cite as:
İlhanlı N, Park SY, Kim J, Ryu JA, Yardımcı A, Yoon D
Prediction of Antibiotic Resistance in Patients With a Urinary Tract Infection: Algorithm Development and Validation
JMIR Med Inform 2024;12:e51326
URL: https://medinform.jmir.org/2024/1/e51326
doi:10.2196/51326
PMID:38421718

Original Paper

Forecasting Hospital Room and Ward Occupancy Using Static and Dynamic Information Concurrently: Retrospective Single-Center Cohort Study

Hyeram Seo[1], BS; Imjin Ahn[2], MS; Hansle Gwon[2], MS; Heejun Kang[3], MS; Yunha Kim[2], MS; Heejung Choi[2], MS; Minkyoung Kim[1], BS; Jiye Han[1], BS; Gaeun Kee[2], MS; Seohyun Park[2], BS; Soyoung Ko[2], BS; HyoJe Jung[2], BS; Byeolhee Kim[2], BS; Jungsik Oh[4], BS; Tae Joon Jun[5*], PhD; Young-Hak Kim[6*], MD, PhD

[1]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center & University of Ulsan College of Medicine, Seoul, Republic of Korea

[2]Department of Information Medicine, Asan Medical Center, Seoul, Republic of Korea

[3]Division of Cardiology, Asan Medical Center, Seoul, Republic of Korea

[4]Department of Digital Innovation, Asan Medical Center, Seoul, Republic of Korea

[5]Big Data Research Center, Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

[6]Division of Cardiology, Department of Information Medicine, Asan Medical Center & University of Ulsan College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Young-Hak Kim, MD, PhD
Division of Cardiology
Department of Information Medicine
Asan Medical Center & University of Ulsan College of Medicine
88, Olympic-ro 43-gil
Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 0955
Email: mdyhkim@amc.seoul.kr

Abstract

Background: Predicting the bed occupancy rate (BOR) is essential for efficient hospital resource management, long-term budget planning, and patient care planning. Although macro-level BOR prediction for the entire hospital is crucial, predicting occupancy at a detailed level, such as specific wards and rooms, is more practical and useful for hospital scheduling.

Objective: The aim of this study was to develop a web-based support tool that allows hospital administrators to grasp the BOR for each ward and room according to different time periods.

Methods: We trained time-series models based on long short-term memory (LSTM) using individual bed data aggregated hourly each day to predict the BOR for each ward and room in the hospital. Ward training involved 2 models with 7- and 30-day time windows, and room training involved models with 3- and 7-day time windows for shorter-term planning. To further improve prediction performance, we added 2 models trained by concatenating dynamic data with static data representing room-specific details.

Results: We confirmed the results of a total of 12 models using bidirectional long short-term memory (Bi-LSTM) and LSTM, and the model based on Bi-LSTM showed better performance. The ward-level prediction model had a mean absolute error (MAE) of 0.067, mean square error (MSE) of 0.009, root mean square error (RMSE) of 0.094, and R^2 score of 0.544. Among the room-level prediction models, the model that combined static data exhibited superior performance, with a MAE of 0.129, MSE of 0.050, RMSE of 0.227, and R^2 score of 0.600. Model results can be displayed on an electronic dashboard for easy access via the web.

Conclusions: We have proposed predictive BOR models for individual wards and rooms that demonstrate high performance. The results can be visualized through a web-based dashboard, aiding hospital administrators in bed operation planning. This contributes to resource optimization and the reduction of hospital resource use.

(JMIR Med Inform 2024;12:e53400) doi:10.2196/53400

KEYWORDS

hospital bed occupancy; electronic medical records; time series forecasting; short-term memory; combining static and dynamic variables

Introduction

Background

The global health care market continues to grow, but the burden of health care costs on governments and individuals is reaching its limits. Consequently, there is increasing interest in the efficient use of limited resources in health care systems, and hospitals must develop approaches to maximize medical effectiveness within budgetary constraints [1,2]. One approach to this is optimizing the use of medical resources. Medical resources can be broadly categorized into 3 categories: human resources, physical capital, and consumables. The appropriate and optimized use of these resources is critical for improving health care quality and providing care to a larger number of patients [3,4].

Among the 3 medical resources, hospital beds are considered one of the physical capitals provided by hospitals to patients. These beds are allocated for various purposes, such as rest, hospitalization, postsurgical recovery, etc. They constitute one of the factors that can directly influence the patient's internal satisfaction within the hospital. However, owing to limited space, hospitals often have a restricted number of beds. Moreover, the number and functionality of beds are often fixed owing to budgetary or environmental constraints, making it difficult to make changes. Nonetheless, if hospital administrators can evaluate bed occupancy rates (BORs) according to different time periods, they can predict the need for health care professionals and resources. On the basis of this information, hospitals can plan resources efficiently, reduce operational costs, and achieve economic objectives [5]. In addition, excessive BORs can exert a negative effect on the health of staff members and increase the possibility of exposure to infection risks. Hence, emphasizing only maintaining a high BOR may not necessarily lead to favorable outcomes for the hospital [6,7]. Considering these reasons, BOR prediction plays a vital role in hospitals and is recognized as a broadly understood necessity for resource optimization in the competitive medical field.

In the medical field, optimizing resources is crucial in the face of limited bed capacity and intense competition. Therefore, bed planning is a vital consideration aimed at minimizing hospital costs [8]. To achieve this, hospitals need to plan staffing and vacations weeks or months in advance [9]. The use of machine learning (ML) technology for BOR prediction is necessary to address fluctuations in patient numbers due to seasonal variations or infectious diseases, ensuring continuous hospital operations. In the Netherlands, hospitals have already implemented ML-based BOR prediction [10], and Johns Hopkins Hospital uses various metrics to effectively manage bed capacity for optimization. Predicting BORs based on quantitative data contributes to validating the clinical quality and cost-effectiveness of treatments. This, in turn, enhances overall accountability throughout the wards and contributes to improving hospital efficiency [11].

Prior Work

Hospital BOR prediction has been investigated using various approaches recently. From studies predicting bed demand using mathematical statistics or regression equation models based on given data [12-15], the focus has shifted toward modeling approaches using time-series analysis. This approach observes recorded data over time to predict future values.

A previous study has taken an innovative approach using time-series analysis alongside the commonly used regression analysis for bed demand prediction, and the study demonstrated that using time-series prediction for bed occupancy yielded higher performance results than using a simple trend fitting approach [16]. Another study used the autoregressive integrated moving average (ARIMA) model for univariate data and a time-series model for multivariate data to predict BORs [17]. With the advancement of deep learning (DL) models that possess strong long-term memory capabilities, such as recurrent neural network (RNN) and long short-term memory (LSTM), there has been an increase in studies applying these models to time-series data for prediction purposes. For instance, in the study by Kutafina et al [9], hospital BORs were predicted based on dates and public holiday data from government agencies and schools, without involving the personal information of patients. The study used a nonlinear autoregressive exogenous model to predict a short-term period of 60 days, with an aim to contribute to the planning of hospital staff. The model demonstrated good performance, with an average mean absolute percentage error of 6.24%. In emergency situations, such as the recent global COVID-19 pandemic, the sudden influx of infected patients can disrupt the hospitalization plans for patients with pre-existing conditions [18]. Studies have been conducted using DL architectures to design models for predicting the BOR of patients with COVID-19 on a country-by-country basis. Some studies incorporated additional inputs, such as vaccination rate and median age, to train the models [19]. Studies have also been conducted to focus on the short-term prediction of BORs during the COVID-19 period [20,21]. Prior studies are summarized in Table 1.

Although previous research has contributed to BOR prediction and operational planning at the hospital level, more detailed and systematic predictions are necessary for practical application in real-world operations. To address this issue, studies have developed their own computer simulation hospital systems to not only predict bed occupancy but also execute scheduling for admissions and surgeries to enhance resource utilization [22-24]. Nevertheless, existing studies have the limitation of focusing solely on the overall BOR of the hospital. As an advancement to these studies, we aim to propose a strategy for predicting the BOR at the level of each ward and room using various variables

in a time-series manner. Interestingly, to our knowledge, this is the first study to apply DL to predict ward- and room-specific occupancy rates using time-series analysis.

Table 1. Summary of prior studies.

Study	Year	Data set	Method	Prediction target
Mackay and Lee [12]	2007	Deidentified data, the date and time of patient admission and discharge between 1998 and 2000	Comparison of 2 compartment models through cross-validation	Entire hospital bed occupancy (annual average)
Littig and Isken [13]	2007	Historical and real-time data warehouse and hospital information systems (emergency department, financial, surgical scheduling, and inpatient tracking systems)	Computerized model of MLR[a] and LR[b]	Entire hospital short-term occupancy (24 h or 72 h) based on LOS[c]
Kumar and Mo [14]	2010	Bed management between June 1, 2006, and June 1, 2007; Information: (1) In each class based on length of stay and admission data; (2) Historical previous year's same week admission data; (3) Relationship between identified variables to aid bed managers	The 3 methods are: (1) Poisson bed occupancy model; (2) Simulation model; and (3) Regression model	The 3 prediction targets are: (1) Estimation of bed occupancy and optimal bed requirements in each class; (2) Bed occupancy levels for every class for the following week; and (3) Weekly average number of occupied beds
Seematter-Bagnoud et al [15]	2015	Inpatient stay data in 2010 (acute somatic care inpatients and outpatients)	Three models of hypothesis-based statistical forecasting of future trends	The 3 targets are: (1) Number of hospital stays; (2) Hospital inpatient days; and (3) Beds for medical stay
Farmer and Emami [16]	1990	Inpatient stay data for general surgery in the age group of 15-44 years between 1969 and 1982	The 2 methods are: (1) Forecasting from a structural model and (2) The time-series or Box-Jenkins method	Entire hospital short-term daily bed requirements
Kim et al [17]	2014	Data warehouse between January 2009 and June 2012	The 2 methods are: (1) The ARIMA[d] model for univariate data and (2) The time-series model for multivariate data	Entire hospital bed occupancy (1 day and 1 week)
Kutafina et al [9]	2019	Inpatient stay data between October 14, 2002, and December 31, 2015 (patient identifier, time of admission, discharge, and name of the clinic the patient was admitted to; no personal information on the patients or staff was provided)	NARX[e] model, a type of RNN[f]	Entire hospital mid-term bed occupancy (60 days, bed pool in units of 30 beds)
Bouhamed et al [19]	2022	COVID-19 hospital occupancy data in 15 countries between December 2021 and early January 2022	The 3 models are: LSTM[g], GRU[h], and SRNN[i]. Incorporate vaccination percentage and median age of the population to improve performance	Entire hospital bed occupancy
Bekker et al [20]	2021	Historical data publicly available until mid-October 2020	The 2 methods are: (1) Using linear programming to predict admissions and (2) Fitting the remaining LOS and using results from the queuing theory to predict occupancy	The 2 targets are: (1) Patient admission and (2) Entire hospital short-term bed occupancy
Farcomeni et al [21]	2021	Patients admitted to the intensive care unit between January and June 2020	The 2 methods are: (1) Generalized linear mixed regression model and (2) Area-specific nonstationary integer autoregressive methodology	Entire hospital short-term intensive care bed occupancy

[a]MLR: multinomial logistic regression.

[b]LR: linear regression.

[c]LOS: length of stay.

[d]ARIMA: autoregressive integrated moving average.

[e]NARX: nonlinear autoregressive exogenous.

[f]RNN: recurrent neural network.

[g]LSTM: long short-term memory.

[h]GRU: grid recurrent unit.

[i]SRNN: simple recurrent neural network.

Goal of This Study

The aim of this study was to predict the BORs of hospital wards and rooms using time-series data from individual beds. Although overall bed occupancy prediction is useful for macro-level resource management in hospitals, resource allocation based on the prediction of occupancy rates for each ward and room is required for specific hospital scheduling and practicality. Through this approach, we aim to contribute to the efficient operational cost optimization of the hospital and ensure the availability of resources required for patient care.

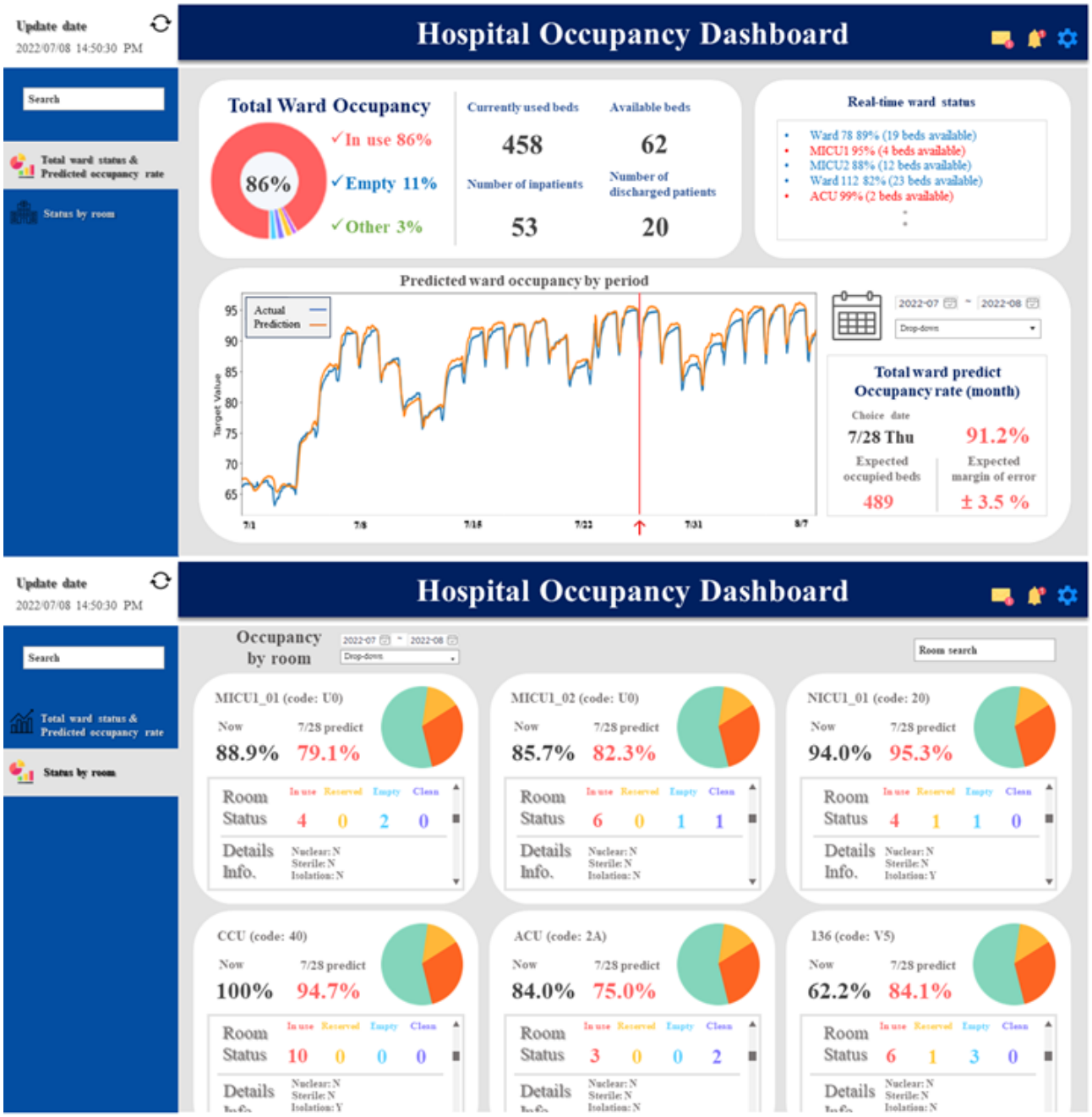
We have developed time-series prediction models based on deep neural network (DNN), among which 1 model combines data representing room-specific features (static data) with dynamic data to enhance the prediction performance for room bed occupancy rates (RBORs). Based on bidirectional long short-term memory (Bi-LSTM), the RBOR prediction model demonstrates a lower mean absolute error (MAE) of 0.049, a mean square error (MSE) of 0.042, a root mean square error (RMSE) of 0.007, and a higher R^2 score of 0.291, indicating the highest performance among all RBOR models.

We developed 6 types of BOR prediction models, of which 2 types were used for predicting ward bed occupancy rates (WBORs), and the other 4 types focused on predicting RBORs. These models use LSTM and Bi-LSTM architectures with strong long-term memory capabilities as their basic structure. We created 6 models for each architecture, resulting in a total of 12 models. The WBOR models were used for predicting weekly and monthly occupancy rates, serving long-term hospital administrative planning purposes. Conversely, the RBOR models were designed for immediate and rapid occupancy planning and were trained with 3- and 7-day intervals. Each RBOR model was enhanced by combining static data, which represent room-specific features, to generate more sophisticated prediction models.

Figure 1 shows the potential application of our model as a form of web software in a hospital setting. Through an online dashboard, it can provide timely information regarding bed availability, enabling intelligent management of patient movements related to admission and discharge. It facilitates shared responsibilities within the hospital and simplifies future resource planning [25].

In the Introduction section, we explored the importance of this research and investigated relevant previous studies, providing a general overview of the direction of our research. In the Methods section, we provide descriptions of the data set used and the structure of the DNN algorithm used, and explain the model architecture and performance. In the Results section, we present the performance and outcomes of this study. Finally, in the Discussion section, we discuss the contributions, limitations, and potential avenues for improvement of the research.

Figure 1. Virtual dashboard of the status and forecast of the ward bed occupancy rate (WBOR) and room bed occupancy rate (RBOR). The first screen presents the overall bed occupancy rate of the hospital, along with the number of beds in use and available. Moreover, a predictive graph displays the anticipated WBOR for selected dates. The second screen presents the WBOR for individual beds, indicating their statuses, such as "in use," "reserved," "empty," and "cleaning." Detailed information about each room is also displayed.



Methods

Overview

We intended to predict the BORs of individual hospital wards and rooms based on the information accumulated in individual bed–level data on an hourly basis, aggregated on a daily basis. For this purpose, we developed 12 time-series models. As the base models, we applied LSTM and Bi-LSTM, which are suitable for sequence data. These models address the limitation of long-term memory loss in traditional RNNs and were chosen because of their suitability for training bed data represented as sequence data.

Based on the model architecture, there were 2 WBOR prediction model types, which were trained at 7- and 30-day intervals to predict the occupancy rate for the next day. Moreover, there were 2 RBOR prediction model types, similar to the ward models, which were trained at 3- and 7-day intervals. Furthermore, as another approach, each RBOR prediction model was augmented with static data, and 2 DL algorithms were proposed for the final comparison of their performances in predicting RBORs.

Ethical Considerations

The study was approved by the Asan Medical Center (AMC) Institutional Review Board (IRB 2021-0321) and was conducted in accordance with the 2008 Declaration of Helsinki.

Materials

Study Setting

This was a retrospective single-center cohort study. Data were collected from AMC, with information on the occupancy status of each bed recorded at hourly intervals between May 27, 2020, and November 21, 2022. The data set comprised a total of 54,632,684 records. This study used ethically preapproved data. Deidentified data used in the study were extracted from ABLE, the AMC clinical research data warehouse.

A total of 57 wards, encompassing specialized wards; 1411 rooms, including private and shared rooms; and 4990 beds were included in this study. Wards and rooms with specific characteristics, such as intensive care unit, newborn room, and nuclear medicine treatment room, were excluded from the analysis as their occupancy prediction using simple and general variables did not align with the direction of this study.

Supporting Data

Supporting data for public holidays were added in our data set. We considered that holidays have both a recurring pattern with specific dates each year and a distinctive characteristic of being nonworking days, which could affect occupancy rates. Based on Korean public holidays, which include Chuseok, Hangeul Proclamation Day, Children's Day, National Liberation Day, Memorial Day, Buddha's Birthday, Independence Movement Day, and Constitution Day, there were 27 days that corresponded to public holidays during the period covered by the data set. We denoted these dates with a value of "1" if they were public holidays and "0" if they were not, based on the reference date.

Preprocessing and Description of Variables

Among the variables representing individual beds, the reference date, ward and room information, patient occupancy status, bed cleanliness status, and detailed room information were available. Based on the recorded date of bed status, we derived additional variables, such as the reference year, reference month, reference week (week of the year), reference day, and reference day of the week.

Room data were derived from the input information representing the cleanliness status of beds. This variable had 2 possible states, namely, "admittable" and "discharge." If neither of these states was indicated, it implied that a patient was currently hospitalized in the bed. As the status of hospitalized patients was indicated by missing values, we replaced them with the number "1" to indicate the presence of a patient in the bed and "0" otherwise. The sum of all "1" values represented the current number of hospitalized patients. The count of beds in each room indicated the capacity of each room. The target variable BOR was calculated by dividing the number of patients in the room by the room capacity, resulting in a room-specific patient occupancy rate variable. The ward data were subjected to a similar process as that of the room data, with the difference being that we generated ward-specific variables, such as ward capacity and WBOR, using the same approach. The static room data consisted of 14 variables, including the title of the room and the detailed information specific to each room.

For the variables in the ward and room data, we disregarded the units of the features and converted them into numerical values for easy comparison, after which we performed normalization. Regarding the variables representing detailed room information, we converted them to numerical values where "yes" was represented as "1" and "no" was represented as "0."

The final set of variables used in this study was categorized into date, ward, room, and detailed room information. Table 2 provides the detailed descriptions of the variables used in our training, including all the administrative data related to beds that are readily available in the hospital.

The explanation of the classification for generating the data sets for training each model is provided in Table 3. The static features of the detailed room information were combined with the room data set, which has sequence characteristics, to generate a separate data set termed Room+Static.

Table 2. Description of variables by category.

Variable	Type	Description
Date		
Year	3 categories	Reference year for bed status
Month	12 categories	Reference month for bed status
Week	53 categories	Reference week for bed status
Day	31 categories	Reference day for bed status
Weekday	7 categories	Reference day of the week for bed status
Holiday	2 categories	Holiday status
Ward		
Ward abbreviation	57 categories	Abbreviations for entire ward names
Ward capacity	Numeric	Number of available ward beds
Ward bed capacity	Numeric	Number of patients currently admitted to the ward
Ward occupancy rate	Numeric	Ward bed capacity divided by ward capacity
Room		
Room abbreviation	1411 categories	Abbreviations for entire room names
Room capacity	Numeric	Number of available room beds
Room bed capacity	Numeric	Number of patients currently admitted to the room
Room occupancy rate	Numeric	Room bed capacity divided by room capacity
Room static feature		
Room code	34 categories	Room grade code
Nuclear	2 categories (N[a]/Y[b])	Nuclear medicine room availability
Sterile	2 categories (N/Y)	Sterile room availability
Isolation	2 categories (N/Y)	Isolation room availability
EEG[c] testing	2 categories (N/Y)	EEG testing room availability
Observation	2 categories (N/Y)	Observation room availability
Kidney	2 categories (N/Y)	Kidney transplant room availability
Liver	2 categories (N/Y)	Liver transplant room availability
Sub-ICU[d]	2 categories (N/Y)	Sub-ICU room availability
Special	2 categories (N/Y)	Special room availability
Small single	2 categories (N/Y)	Small single room availability
Short-term	2 categories (N/Y)	Short-term room availability
Psy-double	2 categories (N/Y)	Psychiatry department double room availability
Psy-open	2 categories (N/Y)	Psychiatry department open room availability

[a]N: No.

[b]Y: Yes.

[c]EEG: electroencephalogram.

[d]ICU: intensive care unit.

Table 3. Data set classification and included variables.

Data set	Variables
Ward data set	Ward abbreviation, year, month, week, day, weekday, holiday, ward capacity, ward bed capacity, and ward occupancy rate
Room data set	Room abbreviation, year, month, week, day, weekday, holiday, room capacity, room bed capacity, and room occupancy rate
Static data set	14 static variables related to detailed room information
Room+Static data set	Room abbreviation, year, month, week, day, weekday, holiday, room capacity, room bed capacity, 14 static variables related to detailed room information, and room occupancy rate

Separation

Each data set was split into training, validation, and test sets for training and evaluation of the model. The training set consisted of 32,153 rows (67.8%), with data from May 27, 2020, to December 2021. The validation set, used for parameter tuning, included 7085 rows (15.0%), with data from January to June 2022. Finally, the test set comprised 8208 rows (17.2%), with data from July 2022 to November 21, 2022.

DL Algorithms

We used various DL algorithms for in-depth learning. In the following subsections, we will provide explanations for each model algorithm used in our research.

LSTM Network

RNN [26] is a simple algorithm that passes information from previous steps to the current step, allowing it to iterate and process sequential data. However, it encounters difficulties in handling long-term dependencies, such as those found in time-series data, owing to the vanishing gradient problem. To address this issue, LSTM [27] was developed. LSTM excels in handling sequence data and is commonly used in natural language processing, machine translation, and time-series data analysis. LSTM consists of an input gate, output gate, and forget gate. The “cell state,” is carefully controlled by each gate to determine whether the memory should be retained or forgotten for the next time step.

Bi-LSTM Network

Although RNN and LSTM possess the ability to remember previous data, they have a limitation in that their results are primarily based on immediate past patterns because the input is processed in a sequential order. This limitation can be overcome through a network architecture known as Bi-LSTM [28]. Bi-LSTM allows end-to-end learning, minimizing the loss on the output and simultaneously training all parameters. It also has the advantage of performing well even with long data sequences. Because of its suitability for models that require knowledge of dependencies from both the past and future, such as LSTM-based time-series prediction, we additionally selected Bi-LSTM as the base model.

Attention Mechanism

Attention mechanism [29,30] refers to the process of incorporating the encoder’s outputs into the decoder at each time step of predicting the output sequence. Rather than considering the entire input sequence, it focuses more on the relevant components that are related to the predicted output, allowing the model to focus on important areas. This mechanism helps minimize information loss in data sets with long sequences, enabling better learning and improving the model’s performance. It has been widely used in areas such as text translation and speech recognition. Nevertheless, as it is still based on RNN models, it has the drawbacks of slower speed and not being completely free from information loss issues.

Combining Static and Dynamic Features

Data can exhibit different characteristics even at the same time. For instance, in data collected at 1-hour intervals for each hospital bed, we can distinguish between “dynamic data,” which include features that change over time, such as the bed condition, date, and patient occupancy, and “static data,” which consist of information that remains constant, such as the ward and room number.

DL allows us to use all the available information for prediction. Therefore, for predicting the RBOR, we investigated an approach that combines dynamic and static data using an LSTM-based method [31]. This approach demonstrated better performance than LSTM alone [32]. Our approach involves adding a layer that incorporates static data as an input to the existing room occupancy prediction model.

Model Architecture

Base Model

Our objective was to predict the intermediate-term occupancy rates of wards and rooms within the hospital to contribute to hospital operation planning. Bi-LSTM was chosen as the base model owing to its improved predictive performance compared with the traditional LSTM model. However, to quantitatively compare these models, we conducted a comparison of the results for each model (6 for each, with a total of 12 models).

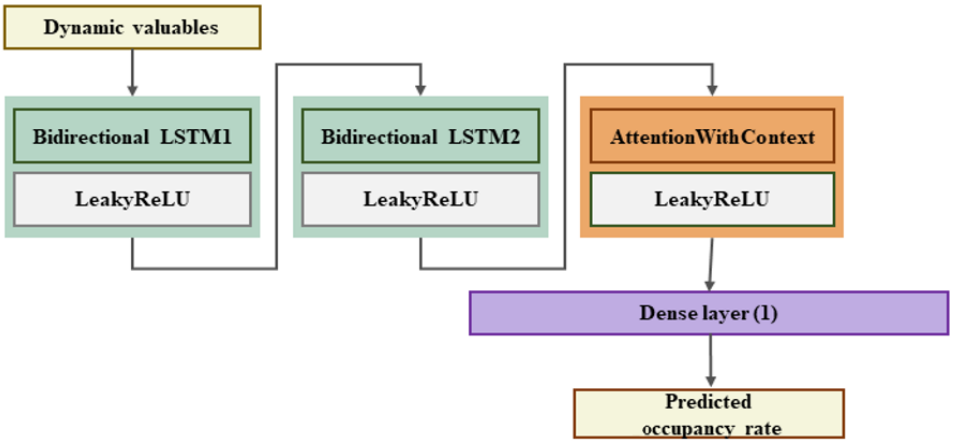
A typical LSTM model processes data sequentially, considering only the information from the past up to the current time step. However, Bi-LSTM, by simultaneously processing data in both forward and backward directions, has a unique feature that allows it to leverage both current and future information for predictions. This bidirectionality helps the model effectively learn temporal dependencies and intricate patterns. However, despite these advantages, Bi-LSTM comes with the trade-off of doubling the number of model parameters, resulting in increased computational costs for training and prediction. While a more complex model can better adapt to the training data, there is an increased risk of overfitting, especially with small

data sets. Nevertheless, the reason for choosing Bi-LSTM for tasks like predicting BORs in hospitals, involving time-series data, lies in its ability to harness the power of bidirectional information. Bi-LSTM processes input data from both past and future directions simultaneously, enabling it to effectively incorporate future information into current predictions. This proves beneficial for handling complex patterns in long time-series data [28].

Moreover, we have enhanced the performance of our models by adding an attention layer to Bi-LSTM. The attention layer assigns higher weights to features that exert a significant impact on the prediction, allowing the model to focus on relevant information and gather necessary input features. This helps improve the accuracy of the prediction. Furthermore, the attention layer reduces the amount of information processed, resulting in improved computational efficiency. Ultimately, this contributes toward enhancing the overall performance of the model.

The window length of the input sequence was divided into 3 different intervals, namely, 3, 7, and 30 days. The WBOR model was trained on sequences with a window length of 7 and 30 days, whereas the RBOR model was trained on sequences with a window length of 3 and 7 days. The first layer of our model consisted of Bi-LSTM, which was followed by the leaky rectified linear unit (LeakyReLU) activation function. LeakyReLU is a linear function that has a small gradient for negative input values, similar to ReLU. It helps the model converge faster. After applying this process once again, the AttentionWithContext layer was applied, which focuses on important components of input sequence data and transforms outputs obtained from the previous layer. After applying the activation function again, a dense layer with 1 neuron was added for generating the final output. The sigmoid function was used to limit the output values between 0 and 1. Finally, our model was compiled using the MSE loss function, Adam optimizer, and MAE metric. The parameters for each layer were selected based on accumulated experience through research. Figure 2 visually represents the above-described structure.

Figure 2. Base bidirectional long short-term memory (Bi-LSTM) model architecture. LeakyReLU: leaky rectified linear unit; LSTM: long short-term memory.



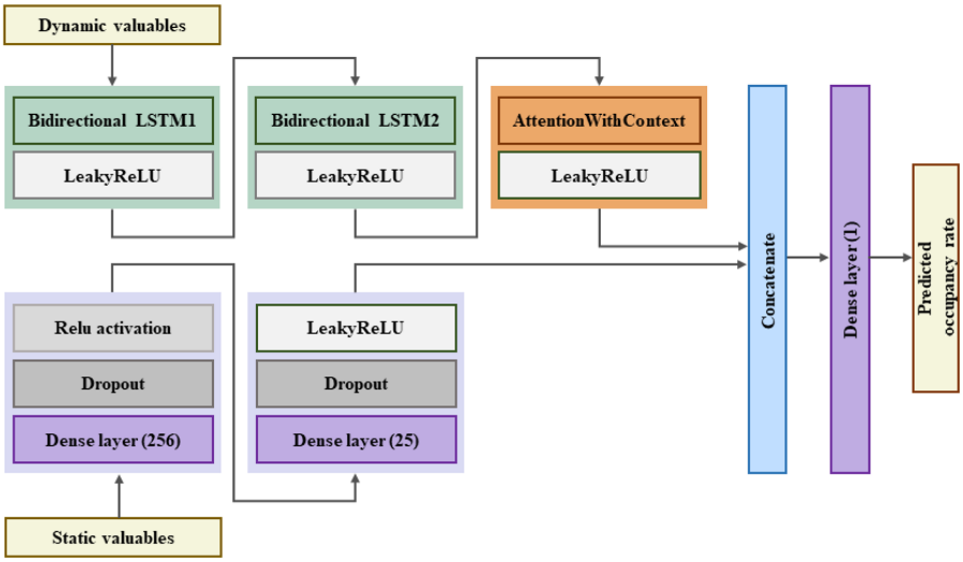
Combining Dynamic and Static Data Using the DL Model

The accumulated bed data, which were collected on a time basis, were divided into dynamic and static data of the rooms, which were then inputted separately. To improve the performance of the BOR prediction model, we designed different DL architectures for the characteristics of these 2 types of data.

We first used a base model based on LSTM and Bi-LSTM to learn the time-series data and then focused the model's attention using the dense layer to process fixed-size inputs. To prevent overfitting, we applied the dropout function to randomly deactivate neurons in 2 dense layers. The hidden states of the 2 networks were combined, and the resulting output was passed to a single layer, combining the time dynamic and static data.

Finally, the hidden states of the 2 networks were combined, and the combined result was passed to a single layer to effectively integrate the dynamic and static data. This allowed us to use the information from both the dynamic and static data for BOR prediction. This architecture is illustrated in Figure 3.

Figure 3. Bidirectional long short-term memory (Bi-LSTM) model architecture combining static and dynamic variables. LeakyReLU: leaky rectified linear unit; LSTM: long short-term memory.



Hyperparameter Tuning

One of the fundamental methods to enhance the performance of artificial intelligence (AI) learning models is the use of hyperparameter tuning. Hyperparameters are parameters passed to the model to modify or adjust the learning process. While hyperparameter tuning may rely on the experience of researchers, there are also functionalities that automatically search for hyperparameters, taking into account the diversity of model structures.

Various methods for search optimization have been proposed [33,34], but we implemented our models using the Keras library. By leveraging Keras Tuner, we automatically searched for the optimal combinations of units and learning rates for each model, contributing to the improvement of their performance.

Time Series Cross-Validation

Time-series data exhibit temporal dependencies between data points, making it crucial to consider these characteristics when validating a model. Commonly used K-fold cross-validation is effective for evaluating models on general data sets [35], providing effectiveness in preventing overfitting and enhancing generalizability by dividing the data into multiple subsets [36,37]. However, for time-series data, shuffling the data randomly is not appropriate owing to the inherent sequential dependency of the observations.

Time series cross-validation is a method that preserves this temporal dependence while dividing the data [38]. It involves splitting the entire hospital bed data set into 5 periods, conducting training and validation for each period, and repeating this process as the periods shift. This approach is particularly effective when observations in the dynamic data set, such as hospital bed data recorded at 1-hour intervals, play a crucial role in predicting future values based on past observations.

Shuffling data randomly using K-fold may disrupt the temporal continuity, leading to inadequate reflection of past and future observations. Therefore, time series cross-validation sequentially partitions the data, ensuring the temporal flow is maintained, and proves to be more effective in evaluating the model's performance. This method enables the model to make more accurate predictions of future occupancy based on past trends.

Evaluation

We selected various metrics to evaluate the performance of time-series data predictions. Among them, MAE represents the absolute difference between the model's predicted values and the actual BOR. We also considered MSE, which is sensitive to outliers. Moreover, to address the limitations of MSE and provide a penalty for large errors, we opted for RMSE. We also used the R^2 score to measure the correlation between the predicted and actual values.

MAE is a commonly used metric to evaluate the performance of time-series prediction models. MAE is intuitive and easy to calculate, making it widely used in practice. Because MAE uses absolute values, it is less sensitive to outliers in the occupancy rate values for specific dates. MAE is calculated using the following formula:

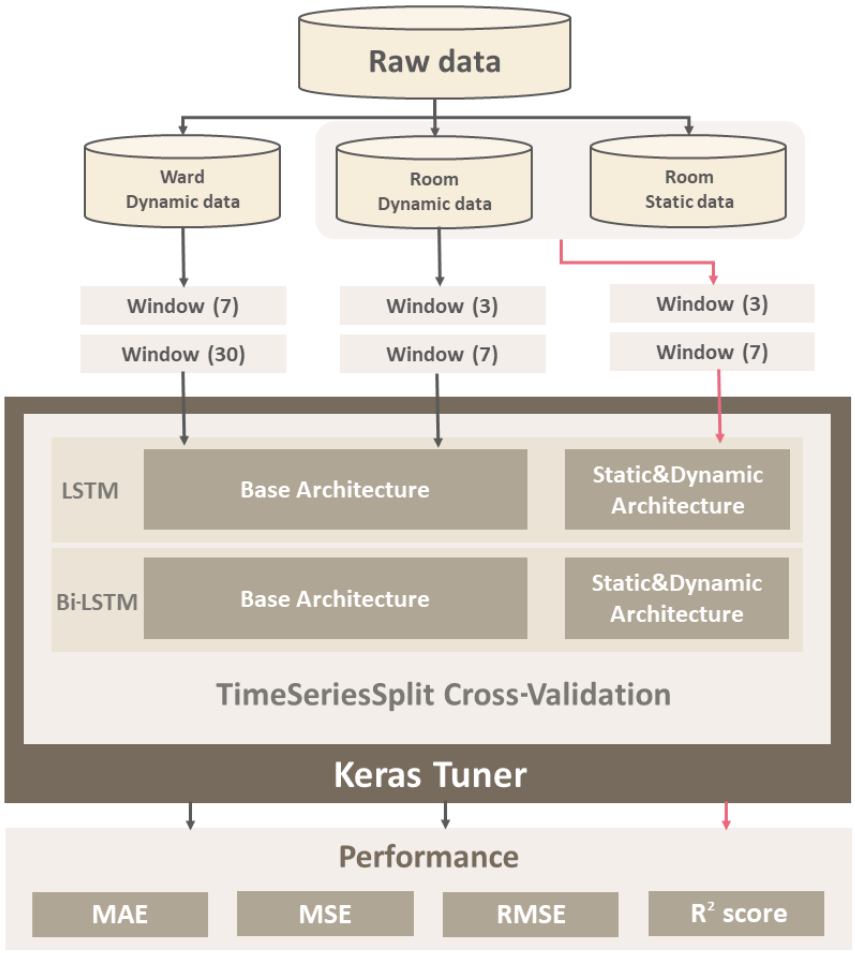
MSE is a metric that evaluates the magnitude of errors by squaring the differences between the predicted and actual values and then taking the average. It is calculated using the following formula:

RMSE is used to address the limitations of MSE where the error scales as a square, providing a more intuitive understanding of the error magnitude between the predicted and actual values. It penalizes large errors, making it less sensitive to outliers. RMSE is calculated using the following formula:

The R^2 score is used to measure the explanatory potential of the prediction model, and it is calculated using the following formula:

Here, SSR represents the sum of squared differences between the predicted and actual values, and SST represents the sum of squared differences between the actual values and the mean value of actual values. Figure 4 shows the prediction method and overall flow in this study.

Figure 4. Overall flow in this study. Bi-LSTM: bidirectional long short-term memory; LSTM: long short-term memory; MAE: mean absolute error; MSE: mean square error; RMSE: root mean square error.



Results

We used 2 DL models, LSTM and Bi-LSTM, and compared the performance of 12 different prediction models. These models have been denoted as ward 7 days (W7D), ward 30 days (W30D), room 3 days (R3D), room 7 days (R7D), room static 3 days (RS3D), and room static 7 days (RS7D). Using Keras Tuner, we adjusted the hyperparameters of the models and subsequently validated the models through a 5-fold time series cross-validation.

The prediction performances of the models for WBOR and RBOR were compared, which showed that they were more accurate at predicting WBOR, with MAE values of 0.06 to 0.07. The W7D model based on Bi-LSTM, which used 7 days of ward data to predict the next day's ward occupancy, had a MAE value of 0.067, MSE value of 0.009, and RMSE value of 0.094, showing high accuracy. The R^2 score was also 0.544, which was approximately 0.240 higher than that of the W30D model (0.304), indicating that the variables in that model explained occupancy reasonably well.

We next compared the performances of the 8 models for RBOR prediction, and among them, the RS7D model based on Bi-LSTM, which was trained on a 7-day time step by integrating static and dynamic data, showed the best performance. It achieved a MAE value of 0.129, MSE value of 0.050, RMSE value of 0.227, and R^2 score of 0.260. In particular, the R^2 score outperformed that of the R3D model by 0.014. These data are summarized in Table 4. Regarding the WBOR prediction model, the model with a shorter training unit, W7D, demonstrated better performance. However, regarding the RBOR prediction model, the model with a longer training unit of 7 days, which incorporated detailed room-specific information, exhibited slightly higher performance than the model with a shorter

training unit of 3 days. The model with the added room-specific information still demonstrated superior performance overall.

We visualized the predicted and actual occupancy for Bi-LSTM models and investigated the occupancy trends since July 2022 on our test data set. First, we selected a specific ward in W7D to demonstrate the change in the WBOR over 2 months. The right panel of Figure 5 shows the WBOR change over 5 months from July 2022 in W30D. The blue line represents the actual occupancy value, and the red line represents the predicted occupancy value by the model. This provides an at-a-glance view of the overall predicted occupancy level for each month and allows hospital staff to observe trends to obtain a rough understanding of the WBOR.

Figure 6 shows graphs of occupancy rate values for a randomized specific room, displaying the predicted and actual values for the 4 RBOR prediction models, with 2 graphs for each model. The left graph shows the occupancy rate change over 5 months from July to November 2022, and the right graph shows the occupancy rate for the months of July and August, providing a detailed view of the RBOR. By examining the trends of the predicted and actual values for the 4 models in this period for a specific room, we can observe that the models maintain a similar trend to the actual occupancy rate.

Table 4. Performances of the occupancy prediction models.

Model and fold	MAE[a]		MSE[b]		RMSE[c]		R^2 score	
	LSTM[d]	Bi-LSTM[e]	LSTM	Bi-LSTM	LSTM	Bi-LSTM	LSTM	Bi-LSTM
Ward								
W30D[f]								
1	0.081	0.097	0.014	0.015	0.117	0.121	0.040	−0.081
2	0.074	0.064	0.011	0.007	0.107	0.085	0.106	0.430
3	0.118	0.109	0.031	0.025	0.175	0.161	−0.130	0.086
4	0.150	0.087	0.033	0.013	0.182	0.113	−0.572	0.399
5	0.087	0.061	0.019	0.008	0.139	0.089	0.212	0.678
Mean	0.102	0.084	0.021	0.014	0.144	0.114	−0.068	0.304
W7D[g]								
1	0.071	0.063	0.011	0.007	0.103	0.086	0.263	0.479
2	0.067	0.054	0.009	0.005	0.094	0.071	0.302	0.606
3	0.119	0.091	0.033	0.016	0.183	0.126	−0.241	0.408
4	0.116	0.068	0.021	0.009	0.145	0.098	−0.009	0.537
5	0.083	0.060	0.015	0.007	0.123	0.087	0.380	0.690
Mean	0.091	0.067	0.018	0.009	0.130	0.094	0.139	0.544
Room								
R7D[h]								
1	0.120	0.111	0.057	0.045	0.238	0.212	0.026	0.226
2	0.127	0.108	0.057	0.047	0.238	0.216	0.054	0.222
3	0.190	0.148	0.167	0.072	0.327	0.269	0.018	0.336
4	0.209	0.162	0.068	0.055	0.261	0.234	−0.089	0.125
5	0.158	0.124	0.069	0.048	0.263	0.220	0.102	0.370
Mean	0.161	0.131	0.071	0.053	0.265	0.230	0.022	0.256
R3D[i]								
1	0.134	0.115	0.058	0.045	0.242	0.212	0.001	0.229
2	0.130	0.097	0.060	0.048	0.245	0.220	0.006	0.195
3	0.178	0.147	0.118	0.080	0.344	0.283	−0.084	0.266
4	0.210	0.204	0.078	0.075	0.280	0.275	−0.247	−0.201
5	0.161	0.120	0.064	0.048	0.254	0.220	0.168	0.377
Mean	0.163	0.167	0.076	0.059	0.273	0.242	−0.031	0.173
RS7D[j]								
1	0.147	0.114	0.057	0.045	0.238	0.212	0.027	0.228
2	0.151	0.099	0.057	0.046	0.240	0.215	0.042	0.227
3	0.216	0.160	0.104	0.063	0.322	0.267	0.048	0.260
4	0.194	0.152	0.064	0.050	0.252	0.224	−0.016	0.198
5	0.181	0.120	0.068	0.047	0.261	0.217	0.112	0.385
Mean	0.178	0.129	0.070	0.050	0.262	0.227	0.043	0.260
RS3D[k]								
1	0.109	0.116	0.056	0.046	0.237	0.215	0.039	0.213

Model and fold	MAE[a]		MSE[b]		RMSE[c]		R^2 score	
	LSTM[d]	Bi-LSTM[e]	LSTM	Bi-LSTM	LSTM	Bi-LSTM	LSTM	Bi-LSTM
2	0.118	0.092	0.061	0.048	0.246	0.219	−0.009	0.203
3	0.182	0.160	0.116	0.090	0.340	0.300	−0.062	0.172
4	0.278	0.191	0.152	0.065	0.389	0.255	−1.410	−0.039
5	0.159	0.116	0.074	0.047	0.272	0.218	0.043	0.387
Mean	0.169	0.135	0.092	0.059	0.297	0.241	−0.028	0.187

[a]MAE: mean absolute error.
[b]MSE: mean square error.
[c]RMSE: root mean square error.
[d]LSTM: long short-term memory.
[e]Bi-LSTM: bidirectional long short-term memory.
[f]W30D: ward 30 days.
[g]W7D: ward 7 days.
[h]R7D: room 7 days.
[i]R3D: room 3 days.
[j]RS7D: room static 7 days.
[k]RS3D: room static 3 days.

Figure 5. Examples of the predicted and actual bed occupancy rates for the 2-month period from July to August 2022 for ward 7 days and the 5-month period from July to November 2022 for ward 30 days.

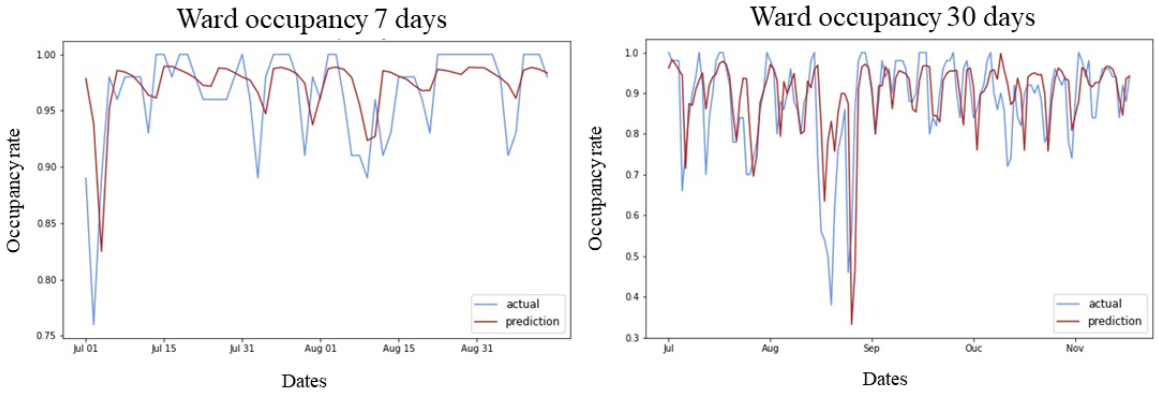
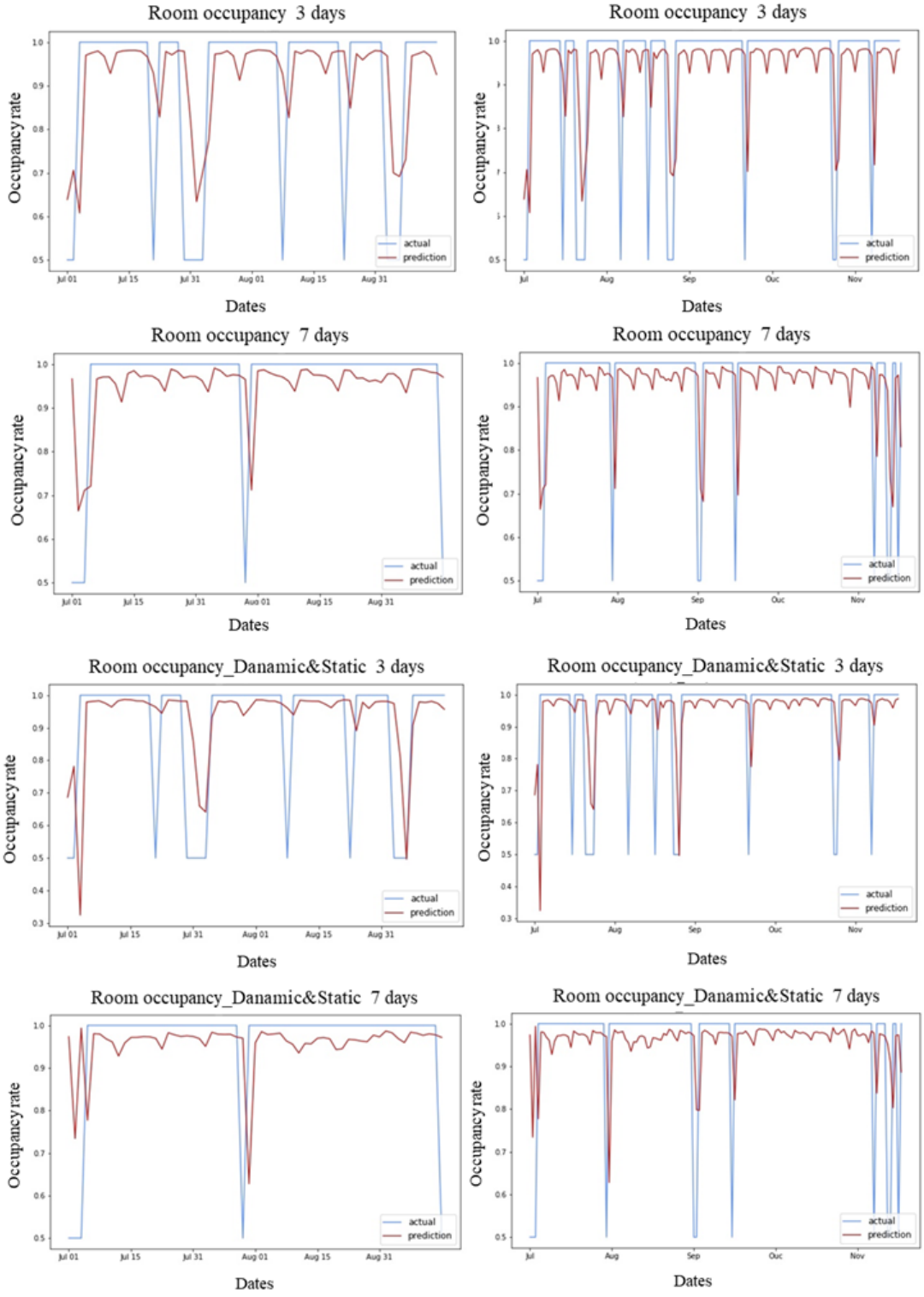


Figure 6. Examples of the predicted and actual bed occupancy rates for the 2-month period from July to August 2022 and the 5-month period from July to November 2022.



Discussion

Principal Findings

The entire data set of this study consisted of administrative data collected at AMC at an hourly interval for each ward from May 27, 2020, to November 21, 2022. To improve the hospital's challenges, we developed a model to predict the occupancy rate of wards and rooms. Our aim was to contribute toward administrative and financial planning for bed management within the hospital.

During the specified period, we compared the results of using DL models to predict the overall BOR for each ward and individual rooms. In the case of WBOR prediction, the MAE of the 7-day window model based on Bi-LSTM was approximately 0.067, demonstrating a remarkably close prediction to the occupancy compared with that of the 30-day

window model based on LSTM, with a difference of approximately 0.035. Furthermore, the MSE and RMSE were 0.009 and 0.094, respectively, indicating high accuracy in the predictions. Moreover, the R^2 score of 0.544 indicated that the model had better explanatory potential than the average. For the individual RBOR prediction, among the 8 models, the RS7D model based on Bi-LSTM performed the best, exhibiting a MAE of approximately 0.129, which was remarkably lower than that of the other models. Moreover, the MSE and RMSE were significantly lower than those of the RBOR models, with differences of 0.042 and 0.07, respectively. The R^2 score of 0.260 indicated that it had higher explanatory potential than the RS3D models based on LSTM, with the value being higher by 0.291.

Finally, we visualized the predicted and actual values on a graph for a specific period and observed that each model captured the trend of the actual BOR quite well. Although the models were less accurate in predicting low occupancy periods, they followed the general trend closely. Overall, these findings demonstrate that our DL models effectively predicted BORs for both wards and individual rooms, with certain models demonstrating superior performance in different scenarios.

Strengths and Limitations

Although the models in this study demonstrated good performance in following the trends of BORs and achieved good results, there were several limitations in this research. First, there were limitations in the data. Although we used administrative data and detailed room information available from the hospital to enable the models to capture occupancy trends, the relationship between the variables and the model's explanatory potential showed room for improvement, as indicated by the R^2 score. To achieve higher prediction accuracy, it would be beneficial to incorporate diverse data sources and real-time updated information.

Second, there was variability in external factors. Hospital BORs are heavily influenced by external environmental factors. Sudden events, such as environmental factors and outbreaks of infectious diseases like COVID-19, can render accurate prediction of bed occupancy challenging [18,32]. Furthermore, seasonal effects and accidents can increase the number of patients. Sufficient collection of long-term data on these external factors would be necessary, but such uncertainties can reduce the accuracy of predictions.

Despite these limitations, our study demonstrated a significant level of adherence to trends in the prediction of individual ward and room occupancy. More detailed variables and a longer period of data accumulation would be required to predict the specific number of beds.

Conclusion

We presented models that can predict the occupancy rates of wards and individual hospital rooms using artificial neural networks based on time-series data. The predicted results of these models demonstrated a high level of accuracy in capturing the future trends of the BOR. In particular, we presented 8 RBOR models with structure and window changes to compare their performance and found that the RS7D model showed the best performance. Our results can be implemented as a web application on hospital online dashboards, as depicted in Figure 1 [25]. In fact, Johns Hopkins University has been applying these methods in their command center to monitor hospital capacity and achieve effectiveness in patient management planning [39].

Furthermore, predicting BORs supports patient admission and discharge planning, helping to alleviate overcrowding in emergency departments and reduce patient waiting times. Staff members can effectively schedule patient admission and discharge, and minimize waiting times by understanding the BOR, providing urgent treatment to emergency patients. Moreover, providing appropriate information to patients waiting in the emergency department can increase patient satisfaction and facilitate efficient transition to hospital admission [40,41]. By applying AI models that combine BOR prediction, which contributes toward reducing emergency department waiting times with individual patient admission and discharge prediction, hospitals can achieve resource optimization and cost savings, resulting in improved patient satisfaction.

Acknowledgments

This work was supported by a Korea Medical Device Development Fund grant funded by the Korean government (the Ministry of Science and ICT; the Ministry of Trade, Industry and Energy; the Ministry of Health & Welfare, Republic of Korea; the Ministry of Food and Drug Safety) (project number: 1711195603, RS-2020-KD000097, 50%) and by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0026).

Conflicts of Interest

None declared.

References

1. Reuben DB, Cassel CK. Physician stewardship of health care in an era of finite resources. JAMA 2011 Jul 27;306(4):430-431. [doi: 10.1001/jama.2011.999] [Medline: 21791692]
2. National Health Expenditure Projections 2011-2021. Centers for Medicare and Medicaid Services. URL: https://www.cms.gov/files/document/forecastsummaryandtablespdf [accessed 2024-02-21]

3. The world health report 2000 - Health systems: improving performance. World Health Organisation. URL: https://cdn.who.int/media/docs/default-source/health-financing/whr-2000.pdf?sfvrsn=95d8b803_1&download=true [accessed 2024-02-21]
4. Kabene SM, Orchard C, Howard JM, Soriano MA, Leduc R. The importance of human resources management in health care: a global context. Hum Resour Health 2006 Jul 27;4(1):20 [FREE Full text] [doi: 10.1186/1478-4491-4-20] [Medline: 16872531]
5. Page K, Barnett AG, Graves N. What is a hospital bed day worth? A contingent valuation study of hospital Chief Executive Officers. BMC Health Serv Res 2017 Feb 14;17(1):137 [FREE Full text] [doi: 10.1186/s12913-017-2079-5] [Medline: 28196489]
6. Keegan AD. Hospital bed occupancy: more than queuing for a bed. Med J Aust 2010 Sep 06;193(5):291-293. [doi: 10.5694/j.1326-5377.2010.tb03910.x] [Medline: 20819049]
7. Kaier K, Mutters N, Frank U. Bed occupancy rates and hospital-acquired infections--should beds be kept empty? Clin Microbiol Infect 2012 Oct;18(10):941-945 [FREE Full text] [doi: 10.1111/j.1469-0691.2012.03956.x] [Medline: 22757765]
8. Anderson D. The impact of resource management on hospital efficiency and quality of care. University of Maryland. 2013. URL: https://api.drum.lib.umd.edu/server/api/core/bitstreams/7ec54849-e2d2-449b-9a9a-f506b429834b/content [accessed 2024-02-21]
9. Kutafina E, Bechtold I, Kabino K, Jonas SM. Recursive neural networks in hospital bed occupancy forecasting. BMC Med Inform Decis Mak 2019 Mar 07;19(1):39 [FREE Full text] [doi: 10.1186/s12911-019-0776-1] [Medline: 30845940]
10. Baas S, Dijkstra S, Braaksma A, van Rooij P, Snijders FJ, Tiemessen L, et al. Real-time forecasting of COVID-19 bed occupancy in wards and Intensive Care Units. Health Care Manag Sci 2021 Jun 25;24(2):402-419 [FREE Full text] [doi: 10.1007/s10729-021-09553-5] [Medline: 33768389]
11. Esteban C, Staeck O, Baier S, Yang Y, Tresp V. Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks. 2016 Presented at: 2016 IEEE International Conference on Healthcare Informatics (ICHI); October 4-7, 2016; Chicago, IL p. 93-101. [doi: 10.1109/ICHI.2016.16]
12. Mackay M, Lee M. Using Compartmental Models to Predict Hospital Bed Occupancy. Semantic Scholar. URL: https://www.semanticscholar.org/paper/Using-Compartmental-Models-to-Predict-Hospital-Bed-Mackay-Lee/f2b32e60df7dd80bd48e8ccd0af920134d1452c5?p2df [accessed 2024-02-21]
13. Littig SJ, Isken MW. Short term hospital occupancy prediction. Health Care Manag Sci 2007 Feb 28;10(1):47-66. [doi: 10.1007/s10729-006-9000-9] [Medline: 17323654]
14. Kumar A, Mo J. Models for Bed Occupancy Management of a Hospital in Singapore. In: Proceedings of the 2010 International Conference on Industrial Engineering and Operations Management. 2010 Presented at: 2010 International Conference on Industrial Engineering and Operations Management; January 9-10, 2010; Dhaka, Bangladesh.
15. Seematter-Bagnoud L, Fustinoni S, Dung D, Santos-Eggimann B, Koehn V, Bize R, et al. Comparison of different methods to forecast hospital bed needs. European Geriatric Medicine 2015 Jun;6(3):262-266. [doi: 10.1016/j.eurger.2014.09.004]
16. Farmer RD, Emami J. Models for forecasting hospital bed requirements in the acute sector. J Epidemiol Community Health 1990 Dec 01;44(4):307-312 [FREE Full text] [doi: 10.1136/jech.44.4.307] [Medline: 2277253]
17. Kim K, Lee C, O'Leary KJ, Rosenauer S, Mehrotra S. Predicting Patient Volumes in Hospital Medicine: A Comparative Study of Different Time Series Forecasting Methods. Northwestern University. URL: https://www.mcs.anl.gov/~kibaekkim/ForecastingHospitalMedicine.pdf [accessed 2024-02-21]
18. Rosenbaum L. Facing Covid-19 in Italy - Ethics, Logistics, and Therapeutics on the Epidemic's Front Line. N Engl J Med 2020 May 14;382(20):1873-1875. [doi: 10.1056/NEJMp2005492] [Medline: 32187459]
19. Bouhamed H, Hamdi M, Gargouri R. Covid-19 Patients' Hospital Occupancy Prediction During the Recent Omicron Wave via some Recurrent Deep Learning Architectures. Int. J. Comput. Commun. Control 2022 Mar 14;17(3):4697. [doi: 10.15837/ijccc.2022.3.4697]
20. Bekker R, Uit Het Broek M, Koole G. Modeling COVID-19 hospital admissions and occupancy in the Netherlands. Eur J Oper Res 2023 Jan 01;304(1):207-218 [FREE Full text] [doi: 10.1016/j.ejor.2021.12.044] [Medline: 35013638]
21. Farcomeni A, Maruotti A, Divino F, Jona-Lasinio G, Lovison G. An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. Biom J 2021 Mar 30;63(3):503-513 [FREE Full text] [doi: 10.1002/bimj.202000189] [Medline: 33251604]
22. Caro JJ, Möller J, Santhirapala V, Gill H, Johnston J, El-Boghdadly K, et al. Predicting Hospital Resource Use During COVID-19 Surges: A Simple but Flexible Discretely Integrated Condition Event Simulation of Individual Patient-Hospital Trajectories. Value Health 2021 Nov;24(11):1570-1577 [FREE Full text] [doi: 10.1016/j.jval.2021.05.023] [Medline: 34711356]
23. Schmidt R, Geisler S, Spreckelsen C. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. BMC Med Inform Decis Mak 2013 Jan 07;13:3 [FREE Full text] [doi: 10.1186/1472-6947-13-3] [Medline: 23289448]
24. Hancock WM, Walter PF. The use of computer simulation to develop hospital systems. SIGSIM Simul. Dig 1979 Jul;10(4):28-32. [doi: 10.1145/1102815.1102819]

25. Shahpori R, Gibney N, Guebert N, Hatcher C, Zygun D. An on-line dashboard to facilitate monitoring of provincial ICU bed occupancy in Alberta, Canada. JHA 2013 Oct 10;3(1):47. [doi: 10.5430/jha.v3n1p47]
26. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct 9;323(6088):533-536. [doi: 10.1038/323533a0]
27. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
28. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans. Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093]
29. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv. 2014. URL: https://arxiv.org/abs/1409.0473 [accessed 2024-02-21]
30. Luong MT, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. arXiv. 2015. URL: https://arxiv.org/abs/1508.04025 [accessed 2024-02-21]
31. Leontjeva A, Kuzovkin I. Combining Static and Dynamic Features for Multivariate Sequence Classification. 2016 Presented at: 2016 IEEE 3rd International Conference on Data Science and Advanced Analytics (DSAA); October 17-19, 2016; Montreal, QC p. 21-30. [doi: 10.1109/DSAA.2016.10]
32. Vincent J, Creteur J. Ethical aspects of the COVID-19 crisis: How to deal with an overwhelming shortage of acute beds. Eur Heart J Acute Cardiovasc Care 2020 Apr 29;9(3):248-252 [FREE Full text] [doi: 10.1177/2048872620922788] [Medline: 32347745]
33. Vakharia V, Shah M, Nair P, Borade H, Sahlot P, Wankhede V. Estimation of Lithium-ion Battery Discharge Capacity by Integrating Optimized Explainable-AI and Stacked LSTM Model. Batteries 2023 Feb 09;9(2):125. [doi: 10.3390/batteries9020125]
34. Joshi S, Owens JA, Shah S, Munasinghe T. Analysis of Preprocessing Techniques, Keras Tuner, and Transfer Learning on Cloud Street image data. 2021 Presented at: IEEE International Conference on Big Data (Big Data); December 15-18, 2021; Orlando, FL. [doi: 10.1109/BigData52589.2021.9671878]
35. Jung Y. Multiple predicting K-fold cross-validation for model selection. Journal of Nonparametric Statistics 2017 Nov 21;30(1):197-215. [doi: 10.1080/10485252.2017.1404598]
36. Nair P, Vakharia V, Borade H, Shah M, Wankhede V. Predicting Li-Ion Battery Remaining Useful Life: An XDFM-Driven Approach with Explainable AI. Energies 2023 Jul 31;16(15):5725. [doi: 10.3390/en16155725]
37. Seo H, Ahn I, Gwon H, Kang HJ, Kim Y, Cho HN, et al. Prediction of hospitalization and waiting time within 24 hours of emergency department patients with unstructured text data. Health Care Manag Sci 2023 Nov 03:09660-5. [doi: 10.1007/s10729-023-09660-5] [Medline: 37921927]
38. Deng A. Time series cross validation: A theoretical result and finite sample performance. Economics Letters 2023 Dec;233:111369. [doi: 10.1016/j.econlet.2023.111369]
39. Martinez DA, Kane EM, Jalalpour M, Scheulen J, Rupani H, Toteja R, et al. An Electronic Dashboard to Monitor Patient Flow at the Johns Hopkins Hospital: Communication of Key Performance Indicators Using the Donabedian Model. J Med Syst 2018 Jun 18;42(8):133. [doi: 10.1007/s10916-018-0988-4] [Medline: 29915933]
40. Gartner D, Padman R. Machine learning for healthcare behavioural OR: Addressing waiting time perceptions in emergency care. Journal of the Operational Research Society 2019 Apr 15;71(7):1087-1101. [doi: 10.1080/01605682.2019.1571005]
41. Welch SJ. Twenty years of patient satisfaction research applied to the emergency department: a qualitative review. Am J Med Qual 2010 Dec 04;25(1):64-72. [doi: 10.1177/1062860609352536] [Medline: 19966114]

Abbreviations

AI: artificial intelligence
AMC: Asan Medical Center
Bi-LSTM: bidirectional long short-term memory
BOR: bed occupancy rate
DL: deep learning
DNN: deep neural network
LeakyReLU: leaky rectified linear unit
LSTM: long short-term memory
MAE: mean square error
ML: machine learning
R3D: room 3 days
R7D: room 7 days
RBOR: room bed occupancy rate
RMSE: root mean square error
RNN: recurrent neural network
RS3D: room static 3 days

RS7D: room static 7 days
W7D: ward 7 days
W30D: ward 30 days
WBOR: ward bed occupancy rate

Edited by C Lovis; submitted 05.10.23; peer-reviewed by V Vakharia, T Leili; comments to author 10.11.23; revised version received 20.12.23; accepted 16.02.24; published 21.03.24.

Please cite as:
Seo H, Ahn I, Gwon H, Kang H, Kim Y, Choi H, Kim M, Han J, Kee G, Park S, Ko S, Jung H, Kim B, Oh J, Jun TJ, Kim YH
Forecasting Hospital Room and Ward Occupancy Using Static and Dynamic Information Concurrently: Retrospective Single-Center Cohort Study
JMIR Med Inform 2024;12:e53400
URL: https://medinform.jmir.org/2024/1/e53400
doi:10.2196/53400
PMID:38513229

Original Paper

Explainable AI Method for Tinnitus Diagnosis via Neighbor-Augmented Knowledge Graph and Traditional Chinese Medicine: Development and Validation Study

Ziming Yin[1*], Prof Dr; Zhongling Kuang[1*], MSc; Haopeng Zhang[2], MD; Yu Guo[2], MD; Ting Li[1], BEng; Zhengkun Wu[1], BEng; Lihua Wang[2], MD

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China

[2]Department of Otolaryngology, Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Lihua Wang, MD
Department of Otolaryngology
Shanghai Municipal Hospital of Traditional Chinese Medicine
Shanghai University of Traditional Chinese Medicine
274 Zhijiang Middle Road, Jing'an District
Shanghai, 200071
China
Phone: 86 18116013561
Email: lihuahanhan@126.com

Abstract

Background: Tinnitus diagnosis poses a challenge in otolaryngology owing to an extremely complex pathogenesis, lack of effective objectification methods, and factor-affected diagnosis. There is currently a lack of explainable auxiliary diagnostic tools for tinnitus in clinical practice.

Objective: This study aims to develop a diagnostic model using an explainable artificial intelligence (AI) method to address the issue of low accuracy in tinnitus diagnosis.

Methods: In this study, a knowledge graph–based tinnitus diagnostic method was developed by combining clinical medical knowledge with electronic medical records. Electronic medical record data from 1267 patients were integrated with traditional Chinese clinical medical knowledge to construct a tinnitus knowledge graph. Subsequently, weights were introduced, which measured patient similarity in the knowledge graph based on mutual information values. Finally, a collaborative neighbor algorithm was proposed, which scored patient similarity to obtain the recommended diagnosis. We conducted 2 group experiments and 1 case derivation to explore the effectiveness of our models and compared the models with state-of-the-art graph algorithms and other explainable machine learning models.

Results: The experimental results indicate that the method achieved 99.4% accuracy, 98.5% sensitivity, 99.6% specificity, 98.7% precision, 98.6% F_1-score, and 99% area under the receiver operating characteristic curve for the inference of 5 tinnitus subtypes among 253 test patients. Additionally, it demonstrated good interpretability. The topological structure of knowledge graphs provides transparency that can explain the reasons for the similarity between patients.

Conclusions: This method provides doctors with a reliable and explainable diagnostic tool that is expected to improve tinnitus diagnosis accuracy.

(JMIR Med Inform 2024;12:e57678) doi:10.2196/57678

KEYWORDS

knowledge graph; syndrome differentiation; tinnitus; traditional Chinese medicine; explainable; ear; audiology; TCM; algorithm; diagnosis; AI; artificial intelligence

Introduction

Tinnitus is a common refractory disease in the field of otolaryngology, and its diagnosis has always been a cutting-edge research topic in audiology. With changes in the social environment and an accelerated pace of life, an increasing number of patients, particularly among the younger generation, have sought medical assistance for tinnitus as their primary complaint in the last decade. Globally, approximately 14% (95% CI 0.8%-1.6%) of adults are affected by tinnitus [1,2], which can cause stress, anxiety, and depression [3]. Distress and hearing impairment brought on by the disease can affect cognitive abilities and lead to suicidal tendencies in severe cases, greatly affecting the work and daily lives of patients [4].

The pathogenesis of tinnitus is extremely complex and not fully understood. Currently, no effective objectification methods are available. Traditional Chinese medicine (TCM) classifies tinnitus into 5 different syndrome patterns: wind fire attacking internally (WFAI), liver fire bearing upward (LFBU), phlegm fire stagnation internally (PFSI), Qi deficiency of the spleen and stomach (QDSS), and kidney essence deficiency (KED). The diagnosis of tinnitus remains a challenge in medical science because it is influenced by several complex factors [5,6], including individual differences among patients and atypical symptom presentations. Clinical diagnosis relies heavily on the personal knowledge and clinical experience of doctors, thereby introducing subjectivity, uncertainty, and ambiguity. Consequently, achieving a high tinnitus diagnostic accuracy becomes difficult. Therefore, tinnitus diagnosis remains an urgent issue requiring further exploration and resolution by medical researchers.

Previous studies have focused on the use of artificial intelligence (AI) to assist doctors in diagnosing tinnitus and improving diagnostic accuracy. Liu et al [7] proposed a meta-learning method based on lateral perception for cross–data set tinnitus diagnosis. Sun et al [8] used a support vector machine classifier to distinguish between patients with tinnitus and healthy individuals. Shoushtarian et al [9] used a naive Bayes algorithm to classify patients with tinnitus and control groups. Sanders et al [10] used a spiking neural network model to classify patients with tinnitus into 2 groups based on different classification criteria. Manta et al [11] used clinical data and patient features to build a machine learning (ML) model for classifying the degree of tinnitus-related distress in individuals and their ears. Allgaier et al [12] used a gradient-boosting engine to classify transient tinnitus. Rodrigo et al [13] used a decision tree model to identify variables related to the success of internet-based cognitive behavioral therapy for tinnitus. Liu et al [14] used a support vector machine model to explore cortical or subcortical morphological neuroimaging biomarkers that effectively distinguished patients with tinnitus from healthy individuals. Niemann et al [15] proposed a LASSO model to predict the severity of depression in patients with tinnitus. Although previous studies have achieved success using their respective data sets, the developed ML- or deep learning–based methods are entirely data-driven modeling approaches that do not make full use of existing medical knowledge. Models built using such methods are equivalent to “black boxes” for doctors, lack interpretability, and are not conducive to clinical promotion and application.

In this study, the aim is to incorporate clinical medical knowledge into a diagnostic model, enabling the integration of knowledge and data for interpretable results. Knowledge graph–based modeling methods offer solutions to such issues by using a novel knowledge representation format that connects entities and concepts in an objective world using semantic relationships. Such methods offer reasoning and interpretability that are highly sought after by both medical practitioners and academia. Li et al [16] used a knowledge graph to predict diabetic macular edema, overcoming the limitations of traditional ML and data-mining techniques that deal with missing feature values. Zhou et al [17] used 124 medical records to construct a knowledge graph for recommending hypertension medication. Lyu et al [18] created a knowledge graph for diabetic nephropathy diagnosis using patient data. Lin et al [19] extracted knowledge from medical texts and historical prescription data to construct a medical knowledge graph and accurately detect clinical prescription risks. Recently, knowledge graph applications have expanded to TCM; for instance, Yang et al [20] built a knowledge graph to extract medical information from TCM case records. Xie et al [21] constructed a knowledge graph using ancient Chinese medical books to infer symptoms and syndromes. Yang et al [22] used electronic medical records (EMRs) to build a knowledge graph, transforming TCM diagnostic issues into multilabel classification problems. Lan et al [23] integrated knowledge graphs with graph neural networks to introduce graph-based supervised contrastive learning, effectively enabling the classification of TCM texts. However, no previous studies have used knowledge graphs in the complex medical field of tinnitus diagnosis. Therefore, this study focuses on knowledge graph technology to assist doctors in tinnitus diagnosis and improve diagnostic accuracy.

This paper aims to establish a comprehensive knowledge graph in TCM specifically tailored for tinnitus. Leveraging this knowledge graph, we propose a novel method for calculating patient similarity. This method takes into account the weighting of symptom-syndrome type relationships, thereby facilitating the inference of syndrome types in patients with tinnitus according to TCM principles. By implementing this approach, clinicians can increase the accuracy of tinnitus diagnosis within the realm of TCM.

In general, we make several noteworthy contributions as follows:

- We propose a method for tinnitus knowledge graph construction based on heterogeneous patient EMRs and TCM clinical knowledge.
- We introduce weights to measure patient similarity into the tinnitus knowledge graph using a method based on prior probabilities and mutual information values.
- A collaborative neighbor algorithm that uses patient similarity scores to obtain recommended diagnostic results is proposed to assist doctors in understanding the model-generated conclusions, thereby improving the accuracy of tinnitus diagnosis.

Methods

Patients

For this study, we collected the EMRs of 1267 patients with tinnitus who visited the ear, nose, and throat departments of 11 medical institutions in Shanghai, China, from November 2019 to July 2023. The inclusion criteria included (1) tinnitus as the primary complaint and (2) the ability to communicate normally. The exclusion criteria included (1) objective tinnitus, (2) nonotogenic tinnitus caused by factors such as endocrine and blood disorders, (3) tinnitus caused by head or ear trauma, and (4) difficulties in communication or severe psychiatric history that could hinder follow-up compliance. After screening the data for quality, 1265 cases were included for further analysis.

The clinical EMR data set recorded medical data of real patients including the relationship between patient symptoms and disease, which was crucial for disease diagnosis. The data set contained patient information such as age, sex, inducement, medical history, tinnitus sound, accompanying symptoms, tongue coating, pulse condition, TCM syndrome differentiation, and sleep status. Each patient had a clear diagnosis that could be classified into 1 of 5 categories: WFAI, LFBU, PFSI, QDSS, and KED. Statistical data are presented in Figures 1-4 .

Figure 1. Age distribution of different syndromes by sex. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.

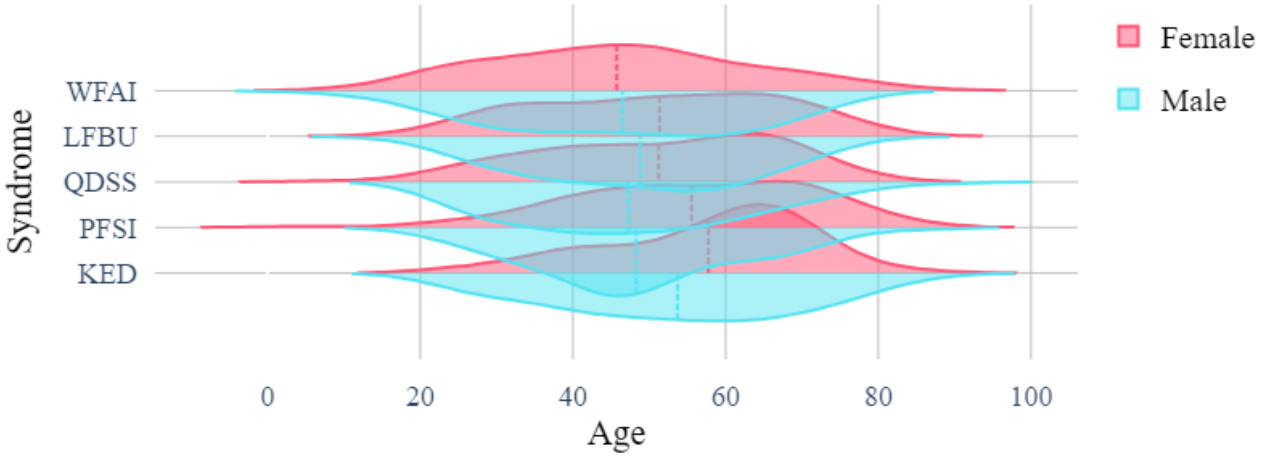


Figure 2. The tongue body distribution of different syndrome types. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.

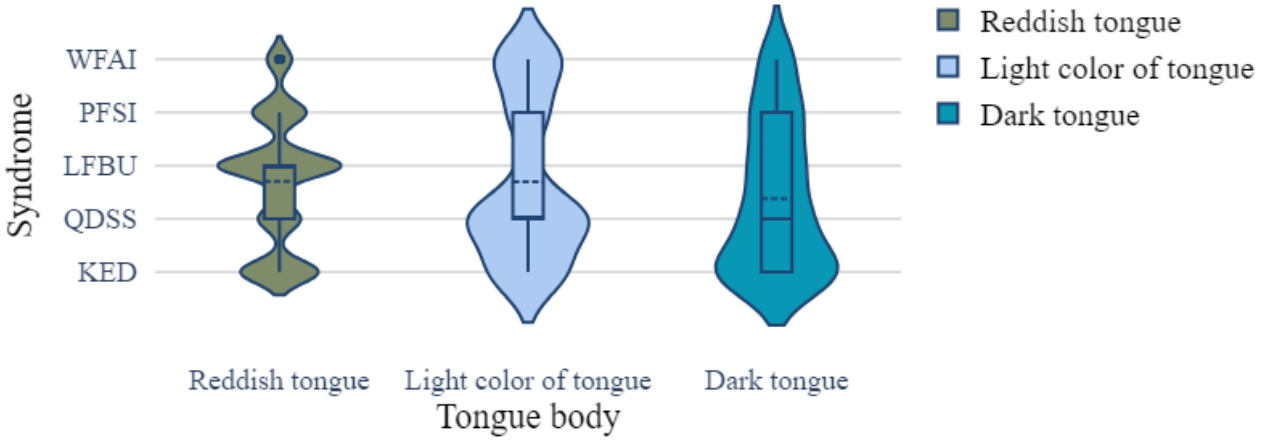


Figure 3. The tongue fur distribution of different syndrome types. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.

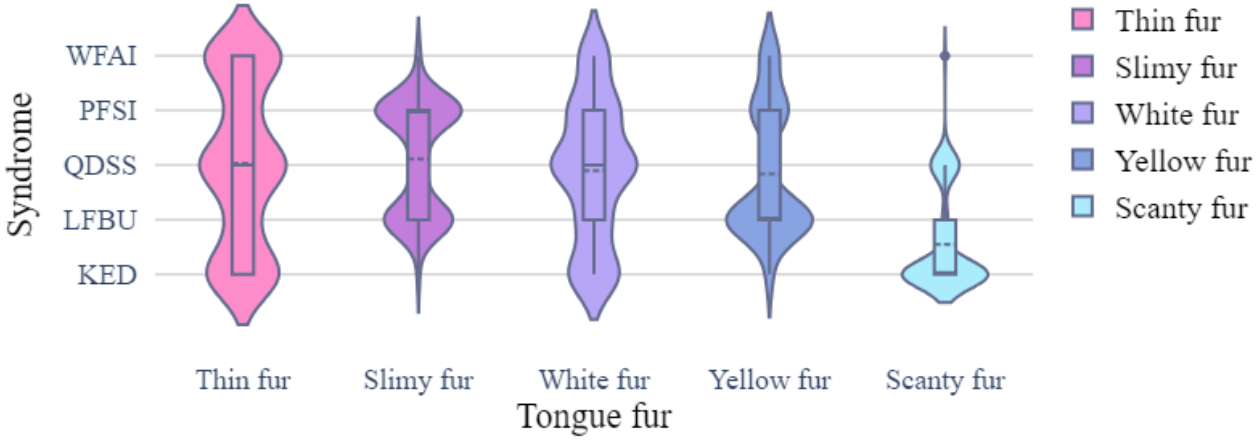
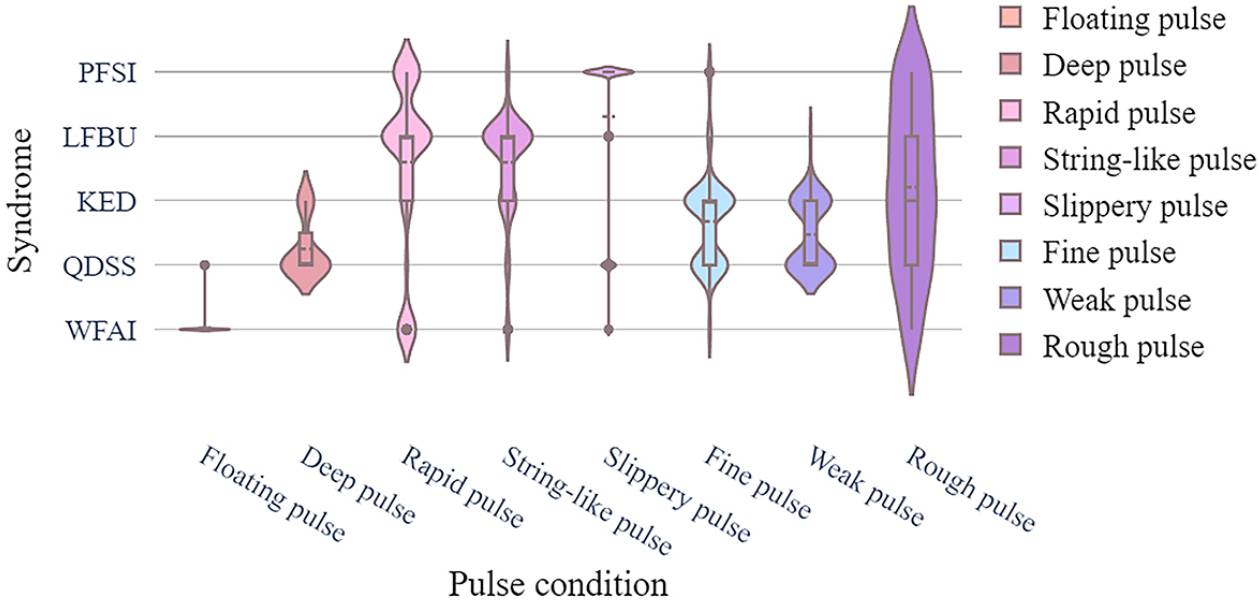


Figure 4. The pulse condition distribution of different syndrome types. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.



Ethical Considerations

This study's protocol was approved by the ethics committee of the Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai, China (2021SHL-KY-70).

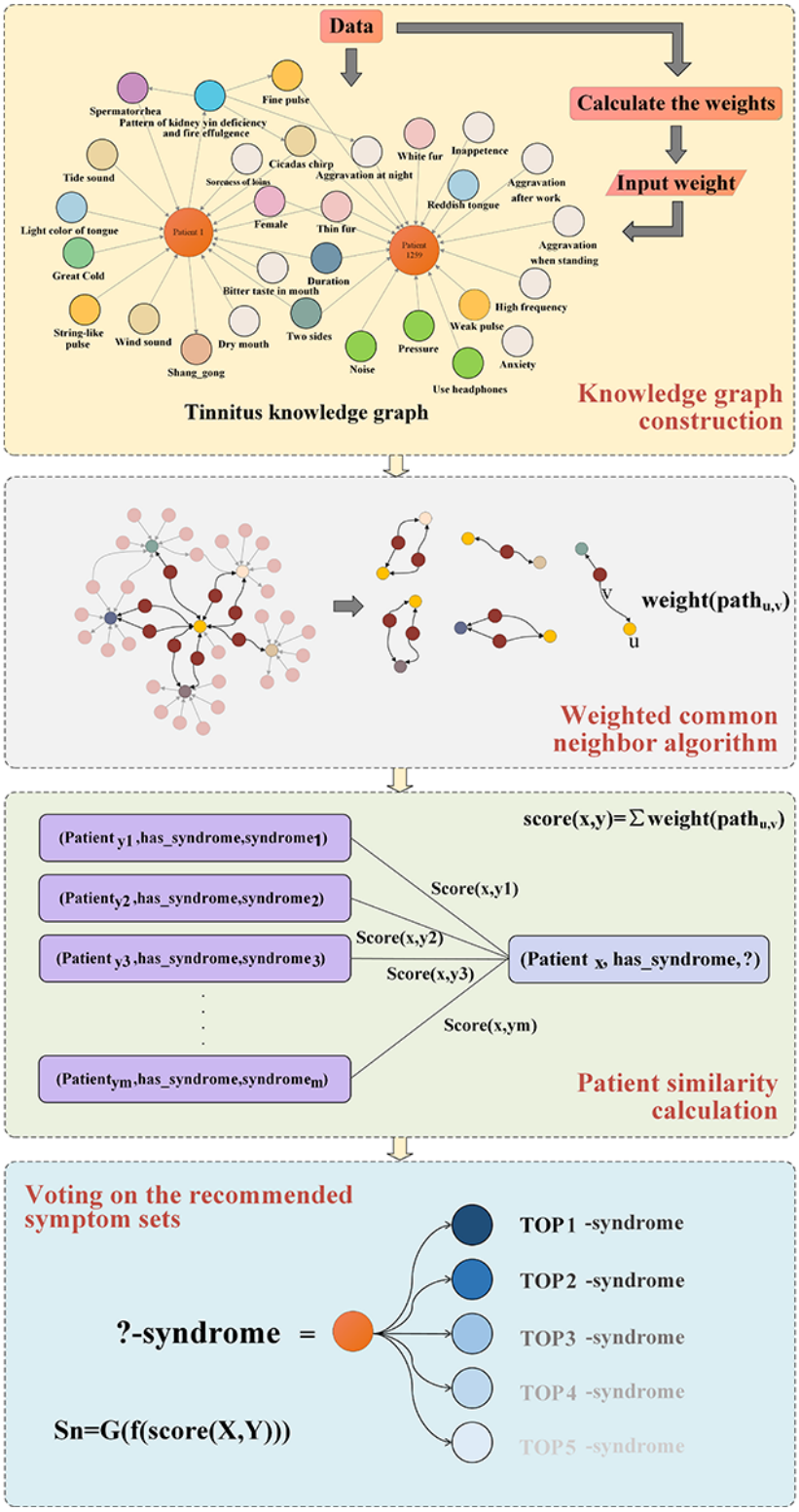
The data was anonymized in order to protect patient privacy. Patients could receive free examinations and treatments throughout the entire process, so no compensation was provided.

Clinical Decision Support for Tinnitus

Overview

To integrate patient EMRs with diagnostic knowledge from TCM textbooks, we constructed a knowledge graph using a combined "top-down" and "bottom-up" approach [24]. First, a patient-centered knowledge graph was developed using EMRs. Then, the knowledge graph was enriched with tinnitus diagnostic knowledge from TCM textbooks. Finally, we used a mutual information–based weight calculation method to enhance the knowledge graph by fusing patient case data with diagnostic knowledge. The resulting knowledge graph simulated the diagnostic reasoning processes of experienced physicians. The entire method consisted of three steps: (1) building a weighted tinnitus knowledge graph, (2) finding and scoring common neighbors, and (3) predicting syndrome patterns based on patient similarity. The overall framework is illustrated in Figure 5.

Figure 5. Overall framework of the proposed method.



Knowledge Graph of Tinnitus Based on Heterogeneous Sources

In response to the diagnostic needs of tinnitus in TCM, the ontology structure of a tinnitus medical knowledge graph should revolve around symptoms, syndrome patterns, diseases, drugs, and treatment methods. For this study, we extracted such common concepts from expert-reviewed EMRs and classic medical textbooks, constructed a conceptual knowledge system, and built a top-level ontology structure. Natural language processing techniques [25] were used to extract entities and relationships from the patient EMRs based on a defined conceptual knowledge system for tinnitus. By applying certain rules and conducting string matching within the text, we extracted 15 and 10 categories of entities and relationships from the 1265 EMR records, respectively. Once the entity types and hierarchy were determined, we embedded the data into the conceptual knowledge system and established a patient-centric

tinnitus knowledge graph in the form of a triple, which maximized the retention of both explicit and implicit diagnostic information.

Furthermore, we enhanced the constructed tinnitus knowledge graph using knowledge extracted from authoritative medical textbooks to supplement tinnitus knowledge information that was not fully expressed in EMRs. Together with the EMR knowledge graph, a complete tinnitus knowledge graph was developed. The knowledge we selected came from 2 classic Chinese medicine textbooks [26,27], from which we extracted basic concepts related to tinnitus including TCM syndromes, prescriptions, Chinese medicinal herbs, and treatment methods to construct the TCM knowledge graph.

Heterogeneous Knowledge Fusion

Redundancy in the entities and relationships extracted from heterogeneous sources was observed owing to the different sources of data and knowledge. Therefore, knowledge fusion was required. First, data normalization and entity alignment were performed to standardize the named entities extracted from multiple data sources. The entities were associated using string-matching and similarity-calculation methods. As entity and attribute texts were relatively short, a lower similarity threshold was more appropriate; therefore, the similarity judgment threshold was set as 0.6 to prevent errors and omissions. The entity similarity calculation results are listed in Table 1. As the knowledge graph was established in Chinese, we calculated the similarity of the Chinese strings.

Then, a matching path was built from the tinnitus ontology–based knowledge graph entity to the EMR-based knowledge graph entity. Patient data were linked to diagnostic knowledge through an ontology. The 2 knowledge graphs were linked by unifying entities with duplicate meanings in the 2 graphs. Manual verification was performed to ensure the accuracy of the knowledge graph. The specific method is illustrated in Figure 6. Finally, the tinnitus knowledge graph consisted of 1247 entities and 9234 relationships.

Table 1. Entity similarity calculation results.

Standardized and ambiguous entities (Chinese)	Similarity
WFAI[a]	
风热外侵证 (wind-heat invasion syndrome)	0.8
风热外犯证 (wind-heat exterior syndrome)	0.6
风热外侵证 (wind-heat exterior assault syndrome)	0.8
LFBU[b]	
肝火上炎证 (liver fire flaming upward syndrome)	0.8
肝热上扰证 (liver heat disturbing upward syndrome)	0.8
肝火上扰清窍证 (liver fire disturbing upward and disturbing clearing orifices syndrome)	0.83
QDSS[c]	
脾胃虚证 (spleen and stomach deficiency syndrome)	0.89
脾胃虚弱证 (spleen and stomach weakness syndrome)	0.8
PFSI[d]	
痰火壅结证 (phlegm-fire concretions syndrome)	0.8
KED[e]	
肾精不足证 (kidney essence insufficiency syndrome)	0.6
肾精亏虚证 (kidney essence deficiency syndrome)	0.8
肾虚精亏证 (kidney deficiency and essence deficiency syndrome)	0.99
肾精亏耗证 (kidney essence consumption syndrome)	0.8

[a]WFAI: wind fire attacking internally.

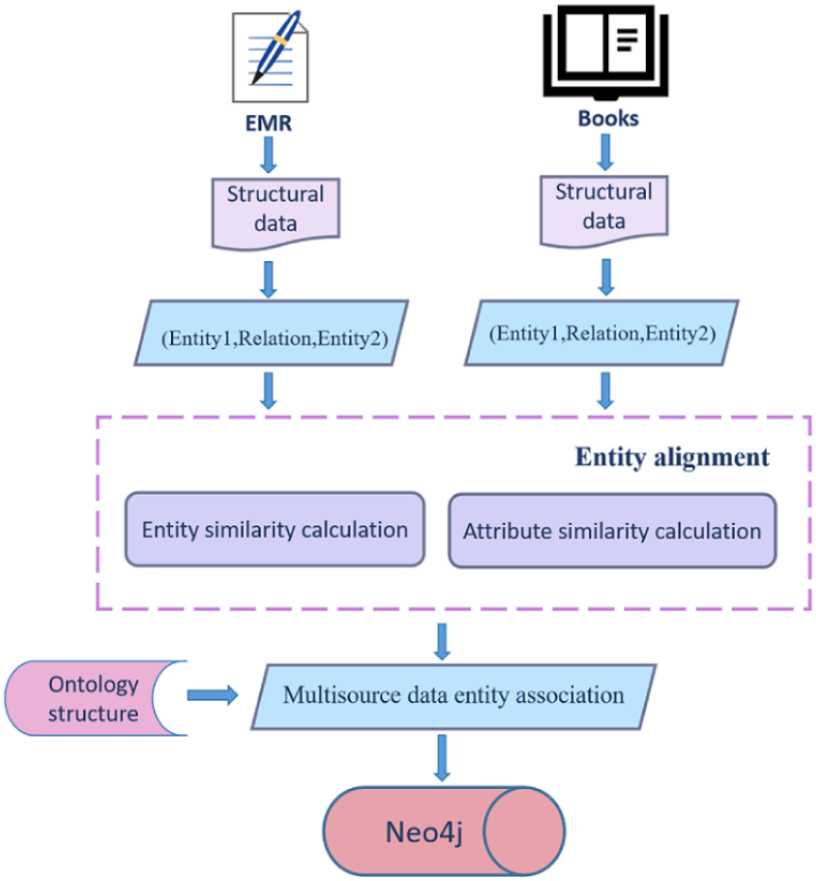
[b]LFBU: liver fire bearing upward.

[c]QDSS: Qi deficiency of the spleen and stomach.

[d]PFSI: phlegm fire stagnation internally.

[e]KED: kidney essence deficiency.

Figure 6. Tinnitus knowledge graph fusion flowchart. EMR: electronic medical record.



Calculation of Knowledge Graph Relationship Weights Based on Mutual Information

Considering the varying importance of different entities for different syndrome patterns, the imbalance in data categories, and the varying amount of information carried by symptoms, the calculation of weights required consideration of entities' importance for diagnostic pattern identification and information content carried by the entities themselves. The data used for weight calculation were derived from real clinical case data used for constructing the knowledge graph. First, the mutual information value (w_{if}) possessed by each entity was obtained using the mutual information method. The obtained value represented the extent to which a variable could acquire diagnostic pattern information.

For a given set of entities $X = \{x_1, x_2, ..., x_n\}$ with corresponding probabilities $P = \{p_1, p_2, ..., p_n\}$, the target variable to be measured was the diagnostic pattern Y. By calculating the overall entropy $H()$, conditional entropy $H(Y|X)$, and mutual information value $Gain(S,x)$, the degree to which the diagnostic pattern was determined based on the entity values or the weight value w_{if} of the entity was calculated. The calculations were performed using equations 1-3.

(1)

(2)

$$w_{if} = Gain(Y,X) = H(Y) - H(Y|X) \quad (3)$$

Further, the feature weights were calculated based on the syndrome patterns under the prior conditions. The probability of each symptom appearing under different syndrome patterns was obtained using statistical methods such as:

$$w_{sd} = p(sym_i|sd_j) \quad (4)$$

where $sym = \{sym_1, sym_2, ..., sym_n\}$ represents the symptom set and $sd = \{sd_1, sd_2, ..., sd_m\}$ represents the diagnostic pattern set. Finally, the edge weight from node u to node v was defined using equation 5.

$$Weight(u,v) = w_{if} + w_{sd} \quad (5)$$

The weights of various symptoms under different syndrome patterns are presented in Table 2.

Table 2. Partial weight value of symptom-syndrome type.

Symptom	Weight
KED[a]	
Spermatorrhea	1.435
Soreness of loins	1.4213
Dreaminess	1.4104
Wake up early in the morning	1.3868
Deficiency and insomnia	1.3856
Aggravation at night	1.167
Cicadas chirp	1.1559
Fine pulse	1.1448
Scanty fur	0.7142
Duration	0.6991
LFBU[b]	
Irritable	1.2376
Restlessness and insomnia	1.1196
Wind sound	1.0271
String-like pulse	1.0056
Tide sound	1.0030
Yellow fur	0.9118
Reddish tongue	0.8992
Duration	0.7036
Dry mouth	0.6855
Bitter taste in mouth	0.6558
PFSI[c]	
Tastelessness	1.1953
Dizziness and heaviness	1.1488
Aural fullness	1.1216
Ear distension	1.0899
Slippery pulse	0.9121
Slimy fur	0.8342
Duration	0.7113
Yellow fur	0.6895
Hearing loss	0.6495
Reddish tongue	0.6440
WFAI[d]	
Cold or rhinitis	1.2089
Tinnitus onset within a month	1.1398
Low voice	1.1398
Thin fur	1.0286
Floating pulse	0.9563
Duration	0.6903
Light color of tongue	0.6664

Symptom	Weight
Yellow fur	0.5082
Hearing loss	0.5032
Dreaminess	0.4993
QDSS[e]	
Feeling emptiness in ear	1.2615
Aggravation after work	1.1813
Aggravation when standing up	1.1562
Fine pulse	1.0782
Duration	0.7370
Thin fur	0.7022
Light color of tongue	0.6745
Anxiety	0.6596
Hearing loss	0.6444
Dreaminess	0.4865

[a]KED: kidney essence deficiency.

[b]LFBU: liver fire bearing upward.

[c]PFSI: phlegm fire stagnation internally.

[d]WFAI: wind fire attacking internally.

[e]ODSS: Qi deficiency of the spleen and stomach.

Patient Similarity Scoring Based on Weighted Common Neighbor Algorithm

By transforming the TCM syndrome diagnostic problem into a prediction problem of linked patient nodes to TCM syndrome nodes, the similarity between 2 patients was calculated to obtain TCM syndrome similarity. For 2 patients, the higher the similarity, the greater the likelihood of having the same diagnostic result. This study measured the similarity using common features. In the knowledge graph, the higher the number of common neighbors to 2 patient nodes, the greater the likelihood of them belonging to the same community (linked to the same TCM syndrome node). The common neighbor graph of patients with different TCM syndromes is shown in Figure 7, where fewer common neighbors were observed. The common neighbor graph of patient 1 and patient 2 with the same TCM syndrome is shown in Figure 8, where more common neighbors were observed; however, different nodes had different importance. In TCM, the importance of pulse condition is greater than that of tinnitus duration while diagnosing tinnitus. The edge weight values of continuous tinnitus and thin pulse-to-kidney deficiency syndrome were 0.6991 and 1.1448, respectively, as shown in Figure 7; however, even for the same pulse condition, the importance varied for different TCM syndromes. In Figure 8, the edge weight values of thin pulse to QDSS and KED syndromes were 1.078 and 1.1447, respectively. Therefore, considering the edge weights of common neighbors to the patient nodes and calculating the score of common neighbors based on the edge weight values were essential when counting the number of common neighbors between patient nodes.

The similarity scoring function between patients x and y was defined by equation 6.

(6)

where $X = \{u_1, u_2, ..., u_m\}$ and $Y = \{v_1, v_2, ..., v_n\}$ represent the sets of neighboring nodes for patients x and y, respectively; $Path_{u,h,v} = (u, h, v)$ denotes the 2-hop path from node u to node v, where h represents the common neighbor of nodes u and v; $Path_{u,h} = (u, h)$ represents the path from node u to the common neighbor h; and $weight(path_{u,h})$ indicates the weight of the path.

When 2 paths with a hop count of 2 between the patient nodes existed, the weights of the paths were calculated to obtain a similarity score list for the patients. The list was then sorted in descending order, and the top 20 patient node syndromes with the highest scores were counted, which represented the most frequently occurring syndrome. Finally, the recommended syndrome was obtained.

$$S_n = G(f_{20}(\text{score}(X,Y))) \quad (7)$$

where G denotes a frequency-counting method in which X and Y represent sets of patient nodes. $f_{20}()$ was used to obtain the top 20 patient syndromes based on the scores.

Figure 7. Sketch map of common neighbors between different syndromes.

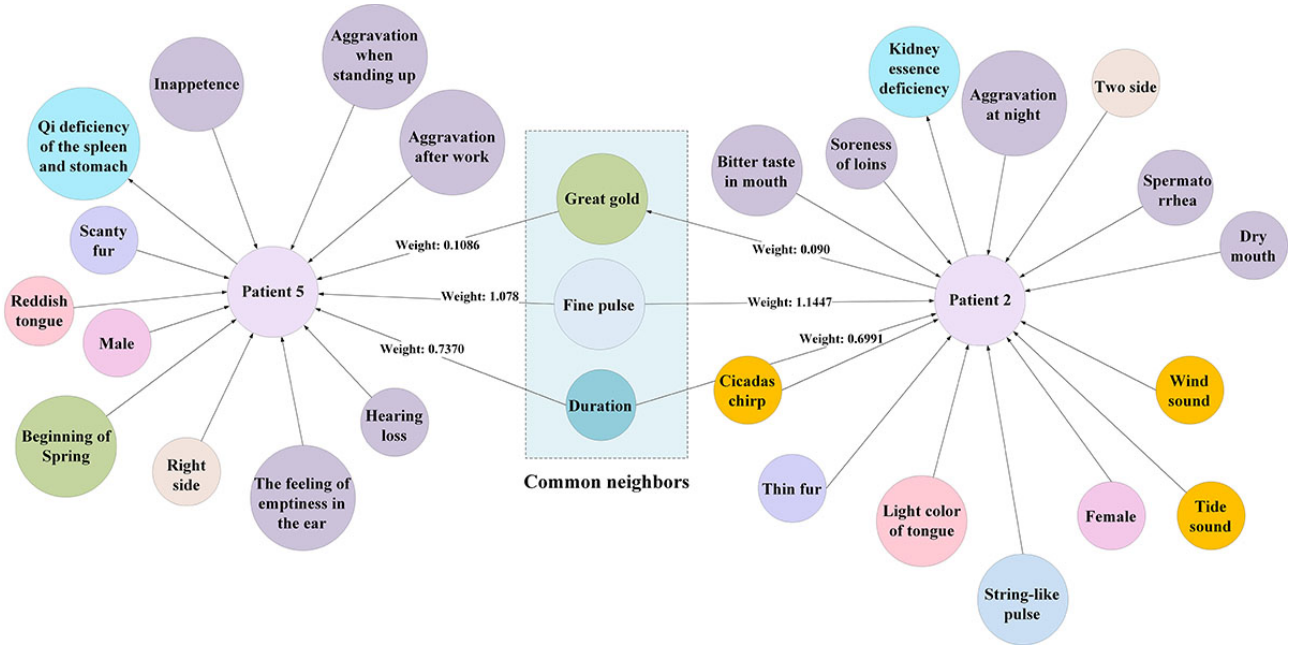
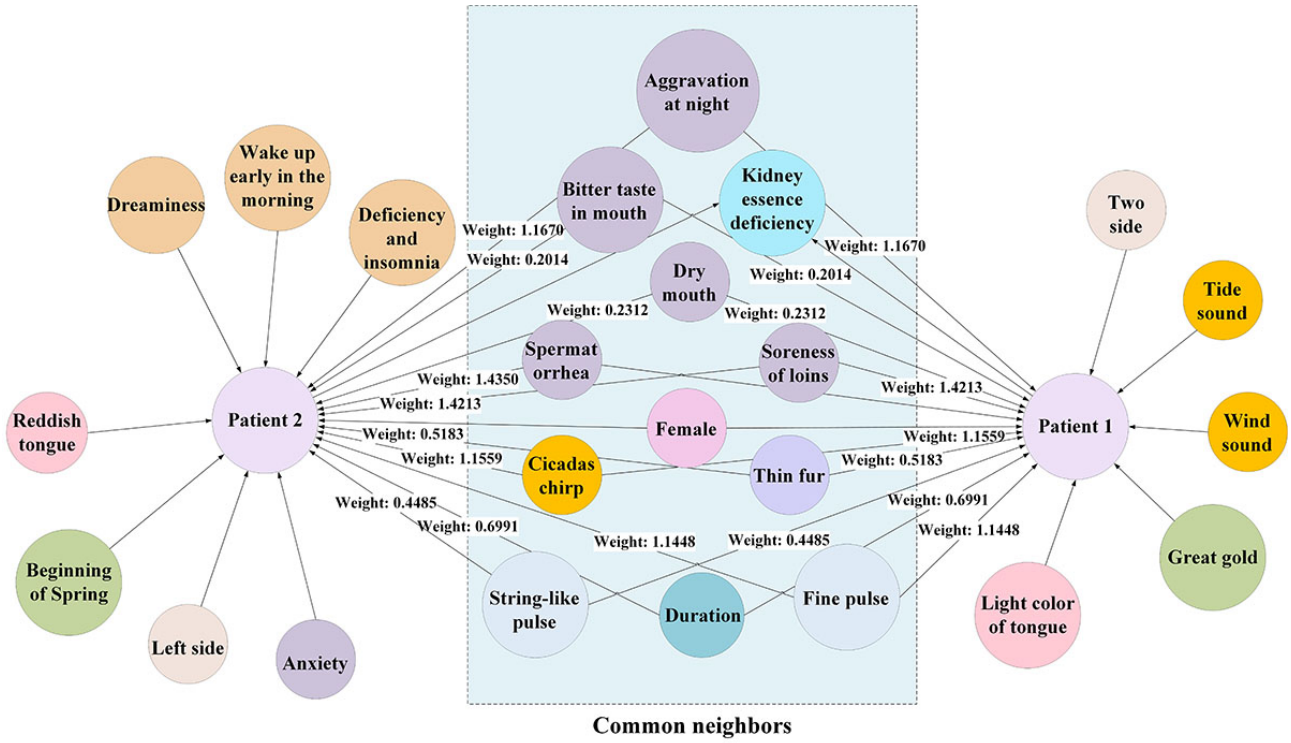


Figure 8. Sketch map of common neighbors between same syndromes.



Experimental Design

In total, 2 experiments were conducted to verify the effectiveness of the proposed method. The first experiment was performed to compare the proposed method with similar graph algorithms, while the second experiment was performed to compare the proposed method with other common explainable ML methods. The evaluation metrics of the algorithm are accuracy, precision, sensitivity, specificity, F_1-score, area under receiver operating characteristic curve (AUC), etc. To demonstrate the interpretability of our method, we selected a tinnitus case for result interpretation to showcase the inference process and interpretability of our method.

Results

Performance Verification

For a given knowledge graph, we extracted the patient nodes and their neighboring nodes to form a knowledge network. The node and edge sets in the knowledge network were divided into training and testing sets. The testing set did not contain syndrome entities. To reasonably divide the training and testing sets, we used a stratified sampling cross-validation method of

randomly dividing the network node and edge sets into 5 subsets: 1 subset as the testing set, and the other 4 subsets as the training set. The training set served as a known network, whereas the testing set was used to verify the syndrome prediction results and evaluate the accuracy of the syndrome prediction algorithm.

Evaluation Outcomes

Comparison With Similar Graph Algorithms

The proposed method was compared with similar graph algorithms such as CommonNeighbors and Adamic-Adar. CommonNeighbors is a common graph algorithm used to infer the potential relationships and proximity between 2 nodes [28]; however, the differences between common neighbors are not considered. Adamic-Adar is a typical algorithm for determining the closeness of 2 points by measuring the outdegree of common neighbors [29]. ResourceAllocation calculates the closeness between 2 nodes using a set of neighboring nodes near the target node [30]. We added common neighbor edge weights based on CommonNeighbors. Unlike Adamic-Adar and ResourceAllocation, our weight calculation method considered each syndrome, which had a higher adaptability to TCM diagnosis by the doctors. The experimental results are listed in Table 3; our method outperformed similar graph algorithms in diagnosing each syndrome.

Table 3. Experimental results of graph algorithm comparison.

Evaluation indicators and models	KED[a] (n=339)	LFBU[b] (n=307)	PFSI[c] (n=194)	QDSS[d] (n=270)	WFAI[e] (n=155)	Value, mean (SD)
Average accuracy						
Common neighbors	0.978	0.978	0.982	0.983	0.988	0.982 (0.004)
Adamic-Adar	0.979	0.979	0.978	0.983	0.989	0.982 (0.004)
Resource allocation	0.918	0.944	0.961	0.936	0.974	0.947 (0.019)
WeightedCommonNeighbors	0.990	0.994	0.995	0.992	0.998	0.994 (0.003)
Average precision						
Common neighbors	0.939	0.941	0.952	0.982	0.971	0.957 (0.017)
Adamic-Adar	0.940	0.949	0.932	0.981	0.971	0.955 (0.019)
Resource allocation	0.794	0.893	0.930	0.860	0.948	0.885 (0.055)
WeightedCommonNeighbors	0.970	0.987	0.993	0.986	1.000	0.987 (0.010)
Average sensitivity						
Common neighbors	0.981	0.971	0.922	0.943	0.929	0.949 (0.023)
Adamic-Adar	0.984	0.965	0.917	0.942	0.935	0.949 (0.023)
Resource allocation	0.933	0.877	0.801	0.840	0.837	0.857 (0.045)
WeightedCommonNeighbors	0.990	0.990	0.976	0.979	0.987	0.985 (0.006)
Average F_1-score						
Common neighbors	0.959	0.956	0.936	0.961	0.949	0.952 (0.009)
Adamic-Adar	0.961	0.957	0.924	0.961	0.952	0.951 (0.014)
Resource allocation	0.856	0.884	0.859	0.849	0.885	0.866 (0.015)
WeightedCommonNeighbors	0.980	0.989	0.984	0.982	0.994	0.986 (0.005)
Average specificity						
Common neighbors	0.978	0.980	0.993	0.995	0.996	0.988 (0.008)
Adamic-Adar	0.978	0.983	0.989	0.995	0.996	0.988 (0.007)
Resource allocation	0.914	0.966	0.990	0.963	0.994	0.965 (0.029)
WeightedCommonNeighbors	0.989	0.996	0.999	0.996	1.000	0.996 (0.004)
Average AUC[f]						
Common neighbors	0.979	0.976	0.958	0.969	0.963	0.969 (0.008)

Evaluation indicators and models	KED[a] (n=339)	LFBU[b] (n=307)	PFSI[c] (n=194)	QDSS[d] (n=270)	WFAI[e] (n=155)	Value, mean (SD)
Adamic-Adar	0.981	0.974	0.953	0.969	0.966	0.968 (0.009)
Resource allocation	0.923	0.922	0.895	0.901	0.915	0.911 (0.011)
WeightedCommonNeighbors	0.990	0.993	0.987	0.988	0.994	0.990 (0.003)

[a]KED: kidney essence deficiency.

[b]LFBU: liver fire bearing upward.

[c]PFSI: phlegm fire stagnation internally.

[d]QDSS: Qi deficiency of the spleen and stomach.

[e]WFAI: wind fire attacking internally.

[f]AUC: area under receiver operating characteristic curve.

Comparison With Other Interpretable ML Methods

The proposed method was compared with common ML classification algorithms including decision tree, random forest, naive Bayes, logistic regression, and k-nearest neighbors algorithms. The results are presented in Table 4. The graph algorithm based on WightedCommonNeighbor outperformed other models in the comprehensive diagnosis of each syndrome on the same data set but was lower than the random forest model in terms of the AUC metric. Although the random forest model had a certain degree of interpretability, the overall complexity of model interpretation increased when a large number of decision trees were included. The higher the number of decision trees in the random forest model, the greater the difficulty of interpreting the relationships and decision processes within the model. Compared to the random forest model, our proposed method had higher interpretability and was more readily accepted by doctors.

Table 4. Experimental results of machine learning classification algorithm comparison.

Evaluation indicators and models	KED[a]	LFBU[b]	PFSI[c]	QDSS[d]	WFAI[e]	Value, mean (SD)
Average accuracy						
WeightedCommonNeighbors	0.990	0.994	0.995	0.992	0.998	0.994 (0.003)
Decision tree	0.975	0.975	0.978	0.970	0.984	0.976 (0.005)
Random forest	0.987	0.982	0.985	0.987	0.994	0.987 (0.004)
Naive Bayes	0.979	0.976	0.979	0.981	0.991	0.981 (0.005)
Logistic regression	0.986	0.983	0.983	0.984	0.994	0.986 (0.004)
KNN[f]	0.986	0.980	0.982	0.986	0.994	0.985 (0.005)
Average precision						
WeightedCommonNeighbors	0.970	0.987	0.993	0.986	1.000	0.987 (0.010)
Decision tree	0.950	0.951	0.917	0.943	0.937	0.939 (0.012)
Random forest	0.974	0.950	0.970	0.982	0.963	0.968 (0.011)
Naive Bayes	0.971	0.923	0.953	0.956	0.980	0.957 (0.019)
Logistic regression	0.971	0.961	0.950	0.964	0.981	0.965 (0.010)
KNN	0.974	0.938	0.958	0.978	0.980	0.966 (0.016)
Average sensitivity						
WeightedCommonNeighbors	0.990	0.990	0.976	0.979	0.987	0.985 (0.006)
Decision tree	0.959	0.945	0.943	0.915	0.936	0.939 (0.014)
Random forest	0.976	0.977	0.933	0.956	0.987	0.966 (0.019)
Naive Bayes	0.953	0.981	0.912	0.956	0.948	0.950 (0.022)
Logistic regression	0.976	0.967	0.938	0.963	0.968	0.963 (0.013)
KNN	0.973	0.984	0.923	0.956	0.968	0.961 (0.021)
Average F_1-score						
WeightedCommonNeighbors	0.980	0.989	0.984	0.982	0.994	0.986 (0.005)
Decision tree	0.953	0.948	0.929	0.928	0.936	0.939 (0.010)
Random forest	0.975	0.963	0.950	0.968	0.975	0.966 (0.009)
Naive Bayes	0.961	0.951	0.932	0.955	0.964	0.953 (0.011)
Logistic regression	0.974	0.964	0.943	0.963	0.974	0.964 (0.011)
KNN	0.973	0.960	0.940	0.966	0.974	0.963 (0.012)
Average specificity						
WeightedCommonNeighbors	0.989	0.996	0.999	0.996	1.000	0.996 (0.004)
Decision tree	0.981	0.984	0.984	0.985	0.991	0.985 (0.003)
Random forest	0.990	0.983	0.994	0.995	0.995	0.992 (0.005)
Naive Bayes	0.989	0.974	0.992	0.988	0.997	0.988 (0.008)
Logistic regression	0.989	0.988	0.991	0.990	0.997	0.991 (0.003)
KNN	0.990	0.979	0.993	0.994	0.997	0.991 (0.006)
Average AUC[g]						
WeightedCommonNeighbors	0.990	0.993	0.987	0.988	0.994	0.990 (0.003)
Decision tree	0.970	0.964	0.964	0.950	0.963	0.962 (0.007)
Random forest	0.995	0.998	0.996	0.997	1.000	0.997 (0.002)
Naive Bayes	0.996	0.996	0.993	0.995	0.997	0.995 (0.001)
Logistic regression	0.997	0.997	0.994	0.995	0.997	0.996 (0.001)

Evaluation indicators and models	KED[a]	LFBU[b]	PFSI[c]	QDSS[d]	WFAI[e]	Value, mean (SD)
KNN	0.993	0.993	0.977	0.988	0.993	0.989 (0.006)

[a]KED: kidney essence deficiency.

[b]LFBU: liver fire bearing upward.

[c]PFSI: phlegm fire stagnation internally.

[d]QDSS: Qi deficiency of the spleen and stomach.

[e]WFAI: wind fire attacking internally.

[f]KNN: k-nearest neighbor.

[g]AUC: area under receiver operating characteristic curve.

Discussion

Principal Findings

The experimental results show that the accuracy, sensitivity, specificity, precision, F_1-score, and AUC of our proposed method all exceed 98% for 5 tinnitus subtypes. Compared to the traditional graph algorithm, our method comprehensively considers the number of neighboring nodes and the weight of edges for patient nodes. This method of calculating the strength of node connections and feature importance can more comprehensively measure the similarity between patient nodes. Further, by calculating the common neighbor score, the similarity between patient nodes can be quantitatively measured, providing a reliable quantitative indicator for the prediction problem of patient-to-syndrome node links. In addition, in the field of TCM, the impact of different features on diagnostic results may vary. This method considers the importance of features through edge weight values, making similarity calculations more realistic. By considering the edge weight values, the reasons for the formation of similarity between patient nodes and the importance of features can be explained, enhancing the interpretability of the model results. This method is not only applicable to the diagnosis of syndrome types in the field of TCM but can also be applied in other fields, especially in the similarity calculation problem that needs to consider feature importance and node correlation strength, which has universality.

In terms of interpretability, the proposed method integrated the knowledge of TCM differential diagnosis and clinical experience into a knowledge graph, which made the method more interpretable. To illustrate the explainability of our method, we randomly selected a patient from the patient records and used their medical information as input to the syndrome diagnosis algorithm, as shown in Figure 9. The patient information was input to the knowledge graph, where we searched for other patients who shared common neighbors with the selected patient. We calculated the common neighbor scores and returned the top k (k=20) patients with the highest scores. The results are summarized in Table 5. Based on the syndromes of the top k patients that were most similar to the target patient, we deduced that the predicted syndrome of the target patient was KED, which was consistent with the actual syndrome of the patient.

Figure 9. The inference process of patient syndrome patterns. KED: kidney essence deficiency.

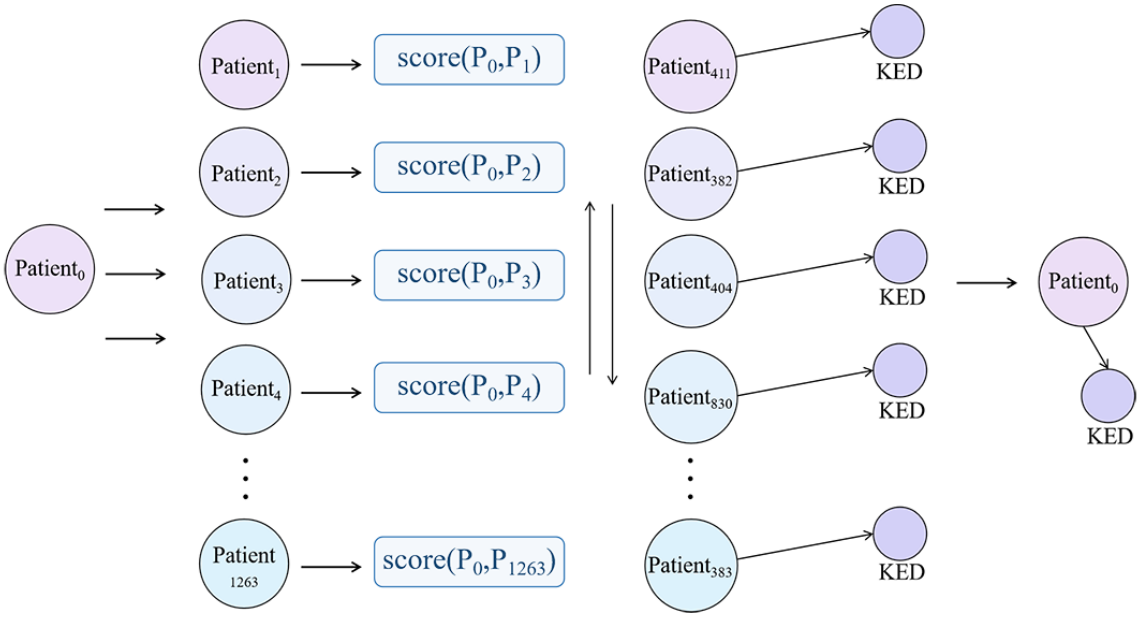


Table 5. Inference results of patient syndrome patterns.

Patient ID	Neighbors	Neighbors score
411	19	14.66
382	16	14.23
404	17	14.23
830	17	14.04
856	16	14.04
395	16	13.97
365	16	13.93
372	15	13.93
386	15	13.93
390	15	13.93
396	16	13.93
400	16	13.93
403	16	13.93
407	15	13.93
410	16	13.93
413	15	13.93
375	17	13.91
389	17	13.91
381	16	13.78
383	16	13.78

Limitations

The proposed method considered the weight of common neighbors and the importance of different symptoms for different syndrome types, but this makes similarity calculation more complex, requiring more computing resources and time. Meanwhile, the calculation of edge weight values requires relatively rich and accurate feature data. If the data quality is not high or features are missing, it will affect the accuracy of similarity calculation. However, compared to large-scale knowledge graphs, our research has a smaller sample size and requires continuous data collection to enrich the knowledge base.

From the experimental results, our method achieved good results in the diagnosis of WFAI, LFBU, PFSI, and QDSS. However, some deficiencies existed in the differential diagnosis of QDSS and KED syndrome types, which could create confusion between the two. The analysis of 3 patients who were misclassified with KED instead of QDSS revealed common entities between them and the top 5 most similar patients among their neighbors (Textbox 1). The common entities between patient 1 (ID 415) and the top 5 most similar patients among their neighbors, who were all patients with QDSS but were misclassified with KED, are listed in Textbox 1. The common entities included worsening conditions when standing up, empty feeling in the ears, left side, worsening condition after physical exertion, hypertension, red tongue, anxiety, thin pulse, hearing loss, continuous symptoms, female sex, and dizziness. Similarly, patient 2 (ID 601) and the top 5 most similar patients among their neighbors shared common entities including worsening condition when standing up, empty feeling in the ears, left side, worsening condition after physical exertion, thin and white coating on the tongue, red tongue, anxiety, thin pulse, and continuous symptoms. Patient 3 (ID 423) and the top 5 most similar patients among their neighbors shared common entities including worsening condition after physical exertion, worsening condition at night, left side, use of headphones, exercise, pale tongue, thin coating on the tongue, tinnitus, middle to low frequency, and intermittent symptoms. By comparing the common entities between the patients and their top 5 most similar neighbors, we found that entities such as worsening condition after physical exertion and left side had higher scores in the differential diagnosis of the 2 syndrome types. However, ML algorithms were prone to confusion in the differential diagnosis because both QDSS and KED could be present in patients with these symptoms.

Textbox 1. Misclassified patient entity.

Patient 1 (ID 415)

- Aggravation when standing up, ear emptiness, left side, aggravation after work, hypertension, tongue redness, anxiety, fine vein, hearing loss, duration, male, and dizziness.

Patient 2 (ID 601)

- Aggravation when standing up, ear emptiness, left side, aggravation after work, thin fur, white fur, tongue redness, anxiety, fine vein, and duration.

Patient 3 (ID 423)

- Aggravation after work, nighttime aggravation, left side, use headphones, exercise, tongue dullness, thin fur, cicada chirping, and interval.

Conclusions

Tinnitus is a complex ear disease that poses challenging issues in clinical diagnosis due to the lack of specific indicators and the reliance on patient complaints. In this study, we constructed a medical knowledge graph based on EMRs and authoritative knowledge of patients with tinnitus and proposed an explainable tinnitus-assisted diagnosis model. The experimental results showed that our proposed method not only performed better in diagnostic performance with a diagnostic accuracy of over 98% for all syndromes but also offered better interpretability compared to general ML algorithms owing to the natural interpretability of the knowledge graph. Thus, the effectiveness of the proposed method was demonstrated to assist Chinese medicine doctors in diagnosing tinnitus during clinical practice.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82074581).

Authors' Contributions

ZY and YG contributed to the conceptualization of this study and the funding acquisition. HZ and LW were responsible for data curation. ZY and ZK designed and implemented the algorithms and conducted the experiments. HZ, LW, ZK, TL, and ZW analyzed the experimental results. ZY wrote this paper with revision assistance from HZ and LW. YG reviewed and edited this paper. All authors have read and approved this paper.

Conflicts of Interest

None declared.

References

1. Jarach CM, Lugo A, Scala M, van den Brandt PA, Cederroth CR, Odone A, et al. Global prevalence and incidence of tinnitus: a systematic review and meta-analysis. JAMA Neurol 2022;79(9):888-900 [FREE Full text] [doi: 10.1001/jamaneurol.2022.2189] [Medline: 35939312]
2. Dawood F, Khan N, Bagwandin V. Management of adult patients with tinnitus: preparedness, perspectives and practices of audiologists. S Afr J Commun Disord 2019;66(1):e1-e10 [FREE Full text] [doi: 10.4102/sajcd.v66i1.621] [Medline: 31793315]
3. Ahmed A, Aqeel M, Akhtar T, Salim S, Ahmed B. Moderating role of stress, anxiety, and depression in the relationship between tinnitus and hearing loss among patients. Pak J Psychol Res 2019;34(4):753-772. [doi: 10.33824/pjpr.2019.34.4.41]
4. Neff P, Simões J, Psatha S, Nyamaa A, Boecking B, Rausch L, et al. The impact of tinnitus distress on cognition. Sci Rep 2021;11(1):2243. [doi: 10.1038/s41598-021-81728-0] [Medline: 33500489]
5. Piccirillo JF, Rodebaugh TL, Lenze EJ. Tinnitus. JAMA 2020;323(15):1497-1498. [doi: 10.1001/jama.2020.0697] [Medline: 32176246]
6. Özbey-Yücel Ü, Uçar A. The role of obesity, nutrition, and physical activity on tinnitus: a narrative review. Obesity Med 2023 Jun;40:100491. [doi: 10.1016/j.obmed.2023.100491]
7. Liu Z, Li Y, Yao L, Lucas M, Monaghan JJM, Zhang Y. Side-aware meta-learning for cross-dataset listener diagnosis with subjective tinnitus. IEEE Trans Neural Syst Rehabil Eng 2022;30:2352-2361. [doi: 10.1109/TNSRE.2022.3201158] [Medline: 35998167]
8. Sun ZR, Cai YX, Wang SJ, Wang C, Zheng Y, Chen Y, et al. Multi-view intact space learning for tinnitus classification in resting state EEG. Neural Process Lett 2019;49(2):611-624. [doi: 10.1007/s11063-018-9845-1]
9. Shoushtarian M, Alizadehsani R, Khosravi A, Acevedo N, McKay CM, Nahavandi S, et al. Objective measurement of tinnitus using functional near-infrared spectroscopy and machine learning. PLoS One 2020;15(11):e0241695 [FREE Full text] [doi: 10.1371/journal.pone.0241695] [Medline: 33206675]

10. Sanders PJ, Doborjeh ZG, Doborjeh MG, Kasabov NK, Searchfield GD. Prediction of acoustic residual inhibition of tinnitus using a brain-inspired spiking neural network model. Brain Sci 2021;11(1):52. [doi: 10.3390/brainsci11010052] [Medline: 33466500]
11. Manta O, Sarafidis M, Schlee W, Mazurek B, Matsopoulos GK, Koutsouris DD. Development of machine-learning models for tinnitus-related distress classification using wavelet-transformed auditory evoked potential signals and clinical data. J Clin Med 2023;12(11):3843 [FREE Full text] [doi: 10.3390/jcm12113843] [Medline: 37298037]
12. Allgaier J, Schlee W, Probst T, Pryss R. Prediction of tinnitus perception based on daily life mHealth data using country origin and season. J Clin Med 2022 Jul 22;11(15):4270 [FREE Full text] [doi: 10.3390/jcm11154270] [Medline: 35893370]
13. Rodrigo H, Beukes E, Andersson G, Manchaiah V. Exploratory data mining techniques (decision tree models) for examining the impact of internet-based cognitive behavioral therapy for tinnitus: machine learning approach. J Med Internet Res 2021;23(11):e28999 [FREE Full text] [doi: 10.2196/28999] [Medline: 34726612]
14. Liu Y, Niu H, Zhu J, Zhao P, Yin H, Ding H, et al. Morphological neuroimaging biomarkers for tinnitus: evidence obtained by applying machine learning. Neural Plast 2019;2019:1712342. [doi: 10.1155/2019/1712342] [Medline: 31915431]
15. Niemann U, Brueggemann P, Boecking B, Mazurek B, Spiliopoulou M. Development and internal validation of a depression severity prediction model for tinnitus patients based on questionnaire responses and socio-demographics. Sci Rep 2020;10(1):4664 [FREE Full text] [doi: 10.1038/s41598-020-61593-z] [Medline: 32170136]
16. Li Z, Fu Z, Li W, Fan H, Li S, Wang X, et al. Prediction of diabetic macular edema using knowledge graph. Diagnostics (Basel) 2023;13(11):1858 [FREE Full text] [doi: 10.3390/diagnostics13111858] [Medline: 37296709]
17. Zhou G, Kuang Z, Tan L, Xie X, Li J, Luo H. Clinical decision support system for hypertension medication based on knowledge graph. Comput Methods Programs Biomed 2022;227:107220. [doi: 10.1016/j.cmpb.2022.107220] [Medline: 36371975]
18. Lyu K, Tian Y, Shang Y, Zhou T, Yang Z, Liu Q, et al. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. J Biomed Inform 2023;139:104298 [FREE Full text] [doi: 10.1016/j.jbi.2023.104298] [Medline: 36731730]
19. Lin Z, Hong L, Cai X, Chen S, Shao Z, Huang Y, et al. Risk detection of clinical medication based on knowledge graph reasoning. CCF Trans Pervasive Comput Interact 2023;5(1):82-97. [doi: 10.1007/s42486-022-00114-5]
20. Yang YM, Li Y, Zhong X. Research on entity recognition and knowledge graph construction based on TCM medical records. J Artif Intell Pract 2021;47(1):1-15. [doi: 10.23977/jaip.2020.040105]
21. Xie Y, Hu L, Chen X. Auxiliary diagnosis based on the knowledge graph of TCM syndrome. Comput Mater Contin 2020;65:481-494. [doi: 10.32604/cmc.2020.010297]
22. Yang R, Ye Q, Cheng C, Zhang S, Lan Y, Zou J. Decision-making system for the diagnosis of syndrome based on traditional Chinese medicine knowledge graph. Evid Based Complement Alternat Med 2022;2022:8693937 [FREE Full text] [doi: 10.1155/2022/8693937] [Medline: 35186106]
23. Lan G, Hu M, Li Y, Zhang Y. Contrastive knowledge integrated graph neural networks for Chinese medical text classification. Eng Appl Artif Intell 2023;122:106057. [doi: 10.1016/j.engappai.2023.106057]
24. Liu D, Wei C, Xia S, YAN J. Construction and application of knowledge graph of Treatise on Febrile Diseases. Digital Chin Med 2022;5(4):394-405. [doi: 10.1016/j.dcmed.2022.12.006]
25. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Jun 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
26. Liu P. Traditional Chinese Otorhinolaryngology. Beijing, China: China Traditional Chinese Medicine Press; 2021:90-94.
27. Wang DJ, Gan ZW. Traditional Chinese Otorhinolaryngology. Shanghai, China: Shanghai Scientific and Technical Publishers; 1985:26-28.
28. Freeman LC. A set of measures of centrality based on betweenness. Sociometry 1977;40(1):35-41. [doi: 10.2307/3033543]
29. Adamic LA, Adar E. Friends and neighbors on the web. Soc Networks 2003;25(3):211-230. [doi: 10.1016/s0378-8733(03)00009-1]
30. Mastrandrea R, Squartini T, Fagiolo G, Garlaschelli D. Enhanced reconstruction of weighted networks from strengths and degrees. New J Phys 2014;16(4):043022. [doi: 10.1088/1367-2630/16/4/043022]

Abbreviations

AI: artificial intelligence
AUC: area under receiver operating characteristic curve
EMR: electronic medical record
KED: kidney essence deficiency
LFBU: liver fire bearing upward
ML: machine learning
PFSI: phlegm fire stagnation internally

QDSS: Qi deficiency of the spleen and stomach
TCM: traditional Chinese medicine
WFAI: wind fire attacking internally

Edited by G Eysenbach, A Benis; submitted 05.03.24; peer-reviewed by L Wang, A Tomar, SN Mohanty; comments to author 20.04.24; revised version received 10.05.24; accepted 15.05.24; published 10.06.24.

Please cite as:
Yin Z, Kuang Z, Zhang H, Guo Y, Li T, Wu Z, Wang L
Explainable AI Method for Tinnitus Diagnosis via Neighbor-Augmented Knowledge Graph and Traditional Chinese Medicine: Development and Validation Study
JMIR Med Inform 2024;12:e57678
URL: https://medinform.jmir.org/2024/1/e57678
doi:10.2196/57678
PMID:38857077

Original Paper

Retrieval-Based Diagnostic Decision Support: Mixed Methods Study

Tassallah Abdullahi[1], MSc; Laura Mercurio[2], MD; Ritambhara Singh[1,3], PhD; Carsten Eickhoff[4], PhD

[1]Department of Computer Science, Brown University, Providence, RI, United States

[2]Departments of Pediatrics & Emergency Medicine, Alpert Medical School, Brown University, Providence, RI, United States

[3]Center for Computational Molecular Biology, Brown University, Providence, RI, United States

[4]School of Medicine, University of Tübingen, Tübingen, Germany

Corresponding Author:
Carsten Eickhoff, PhD
School of Medicine
University of Tübingen
Schaffhausenstr, 77
Tübingen, 72072
Germany
Phone: 49 7071 29 843
Email: carsten.eickhoff@uni-tuebingen.de

Abstract

Background: Diagnostic errors pose significant health risks and contribute to patient mortality. With the growing accessibility of electronic health records, machine learning models offer a promising avenue for enhancing diagnosis quality. Current research has primarily focused on a limited set of diseases with ample training data, neglecting diagnostic scenarios with limited data availability.

Objective: This study aims to develop an information retrieval (IR)–based framework that accommodates data sparsity to facilitate broader diagnostic decision support.

Methods: We introduced an IR-based diagnostic decision support framework called CliniqIR. It uses clinical text records, the Unified Medical Language System Metathesaurus, and 33 million PubMed abstracts to classify a broad spectrum of diagnoses independent of training data availability. CliniqIR is designed to be compatible with any IR framework. Therefore, we implemented it using both dense and sparse retrieval approaches. We compared CliniqIR's performance to that of pretrained clinical transformer models such as Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT) in supervised and zero-shot settings. Subsequently, we combined the strength of supervised fine-tuned ClinicalBERT and CliniqIR to build an ensemble framework that delivers state-of-the-art diagnostic predictions.

Results: On a complex diagnosis data set (DC3) without any training data, CliniqIR models returned the correct diagnosis within their top 3 predictions. On the Medical Information Mart for Intensive Care III data set, CliniqIR models surpassed ClinicalBERT in predicting diagnoses with <5 training samples by an average difference in mean reciprocal rank of 0.10. In a zero-shot setting where models received no disease-specific training, CliniqIR still outperformed the pretrained transformer models with a greater mean reciprocal rank of at least 0.10. Furthermore, in most conditions, our ensemble framework surpassed the performance of its individual components, demonstrating its enhanced ability to make precise diagnostic predictions.

Conclusions: Our experiments highlight the importance of IR in leveraging unstructured knowledge resources to identify infrequently encountered diagnoses. In addition, our ensemble framework benefits from combining the complementary strengths of the supervised and retrieval-based models to diagnose a broad spectrum of diseases.

(JMIR Med Inform 2024;12:e50209) doi:10.2196/50209

KEYWORDS

clinical decision support; rare diseases; ensemble learning; retrieval-augmented learning; machine learning; electronic health records; natural language processing; retrieval augmented generation; RAG; electronic health record; EHR; data sparsity; information retrieval

Introduction

Background

Identifying an accurate and timely cause for a patient's health problem represents a challenging and complex cognitive task. A clinician must consider a complex range of composite information sources, including the patient's medical history, current state, imaging, laboratory test results, and other clinical observations, to formulate an accurate diagnosis. Diagnostic errors are a leading cause of delayed treatment, potentially affecting millions of patients each year. Research suggests that these errors contribute to 6% to 17% of adverse events [1].

Studies [2,3] have shown that, rather than relying on a single physician for a final diagnosis, obtaining recommendations from multiple physicians increases diagnostic accuracy. To improve the diagnostic process while maintaining economic feasibility, different variants of automated assistants, also known as diagnostic decision support systems (DDSSs) and symptom checkers, have been introduced [4]. Early DDSSs [1,5] were driven by structured databases that maintain information about diseases and other medical information in a structured form. Although promising, these systems have yet to be highly successful for several reasons, including limited accessibility, poor flexibility, and scalability issues [6,7]. Hence, the traditional DDSS is gradually being replaced by machine learning and deep learning models.

Recent studies [8-13] highlight the importance of electronic health records for supervised machine learning algorithms in health care. These algorithms use the electronic health record of a patient as input to predict their diagnosis. However, supervised model development has been limited to a select number of diseases with higher prevalence and extensive documentation due to the availability of large amounts of labeled data. As a result, infrequently occurring diagnoses remain poorly studied. In real-world diagnostic scenarios, physicians are faced with the challenge of identifying the correct diagnosis from a plethora of possibilities. Therefore, a system that considers a broad range of diagnoses, including rare conditions, is desirable for improved diagnostic accuracy. However, recent studies [14,15] demonstrate that traditional supervised learning models are challenging to use in such scenarios due to their reliance on large, labeled data sets with many examples per diagnosis. However, most clinical cohorts exhibit imbalanced class distributions, characterized by a long-tailed pattern [15,16] in which certain diagnostic classes represent most training samples whereas others exhibit few or even 0 data points. In such scenarios, most traditional supervised models overfit the majority class, resulting in poor performance for the minority classes. As such, large labeled data sets may not be a straightforward solution for achieving an efficient supervised classifier that supports diverse diagnoses.

In response, researchers have leveraged a technique called transfer learning, which is a widely used method for building classifiers that enables generalization to classes with limited labeled data. A common transfer learning technique involves fine-tuning pretrained models—models trained on large and diverse data sets—on a smaller, domain-specific corpus to enhance model performance. However, the effectiveness of this approach still relies on the size of the data set available for fine-tuning. Zero-shot learning and few-shot learning [17,18] represent promising alternatives for fine-tuning large models with limited labeled data. In zero-shot learning, the model can classify samples from classes without labeled training data. Few-shot learning requires at least one labeled example per class to enable the model to make accurate predictions. Although some studies [19,20] have shown that pretrained language models have zero-shot and few-shot learning capabilities, their performance remains inferior to that of models trained on extensive labeled data. While zero-shot and few-shot approaches have demonstrated success in the vision domain [21,22], their application to language models remains an ongoing area of research.

Leveraging external knowledge resources can improve predictive performance, especially with a limited training sample size, as shown in previous work by Prakash et al [7] and Müller et al [6]. Classical information retrieval (IR) systems can use a vast collection of resources for various applications with low computational complexity and no need for labeled data. In the medical setting, studies [23-25] and competitions such as the text retrieval conference (TREC) clinical decision support track [26] have focused on developing and evaluating IR systems to support clinician decision-making. Typically, these IR systems have been applied to biomedical literature retrieval to aid in clinical decision support. However, these systems can also be adapted for other downstream clinical tasks. For example, Naik et al [27] trained a model to predict patient admission outcomes (ventilation need, mortality, and length of stay) by integrating relevant medical literature with patient notes, leaving an open question of how IR systems would fare in directly predicting the underlying diagnosis. Therefore, our study applied IR techniques to perform literature-guided diagnostic prediction.

Objectives

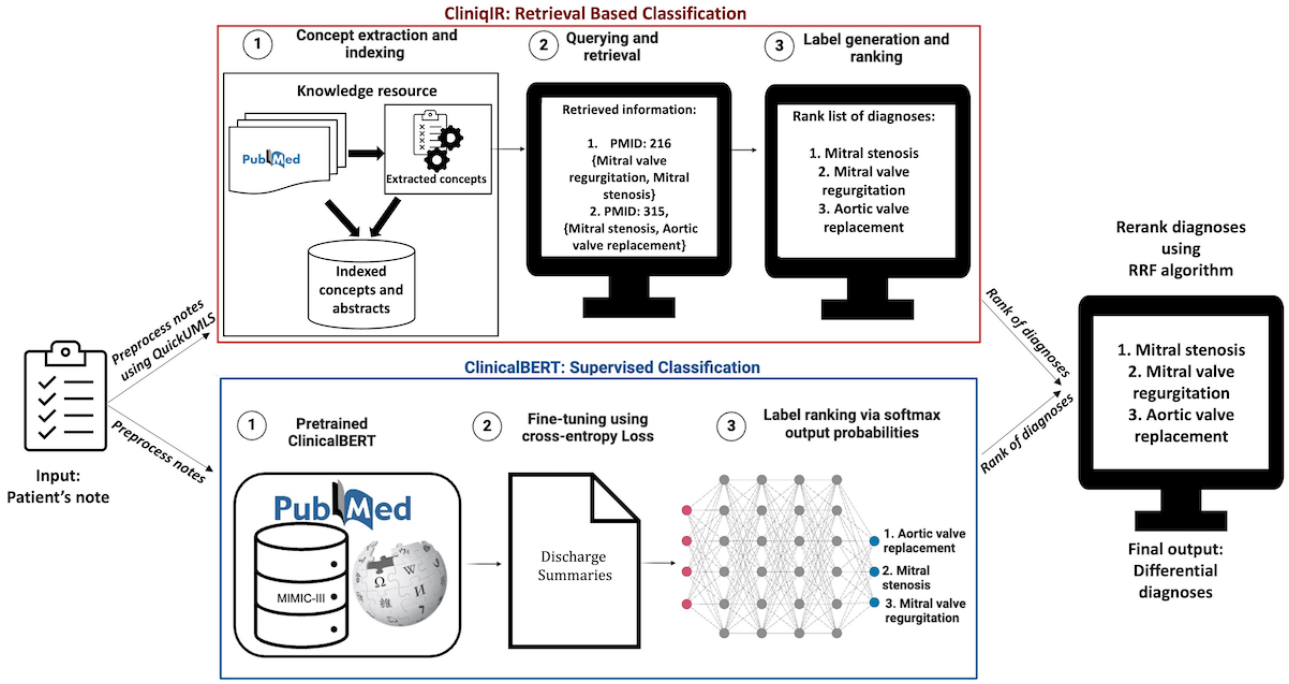
We introduce "CliniqIR," a novel clinical decision support algorithm that uses an IR system to match a patient's medical record to a specific diagnosis from a large pool of possible diagnoses. Our study aimed to improve the current state of predictive modeling and diagnostic decision support for a broad range of diagnoses regardless of their training data availability. By using clinical text records and external knowledge sources, including the Unified Medical Language System (UMLS) Metathesaurus [28] and PubMed abstracts [29], we demonstrated that "CliniqIR" successfully generalizes to less common diagnostic categories with heavily skewed data distributions. Our work also shows CliniqIR to be highly adaptable, allowing for easy integration with any IR system. This flexibility ensures the model's ability to adapt to available resources and work across various retrieval methods.

To assess CliniqIR's ability to predict diagnoses with no available training samples, we evaluated its performance on the DC3 data set [30]. We compared its performance to that of pretrained clinical models in a zero-shot setting, and our results showed that CliniqIR has the capability to recognize a broad spectrum of rare and complex diseases without relying on labeled training data. We also compared the performance of

CliniqIR with that of supervised fine-tuned pretrained biomedical large language models and found that supervised models have limitations when used on highly imbalanced data, especially for diagnoses with limited training samples. Then, we leveraged an ensemble strategy combining CliniqIR and a fine-tuned Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT) to make predictions for a wide range of diagnoses that include frequent and infrequent conditions, summarized in Figure 1.

Our study highlights the valuable synergy between retrieval-based systems and supervised learning models, showcasing how their combination can achieve state-of-the-art performance, particularly in data sets characterized by a long-tailed distribution. This finding holds significant promise and offers new avenues to address the challenges of imbalanced data in various domains.

Figure 1. CliniqIR and Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT), classify patient notes and generate ranked lists of potential diagnoses. The reciprocal rank fusion (RRF) ensemble reranks the lists from both models to provide clinicians with a more accurate final ranking of differential diagnoses to aid the diagnostic process. MIMIC-III: Medical Information Mart for Intensive Care III; PMID: PubMed ID.



Methods

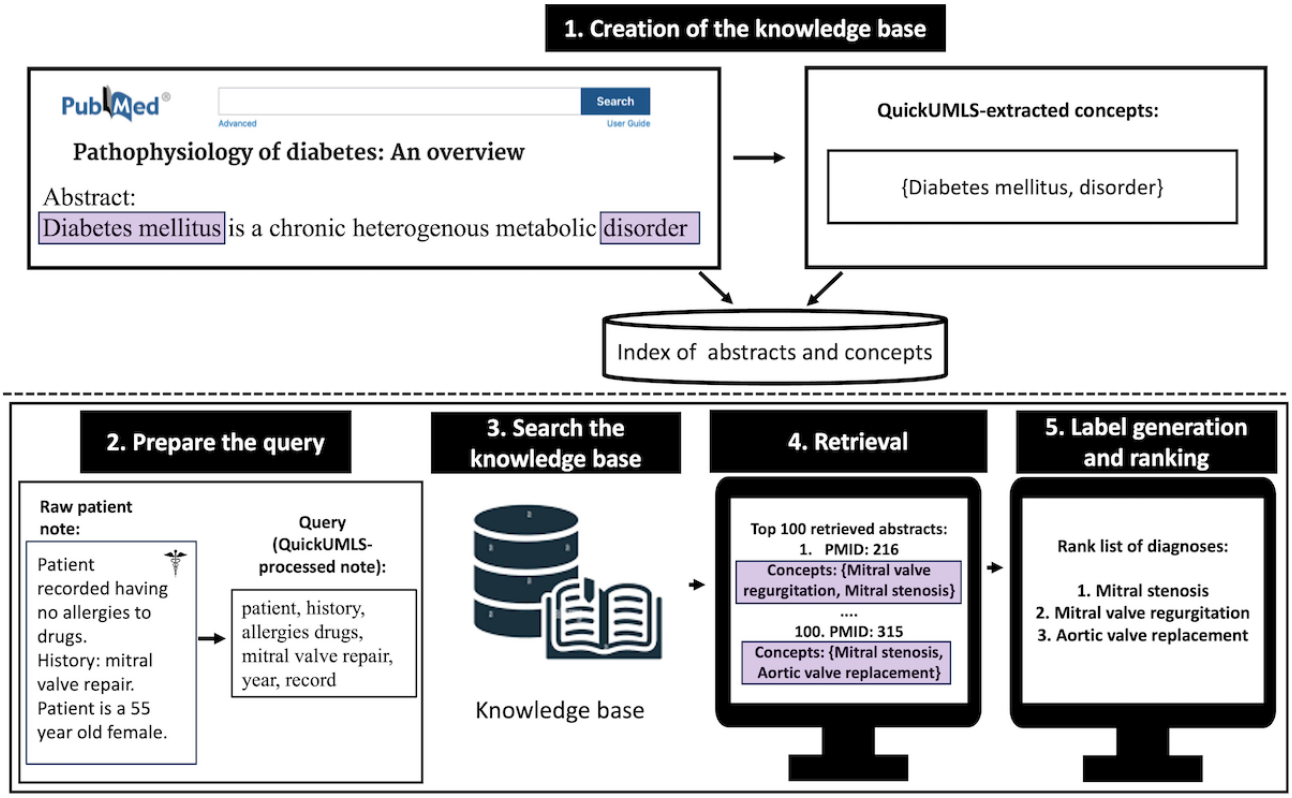
CliniqIR: The Retrieval-Based Model

Overview

We present CliniqIR, a novel literature-guided system that maps a patient's note to a specific diagnosis. By leveraging unlabeled external knowledge sources, CliniqIR uses an IR system to classify a wide range of diagnoses without relying on the availability of notes for each individual diagnosis (labeled training data). As a result, CliniqIR represents a valuable disease classification tool when labeled training data are limited or unavailable.

An overview of our method is shown in Figure 2. The backbone of CliniqIR is its knowledge base. Once the knowledge base is built, we can query the system to provide a list of probable diagnoses. In this study, a clinical narrative with a patient's medical history or summary was preprocessed and treated as a query. To make inferences given a patient's clinical note, as a preprocessing step, we first used QuickUMLS (Soldani and Goharian [31]) explained in the *Knowledge Extraction Using QuickUMLS* section, to extract medical keywords from the note to obtain a query. Next, we fed the query (preprocessed note) to the retrieval system, which returned a list of matching relevant PubMed abstracts alongside the medical conditions mentioned in each abstract. Afterward, we selected the top 100 items from the list and then computed the frequency of each concept across the list of abstracts. Finally, the model returned a list of concepts ranked according to their term frequency–inverse document frequency (TF-IDF) defined in equation 1. The list returned was similar to a medical differential diagnosis (a ranked list of possible diagnoses that could cause a patient's illness). The medical condition with the highest TF-IDF score was predicted as the most likely diagnosis. We provide a detailed description of the individual processing steps in the following sections.

Figure 2. Overview of CliniqIR, the retrieval-based clinical decision support system. PMID: PubMed ID.



Concept Extraction and Indexing

We extracted unique medical concepts (conditions) from each PubMed abstract using QuickUMLS, described in the *Knowledge Extraction Using QuickUMLS* section. Medical concepts included diseases, symptoms, or any information about a medical procedure. Subsequently, we built the knowledge base of the retrieval system by indexing each abstract and its corresponding article title, article ID, and a concept dictionary that contained all the unique concepts mentioned in that abstract. Indexing involves storing and organizing data to enable efficient IR at search time. Using the index of PubMed abstracts, the model inputs a patient's notes as a query and returns relevant information from the indexed abstracts as an output. Figure 2 provides visual details.

Querying and Retrieval

Once the index was built, we submitted queries to the retrieval system. The *Retrieval System Implementation* section provides more details. After we submitted a query, the system returned a list of abstracts and their corresponding attributes (dictionary of concepts, article title, and article ID number) ranked according to query relevance. For each query, we selected the top 100 abstracts because the top few documents are most likely to contain relevant query information.

Label Generation

After the querying operation, we focused on the extracted concepts of the top 100 abstracts. The previous retrieval phase can potentially return multiple abstracts that contain similar information in response to a given query, resulting in concept dictionaries of ≥2 abstracts containing similar concepts. Multiple occurrences could indicate the relevance of a concept across abstracts. To account for such duplication, we calculated each unique concept's recurrence, or term frequency (TF), across the list. The TF of a concept across a list of abstracts would be 1 if it appeared in only 1 abstract. If it appeared in 2 abstracts, its TF would become 2, and so on. Calculating the recurrence of concepts across the top-100 list resulted in a new list that contained medical concepts and their TFs. These medical concepts were regarded as labels and used for classification purposes. Thus, each unique concept became a potential diagnosis, and the TF of each concept is subsequently used for ranking purposes in equation 1. Textbox 1 describes the concepts the model returned (in no order of importance) after the retrieval stage given a set of queries processed using QuickUMLS. The list was filtered for a simple illustration. As mentioned previously, concepts are biomedical terms that include symptoms, signs, and diseases, among other things. On the other hand, a diagnosis could represent a disease, an injury, a neoplastic process, or a medical term describing a condition a patient is experiencing. Therefore, to account for a wide range of possible diagnoses, we kept all concepts in the label generation phase, and we considered a concept as a diagnosis when it matched the ground truth. Therefore, in this paper, we use *concepts* and *diagnoses* interchangeably.

Textbox 1. The output returned by the retrieval-based model (CliniqIR) given a query.

Query and concepts retrieved (labels)

- Abdominal pain, bloating, rectal bleeding, weight loss, anxiety, disruptive thoughts, and suicidality: "generalized anxiety disorder," "panniculitis," "chronic abdominal pain," "Burkitt's lymphoma," and "Whipple's disease"
- Chest pain, radiation to neck, dyslipidemia, lung crackles, bradycardia, and ST elevation: "acute myocardial infarction," "acute coronary syndrome," "coronary artery disease," "myocardial ischemia," "myopericarditis," and "myocardial infarctions"
- Night sweats, abdominal pain (pleuritic), nausea, loose stools, lymphadenopathy (inguinal), plaques, leucopenia, neutrophilia, and elevated (Angiotensin converting enzyme) ACE: "sarcoidosis," "lymphomas," "lymph node," "tuberculosis," "lupus erythematosus," "Rosai-Dorfman disease," and "Kikuchi-Fujimoto disease"

Ranking and Predictions

It is important to note that our model differs from traditional classification schemes. In our case, the observed mappings between patients' notes and ground-truth diagnoses are not provided for learning purposes. Therefore, a list of relevant diagnoses (a subset of the retrieved concepts) must be generated independently for each query. However, as the diagnosis list is not generated based on ground truth, it may contain information that is not relevant to the data set to be evaluated. For example, given a data set with 3 possible ground-truth diagnoses—*lymphoma*, *coronary artery disease*, and *gastroenteritis*—the model might return concepts such as *coronary artery disease*, *myocardial infarction*, and *chest pain* in the label retrieval phase for a query whose ground truth is *coronary artery disease*. To address this and ensure a fair comparison with other classification models, we filtered the retrieved concepts during the evaluation and only kept diagnoses that were part of the ground truth. Therefore, in the aforementioned example, we filtered out *myocardial infarction* and *chest pain*. Then, we assigned ranks to the remainder of the diagnoses in the list using the TF-IDF function shown in equation 1:

$$\text{TFIDF}(c,a,d) = \text{TF}(c,a).\text{IDF}(c,d) \quad \textbf{(1)}$$

Knowledge Resources: PubMed Abstracts

Over the years, research in predictive modeling for diagnostic decision support has witnessed enormous success in transfer learning, particularly where a model leverages an auxiliary data source (often a knowledge base) to perform several predictive tasks. Some studies [7,22,32] have used resources such as Wikipedia and PubMed [29] to create systems that perform classification tasks or retrieve useful articles with specific information. In contrast, most early DDSSs [1,33] were built on structured knowledge bases; however, most computable knowledge bases are not freely accessible.

Inspired by previous research, we used abstracts from PubMed articles as an unstructured collection of knowledge resources to guide the prediction of diagnoses for all our experiments. An abstract may contain information about a specific condition, its signs, or its symptoms. Some abstracts include medical case reports, whereas others may contain information about a medical device. To build a retrieval system grounded in reliable information, we leveraged the vast collection of abstracts in the PubMed database. PubMed, maintained by the National Library of Medicine, houses >33 million citations for biomedical literature, encompassing life science journals and books dating back to 1946. However, the number of abstracts available per condition varies considerably. Therefore, for our core experiments, we implemented a 100-abstract inclusion threshold for diagnoses (Multimedia Appendix 1).

Knowledge Extraction Using QuickUMLS

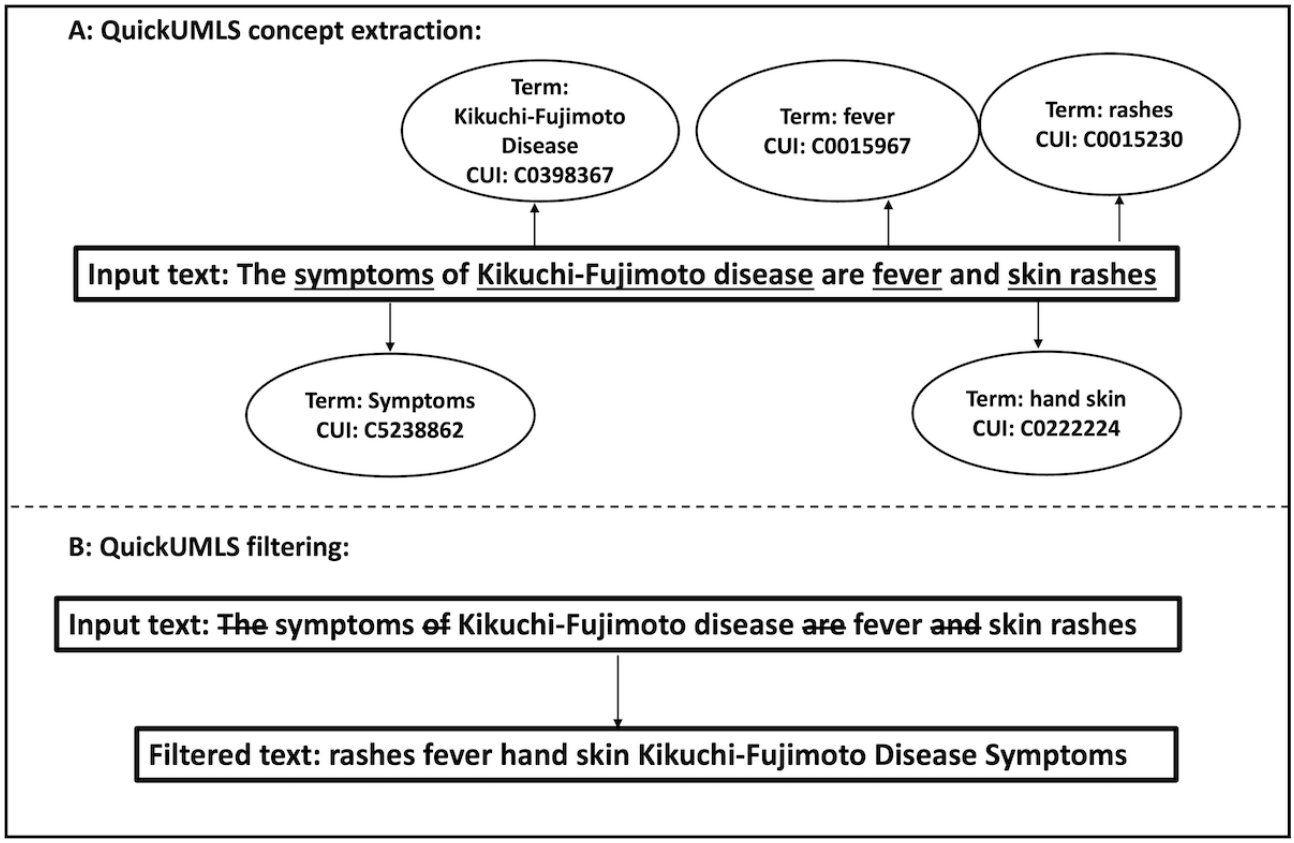
Overview

QuickUMLS [31] is an unsupervised medical concept extraction tool that detects mentions of medical entities such as diseases, symptoms, and other medical concepts from unstructured text. Given a document, QuickUMLS matches each possible token in the document against concepts in the UMLS [28]. In this study, we used QuickUMLS for 2 different purposes.

Extraction

We used QuickUMLS to extract unique biomedical concepts from each PubMed abstract. A concept can be any medical term, including a diagnosis or symptom. As shown in Figure 3, each biomedical term in a text is a concept with a corresponding unique alphanumeric identifier (concept ID) in the UMLS vocabulary. We kept all the concepts associated with each abstract in a dictionary.

Figure 3. Outputs from the QuickUMLS tool developed by Soldani and Goharian [33] showing: (A) a graph of extracted concepts and their concept unique identifier (CUI) for a specific input text; the underlined texts are considered important words, and their corresponding Unified Medical Language System terms and CUIs are returned, (B) a query processing pipeline. Each text marked with a strike-through is filtered out to obtain a query.



Filtering

We also used QuickUMLS as a data preprocessor to filter out noisy, uninformative, and nonclinical terms, such as stop words, from a patient's clinical note, resulting in a query that contained only medical terms. Citing the input text example in Figure 3, the outcome of the filtering operation was "rashes fever hand skin Kikuchi-Fujimoto disease symptoms." This filtering step is equivalent to keeping only the QuickUMLS-recognized medical terms and concepts.

Retrieval System Implementation

Overview

CliniqIR is designed to be highly adaptable to arbitrary IR systems. This flexibility ensures the model's ability to work across various retrieval methods, adapting to the resources available. In this study, we performed experiments on a sparse and a dense retriever.

Sparse Retriever

We built our knowledge base by indexing PubMed abstracts and their concepts using Apache Lucene (Apache Software Foundation [34]), which enables users to search this index with queries ranging from single words to sentences. The relevance of an abstract to a query is determined by a similarity score, with Lucene's default "BM25" [35] function estimating the best-matching abstract.

Dense Retriever

Unlike sparse retrievers, which represent queries as word frequencies, dense retrievers capture the semantic meaning and relationships within the text using dense embedding vectors. This allows for retrieval based on similarity, usually calculated through maximum inner-product search. To implement this approach, we leveraged the Medical Contrastive Pre-trained Transformers (MedCPT) [36], a state-of-the-art biomedical retrieval system in a zero-shot setting using its default parameters. Section S1 and Figure S1 in the Multimedia Appendix 1 provides details on the parameter settings for both retrieval systems.

Pretrained Transformer Models

In this study, we used 2 well-known methods, namely, supervised fine-tuning and zero-shot learning, to harness the benefits of transfer learning from 6 pretrained clinical and biomedical language models. The models we used are ClinicalBERT, PubMed Bidirectional Encoder Representations from Transformers (PubMedBERT), Scientific Bidirectional Encoder Representations from Transformers (SciBERT), Self-alignment Pretrained Bidirectional Encoder Representations from Transformers (SapBERT), cross-lingual knowledge-infused medical term embedding (CODER) and MedCPT. We describe them briefly in the follows:

ClinicalBERT Model

The ClinicalBERT [37] is an extension of Biomedical Bidirectional Encoder Representations from Transformers [38]

trained further on discharge summary notes from the Medical Information Mart for Intensive Care III (MIMIC-III) database [39]. It was designed to handle the complexity and nuances of clinical text.

PubMedBERT Model

PubMedBERT [40] was specifically designed to capture domain-specific knowledge present in biomedical literature. It was initialized from Bidirectional Encoder Representations From Transformers (BERT) and trained further on the collection of PubMed abstracts.

SciBERT Model

SciBERT [41] is a BERT-based language model pretrained on 1.14 million full-text papers from Semantic Scholar. The corpus domain cuts across the field of computer science and the biomedical space.

SapBERT Model

SapBERT [42] is also a BERT-based model initialized from PubMedBERT. SapBERT was further pretrained on UMLS [28], which consists of a wide range of biomedical ontologies for >4 million concepts.

CODER Model

CODER [43] is another BERT-based model formulated to generate biomedical embeddings. It was also initialized from PubMedBERT. CODER was further pretrained using the concepts from the UMLS [28] and optimized to increase the embedding similarities between terms with the same concept unique identifier.

MedCPT Model

MedCPT [36] is a contrastive pretrained PubMedBERT-based model also formulated to generate biomedical text embeddings for multiple tasks.

Supervised Fine-Tuning

Given a set of patients' notes (hereinafter also referred to as *notes*) as inputs and their corresponding diagnoses as outputs, we fine-tuned the pretrained models in a supervised fashion to classify diagnoses by feeding in a series of notes and their corresponding ground-truth diagnoses. Each note was a textual document describing a patient's health condition and medical history. The ground-truth diagnosis of a note was the corresponding health condition of the patient. Multimedia Appendix 1 provides details of the models' parameter settings. After fine-tuning, given a test set of notes, a model assigned probabilities to each ground-truth diagnosis for each note. The diagnosis with the highest probability corresponded to the model's most confident prediction. We assigned ranks to each diagnosis in the order of their decreasing probability score for all our predictions. These ranks were further used to compute the mean reciprocal rank (MRR) for model evaluation (refer to the *Evaluation Metrics* section for details). We justify the use of ranking output probabilities across classes to compute the MRR because the probabilities generated by the classifier represent the classifier's confidence in predicting each incidence. Supervised fine-tuning requires diagnosis-specific training data (availability of historic patient notes for each diagnosis) to deliver state-of-the-art performance. Unfortunately, labeled data are expensive to generate. This requirement makes it impractical to use a supervised fine-tuned model to diagnose those diseases without (many) notes for training. Hence, we used this method to make predictions only when training data were available.

Zero-Shot Learning

Given our focus on predicting diagnoses with few or 0 training samples, we included zero-shot learning methods as baselines. Leveraging the high quality of the aforementioned pretrained transformer embeddings, we adopted a zero-shot strategy by classifying patients' notes based on their semantic similarity to potential diagnoses. This can be achieved by using pretrained models as biencoders [18,44]. Using this approach, we accounted for the diagnosis classes (classes without training samples) that the supervised fine-tuned models could not handle.

Given a patient's note (our query) and the list of candidate diagnoses as labels, we used different variants of BERT as biencoders to encode queries and the full names of all ground-truth diagnoses to produce their respective representation vectors separately. Next, we computed their cosine similarity score and ranked each diagnosis for each query according to this score. The diagnosis with the highest cosine similarity became the model's most confident diagnostic prediction (refer to Multimedia Appendix 1 for more details).

Model Ensemble: Reciprocal Rank Fusion

The label retrieval process allowed the CliniqIR (retrieval-based model) to diagnose unseen conditions regardless of training data availability. This property is beneficial for diagnoses with little or no training data. On the other hand, a supervised fine-tuned model can draw much deeper insights from available historical case data. We adopted an ensemble strategy to combine the advantages of both paradigms.

In IR and general machine learning, ensemble strategies combine results from multiple models to produce a single joint output. Ideally, the ensemble model should produce a new output whose performance is superior to that of the individual constituent models. Several studies [32,45,46] have shown that high-performance gains can be achieved through model ensembling. One of the simplest ways to build such a model is to focus on applying a reranking heuristic to the ranks of each item in a model's output list. Hence, we collected the ranked list of diagnoses from a CliniqIR model and that of the best-performing supervised fine-tuned model, ClinicalBERT, and combined the 2 lists. We then applied a modified version of the reciprocal rank fusion (RRF) [45] algorithm using equation 2 to merge their results and produce a single, final output list. Given a set *C*of concepts(diagnoses) to be ranked and a set of rankings *R*for all concepts obtained from each ensemble member (CliniqIR and ClinicalBERT), we computed the RRF score for each concept ($c \in C$) as follows:

(2)

In the aforementioned equation, "$r \in R$" is the rank of concept *c*according to an ensemble member. We summed up the individual ranks of a concept from each ensemble member "$r(c)$"

with *k* and computed the inverse. Previous work by Cormack et al [45] reported that setting *k*to 60 was the near-optimal choice for most of their experiments. Hence, we set *k* to 60 for all experiments. When concepts (diagnoses) had more than one training sample, we selected their individual ranks *r*from each ensemble member to compute the RRF score; otherwise, we selected ranks from the CliniqIR model. We used the RRF algorithm due to some key advantages: (1) it is a simple unsupervised method that eliminates the need for training samples, and (2) it effectively combines the results from various models without reliance on a weighting or voting mechanism.

Experimental Setup

Data Sources

DC3 Data Set

The DC3 data set [30] was designed specifically for the evaluation of diagnostic support systems. The data set comprises 30 rare and difficult-to-diagnose cases compiled and solved by clinical experts in the *New England Journal of Medicine* Case Challenges. This data set lacks large, labeled training data, but it covers a wide range of diagnostic cases for various specialties. Therefore, we used this data set to determine the applicability of CliniqIR for diagnostic inference when the underlying patient condition is rare. Each case is a patient's note and its corresponding true diagnosis written as free text. We mapped the true diagnoses to their UMLS concept IDs to produce test labels for evaluation consistency. When we did not find an exact matching term for a diagnosis, we considered the closest match returned by the UMLS browser. During the preprocessing step, we found that some cases in the DC3 data set had multiple terms representing a ground-truth diagnosis, making it difficult to find a single UMLS concept ID for such cases. To ensure an accurate mapping with the UMLS concept IDs, we split such cases into separate terms. For example, the case "Acute and chronic cholecystitis and extensive cholelithiasis with transmural gallbladder inflammation" was split into 2 separate terms: "Acute and chronic cholecystitis" and "Extensive cholelithiasis with transmural gallbladder inflammation." Then, we mapped each case to its corresponding UMLS concept ID. Next, we computed the document frequency of all the true diagnoses (now represented as concepts) across all PubMed abstracts. In these cases, either of the concepts could be considered as the ground truth. As the data did not contain sufficient notes to train a model, we formulated this task as a zero-shot multiclassification problem. Specifically, we expected a model to predict the underlying condition given a patient's note without labeled training data.

MIMIC-III Data Set

The MIMIC-III [39] is a freely accessible medical database that contains information on >50,000 intensive care unit patients. The data include laboratory events, vital sign measurements, clinical observations, notes, and diagnoses structured as *ICD-9-CM* (*International Classification of Diseases, Ninth Revision, Clinical Modification*), codes. We worked with the discharge notes for all experiments because they document a free-text synopsis of a patient's hospital stay from admission to discharge. In MIMIC-III, each discharge note is mapped to multiple diagnoses ranked according to priority. We considered the highest-priority diagnosis to be the admission's ground-truth diagnosis (and prediction target). We excluded admissions primarily for birth and pregnancy as they did not represent a primary pathological diagnosis. After preprocessing, the discharge notes contained 2634 unique *ICD-9-CM* diagnoses. We mapped these *ICD-9-CM* diagnoses to their corresponding UMLS concept IDs to calculate their TF across the knowledge resource (PubMed abstracts). The resulting unique diagnoses were associated with notes ranging from thousands of occurrences of frequent conditions, such as coronary atherosclerosis and aortic valve disorders, to rare ones, such as Evans syndrome and ehrlichiosis, with just a single instance forming a long-tailed distribution. A total of 902 diagnoses fell into the singleton category. One discharge note representing a specific diagnosis is insufficient to train and test a model. Thus, we reserved all diagnoses with only 1 available note for model testing. For diagnoses with <5 note samples, we reserved 1 sample for testing, and the rest were included in model training. We split the remainder of the data set (instances of diagnoses with ≥5 associated notes) into training, validation, and testing sets in the ratio 70:15:15; this split resulted in the training set containing notes representing 1732 unique diagnoses and the test set containing notes representing a total of 2634 unique diagnoses (refer to Multimedia Appendix 1 for more details). For models that did not require training (eg, the retrieval model), we used the validation and training sets for hyperparameter tuning purposes and the test set for final model evaluations.

Baselines

Previous studies and competitions, such as the TREC clinical decision support track, have emphasized the development and evaluation of IR systems to aid clinical decision-making. While these systems are commonly used for evidence-based literature searches, our study explored their adaptation for direct literature-guided diagnosis prediction. Although a direct comparison to the systems in the TREC clinical decision support track was not possible, insights gained from these competitions informed the engineering of our retrieval system. To evaluate our model, we used 2 transfer learning techniques—supervised fine-tuning and zero-shot classification methods (refer to the *Pretrained Transformer Models* section)—because of their performance in scenarios where labeled data are limited or unavailable. In addition, some studies [47-49] have shown pretrained language models to attain superior performance to that of count vector–based models and traditional supervised methods in various medical tasks. We used "Z" to identify when models were used in a zero-shot classification setting, an "S" for supervised fine-tuning, and "CliniqIR" when models were used in a retrieval setting. Table 1 provides details.

Table 1. Overview of the experiments conducted using the different models and their task description.

Experiment	Models used	Task description
Retrieval-based experiments (CliniqIR)	BM25[a] and MedCPT[b]	Models retrieved relevant abstracts to inform diagnostic predictions.
Zero-shot experiments (Z)	ClinicalBERT[c], PubMedBERT[d], CODER[e], SapBERT[f], and MedCPT	Models classified diseases in a zero-shot setting without previous task-specific training.
Supervised experiments (S)	ClinicalBERT	Models were fine-tuned using labeled data for enhanced disease prediction accuracy.

[a]BM25: Best Match 25.

[b]MedCPT: Medical Contrastive Pretrained Transformers.

[c]ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers.

[d]PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers.

[e]CODER: cross-lingual knowledge-infused medical term embedding.

[f]SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers.

Evaluation Metrics

In our experiments, each model returned a ranked list of diagnoses analogous to a ranked list of differential diagnoses formulated by a medical expert. Given a query and a list of ranked items produced by a model, a simple classification accuracy metric tracks whether the model made the correct prediction at the top of the list. Instead, we used the MRR [50] because it told us *where* the true diagnosis was placed in the list in equation 3. If a model returned the reference diagnosis at rank 1 (ie, at the top of the list), the reciprocal rank (RR) was 1; if the most appropriate item was at rank 2, then the RR was 0.5. The RR decreases as the relevant item moves farther down the list. We calculated the MRR by computing the average RR across admissions. An MRR of 1 meant that the model returned the correct diagnosis at the top of its list for every patient, and an MRR of 0 implied that the model never produced a correct diagnosis. Mathematically, the MRR can be represented as follows:

(3)

where $|Q|$ denotes the total number of queries and denotes the rank of the correct diagnosis. We also calculated the mean average precision (MAP) to evaluate the balance between precision and recall of the retrieval systems (for details, refer to Multimedia Appendix 1).

Ethical Considerations

No ethics approval was pursued for this research, given that the data were publicly accessible and deidentified.This aligns with the guidelines outlined by the US Department of Health and Human Services, Office for Human Research Protections, §46.101 (b)(4) [51].

Results

CliniqIR Models Retrieved Useful Literature and Meaningful Concepts

Table 2 showcases qualitative results for 3 selected queries, displaying the top 3 documents retrieved by the CliniqIR model. Notably, the retrieved articles and their corresponding concepts demonstrated clear relevance to the ground-truth diagnoses of the respective queries.

Table 2. Qualitative overview of the top documents and concepts retrieved for 3 selected queries along with their respective correct concepts. This table illustrates the types of results our system generates for each query, showing the alignment with the ground-truth concepts.

Ground-truth diagnosis, and top 3 documents	Retrieved concepts
1. Viral pneumonia	
Relevant article 1—PMID[a] 15336585: Cases from the Osler Medical Service at Johns Hopkins University. Diagnosis: P. carinii pneumonia and primary pulmonary sporotrichosis	{“C1956415”: [“paroxysmal nocturnal dyspnea”], “C0239295”: [“esophageal candidiasis”], “C0236053”: [“mucosal ulcers”], “C1535939”: [“Pneumocystis”], “C0031256”: [“petechiae”], “C0006849”: [“thrush”], “C0011168”: [“dysphagia”]}
Relevant article 2—PMID 32788269: A 16-Year-Old Boy with Cough and Fever in the Era of COVID-19	{“C0746102”: [“chronic lung disease”], “C0004096”: [“asthma”], “C0009443”: [“cold”], “C0206750”: [“Coronavirus”], “C0018609”: [“h disease”]}
Relevant article 3—PMID 30225154: Meningococcal Pneumonia in a Young Healthy Male	{“C3714636”: [“pneumonias”], C1535950”: [“GI inflammation”]}
2. Hypoparathyroidism	
Relevant article 1—PMID 34765380: A Challenging Case of Persisting Hypokalemia Secondary to Gitelman Syndrome	{“C0220983”: [“metabolic alkalosis”], “C0151723”: [“hypomagnesemia”], “C0020599”: [“hypocalciuria”], “C0014335”: [“enteritis”], “C0012634”: [“Diagnosis”], “C0235394”: [“wasting”], “C0271728”: [“Hyperreninemic hyperaldosteronism”], “C0268450”: [“gitelman syndrome”], “C3552462”: [“Tubulopathy”]}
Relevant article 2—PMID 27190662: Suppression of Parathyroid Hormone in a Patient with Severe Magnesium Depletion	{“C0151723”: [“hypomagnesemia”], “C0030554”: [“paresthesias”], “C0020598”: [“hypocalcemias”], “C0020626”: [“Low parathyroid hormone”], “C0030517”: [“Parathyroid”], “C0033806”: [“pseudo hypoparathyroidism”]}
Relevant article 3—PMID 28163524: Afebrile Seizures as Initial Symptom of Hypocalcemia Secondary to Hypoparathyroidism	{“C0020626”: [“Hypoparathyroidism”], “C0012236”: [“DiGeorge syndrome”], “C0863106”: [“afebrile seizures”], “C0030353”: [“papilledema”], “C0020598”: [“Hypocalcemias”], “C0012634”: [“Diagnosis”], “C0042870”: [“Vitamin D deficiency”]}
3. Intracerebral hemorrhage	
Relevant article 1—PMID 9125737: A 36-year-old woman with acute onset left hemiplegia and anosognosia	{“C0020564”: [“enlargement”], “C0019080”: [“hemorrhage”]}
Relevant article 2—PMID 25830084: Multiple extra-ischemic hemorrhages following intravenous thrombolysis in a patient with Trousseau syndrome: case study.	{“C2937358”: [“Intracerebral hemorrhage”], “C0151699”: [“intracranial hemorrhage”], “C0019080”: [“hemorrhages”], “C0020564”: [“enlargement”], “C0021308”: [“infarct”], “C0022116”: [“ischemia”]}
Relevant article 3—PMID 1434057: A case of recurrent cerebral hemorrhage considered to be cerebral amyloid angiopathy by cerebrospinal fluid examination.	{“C0472376”: [“thalamic hemorrhage”], “C2937358”: [“cerebral hemorrhage”], “C0019080”: [“bleeding”], “C0023182”: [“cerebrospinal fluid leak”]}

[a]PMID: PubMed ID.

CliniqIR Models Yielded State-of-the-Art Performance for Rare and Complex Diagnoses

We examined the retrieval-based models’ (CliniqIR) performance on the DC3 data set to show their applicability for rare and complex diagnostic cases. The absence of training data for this data set implied that supervised learning would not be applicable and the models could only make predictions in an unsupervised or zero-shot setting. Hence, on this data set, we compared the CliniqIR models’ performance to that of pretrained transformers in a zero-shot setting. In contrast to the CliniqIR model, which creates its own set of labels, we supplied the pretrained transformers with a range of potential diagnoses for each query to enable zero-shot predictions. This gave the models a significant advantage over their use in a real-world setting, where such information would not be readily available. Table 3 shows the MRR of the chosen models on the DC3 data set. Even with the supporting assumption that the range of possible diagnoses was known to the pretrained models, the CliniqIR models outperformed them with an MRR of 0.35 and 0.32 for CliniqIR_BM25 and CliniqIR_MedCPT, respectively. This means that, on average, CliniqIR_BM25 and CliniqIR_MedCPT were more likely to return the correct diagnosis within the top 3 predictions for a case.

The MRR scores of the pretrained zero-shot methods were similar to one another but markedly lower; the scores were 0.15, 0.22, 0.25, 0.25, 0.24, and 0.18 for ClinicalBERT, PubMedBERT, SciBERT, CODER, SapBERT, and MedCPT, respectively. Our results show that the CliniqIR models are capable of making useful predictions in the case of rare and complex diagnoses with limited or no training data availability.

Table 3. Performance evaluation of the models on the DC3 data sets across all case. The retrieval-based models, denoted using "CliniqIR" gave the best overall performance compared to the zero-shot models, denoted using "Z."

Model used	Mean reciprocal rank
ClinicalBERT[a] (Z)	0.15
PubMedMERT[b] (Z)	0.22
SciBERT[c] (Z)	0.25
CODER[d] (Z)	0.25
SapBERT[e] (Z)	0.24
CliniqIR_BM25	*0.35* [f]
MedCPT[g] (Z)	0.18
CliniqIR_MedCPT	*0.32*

[a]ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers.

[b]PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers.

[c]SciBERT: Scientific Bidirectional Encoder Representations from Transformers.

[d]CODER: cross-lingual knowledge-infused medical term embedding.

[e]SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers.

[f]Highest mean reciprocal rank is italicized.

[g]MedCPT: Medical Contrastive Pre-trained Transformers.

Performance on MIMIC-III

Supervised Prediction Models Failed at Making Rare Diagnoses

When training data are available, supervised models are preferred. Thus, we investigated the effectiveness of a supervised learning approach for a highly imbalanced data set such as MIMIC-III. We fine-tuned the pretrained models to predict diagnoses using available clinical notes. Diagnoses were categorized based on the frequency of associated notes to show how training data availability affects a supervised model's predictive capacity. A total of 902 diagnoses had no training data (only 1 note representative in MIMIC-III), whereas 1732 had at least one training sample (≥2 note representatives in MIMIC-III). Predictions were made only for the 1732 diagnoses, excluding those without training data. We introduced sample weights in the loss function to handle the enormous data imbalance. This approach weighs the loss computed for samples differently depending on their class training size. Our results in Figure S2 in Multimedia Appendix 1 show that ClinicalBERT performed best among all pretrained models. Hence, we used ClinicalBERT as our supervised baseline for the remainder of our experiments.

After training ClincalBERT, we tested it on different clinical note frequency–based categories (Table 4). In Table 4, we observed that the MRR score of the ClinicalBERT model was higher for diagnosis categories with many training examples (>10 notes). In addition, in the data set category with 1 to 10 notes available per diagnosis (1<notes≤10), ClinicalBERT obtained a low MRR score of 0.07. An MRR of 0.08 indicates that, on average, ClinicalBERT returned the correct diagnosis for a case among its top 13 predictions for these diagnoses. While the model could not perform predictions for 902 diagnoses due to the lack of training data, the drastic decline in ClinicalBERT's performance also indicates that the model is not suitable for making predictions for diagnoses with <10 clinical notes available for training. We also noticed a decline in performance for diagnoses with training samples between 500 and 750. This was likely due to many diagnoses having similar symptoms and manifestations. Therefore, the supervised learning approaches struggle to find a fine delineation of boundaries between similar classes without sufficient training data.

Table 4. Performance of the best-performing fine-tuned supervised model, Clinical Bidirectional Encoder Representations from Transformers on the Medical Information Mart for Intensive Care III data set. We categorized the results by the frequency of training note representation per diagnosis.

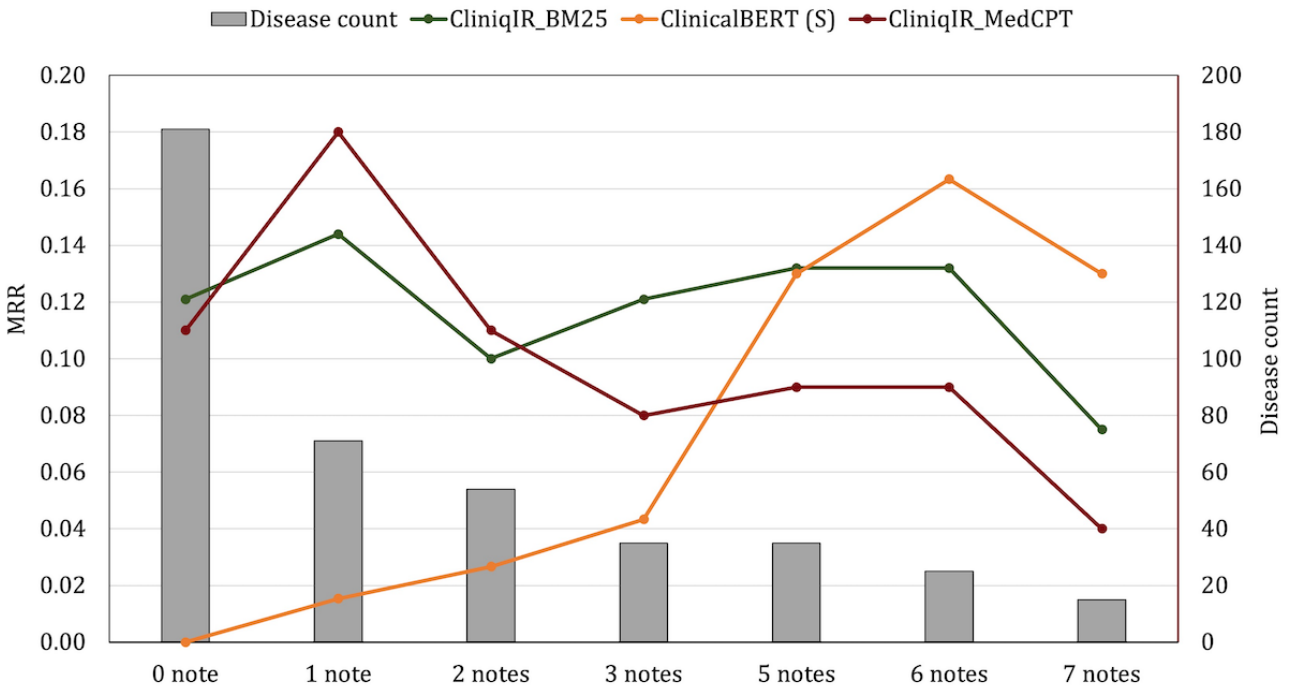
Data set category	Mean reciprocal rank
0 note	—[a]
1≤Notes≤10	0.08
10<Notes≤50	0.33
50<Notes≤100	0.49
100<Notes≤250	0.52
250<Notes≤500	0.57
500<Notes≤750	0.44
750<Notes	0.41
0<Notes	0.37

[a]Not applicable.

CliniqIR Models Outperformed ClinicalBERT for Rare Diagnoses

The objective of this experiment was to determine to what extent the CliniqIR models can be used in place of a supervised model when the training sample size is small. Results in Figure 4 show that CliniqIR-based models performed better than ClinicalBERT for diagnoses with up to 3 training samples. In addition, CliniqIR_BM25 and ClinicalBERT had similar MRR scores for diagnoses with 5 training samples. The average MRR score for the CliniqIR-based models was approximately 0.1 across most categories except for diagnoses with at least 7 training samples. This result indicates that, on average, their correct prediction for a query was ranked 10th on the list. The disease count bars in Figure 4 (in gray) also show that the number of diseases with <5 training samples was more than twice the number of diseases with >5 training samples. Thus, CliniqIR allows for more disease coverage and also generalizes well for cases with low note availability. This result confirms that, while supervised models may perform well with sufficient labeled training data, CliniqIR-based models' performance stands out as remarkable for diagnoses in the low-data regime.

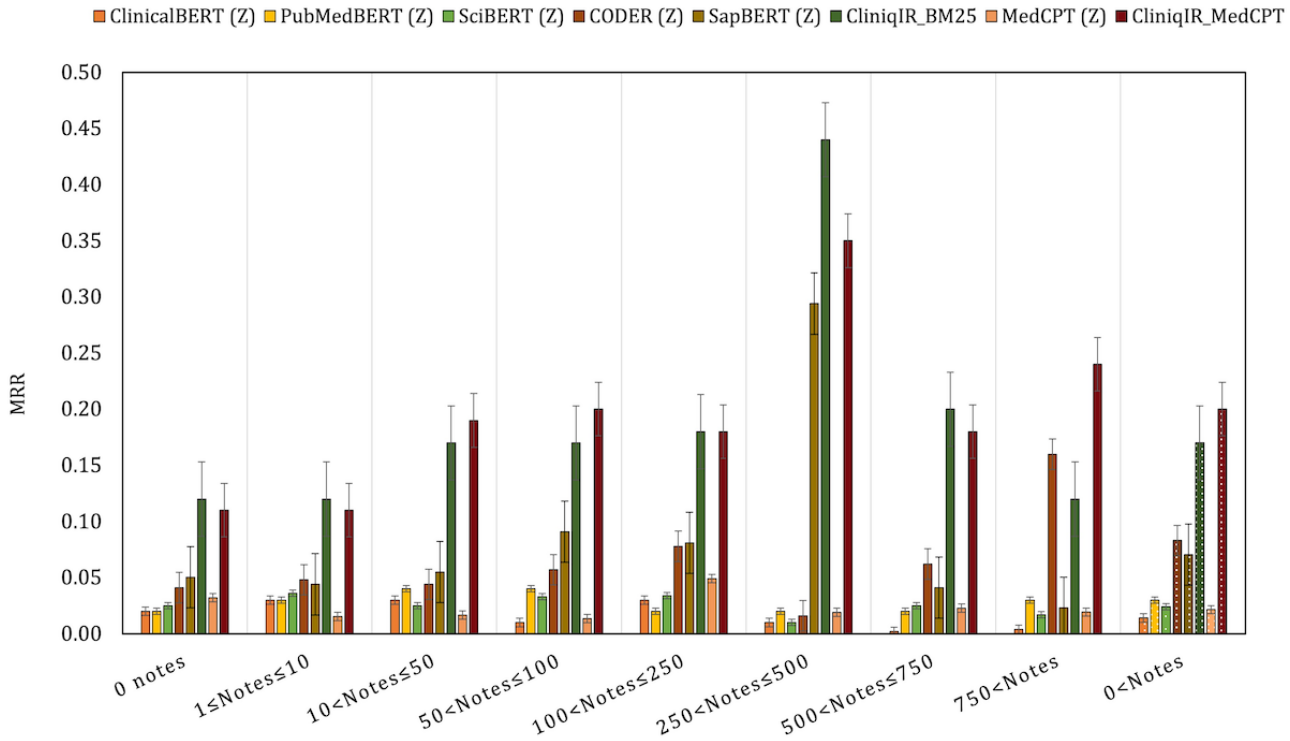
Figure 4. Mean reciprocal rank (MRR) results for CliniqIR-based models and Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT) when predicting diagnoses with training sample sizes of 0, 1, 2, 3, 5, 6, and 7. Results indicate that the CliniqIR-based models perform best when the training sample size is between 0 and 5. However, ClinicalBERT performs best as training data size increases. "S" denotes that the ClinicalBERT model was used in a supervised setting.



CliniqIR Models Outperformed Zero-Shot Baselines for Rare Diagnoses

To further demonstrate the utility of CliniqIR models for diagnoses with little or no training samples, we compared its performance to that of the pretrained models in a zero-shot setting. As mentioned previously, the MIMIC-III data set comprises >2634 diagnoses, but the supervised fine-tuned models were effective only for a subset of diagnoses with training data; 902 diagnoses had no training samples at all. In zero-shot settings, pretrained models can make predictions without reliance on training data. In Figure 5, we observe that CliniqIR models outperformed the zero-shot pretrained models across most data set categories, especially when diagnoses had a low number of associated training notes. The highest and lowest MRR scores obtained by CliniqIR_BM25 were 0.44 and 0.12, respectively, whereas CliniqIR_MedCPT's highest and lowest scores were 0.35 and 0.11.

Figure 5. Performance evaluation of CliniqIR models and each pretrained zero-shot baseline on the Medical Information Mart for Intensive Care III data set. We categorized the results by the frequency of note representative per diagnosis. "Z" represents models used in a zero-shot setting. The CliniqIR models performed best across data set categories in the low-resource regime. ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers; CODER: cross-lingual knowledge-infused medical term embeddin; MedCPT: Medical Contrastive Pre-trained Transformers; MRR: mean reciprocal rank; PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers; SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers; SciBERT: Scientific Bidirectional Encoder Representations from Transformers.



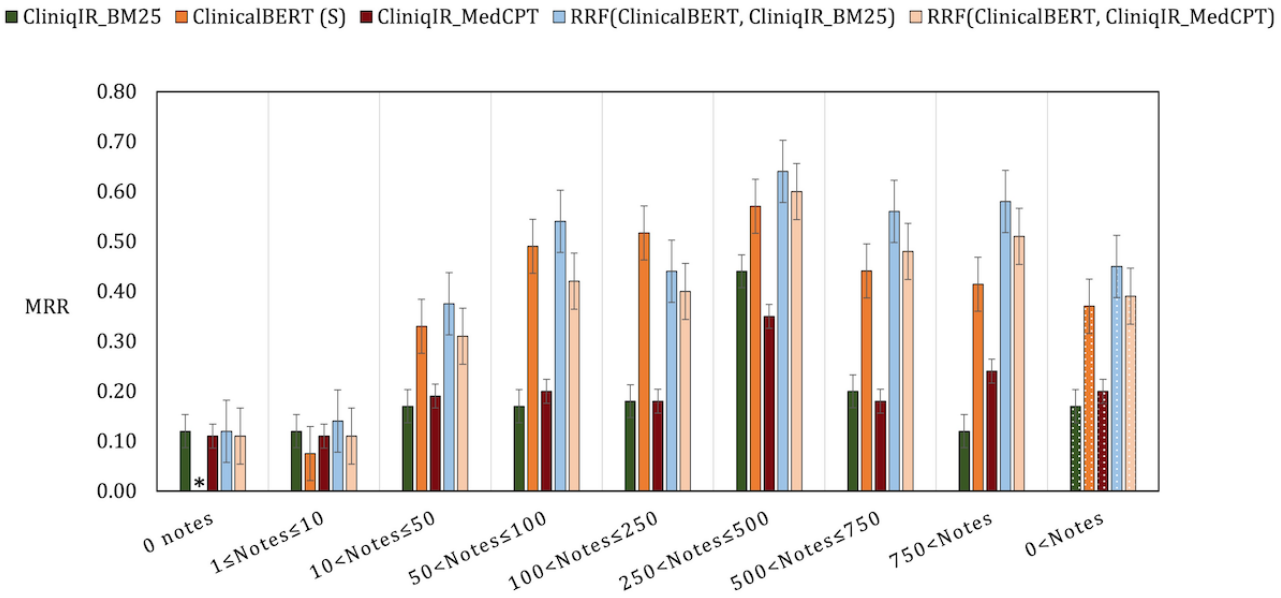
Among the zero-shot baseline, CODER and SapBERT's performances were better across most data set categories. However, in the category in which all diagnoses were considered (diagnoses with >0 notes), CODER outperformed SapBERT, obtaining a maximum and minimum MRR score of 0.16 and 0.02, respectively. These MRR scores indicate that, on average, both CliniqIR models returned the correct diagnosis for a case among their top 5 predictions. In contrast, the best-performing pretrained zero-shot baselines, CODER and SapBERT, returned an accurate diagnosis for a query among their top 15 and 12 predictions, respectively.

Ensemble Models Yielded State-of-the-Art Performance

We have shown that CliniqIR models deliver valuable diagnostic decision support in the setting of limited or unavailable training data. On the other hand, a supervised pretrained model such as ClinicalBERT is an efficient alternative when training data are abundant. To combine the strengths of both models, we used the RRF algorithm as an ensemble strategy. The RRF algorithm combines the ranks of all the ensemble members (a CliniqIR model and a supervised model) to produce a new ranked list of diagnoses for a given patient's clinical note. We hypothesized that creating an ensemble with both models would boost predictive performance across various diagnoses regardless of the availability of associated clinical notes.

To implement the RRF algorithm introduced in the *Model Ensemble: Reciprocal Rank Fusion* section, we used ClinicalBERT and a CliniqIR model to obtain separate ranked lists for each diagnosis and concept across queries. We compared the predictive performance of each individual model to that of the ensemble in terms of MRR for each note availability category. Figure 6 shows the output of the experiments before and after fusing the predicted ranks from both models.

Figure 6. Performance evaluation of the models on the Medical Information Mart for Intensive Care III data set before and after the ensemble. Adopting the reciprocal rank fusion (RRF) algorithm as an ensemble strategy boosted predictive performance across the data set. The Clinical Bidirectional Encoder Representations from Transformer (ClinicalBERT) model cannot directly make predictions for diagnoses with no training samples. Hence, we used "*" to mark such data set categories. The letter "S" denotes that ClinicalBERT was used as a supervised model. MedCPT: Medical Contrastive Pre-trained Transformers; MRR: mean reciprocal rank.



Interestingly, for either of the CliniqIR models used, the ensemble model improved the overall average performance for predicting a wide range of diagnoses (>0 notes) in the MIMIC-III data set. We also found that the RRF ensemble successfully boosted performance across diagnosis categories with both high and low note availability. On average, the RRF ensemble model performed better than either of its constituent models. Notable exceptions include the categories in which the individual mean average precision of both CliniqIR_BM25 and CliniqIR_MedCPT was <0.50 (refer to Multimedia Appendix 1 for details) and in the 100 to 250 training example range, in which the ensemble was slightly worse than the supervised model. In all other conditions, the interaction between both models (the ensemble) led to better performance.

Discussion

Principal Findings

With thousands of known diseases potentially causing a patient's condition, it is often difficult—even for experienced clinicians—to accurately diagnose every disease. Unlike the pretrained models that require user input of possible diagnoses before predictions can be made, CliniqIR represents a potential decision support tool that takes advantage of the wealth of medical literature in PubMed to generate a differential diagnosis. Our study evaluated CliniqIR's ability to formulate differentials and predict uncommon diagnoses with few or no training examples, reflecting conditions easily missed in real-life practice. Results comparing CliniqIR's performance to those of pretrained biomedical transformers in supervised and zero-shot settings highlight CliniqIR's ability to operate successfully as an unsupervised model. Therefore, our model's strength is not limited to rare and infrequent diagnosis prediction, and our model is also a useful tool for generating a first-stage differential diagnosis list. As such, a diagnostic decision support tool such as CliniqIR can enhance physician differential diagnoses and facilitate more efficient diagnoses by providing literature-guided suggestions. Beyond disease prediction, CliniqIR also demonstrates relevance in medical education as a patient-centric literature search tool. Our study demonstrated its ability to accurately cultivate a list of PubMed literature relevant to a patient's clinical narrative. This functionality could greatly improve physician researcher efficiency in performing dedicated literature reviews on behalf of their patients.

In the era of large complex neural models, it is critically important that diagnostic support tools remain simple and interpretable. In health care, where decision-making is critical and patient outcomes are at stake, clinicians' ability to understand and trust the inner workings of a diagnostic tool is paramount. In response, CliniqIR is built on retrieval systems that use simple and transparent weighting schemes to retrieve and rank important terms in a collection of documents. This transparency fosters trust in the tool's accuracy and facilitates collaboration between the tool and the health care professionals, leading to ongoing model refinement as well as enhanced clinical decision-making.

Limitations

The medical field is witnessing a growing trend in applications built on generative large language models [52]. While our work used a simpler approach, it remains valuable in scenarios with limited access to significant computing resources. In addition, it serves as a proof of concept for a retrieval-augmented medical model, potentially leading to enhanced explainability and accuracy for large language models in the health care domain.

Our study has 3 potential limitations. First, CliniqIR's knowledge source is limited to abstracts in PubMed, which has well-known publication biases toward certain diagnoses [53,54].

Therefore, the use of a single knowledge resource limits CliniqIR's generalizability to diseases and conditions not represented in the PubMed corpus. For instance, conditions such as COVID-19 and Alzheimer disease or rare diseases such as sarcoidosis and cholangitis are covered in thousands of published literature entries, whereas other conditions such as "cellulitis and abscess of the leg" or "closed fracture of the sternum" may receive less attention. Future studies will involve a review of seemingly unrepresented diagnostic codes by linking them back to their parent diagnostic codes to ensure appropriate mapping between diagnosis codes and PubMed.

Second, our main experimental results were restricted to diagnoses with at least 100 PubMed abstract representatives. We identified a significant number of *ICD-9-CM* codes in MIMIC-III with no associated medical literature among the 33 million PubMed abstracts (an overview of MIMIC-III diagnosis distribution classes can be found in Multimedia Appendix 1). We also found that CliniqIR's predictive performance improved with increasing PubMed coverage of the diagnosis, guiding our decision to establish the 100-abstract inclusion criterion for diagnoses (Multimedia Appendix 1). Future work will combine information from biomedical journals, medical textbooks, and Wikipedia for wider disease coverage.

Third, our MIMIC-III experiments limited the input to patient discharge summaries containing a succinct synopsis of a patient's hospital stay, including symptoms, diagnostic evaluation, clinical progression, and treatment information. In real-world clinical situations, such complete retrospective information would not be available during the initial diagnostic process. Therefore, the results presented in this paper represent a first feasibility study of CliniqIR and highlight some of the difficulties involved in developing diagnostic support tools.

Conclusions

In this study, we presented CliniqIR, an unsupervised retrieval-based model that leverages unstructured knowledge resources to aid in the diagnostic process. We showed that the CliniqIR models outperformed a supervised fine-tuned pretrained clinical transformer model in predicting diagnoses with <5 training samples. We also demonstrated that CliniqIR outperformed pretrained clinical transformers in making predictions for rare and complex conditions in a zero-shot setting. While many existing research studies on diagnostic prediction have focused on one disease at a time or only on highly prevalent conditions, we combined the strengths of CliniqIR and supervised learning to build a single ensemble model that aids in diagnosing a broad spectrum of conditions regardless of training data availability. Overall, our study reveals the potential of IR-based models in aiding diagnostic decision-making in an efficient, transparent, and educational manner. This work will direct future studies to facilitate successful application of machine learning and IR to building robust and accurate clinical diagnostic decision support tools.

Acknowledgments

LM's research is supported in part by a grant (T32DA013911) from the National Institutes of Health. We acknowledge support from the Open Access Publication Fund of the University of Tübingen.

Data Availability

The datasets analyzed during this study are available. The Medical Information Mart for Intensive Care III data set can be accessed with permission via the work by Johnson et al [39]. The DC3 data set can be accessed via the work by Eickhoff et al [30]. The code to implement this work is also available [55].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on model implementation, configuration, and performance comparisons across various data sets and configurations.
[DOCX File , 5518 KB - medinform_v12i1e50209_app1.docx]

References

1. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. JAMA 1987 Jul 03;258(1):67-74. [doi: 10.1001/jama.1987.03400010071030]
2. Khoong EC, Nouri SS, Tuot DS, Nundy S, Fontil V, Sarkar U. Comparison of diagnostic recommendations from individual physicians versus the collective intelligence of multiple physicians in ambulatory cases referred for specialist consultation. Med Decis Making 2022 Apr;42(3):293-302 [FREE Full text] [doi: 10.1177/0272989X211031209] [Medline: 34378444]
3. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. JAMA Netw Open 2019 Mar 01;2(3):e190096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0096] [Medline: 30821822]
4. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 06;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]

5. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. Arch Dis Child 2003 May;88(5):408-413 [FREE Full text] [doi: 10.1136/adc.88.5.408] [Medline: 12716712]
6. Müller L, Gangadharaiah R, Klein SC, Perry J, Bernstein G, Nurkse D, et al. An open access medical knowledge base for community driven diagnostic decision support system development. BMC Med Inform Decis Mak 2019 Apr 27;19(1):93 [FREE Full text] [doi: 10.1186/s12911-019-0804-1] [Medline: 31029130]
7. Prakash A, Zhao S, Hasan S, Datla V, Lee K, Qadir A, et al. Condensed memory networks for clinical diagnostic inferencing. Proc AAAI Conf Artif Intell 2017 Feb 12;31(1). [doi: 10.1609/aaai.v31i1.10964]
8. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep 2021 Feb 05;11(1):3254 [FREE Full text] [doi: 10.1038/s41598-020-74399-w] [Medline: 33547343]
9. Liu J, Zhang Z, Razavian N. Deep EHR: chronic disease prediction using medical notes. arXiv Preprint posted online August 15, 2018 [FREE Full text]
10. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One 2018 Feb 15;13(2):e0192360 [FREE Full text] [doi: 10.1371/journal.pone.0192360] [Medline: 29447188]
11. Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: a review. Egypt Inform J 2018 Nov;19(3):179-189. [doi: 10.1016/j.eij.2018.03.002]
12. Singh Kohli P, Arora S. Application of machine learning in disease prediction. In: Proceedings of the 4th International Conference on Computing Communication and Automation. 2018 Presented at: ICCCA 2018; December 14-15, 2018; Greater Noida, India. [doi: 10.1109/ccaa.2018.8777449]
13. Abdullahi T, Nitschke G, Sweijd N. Predicting diarrhoea outbreaks with climate change. PLoS One 2022 Apr 19;17(4):e0262008 [FREE Full text] [doi: 10.1371/journal.pone.0262008] [Medline: 35439258]
14. Alon G, Chen E, Savova G, Eickhoff C. Diagnosis prevalence vs. efficacy in machine-learning based diagnostic decision support. arXiv Preprint posted online June 24, 2020 [FREE Full text]
15. Rios A, Kavuluru R. Few-shot and zero-shot multi-label learning for structured label spaces. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP 2018; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1352]
16. Zhao Y, Wong ZS, Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. J Healthc Eng 2018 May 22;2018:6275435 [FREE Full text] [doi: 10.1155/2018/6275435] [Medline: 29951182]
17. Romera-Paredes B, Torr PH. An embarrassingly simple approach to zero-shot learning. In: Feris R, Lampert C, Parikh D, editors. Visual Attributes. Cham, Switzerland: Springer; Mar 22, 2017.
18. Veeranna SP, Nam J, Mencía EL, Furnkranz J. Using semantic similarity for multi-label zero-shot classification of text documents. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. 2016 Presented at: ESANN 2016; April 27-29, 2016; Bruges, Belgium.
19. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online May 28, 2020 [FREE Full text]
20. Sanh V, Webson A, Raffel C, Bach SH, Sutawika L, Alyafeai Z, et al. Multitask prompted training enables zero-shot task generalization. arXiv Preprint posted online October 15, 2021 [FREE Full text]
21. Li Y, Chao X. Semi-supervised few-shot learning approach for plant diseases recognition. Plant Methods 2021 Jun 27;17(1):68 [FREE Full text] [doi: 10.1186/s13007-021-00770-1] [Medline: 34176505]
22. Zhao Y, Lai H, Yin J, Zhang Y, Yang S, Jia Z, et al. Zero-shot medical image retrieval for emerging infectious diseases based on meta-transfer learning - worldwide, 2020. China CDC Wkly 2020 Dec 25;2(52):1004-1008 [FREE Full text] [doi: 10.46234/ccdcw2020.268] [Medline: 34594825]
23. Soldaini L, Cohan A, Yates A, Goharian N, Frieder O. Retrieving medical literature for clinical decision support. In: Proceedings of the 37th European Conference on IR Research. 2015 Presented at: ECIR 2015; March 29-April 2, 2015; Vienna, Austria. [doi: 10.1007/978-3-319-16354-3_59]
24. Hasan SA, Ling Y, Liu J, Farri O. Using neural embeddings for diagnostic inferencing in clinical question answering. In: Proceedings of the Twenty-Fourth Text REtrieval Conference. 2015 Presented at: TREC 2015; November 17-20, 2015; Gaithersburg, MD.
25. Hasan SA, Zhu X, Dong Y, Liu J, Farri O. A hybrid approach to clinical question answering. In: Proceedings of the Twenty-Third Text REtrieval Conference 2014. 2014 Presented at: TREC 2014; November 19-21, 2014; Gaithersburg, MD.
26. Text REtrieval Conference (TREC) home page. Text REtrieval Conference (TREC). URL: https://trec.nist.gov/ [accessed 2024-06-03]
27. Naik A, Parasa S, Feldman S, Wang LL, Hope T. Literature-augmented clinical outcome prediction. arXiv Preprint posted online November 16, 2021 [FREE Full text] [doi: 10.18653/v1/2022.findings-naacl.33]
28. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]

29. White J. PubMed 2.0. Med Ref Serv Q 2020 Oct 21;39(4):382-387. [doi: 10.1080/02763869.2020.1826228] [Medline: 33085945]
30. Eickhoff C, Gmehlin F, Patel AV, Boullier J, Fraser H. DC3 -- a diagnostic case challenge collection for clinical decision support. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. 2019 Presented at: ICTIR '19; October 2-5, 2019; Santa Clara, CA. [doi: 10.1145/3341981.3344239]
31. Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: Proceedings of the 2nd SIGIR Workshop on Medical Information Retrieval (MedIR). 2016 Presented at: SIGIR 2016; July 21, 2016; Pisa, Italy.
32. Zhao Z, Jin Q, Chen F, Peng T, Yu S. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. Sci Data 2023 Dec 18;10(1):909 [FREE Full text] [doi: 10.1038/s41597-023-02814-8] [Medline: 38110415]
33. Vardell E, Moore M. Isabel, a clinical decision support system. Med Ref Serv Q 2011 Apr 25;30(2):158-166. [doi: 10.1080/02763869.2011.562800] [Medline: 21534115]
34. Welcome to Apache Lucene. Apache Lucene. URL: https://lucene.apache.org/ [accessed 2024-06-03]
35. Robertson S, Walker S, Hancock-Beaulieu MM, Gatford M, Payne A. Okapi at TREC-4. In: Proceedings of the Fourth Text REtrieval Conference. 1995 Presented at: TREC-4 1995; November 1-3, 1995; Gaithersburg, MD.
36. Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics 2023 Nov 01;39(11):2023 [FREE Full text] [doi: 10.1093/bioinformatics/btad651] [Medline: 37930897]
37. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP 2019; June 7, 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1909]
38. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
39. Johnson AE, Pollard TJ, Shen LW, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
40. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
41. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv Preprint posted online March 26, 2019 [FREE Full text] [doi: 10.18653/v1/d19-1371]
42. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. arXiv Preprint posted online October 22, 2020 [FREE Full text] [doi: 10.18653/v1/2021.naacl-main.334]
43. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform 2022 Feb;126:103983 [FREE Full text] [doi: 10.1016/j.jbi.2021.103983] [Medline: 34990838]
44. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv Preprint posted online August 27, 2019. [doi: 10.18653/v1/d19-1410]
45. Cormack GV, Clarke CL, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009 Presented at: SIGIR '09; July 19-23, 2009; Boston, MA. [doi: 10.1145/1571941.1572114]
46. Kurland O, Culpepper J. Fusion in information retrieval: SIGIR 2018 half-day tutorial. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018 Presented at: SIGIR '18; July 8-12, 2018; Ann Arbor, MI. [doi: 10.1145/3209978.3210186]
47. Mugisha C, Paik I. Comparison of neural language modeling pipelines for outcome prediction from unstructured medical text notes. IEEE Access 2022;10:16489-16498. [doi: 10.1109/access.2022.3148279]
48. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inform Assoc 2019 Dec 01;26(12):1632-1636 [FREE Full text] [doi: 10.1093/jamia/ocz164] [Medline: 31550356]
49. Shen Z, Schutte D, Yi Y, Bompelli A, Yu F, Wang Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. BMC Med Inform Decis Mak 2022 Jul 07;22(Suppl 1):88 [FREE Full text] [doi: 10.1186/s12911-022-01819-4] [Medline: 35799294]
50. Craswell N. Mean reciprocal rank. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer; 2009.
51. Basic HHS Policy for Protection of Human Research Subjects. US Department of Health and Human Services, Office for Human Research Protections. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/regulatory-text/index.html#46.101
52. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. JMIR Med Educ 2024 Feb 13;10:e51391 [FREE Full text] [doi: 10.2196/51391] [Medline: 38349725]

53. McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research - a narrative review. Trials 2010 Apr 13;11(1):37 [FREE Full text] [doi: 10.1186/1745-6215-11-37] [Medline: 20388211]
54. Montori VM, Smieja M, Guyatt GH. Publication bias: a brief review for clinicians. Mayo Clin Proc 2000 Dec;75(12):1284-1288. [doi: 10.4065/75.12.1284] [Medline: 11126838]
55. CliniqIR: retrieval-based diagnostic decision support. GitHub. URL: https://github.com/rsinghlab/CliniqIR [accessed 2024-05-31]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers
CODER: cross-lingual knowledge-infused medical term embedding
DDSS: diagnostic decision support system
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
IR: information retrieval
MAP: mean average precision
MedCPT: Medical Contrastive Pre-trained Transformers
MIMIC-III: Medical Information Mart for Intensive Care III
MRR: mean reciprocal rank
PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers
RR: reciprocal rank
RRF: reciprocal rank fusion
SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers
SciBERT: Scientific Bidirectional Encoder Representations from Transformers
TF: term frequency
TF-IDF: term frequency–inverse document frequency
TREC: text retrieval conference
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 25.06.23; peer-reviewed by J Zheng, Q Jin; comments to author 06.02.24; revised version received 10.03.24; accepted 17.04.24; published 19.06.24.

Please cite as:
Abdullahi T, Mercurio L, Singh R, Eickhoff C
Retrieval-Based Diagnostic Decision Support: Mixed Methods Study
JMIR Med Inform 2024;12:e50209
URL: https://medinform.jmir.org/2024/1/e50209
doi:10.2196/50209
PMID:38896468

Implementation Report

Hjernetegn.dk—The Danish Central Nervous System Tumor Awareness Initiative Digital Decision Support Tool: Design and Implementation Report

Kathrine Synne Weile[1,2], MD; René Mathiasen[3,4], MD, PhD; Jeanette Falck Winther[2,5], MD, DMSc; Henrik Hasle[1,2], MD, PhD; Louise Tram Henriksen[1,2], MD, PhD

[1]Department of Pediatric and Adolescent Medicine, Aarhus University Hospital, Aarhus, Denmark

[2]Department of Clinical Medicine, Faculty of Health, Aarhus University, Aarhus, Denmark

[3]Department of Pediatric and Adolescent Medicine, Copenhagen University Hospital, Copenhagen, Denmark

[4]Department of Clinical Medicine, Faculty of Medicine, University of Copenhagen, Copenhagen, Denmark

[5]Danish Cancer Institute, Danish Cancer Society, Copenhagen, Denmark

Corresponding Author:
Kathrine Synne Weile, MD
Department of Pediatric and Adolescent Medicine
Aarhus University Hospital
Palle Juul-Jensens Boulevard 99
Aarhus, 8200
Denmark
Phone: 45 2927 9265
Email: kathrineweile@clin.au.dk

Abstract

Background: Childhood tumors in the central nervous system (CNS) have longer diagnostic delays than other pediatric tumors. Vague presenting symptoms pose a challenge in the diagnostic process; it has been indicated that patients and parents may be hesitant to seek help, and health care professionals (HCPs) may lack awareness and knowledge about clinical presentation. To raise awareness among HCPs, the Danish CNS tumor awareness initiative hjernetegn.dk was launched.

Objective: This study aims to present the learnings from designing and implementing a decision support tool for HCPs to reduce diagnostic delay in childhood CNS tumors. The aims also include decisions regarding strategies for dissemination and use of social media, and an evaluation of the digital impact 6 months after launch.

Methods: The phases of developing and implementing the tool include participatory co-creation workshops, designing the website and digital platforms, and implementing a press and media strategy. The digital impact of hjernetegn.dk was evaluated through website analytics and social media engagement.

Implementation (Results): hjernetegn.dk was launched in August 2023. The results after 6 months exceeded key performance indicators. The analysis showed a high number of website visitors and engagement, with a plateau reached 3 months after the initial launch. The LinkedIn campaign and Google Search strategy also generated a high number of impressions and clicks.

Conclusions: The findings suggest that the initiative has been successfully integrated, raising awareness and providing a valuable tool for HCPs in diagnosing childhood CNS tumors. The study highlights the importance of interdisciplinary collaboration, co-creation, and ongoing community management, as well as broad dissemination strategies when introducing a digital support tool.

(JMIR Med Inform 2024;12:e58886) doi:10.2196/58886

KEYWORDS

digital health initiative; digital health initiatives; clinical decision support; decision support; decision support system; decision support systems; decision support tool; decision support tools; diagnostic delay; awareness initiative; pediatric neurology; pediatric neurology; pediatric CNS tumors; CNS tumor; CNS tumour; CNS tumours; co-creation; health systems and services; communication; central nervous system

Introduction

Aim and Context

Primary tumors of the central nervous system (CNS) are the second most prevalent childhood cancer, constituting approximately 20% of all cases [1,2]. The 5-year survival has risen to 75% [3], but survivors face severe late effects [4]. In Denmark, approximately 50 patients younger than 18 years are diagnosed every year [3]. Early diagnosis is crucial for the quality of life of survivors, and early detection and diagnostic delay have been in focus for decades [5-7]. The presenting symptoms at the time of onset can be vague, adding difficulty to diagnostics [8]. Time from first symptom to diagnosis can be divided into intervals, with the potential to identify inequities in the diagnostic process [9]. The total diagnostic interval (TDI) is the sum of the patient interval (PI) and the diagnostic interval [DI] PI: the time from symptom onset to first contact to a health care professional (HCP). DI: time from first contact to an HCP to diagnosis. Intervals are shown in Figure 1).

Over the past decade, research has reported TDI ranging from 28 to 123 days [10-26]. Recent studies indicate a knowledge gap in HCPs, causing a delay in the DI, and advocate for interventions specifically aimed toward HCPs to reduce the DI [27-29]. To map the trajectory for Danish patients, a questionnaire study was undertaken in 2022. The results showed an elongated median TDI of 106 days, with DI providing the larger part of the delay, thus displaying that challenging diagnostic processes are present in Denmark too. The results in detail have been reported separately [30].

In 2017, The Danish Collaborative Comprehensive Childhood CNS Tumor Consortium (5C) was established, providing a long-term strategic research platform, to accelerate diagnostics and reduce late effects specifically in patients with CNS tumors [31]. The decision was made to create a Danish childhood CNS tumor awareness initiative, with the aim of reducing diagnostic delay. hjernetegn.dk was launched after 4 years of preparation. Figure 2 shows the phases of progression from 2020 onward.

Figure 1. The diagnostic route: visualization of diagnostic intervals. HCP: health care professional.

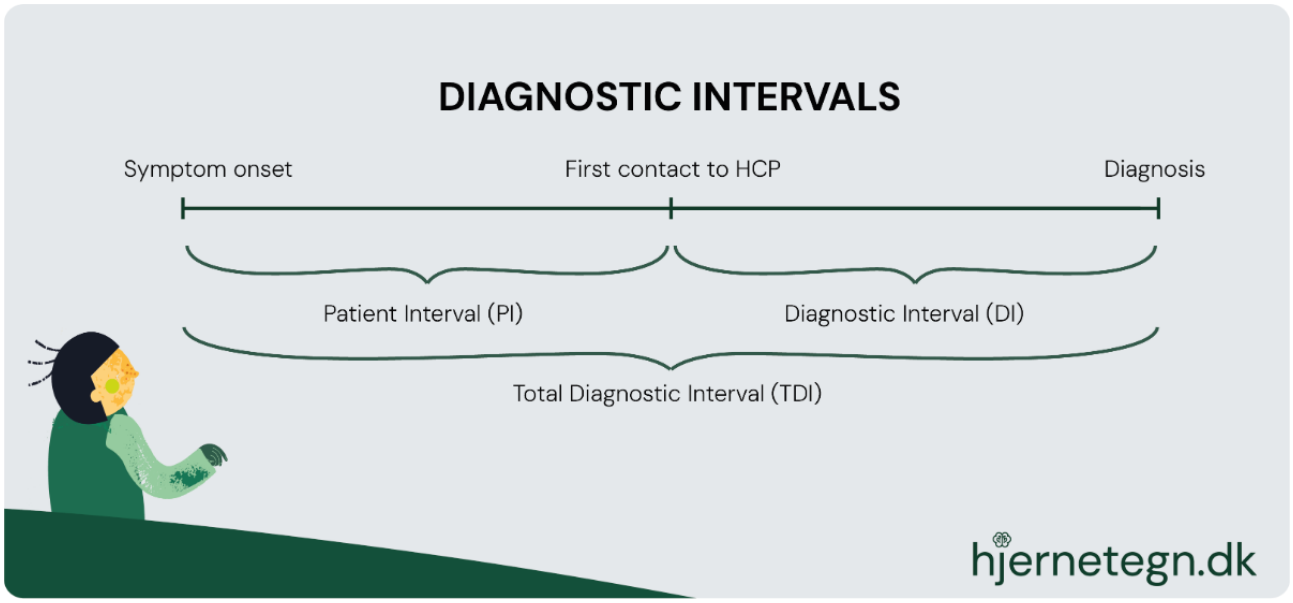
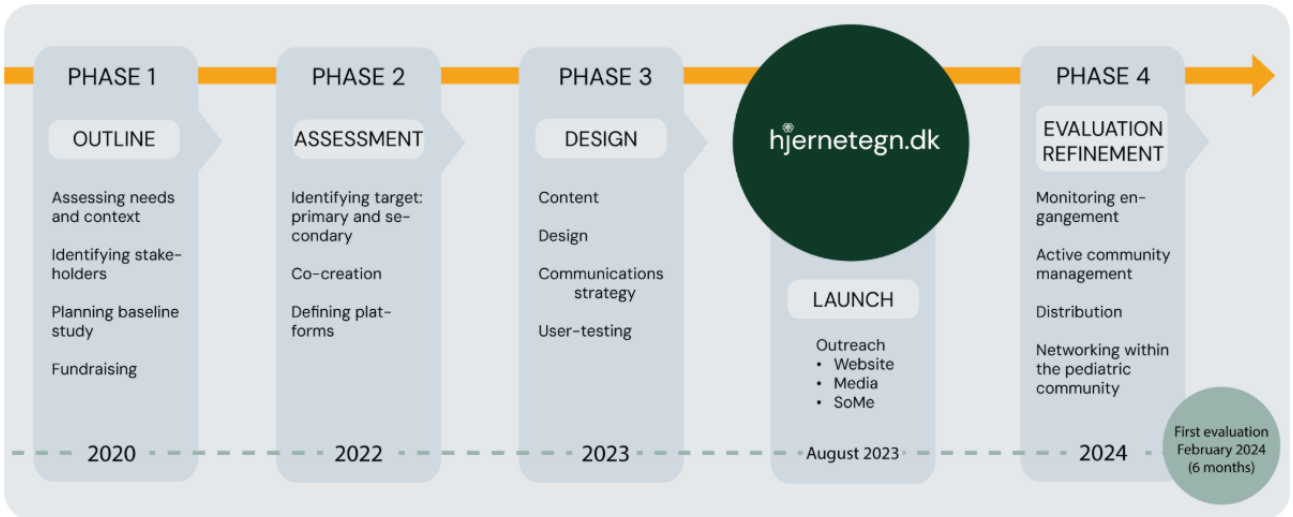


Figure 2. Workflow creating hjernetegn.dk. The initiative was launched on August 21, 2023. The evaluation of the digital impact was initiated in February 2024, 6 months after launch. SoMe: social media.



Similar Interventions

Efforts to accelerate timely diagnosis to enhance survival and the quality of life after cancer are advocated for worldwide. Childhood cancer has been a World Health Organization priority since 2018 [32]. In the United Kingdom, the HeadSmart campaign was initiated in 2011; actively reducing the time from symptom onset to diagnosis by several weeks, with a reduction in the time from onset until families approached a HCP, as well as an accelerated referral within the primary and secondary sector [18,33,34]. The HeadSmart initiative was set up as a nationwide public campaign, creating awareness in the public and among HCPs on alarm symptoms of CNS tumors in children.

Different organization of, and access to, the health care system may influence the differences in delay between countries. For hjernetegn.dk, we found inspiration in the methods of the HeadSmart campaign, but the Danish medical culture and health care organization, as well as being in an era of digitization, required designing a novel approach for the Danish initiative including constricting our target audience to HCPs.

The implementation is reported in accordance with iCHECK-DH (Guidelines for the Reporting on Digital Health Implementation) [35].

Methods

We report the approach and implementation of the Danish CNS tumor initiative hjernetegn.dk, a tool to support the diagnostic process of childhood CNS tumors in the interface between primary and secondary care.

Target

HCPs targeted for the initiative and tool were determined based on their involvement in clinical diagnostics, thus covering all HCPs who would first assess a child presenting with complaints that might be caused by a CNS tumor. Primary targets were general practitioners (GPs), ophthalmologists, and pediatricians. Secondary targets were nurse practitioners; neurologists; psychiatrists; ear, nose, and throat doctors; optometrists; child physiotherapists; and chiropractors.

We approached all specialists with a vested interest in contributing to the content and guidelines provided. We also included associations from specialties identified as recipients, involving them in the early process. National health authorities were informed prior to launch, timely enough for them to be able to respond but not with the intent to consult. Patient associations and advocacy groups were informed as well.

Participating Entities

Setting up hjernetegn.dk required interdisciplinary experts, covering academic and clinical knowledge, communication strategy skills, and digital know-how. A qualified team of pediatric oncologists formulated the idea, nested within the 5C, and then included the Danish Cancer Society and the Danish Childhood Cancer Foundation in the collaboration to support the commercial and communications strategical part of the initiative.

Hjernetegn.dk: Web Use, Website, and Social Media

We invited specialists from the primary target groups to participate in a co-creation workshop, encouraging "a collaborative approach of creative problem solving between stakeholders" [36]. The approach was set up to create a user journey map to identify the workflow in general practice and pediatric clinics, and to define what would be required in content and design to make the decision support tool feasible and successful.

The established key points were carried into the process of defining which platforms to make our tool accessible. In a postpandemic online era, printed material is obsolete in a Danish setting. To enable timely updated content, online platforms were used: a website as a nest for the tool and social media (SoMe) platforms to nest the outreach and network. We chose LinkedIn as the platform for SoME outreach based on its profile as a professional community for audience augmentation, with an algorithm that enables outreach by profession, favoring peer-to-peer communication [37]. Furthermore, funds were allocated to develop a LinkedIn campaign and ads on Google Search.

Blueprint Summary: Hjernetegn.dk

The hjernetegn.dk website offers a list of alarm symptoms that require assessment to rule out a CNS tumor; a checklist for examination, listing what to include in the primary assessment; and finally, a decision support tool, offering a hands-on algorithm to decide whether to approach with watchful waiting, reevaluate within a certain timeframe, or refer directly for further evaluation including neuroimaging.

Content was provided by pediatric oncologists in collaboration with stakeholders from relevant specialties. Using national Danish guidelines and the National Institute of Health and Care Excellence–accredited Brain Pathways guideline [38] from the HeadSmart initiative as a backbone, website material applicable to a Danish setting was developed for the digital era, framing dissemination in a primarily digital strategy.

An example of the assessment tool is shown in Figures 3-5. Figure 3 displays alarm symptoms, Figure 4 displays points for examination and assessment, and Figure 5 displays the decision support tool.

A communications bureau provided design and graphical solutions.

The design accommodated the reported timeframe from the co-creation, allowing only 1 minute to unravel the algorithm. The aim was to facilitate an overview and to promote awareness of other symptoms to consider while using the tool for a specific assessment.

Figure 3. Alarm symptoms on hjernetegn.dk. CNS: central nervous system.

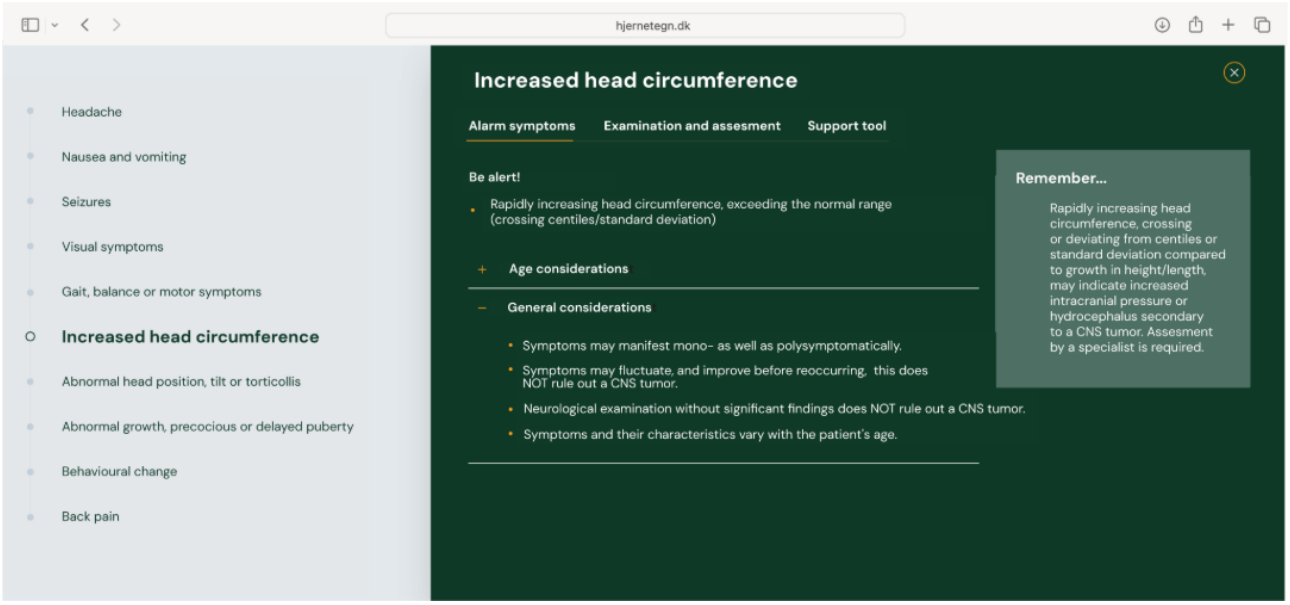


Figure 4. Examination and assessment on hjernetegn.dk. CNS: central nervous system.

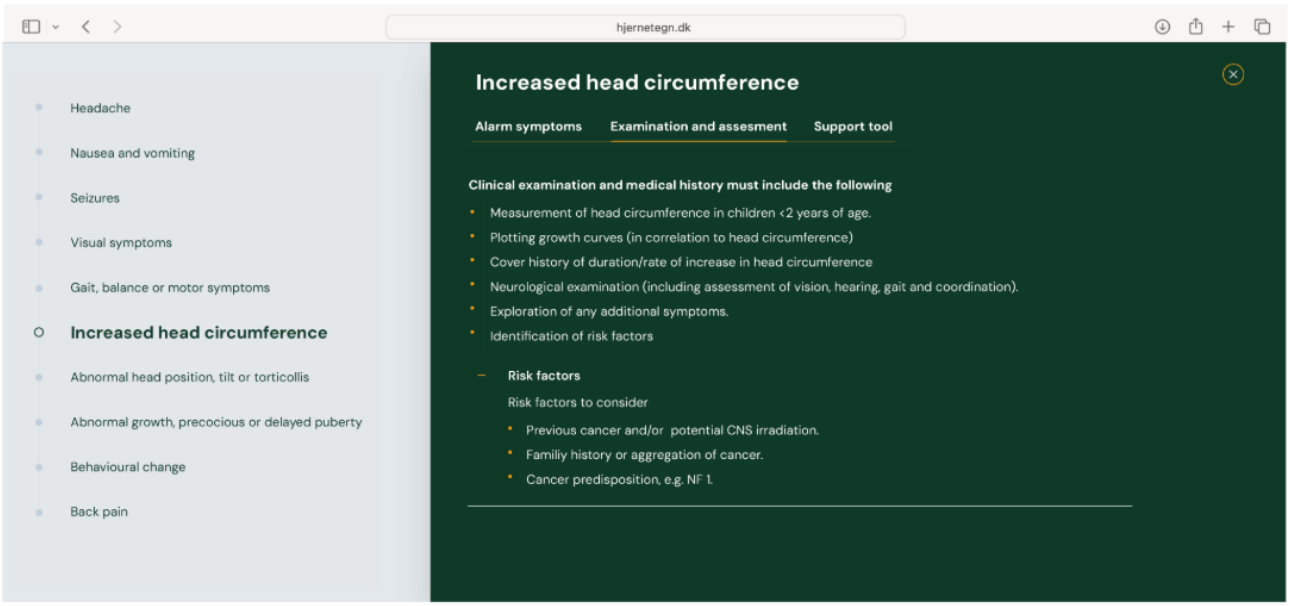
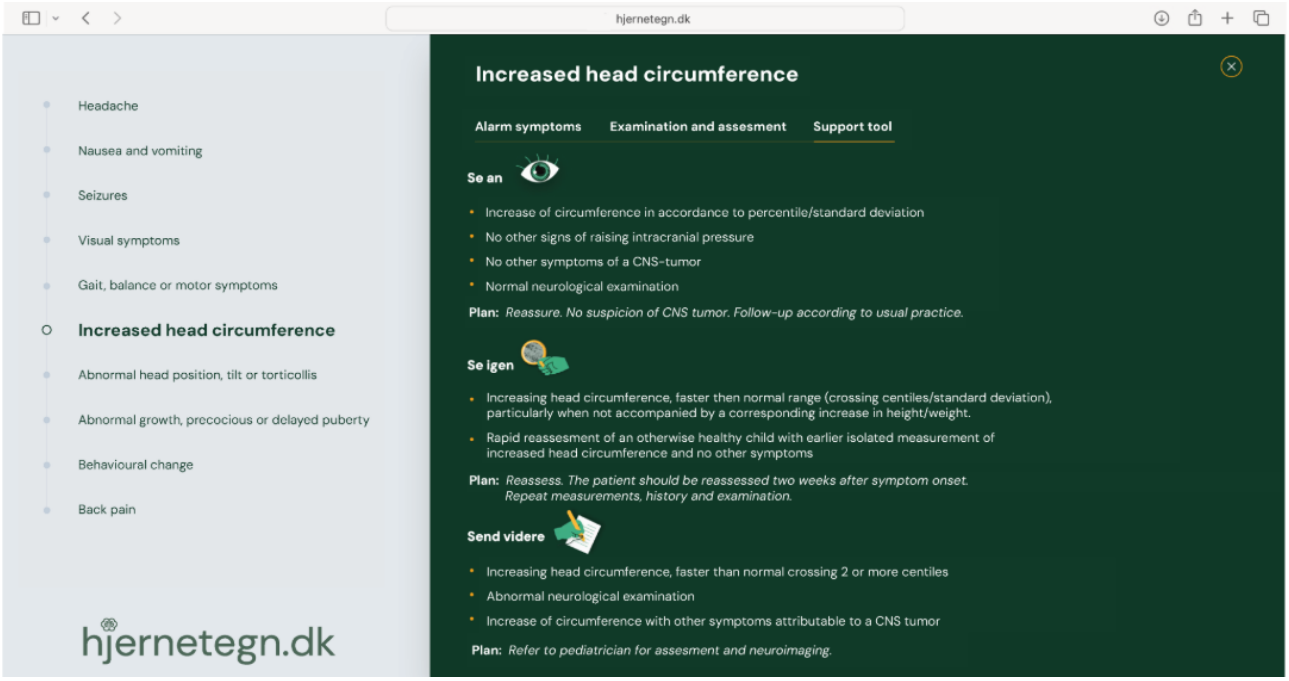


Figure 5. Support tool on hjernetegn.dk. CNS: central nervous system.



Press/Media and Digital Strategy

The press/media strategy aimed to create awareness as well as drive users to the website when launched. Interviews were set up with major Danish health news outlets to be released on the day of launch. Within the first 3 months after release, 8 written articles were published in trade journals.

A written press release carrying information on the initiative and a graphical package was provided to all stakeholders, thereby inviting them to share the launch on their platforms and with their networks.

In the digital strategy, sponsored LinkedIn ads were planned for 3 iterations of advertising, each 3 weeks long (August 21, 2023, and 12 weeks after), and a Google Search strategy and ads first ran for 2 months, with plans to restrategize later. Furthermore, a plan for the organic content for the LinkedIn profile was designed. To embrace the algorithm, weekly posts were planned.

Goals for Digital Performance the First 6 Months After Launch

We expected high activity at launch and in the following months, aligning with press releases and published articles. We presumed we would reach a more indicative plateau after 2 months.

Key performance indicators (KPIs), measurable variables for the evaluation of an initiative, were defined covering the number of visitors to the website, number of impressions on LinkedIn (impressions are defined as the number of exposures to posts or ads of individual feeds on the platform; organic impressions are defined as tailored content managed by hand), and number of searches on Google Search.

Defining the KPIs for using the website, we set a standard estimate of the possible number of assessments of children who visit their general practice, suggesting the number of cases where using the guideline and tool provided in hjernetegn.dk would be relevant. In consensus with experts from general practice, it was estimated that less than 20% of children's visits with their GP require active use of guidelines, and by the same consensus, most likely <1 patient per week would require the GP to use hjernetegn.dk. With 3500 practicing doctors in Denmark, the KPI for visitors to the website was set to 500 visits per month. On LinkedIn, the KPI was set to 180,000 impressions, estimated from the number of possible recipients by the stated occupation on the platform. In Google Search, the KPI was set to 500 clicks per month, estimated from GP usage and searches. For the web page bounce rate KPI, the percentage of users departing from the entry page without further interaction was targeted as <55%.

Data

Data analysis to monitor digital use and traffic was conducted utilizing Piwik PRO Analytics (version 16.26; Piwik Pro) [39]. Details on LinkedIn activity were extracted directly from the platform.

Budget and Resources

The initiative was managed as part of a PhD study at Aarhus University Hospital, endorsed by the pediatric neuro-oncology network.

In the 4-year period, the allocation of work hours varied, with workload increasing and peaking at the time of launch. The project manager and a communications officer from the Danish Cancer Society had 20% of a full-time equivalent (FTE) workload from 2020 to 2022. Starting in 2023, the workload increased to 100% FTE for the project manager and 70% FTE for the communications officer.

Brand and graphic design, technical support, and digital solutions were provided by an external communications bureau. Collaboration and joint strategy were initiated 18 months before launch. The budget for the external bureau was approximately €100,000 (US $108,000), covering all external costs for building

and implementing the initiative, as well as running the project 6 months after launch. Funding of approximately €65,000 (US $70,000) was allocated to refine, recommunicate, and manage the web page and SoMe platforms the second year after launch.

Sustainability

The initiative is intended to become a part of a clinician's toolbox, implying sustained use.

Implementation to stay is highly dependent on ongoing advocacy and a well-planned strategy for communication and up-to-date management of the community and web pages. Dissemination, through advertising, participating in relevant events, and reaching out on digital and real-life platforms are continuously required. Consequently, continuous funding is crucial.

Managing the community day to day requires active surveillance on SoMe channels to enable fast and sufficient responses, totaling 5 hours of work per week, inevitably spread out and not only as office hours.

Ethical Considerations and Potential Barriers

It was important to balance the need to identify patients presenting strong indications for neuroimaging without instigating high numbers of unnecessary referrals and contacts to specialty units in both the primary and secondary sectors.

We acknowledged the risk and barrier that GPs might be reluctant to support the initiative due to worries of added workload in the primary sector.

Introducing an initiative as the one described, does not require ethical approval in the state of Denmark, when initiated by specialist networks as the 5C. Furthermore, when conducting the mentioned baseline studies [31], ethical approval was obtained from the Danish Data Protection Agency (file number: 1-16-02-300-19).

Implementation (Results)

Digital Results and Engagement

hjernetegn.dk had 13,705 visitors from August 21, 2023, to February 25, 2024. Each visit had a mean of 6 engaging actions (click, scrolling, or other interactions). Figure 6 shows the number of visitors and returning visitors to the website in the first 6 months after launch. Access from smartphones and desktops covered 78% and 19%, respectively. The remaining 3% was accessed from other smart devices.

From August 21 to February 25 the overall bounce rate was 63%. On the page level, the bounce rate was <10%, indicating the use of the tool within the site. For desktop users, the bounce rate was 46%. The average time spent on the page was 1 minute and 45 seconds, and the average number of page views was 2.3.

By November, a plateau was met, showing an average of 300 visits per week, with a return rate of 10%. The bounce rate was reduced to 54% accordingly. The average time spent on a web page was 2 minutes, and the average number of web page views was 2.5 web pages.

On the website, 33,075 pages were viewed, with 60% (n=19,713 views, bounce rate 60%) of visits being the home page.

Half of all visitors accessed the web pages through Google Search, and 24% accessed them by URL. SoMe outlets constituted 21% of the visitors, particularly LinkedIn and Facebook. Less than 5% of visitors were directed by referral from other platforms, such as guidelines, health authority outlets, health care–related media, and collaborators' websites.

The sponsored LinkedIn ad campaign constituted the most impressions, covering 201,888 impressions and reaching the set KPI. Organic impressions formed 14,133 impressions in the same period. Figure 7 shows sponsored and organic impressions in the first 6 months.

Figure 6. Website engagement 6 months after launch. Visitors on hjernetegn.dk from launch August 2023 to February 2024. The lower graph shows the number of visitors after the plateau was reached from November 2023 to February 2024.

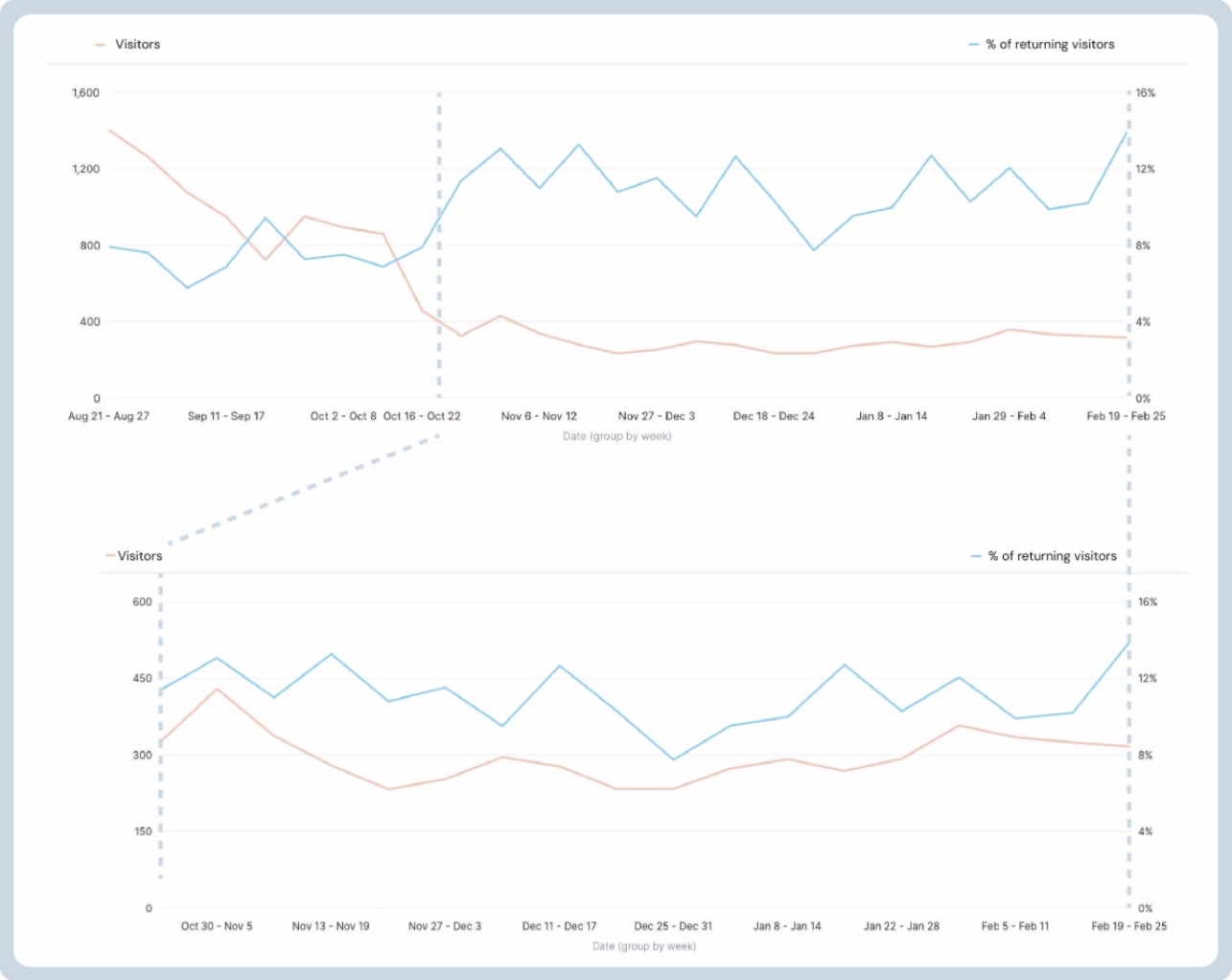
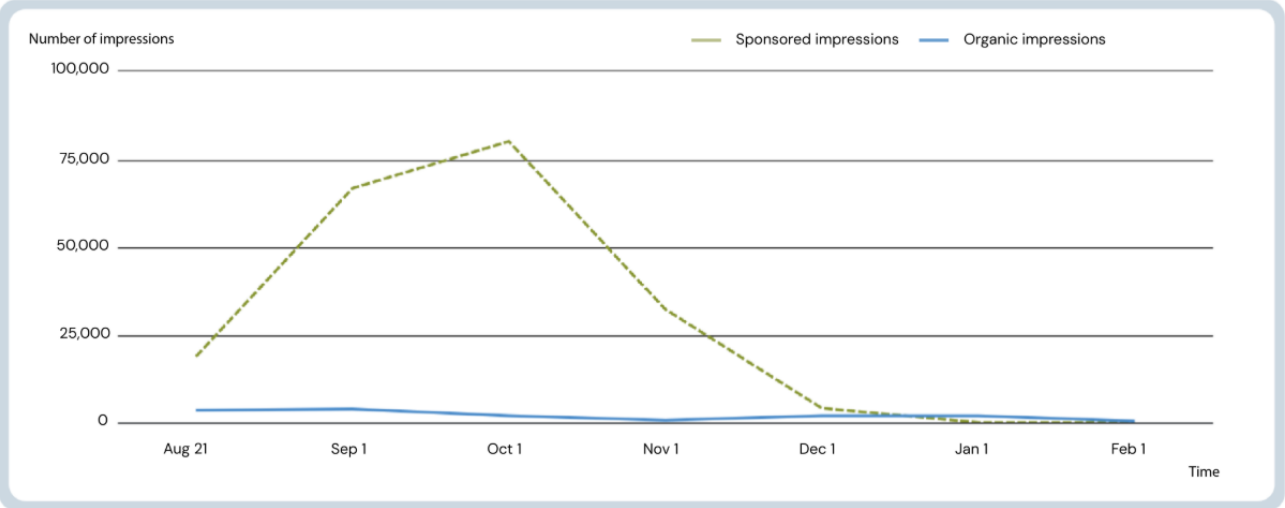


Figure 7. Impressions of active content on LinkedIn 6 months after launch. Impressions indicate number of exposures to individual user feeds. Sponsored impressions: impressions by ads. Organic impressions: tailored content managed directly from the hjernetegn.dk LinkedIn-page.



Implementation to Sustain

hjernetegn.dk is nested within the 5C and the four pediatric oncology centers nationwide. Efforts to maintain and update the website as well as strategizing to reach even more HCPs are ongoing. We are planning to solicit feedback from site users to better target HCPs and enhance the site and tool.

The hjernetegn.dk approach and guidelines are now taught in medical programs and among pediatric fellows as part of the

curriculum, and the tool is continuously being introduced to GPs.

Discussion

This study reports the development and implementation of hjernetegn.dk and the digital impact in the first 6 months after launch. The initiative covered the development of the digital diagnostic decision support tool and dissemination.

Impact of the Digital Dissemination Strategy

Contemplating results from the digital strategy, as anticipated, we saw a very high interest following the main events around the launch, and from August 21 to October 22, the number of visitors peaked in alignment with publications in health media. From November, a plateau was reached, showing a stable number of visitors and returning visitors. When introducing new eHealth innovations, it is proposed in other works that the Gartner Hype Cycle can be applied [40,41]. It shows that reaching a plateau of productivity is natural but will be preceded by a so-called peak of inflated expectations and a trough of inflated expectations; although there was no true trough, the suggested phases in implementation matched our digital results, and the plateau was met with approximately 50 users daily, of whom 10% were returning visitors.

The results met our set KPIs. In fact, the number of visitors to the website far exceeded our expectations. At the plateau, the monthly number of visitors reached 1500, while the KPI was originally set to 500 within the first 6 months. This might have been a low target, but it was estimated based on expert notions of the number of contacts and that new guidelines take time to implement. Our results indicate that the communications strategy in our initiative has some effect. Using a so-called push-pull strategy, SoMe and ads endorsed and directed users to the website, introducing potential users to the tool, and the website as the actual tool had the effect of pulling in viewers. The bounce rate dropping over time indicates the tool is in actual use for supporting diagnostics and not only for visitors viewing the website.

For the LinkedIn campaign, the main focus was to create awareness and curiosity. A recent review on SoMe in public health campaigning presents the modified hierarchy of effect model, suggesting that health outcomes can be improved simply by SoMe exposure changing individuals' behaviors [42]. Engagement leads to further impressions in secondary networks, raising awareness, directing traffic to the website, and supporting the push-pull strategy setup. Access from users via Facebook may be the result of user engagement on LinkedIn.

Moving forward, we will adjust set KPIs to encompass the results from the first 6 months of the initiative. Furthermore, it is being discussed whether to include more SoMe platforms.

Lessons Learned

Inviting key stakeholders and primary targets to identify drivers and solutions and to point out barriers in the process is essential to design for sustainability. Finding a common understanding of the problem of diagnostic delay, especially when approaching GPs, was a time-consuming and delicate process. GPs were hesitant and implied that the motivation for the initiative was a criticism of their efforts in clinical practice, whereas the actual motivation was to provide support in a challenging process. The same issue was raised with the external communications consultants. A clear objective and common vocabulary is imperative for effective communication and collaboration between clinicians in different fields and the design producers. Barnes et al [43] advocate for interdisciplinary team approaches, bringing together specialists across the fields of health sciences, IT, and communications when producing initiatives in a computational health science era. We support this structural mindset as well as stress the importance of co-creation for content, setting the scene for successful health care interventions.

Through our efforts, we have strived to but not succeeded in reaching all GPs. In the interface between the analog and digital eras, substantial funds and resources are required to reach all areas. We realize some individuals will not be reached unless contacted directly.

The multiphase setup allowed stakeholders to participate continuously, which strengthened the implementation and is strongly advised. The design and platforms ensure flexible editing, allowing add ons easily. This enables future changes or revisions, or even more tools, to be embedded when needed.

When initiating a similar initiative, we would recommend the following:

- Establish a multidisciplinary team by involving communication experts, graphic designers, and technical designers from an early stage. Common ground and understanding of the aim during production is of utmost importance.
- Invite and engage stakeholders to define drivers for success and possible positive or negative attitudes toward the effort.
- Willingness to "kill your darlings." As senders and clinical experts, be receptive to feedback from stakeholders if they do not see eye to eye.
- Have sufficient time and funding. Time is money, and fundraising takes time. In our case, it took 4 years from the formation of the project group to the launch of hjernetegn.dk. We had initially planned for 5 to 8 months.
- Develop realistic plans for follow-up and monitoring. Initiatives should be dynamic and easy to edit; create ongoing revision plans. If possible, conduct a baseline study so that both digital impact and impact in the clinic, as well as on the primary goal, namely the patients, can be measured and evaluated, and the initiative improved accordingly.

Perspectives

We believe our study design is applicable to many rare diseases. Initiatives such as the website presented in this study create a hub for all applications necessary to recognize and interpret cardinal symptoms, followed by guidelines to support relevant and timely referrals. In this first edition, the initiative was targeted toward HCPs. In time, we will revisit the strategy and may plan to broaden the format to raise public awareness, thus covering the entire trajectory for patients with a CNS tumor.

Acknowledgments

We owe gratitude to all participants in workshops and tests throughout the process of developing hjernetegn.dk. Our gratitude also goes to colleagues and collaborators across specialties, sectors, and regions for offering time and expertise providing insight to their domain and knowledge to ours. Furthermore, we thank the Department of Prevention and Information at The Danish Cancer Society and The Danish Childhood Cancer Foundation for their support. Finally, we wish to acknowledge Professor David Walker and his team at the Children's Brain Tumour Research Centre, Nottingham, United Kingdom, for their initiative HeadSmart, which laid the ground stones for this and many projects to be built.

hjernetegn.dk was made possible by donations from The Danish Cancer Society and The Danish Childhood Cancer Foundation.

Conflicts of Interest

None declared.

References

1. Baade PD, Youlden DR, Valery PC, Hassall T, Ward L, Green AC, et al. Trends in incidence of childhood cancer in Australia, 1983-2006. Br J Cancer 2010 Feb 02;102(3):620-626 [FREE Full text] [doi: 10.1038/sj.bjc.6605503] [Medline: 20051948]
2. Grabas MR, Kjaer SK, Frederiksen MH, Winther JF, Erdmann F, Dehlendorff C, et al. Incidence and time trends of childhood cancer in Denmark, 1943-2014. Acta Oncol 2020 May;59(5):588-595. [doi: 10.1080/0284186X.2020.1725239] [Medline: 32048526]
3. Helligsoe ASL, Kenborg L, Henriksen LT, Udupi A, Hasle H, Winther JF. Incidence and survival of childhood central nervous system tumors in Denmark, 1997-2019. Cancer Med 2022 Jan;11(1):245-256 [FREE Full text] [doi: 10.1002/cam4.4429] [Medline: 34800006]
4. Kenborg L, Winther JF, Linnet KM, Krøyer A, Albieri V, Holmqvist AS, ALiCCS study group. Neurologic disorders in 4858 survivors of central nervous system tumors in childhood-an Adult Life after Childhood Cancer in Scandinavia (ALiCCS) study. Neuro Oncol 2019 Jan 01;21(1):125-136 [FREE Full text] [doi: 10.1093/neuonc/noy094] [Medline: 29850875]
5. Wilne S, Collier J, Kennedy C, Koller K, Grundy R, Walker D. Presentation of childhood CNS tumours: a systematic review and meta-analysis. Lancet Oncol 2007 Aug;8(8):685-695. [doi: 10.1016/S1470-2045(07)70207-3] [Medline: 17644483]
6. Wilne S, Koller K, Collier J, Kennedy C, Grundy R, Walker D. The diagnosis of brain tumours in children: a guideline to assist healthcare professionals in the assessment of children who may have a brain tumour. Arch Dis Child 2010 Jul;95(7):534-539. [doi: 10.1136/adc.2009.162057] [Medline: 20371594]
7. Walker D, Hamilton W, Walter FM, Watts C. Strategies to accelerate diagnosis of primary brain tumors at the primary-secondary care interface in children and adults. CNS Oncol 2013 Sep;2(5):447-462 [FREE Full text] [doi: 10.2217/cns.13.36] [Medline: 25054667]
8. Crawford J. Childhood brain tumors. Pediatr Rev 2013 Feb;34(2):63-78. [doi: 10.1542/pir.34-2-63] [Medline: 23378614]
9. Weller D, Vedsted P, Rubin G, Walter FM, Emery J, Scott S, et al. The Aarhus statement: improving design and reporting of studies on early cancer diagnosis. Br J Cancer 2012 Mar 27;106(7):1262-1267 [FREE Full text] [doi: 10.1038/bjc.2012.68] [Medline: 22415239]
10. Ramaswamy V, Remke M, Shih D, Wang X, Northcott PA, Faria CC, et al. Duration of the pre-diagnostic interval in medulloblastoma is subgroup dependent. Pediatr Blood Cancer 2014 Jul;61(7):1190-1194. [doi: 10.1002/pbc.25002] [Medline: 24616042]
11. Arnautovic A, Billups C, Broniscer A, Gajjar A, Boop F, Qaddoumi I. Delayed diagnosis of childhood low-grade glioma: causes, consequences, and potential solutions. Childs Nerv Syst 2015 Jul;31(7):1067-1077 [FREE Full text] [doi: 10.1007/s00381-015-2670-1] [Medline: 25742877]
12. Stocco C, Pilotto C, Passone E, Nocerino A, Tosolini R, Pusiol A, et al. Presentation and symptom interval in children with central nervous system tumors. A single-center experience. Childs Nerv Syst 2017 Dec;33(12):2109-2116. [doi: 10.1007/s00381-017-3572-1] [Medline: 28808765]
13. Azizi AA, Heßler K, Leiss U, Grylli C, Chocholous M, Peyrl A, et al. From symptom to diagnosis-the prediagnostic symptomatic interval of pediatric central nervous system tumors in Austria. Pediatr Neurol 2017 Nov;76:27-36. [doi: 10.1016/j.pediatrneurol.2017.08.006] [Medline: 28935367]
14. Boutahar FZ, Benmiloud S, El Kababri M, Kili A, El Khorassani M, Allali N, et al. Time to diagnosis of pediatric brain tumors: a report from the Pediatric Hematology and Oncology Center in Rabat, Morocco. Childs Nerv Syst 2018 Dec;34(12):2431-2440 [FREE Full text] [doi: 10.1007/s00381-018-3927-2] [Medline: 30054805]

15. Coven SL, Stanek JR, Hollingsworth E, Finlay JL. Delays in diagnosis for children with newly diagnosed central nervous system tumors. Neurooncol Pract 2018 Nov;5(4):227-233 [FREE Full text] [doi: 10.1093/nop/npy002] [Medline: 31386013]
16. Gilli IO, Joaquim AF, Tedeschi H, Dos Santos Aguiar S, Morcillo AM, Ghizoni E. Factors affecting diagnosis of primary pediatric central nervous system neoplasias in a developing country. Childs Nerv Syst 2019 Jan;35(1):91-96. [doi: 10.1007/s00381-018-3958-8] [Medline: 30250987]
17. Patel V, McNinch NL, Rush S. Diagnostic delay and morbidity of central nervous system tumors in children and young adults: a pediatric hospital experience. J Neurooncol 2019 Jun;143(2):297-304. [doi: 10.1007/s11060-019-03160-9] [Medline: 30929127]
18. Shanmugavadivel D, Liu J, Murphy L, Wilne S, Walker D, HeadSmart. Accelerating diagnosis for childhood brain tumours: an analysis of the HeadSmart UK population data. Arch Dis Child 2020 Apr;105(4):355-362. [doi: 10.1136/archdischild-2018-315962] [Medline: 31653616]
19. Hirata K, Muroi A, Tsurubuchi T, Fukushima H, Suzuki R, Yamaki Y, et al. Time to diagnosis and clinical characteristics in pediatric brain tumor patients. Childs Nerv Syst 2020 Sep;36(9):2047-2054. [doi: 10.1007/s00381-020-04573-y] [Medline: 32157367]
20. Maaz AUR, Yousif T, Saleh A, Pople I, Al-Kharazi K, Al-Rayahi J, et al. Presenting symptoms and time to diagnosis for pediatric central nervous system tumors in Qatar: a report from Pediatric Neuro-Oncology Service in Qatar. Childs Nerv Syst 2021 Feb;37(2):465-474 [FREE Full text] [doi: 10.1007/s00381-020-04815-z] [Medline: 32710251]
21. Lu P, Raynald, Liu W, Gong J, Sun T, Li C, et al. Factors impacting time to diagnosis in pediatric CNS tumors in Chinese children. Support Care Cancer 2021 Jul;29(7):3633-3642. [doi: 10.1007/s00520-020-05863-6] [Medline: 33179135]
22. Barragán-Pérez EJ, Altamirano-Vergara CE, Alvarez-Amado DE, García-Beristain JC, Chico-Ponce-de-León F, González-Carranza V, et al. The role of time as a prognostic factor in pediatric brain tumors: a multivariate survival analysis. Pathol Oncol Res 2020 Oct;26(4):2693-2701 [FREE Full text] [doi: 10.1007/s12253-020-00875-3] [Medline: 32661835]
23. Yamada Y, Kobayashi D, Terashima K, Kiyotani C, Sasaki R, Michihata N, et al. Initial symptoms and diagnostic delay in children with brain tumors at a single institution in Japan. Neurooncol Pract 2021 Feb;8(1):60-67 [FREE Full text] [doi: 10.1093/nop/npaa062] [Medline: 33664970]
24. Goldman RD, Cochrane DD, Dahiya A, Mah H, Buttar A, Lambert C, et al. Finding the needle in the hay stack: population-based study of prediagnostic symptomatic interval in children with CNS tumors. J Pediatr Hematol Oncol 2021 Nov 01;43(8):e1093-e1098. [doi: 10.1097/MPH.0000000000002012] [Medline: 33235150]
25. Rask O, Nilsson F, Lähteenmäki P, Ehrstedt C, Holm S, Sandström PE, et al. Prospective registration of symptoms and times to diagnosis in children and adolescents with central nervous system tumors: A study of the Swedish Childhood Cancer Registry. Pediatr Blood Cancer 2022 Nov;69(11):e29850 [FREE Full text] [doi: 10.1002/pbc.29850] [Medline: 35727740]
26. Jovanović A, Ilić R, Pudrlja Slović M, Paripović L, Janić D, Nikitović M, et al. Total diagnostic interval in children with brain tumours in a middle-income country: national experience from Serbia. Childs Nerv Syst 2023 Nov;39(11):3169-3177 [FREE Full text] [doi: 10.1007/s00381-023-05958-5] [Medline: 37097460]
27. Rajagopal R, Moreira DC, Faughnan L, Wang H, Naqvi S, Krull L, et al. An international multicenter survey reveals health care providers' knowledge gap in childhood central nervous system tumors. Eur J Pediatr 2023 Feb;182(2):557-565. [doi: 10.1007/s00431-022-04712-4] [Medline: 36383283]
28. Jogendran M, Ronsley R, Goldman RD, Cheng S. Perceived barriers to the time to diagnosis of central nervous system tumors in children: surveying the perspectives from the frontline. J Pediatr Hematol Oncol 2021 Nov 01;43(8):e1262-e1265. [doi: 10.1097/MPH.0000000000002245] [Medline: 34133384]
29. HeadSmart Jordan. URL: https://headsmartjordan.khcc.jo/ [accessed 2024-01-07]
30. Weile KS, Helligsoe ASL, von Holstein SL, Winther JF, Mathiasen R, Hasle H, et al. Patient- and parent-reported diagnostic delay in children with central nervous system tumors in Denmark. Pediatr Blood Cancer 2024 Aug;71(8):e31128. [doi: 10.1002/pbc.31128] [Medline: 38814259]
31. Danish Collaborative Comprehensive Childhood CNS Tumour Consortium. 2023. URL: https://www.5c.nu/ [accessed 2024-07-05]
32. WHO global initiative for childhood cancer: an overview. World Health Organization. 2020. URL: https://www.who.int/docs/default-source/documents/health-topics/cancer/who-childhood-cancer-overview-booklet.pdf [accessed 2024-01-07]
33. Shanmugavadivel D, Walker D, Liu J, Wilne S. HeadSmart: are you brain tumour aware? Paediatr Child Health 2016 Feb;26(2):81-86. [doi: 10.1016/j.paed.2015.10.006]
34. HeadSmart Be Brain Tumour Aware. A new clinical guideline from the Royal College of Paediatrics and Child Health with a national awareness campaign accelerates brain tumor diagnosis in UK children--"HeadSmart: Be Brain Tumour Aware". Neuro Oncol 2016 Mar;18(3):445-454 [FREE Full text] [doi: 10.1093/neuonc/nov187] [Medline: 26523066]
35. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]

36. Vargas C, Whelan J, Brimblecombe J, Allender S. Co-creation, co-design, co-production for public health - a perspective on definition and distinctions. Public Health Res Pract 2022 Jun 15;32(2):3222211. [doi: 10.17061/phrp3222211] [Medline: 35702744]
37. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]
38. The Brain Pathways guideline: a guideline to assist healthcare professionals in the assessment of children who may have a brain tumour. Better Safe Than Tumour. 2017. URL: https://bettersafethantumour.com/app/uploads/2022/07/diagnosis_of_brain_tumours_in_children_guideline_-_full_report.pdf [accessed 2024-01-07]
39. Piwik PRO. 2024. URL: https://piwik.pro/ [accessed 2024-07-05]
40. Linden A, Fenn J. Understanding Gartner's Hype Cycles. Ask-force.org. 2003 May 30. URL: http://ask-force.org/web/Discourse/Linden-HypeCycle-2003.pdf [accessed 2024-07-05]
41. Versluis A, van Luenen S, Meijer E, Honkoop PJ, Pinnock H, Mohr DC, et al. SERIES: eHealth in primary care. Part 4: addressing the challenges of implementation. Eur J Gen Pract 2020 Dec;26(1):140-145 [FREE Full text] [doi: 10.1080/13814788.2020.1826431] [Medline: 33025820]
42. Kite J, Chan L, MacKay K, Corbett L, Reyes-Marcelino G, Nguyen B, et al. A model of social media effects in public health communication campaigns: systematic review. J Med Internet Res 2023 Jul 14;25:e46345 [FREE Full text] [doi: 10.2196/46345] [Medline: 37450325]
43. Barnes M, Hanson C, Giraud-Carrier C. The case for computational health science. J Healthc Inform Res 2018;2(1):99-110 [FREE Full text] [doi: 10.1007/s41666-018-0024-y] [Medline: 29974076]

Abbreviations

5C: The Danish Collaborative Comprehensive Childhood CNS Tumor Consortium
CNS: central nervous system
DI: diagnostic interval
FTE: full-time equivalent
GP: general practitioner
HCP: health care professional
iCHECK-DH: Guidelines for the Reporting on Digital Health Implementation
KPI: key performance indicator
SoMe: social media
TDI: total diagnostic interval

Edited by C Perrin; submitted 15.04.24; peer-reviewed by S Sarbadhikari, G Myreteg; comments to author 13.05.24; revised version received 22.05.24; accepted 02.07.24; published 25.07.24.

Please cite as:
Weile KS, Mathiasen R, Winther JF, Hasle H, Henriksen LT
Hjernetegn.dk—The Danish Central Nervous System Tumor Awareness Initiative Digital Decision Support Tool: Design and Implementation Report
JMIR Med Inform 2024;12:e58886
URL: https://medinform.jmir.org/2024/1/e58886
doi:10.2196/58886
PMID:

Original Paper

Alarm Management in Provisional COVID-19 Intensive Care Units: Retrospective Analysis and Recommendations for Future Pandemics

Maximilian Markus Wunderlich[1*], MSc; Nicolas Frey[1*], MSc; Sandro Amende-Wolf[1]; Carl Hinrichs[2], Dr med; Felix Balzer[1*], Prof Dr med, Dr rer nat; Akira-Sebastian Poncette[1*], PD, Dr med

[1]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

[2]Department of Nephrology and Medical Intensive Care, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

[*]these authors contributed equally

Corresponding Author:
Akira-Sebastian Poncette, PD, Dr med
Institute of Medical Informatics
Charité – Universitätsmedizin Berlin
Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health
Invalidenstraße 90
Berlin, 10115 Berlin
Germany
Phone: 49 030 450 581 018
Email: akira-sebastian.poncette@charite.de

Abstract

Background: In response to the high patient admission rates during the COVID-19 pandemic, provisional intensive care units (ICUs) were set up, equipped with temporary monitoring and alarm systems. We sought to find out whether the provisional ICU setting led to a higher alarm burden and more staff with alarm fatigue.

Objective: We aimed to compare alarm situations between provisional COVID-19 ICUs and non–COVID-19 ICUs during the second COVID-19 wave in Berlin, Germany. The study focused on measuring alarms per bed per day, identifying medical devices with higher alarm frequencies in COVID-19 settings, evaluating the median duration of alarms in both types of ICUs, and assessing the level of alarm fatigue experienced by health care staff.

Methods: Our approach involved a comparative analysis of alarm data from 2 provisional COVID-19 ICUs and 2 standard non–COVID-19 ICUs. Through interviews with medical experts, we formulated hypotheses about potential differences in alarm load, alarm duration, alarm types, and staff alarm fatigue between the 2 ICU types. We analyzed alarm log data from the patient monitoring systems of all 4 ICUs to inferentially assess the differences. In addition, we assessed staff alarm fatigue with a questionnaire, aiming to comprehensively understand the impact of the alarm situation on health care personnel.

Results: COVID-19 ICUs had significantly more alarms per bed per day than non–COVID-19 ICUs (P<.001), and the majority of the staff lacked experience with the alarm system. The overall median alarm duration was similar in both ICU types. We found no COVID-19–specific alarm patterns. The alarm fatigue questionnaire results suggest that staff in both types of ICUs experienced alarm fatigue. However, physicians and nurses who were working in COVID-19 ICUs reported a significantly higher level of alarm fatigue (P=.04).

Conclusions: Staff in COVID-19 ICUs were exposed to a higher alarm load, and the majority lacked experience with alarm management and the alarm system. We recommend training and educating ICU staff in alarm management, emphasizing the importance of alarm management training as part of the preparations for future pandemics. However, the limitations of our study design and the specific pandemic conditions warrant further studies to confirm these findings and to explore effective alarm management strategies in different ICU settings.

(JMIR Med Inform 2024;12:e58347) doi:10.2196/58347

KEYWORDS

patient monitoring; intensive care unit; ICU; alarm fatigue; alarm management; patient safety; alarm system; alarm system quality; medical devices; clinical alarms; COVID-19

Introduction

Background

Patients critically ill with COVID-19 frequently experience multiple organ failure and need life-sustaining measures such as continuous renal replacement therapy, mechanical ventilation, or even extracorporeal membrane oxygenation [1]. Therefore, during major COVID-19 outbreaks (eg, in 2019 and 2020), intensive care units (ICUs) worldwide saw an increase in admission rates of up to 300% [2]. Intensive care capacities became stretched; in some regions, treatment facilities were fully occupied [3]. Hospitals thus had to increase their intensive care capacities to cope with the proliferating number of patients with COVID-19 [4]. As a consequence, in many countries, provisional ICUs were set up [5-9].

In provisional COVID-19 ICUs, as in standard ICU settings, patients require continuous monitoring of their vital signs, which include heart rate, cardiac rhythm, and peripheral oxygen saturation (SpO_2) via pulse oximetry. If an abnormal situation occurs (eg, a vital sign exceeds a predetermined threshold), an alarm is issued to alert clinicians [10]. However, ICU staff today often experience alarm fatigue [11] because of high rates of false positive or nonactionable alarms [12]. Alarm fatigue can be hazardous for patient safety, especially when critical alarms are missed [13,14]. The constant noise can cause high levels of stress among staff and disrupt patients' sleep-wake cycles, increasing the risk of delirium and potentially poorer recovery outcomes [15,16].

In COVID-19 ICUs, the alarm situation might have been even worse: ICU staff, likely overworked due to an overwhelming patient load and a massive workload, had to cope with the psychological burden of the fear of infection and the distressing reality of witnessing numerous patient deaths. ICU staff experienced poor mental health during the COVID-19 pandemic [17-19], and more errors in clinical settings occurred [20].

Moreover, the makeshift nature of the ICUs, which were set up in haste to handle the surge in patients, likely exacerbated these problems. The often improvised and provisional nature of these ICUs likely introduced additional operational complexities and stressors, such as limited space and insufficient technical equipment, further taxing the already strained staff. All these issues may have been compounded by a loud and potentially unreliable alarm system.

Objectives

This study systematically assesses the alarm situations in 2 provisional COVID-19 ICUs and 2 non–COVID-19 ICUs, focusing on both the actual alarm rates and the extent of alarm fatigue experienced by ICU staff. By incorporating expert interviews with medical professionals who have worked in COVID-19 ICUs, alongside quantitative analyses of alarm log data from the bedside monitors [21-23] and results from the alarm fatigue questionnaire [24], this investigation follows a multimodal approach.

Methods

Study Design

In this study, the alarm situations in 2 provisional COVID-19 ICUs were compared with those in 2 non–COVID-19 ICUs; all 4 ICUs were situated on the same campus at a tertiary care university hospital in Berlin, Germany. A retrospective observational analytical study design was used, with the comparison being approached from three perspectives: (1) hypotheses were derived and tested based on interviews with medical experts, exploring potential differences in alarm situations between COVID-19 and non–COVID-19 ICUs; (2) alarm log data from both types of ICUs were analyzed; and (3) the results from an alarm fatigue questionnaire, which was administered as part of another study [24] during the same period, were collected and analyzed.

The non–COVID-19 ICUs consist of an interdisciplinary ICU with patients who have recently undergone surgery or are in the perioperative stage (hereinafter *surgical ICU*); and an ICU with a focus on internal medicine (hereinafter *medical ICU*) where patients with severe infection, cardiac diseases, kidney failure, or other single- or multiple-organ failures are treated.

The medical and surgical ICUs have single and multiple bedrooms and contain 20 and 21 beds, respectively. The COVID-19 ICU was divided into 2 separate units: COVID-19 ICU-A had single bedrooms with 16 beds, and COVID-19 ICU-B had single and multiple bedrooms with 24 beds.

Expert Interviews for Hypothesis Design

To derive hypotheses regarding how alarm situations might differ between COVID-19 and non–COVID-19 ICUs, 5 medical experts (physicians: n=2, 40%; nurses: n=3, 60%) were interviewed from July to December 2021. The interviews were semistructured [25,26]. The foundational questions were grounded in our overarching research question: "What distinguishes the alarm situations in COVID-19 ICUs from those in non–COVID-19 ICUs?" A detailed list of these questions can be found in Table S1 in Multimedia Appendix 1.

In addition, 2 job shadowing sessions were conducted (one on October 1, 2021, in the medical ICU; and another on November 12, 2021, in COVID-19 ICU-B). During the second session, an interview was conducted in COVID-19 ICU-B. The remaining 4 interviews were held online via Microsoft Teams [27]. An integral aspect of the interviews was that beyond the structured questions, the medical experts were proactively asked about potential hypotheses regarding the differences in alarm situations between the ICU types.

Subsequent to the interviews, the collected data, including transcripts and notes, were meticulously examined. The

hypotheses were derived through an in-depth examination and collective discussion within an interdisciplinary team.

Statistical Analysis

Data Analysis

The alarm logs were processed and analyzed as previously described [22], using R (R Foundation for Statistical Computing) [28] with RStudio (Posit Software, PBC) [29] in combination with the following packages: *lubridate* [30], *ggplot2* [31], and *dplyr* [32]. The logs were extracted from the Philips IntelliVue patient monitoring systems (in the surgical and medical ICUs: MX800, software version *m*; in the COVID-19 ICUs: MX750, software version *p*) as CSV files from the non–COVID-19 ICUs and as XML files from the COVID-19 ICUs (structure of the cleaned alarm log data frames is listed in Table S2 Multimedia Appendix 1). Alarm signals from mechanical ventilation devices were not correctly stored in the alarm log data due to transmission errors and were therefore excluded from the analysis. The included alarm signals and the assignment to their medical device are listed in Table S3 in Multimedia Appendix 1. The logs range across 113 days (from November 19, 2020, to March 11, 2021).

Hypothesis Testing

To test for statistical differences between COVID-19 and non–COVID-19 ICUs, the units were grouped accordingly. All tests were 1-tailed, with a significance level of α=.05. *P* values were adjusted using the Bonferroni correction. The first hypothesis (H1) proposed that the alarm load is higher in COVID-19 ICUs than in non–COVID-19 ICUs. This hypothesis was further subdivided into total alarm load, clinical red alarms, clinical yellow alarms, and technical alarms.

The second hypothesis (H2) posited that more alarms are issued from specific medical devices, namely electrocardiogram (ECG) and invasive blood pressure (IBP) devices, in COVID-19 ICUs than in non–COVID-19 ICUs. The alarms from these devices were further categorized based on the alarm color (red: potentially life-threatening events; yellow: vital signs exceed predetermined thresholds).

The third hypothesis (H3) proposed that the alarm duration is longer in COVID-19 ICUs than in non–COVID-19 ICUs. For this hypothesis, the alarm durations from clinical alarms were subdivided into medical devices (non-IBP [NIBP], temperature, SpO_2, ECG, and IBP devices) and alarm colors (red and yellow).

For H1 and H2, the alarm load was quantified in alarms per bed per day. Given that the distributions were skewed to the right and approximately gamma distributed, a dummy-coded no-intercept generalized linear model with a log link function was used. Cohen *d* was used as the effect size measure and was calculated using the package *effsize* developed by Torchiano [33].

For H3, testing was conducted with median alarm durations. The distributions of alarm durations were highly skewed to the right and approximately exponentially distributed; therefore, nonparametric bootstrapping with the *boot* package developed by Canty et al [34] was used with an a priori estimation of the difference in alarm duration with equation 1.

$$H_0: \mu_1 - \mu_2 \leq 3 \quad \textbf{(1)}$$

$$H_A: \mu_1 - \mu_2 > 3$$

Differences of 3 seconds were considered significant, while smaller effects that could already be significant due to the large sample were considered not significant. A median-based estimator for Cohen *d*–type effect size (equation 2) with an estimation of variance (equation 3) was used, where *k* is the number of units and *l* is the number of alarms.

To test the differences in alarm fatigue experienced by ICU staff between COVID-19 and non–COVID-19 ICUs (H4), an unpaired 1-tailed *t* test using Cohen *d* as the effect size measure was conducted.

Exploratory Data Analysis

Metrics defined in our previous study [22] were used for the evaluation of alarm situations in the ICUs: alarms per bed per day, critical alarms, alarms per device, alarm flood conditions (≥10 alarms within 10 minutes), use of the alarm pause function per bed per day, proper pause-to-pause ratio, and concurrent alarm duration per bed per day. The last metric was first introduced by Varisco et al [12] and is calculated by summing the number of active parallel alarms. Specifically, if 2 alarms sound simultaneously within a second, this is counted as 1 second of concurrent alarm duration, and if 3 alarms sound simultaneously within a second, 2 seconds are counted.

The metrics were related to bed occupancy and time period to compare the results between the different ICUs. Due to the absence of information regarding the cause of termination in the data, it was not possible to determine the alarm response time. Therefore, alarm duration was used instead.

In the calculations of alarm flood conditions and concurrent alarm duration, only alarms with an auditory modality, specifically yellow and red alarms, were included. For the alarm duration used in calculating concurrent alarm duration, a cutoff was set at 1800 seconds. Alarm durations exceeding this limit were considered outliers and were therefore excluded from the analysis.

Alarm Fatigue Questionnaire

The questionnaire data were taken from a separate study [24] that coincided with our data collection phase. The questionnaire, distributed as a web-based survey via REDCap (Research Electronic Data Capture; Vanderbilt University) [35] to ICU staff at the same German hospital between April and June 2021, provided responses from COVID-19 ICU-A, COVID-19 ICU-B, a third COVID-19 ICU (COVID-19 ICU-C), and the 2 non–COVID-19 ICUs (medical and surgical ICUs). The original questionnaire consisted of 27 items; however, only those aligning with the 9-item questionnaire developed by Wunderlich et al [36] were included in this analysis. Each item was measured on a 5-point Likert scale, ranging from *I strongly agree* to *I strongly disagree*.

Demographic questions about work experience, place of work, and position (nurses and physicians, as well as support staff, ie, students or nurses from general wards) were also part of the questionnaire. Only responses from participants who consented to data analysis were included in the study. Submissions with 1 or 2 missing items were imputed at random based on the predictive mean matching algorithm using 1 imputation with the *mice* package [37]. To calculate an alarm fatigue score, the items were scored from –2 (*I strongly disagree*) to 2 (*I strongly agree*). *I partly agree* was scored with 0. Four items were scored reversely. The sum of all Likert items results in the alarm fatigue score, which ranges from –18 to 18. A score of –18 would indicate that the staff members are not experiencing alarm fatigue at all, while a score of 18 would mean that the staff members are experiencing extreme alarm fatigue; the midpoint is 0. We report the alarm fatigue in percentage as recommended by Wunderlich et al [24] with equation 4.

Ethical Considerations

Ethics approval for this study was granted by the ethics commission of Charité–Universitätsmedizin Berlin (EA4/218/20). All participants provided consent before the study. Data confidentiality was ensured through anonymization in compliance with General Data Protection Regulation. No compensation was provided to participants.

Results

Beginning with insights from expert interviews to formulate our hypotheses, we proceeded to test them empirically using alarm log data and an alarm fatigue questionnaire, concluding with insights from our exploratory data analysis.

Expert Interviews for Hypothesis Design

Of the 5 expert interviewees, 2 (40%) reported that only approximately one-third of the staff in COVID-19 ICUs had experience in intensive care, while the remaining two-thirds consisted of nurses, who until then had only worked on general wards; or individuals without specific experience, such as medical students; and even untrained personnel. The staff were assigned different tasks depending on their qualifications. Of the 2 physicians, 1 (50%) suggested that staff were not trained on how to properly apply sensors, such as ECG electrodes, potentially leading to additional (medically irrelevant) alarms. Of the 3 nurses, 1 (33%) suggested that alongside the alarm burden, the high fatality rate among patients imposes psychological strains on the staff. According to the interviewees, the patient cohort in the COVID-19 ICUs presented a more or less homogeneous clinical picture with varying COVID-19 severity. Many patients needed mechanical ventilation and underwent continuous renal replacement therapy with dialysis devices that produce very loud and unpleasant alarms. However, these alarms were not recorded in the alarm logs.

Of the 2 physicians, 1 (50%) reported that patients with COVID-19 infection often have multiple organ failure and a high length of stay; therefore, they are often equipped with ≥7 perfusors for medications (eg, antibiotics, catecholamines, sedatives, and parenteral nutrition), all of which trigger additional alarms that are also not recorded in the alarm logs. All 3 nurses reported a higher alarm burden compared with those working in non–COVID-19 ICUs. All interviewees described the removal and reattachment of all patient sensors during transition from prone to supine position or vice versa as a possible cause of false alarms if the alarm pause function was not used. On the basis of this information, we hypothesized as follows: (H1) The alarm load is higher in COVID-19 ICUs than in non–COVID-19 ICUs.

Of the 3 nurses, 2 (67%) reported that they perceived many patients to be multimorbid, overweight, and of older age, with many having a cardiac or pulmonary history and tachycardia. All interviewees reported that the blood circulation of patients with COVID-19 was extremely unstable, which led us to hypothesize that medical devices related to blood circulation (ie, the ECG, NIBP, or IBP devices) issue more alarms in COVID-19 ICUs: (H2) More alarms are issued from ECG, NIBP, and IBP devices in COVID-19 ICUs than in non–COVID-19 ICUs.

Both COVID-19 ICUs featured long corridors with inward-opening doors, which were usually closed to isolate patients who were infectious. The interviewees reported that the corridors, the dispensary room, and the physician's room were not equipped with central monitoring; thus, the health care providers had no overview of the patients, which made it difficult to locate the origin of the auditory alarms. Only the nurses' room was equipped with central monitoring, but alarms could not be turned off remotely. According to the interviewees, temporary arrangements were made by the staff to address this problem, such as leaving the doors to patients' rooms slightly open or placing speakers linked to the monitors in the hallway. However, to respond to or turn off an alarm, staff had to enter the patient's room. This required them to don personal protective equipment (ie, gloves, a protective hood, a polypropylene protective gown, and a face shield or goggles), which took approximately 30 seconds and hindered quick movement between rooms. All 3 nurses reported that the protective equipment did not interfere with turning off the alarms, the usability of the monitor displays, or the adjustment of monitor settings. However, they all described it as strenuous and time consuming. Accordingly, we derived the third hypothesis: (H3) The alarm duration is higher in COVID-19 ICUs than in non–COVID-19 ICUs.

Patients in COVID-19 ICUs often present with severe, complex medical conditions that require close monitoring and interventions. Of the 3 nurses, 1 (33%) suggested that next to the alarm burden and heavy workload, the high fatality rate among patients could strain and distress staff psychologically. This mental and emotional strain could potentially affect staff performance and cognitive abilities, potentially hindering their response to alarms. Combining this information with our reasoning for the previous 3 hypotheses, we formulated the fourth hypothesis as follows: (H4) Staff alarm fatigue is higher in COVID-19 ICUs than in non–COVID-19 ICUs.

Having delineated the hypotheses informed by the expert interviews, we proceeded to empirically test each of them, starting with the alarm load in different ICU settings.

Hypothesis 1: The Alarm Load Is Higher in COVID-19 ICUs Than in Non–COVID-19 ICUs

Significant differences were observed in all 4 tests, which confirmed our hypothesis that the alarm load was higher in provisional COVID-19 ICUs than in non–COVID-19 ICUs (P<.001; Tables 1 and 2). COVID-19 ICUs experienced an average of 23% more alarms in total, 41% more red critical alarms, and 24% more yellow alarms compared with non–COVID-19 ICUs. The alarm load caused by technical alarms was 109% higher in COVID-19 ICUs. The alarm load results, subdivided by medical device and alarm color, are reported in Figure S1 in Multimedia Appendix 1. The technical alarm signal that resulted in the most alarms per bed per day was *ECG lead off*, with mean values recorded as follows: surgical 3.42 (SD 2.10), medical 3.84 (SD 2.87), COVID-19 ICU-A 7.22 (SD 1.93), and COVID-19 ICU-B 7.77 (SD 2.41).

Table 1. Alarm load from all intensive care units (ICUs; n=113).

	Alarm load (alarms per bed per day)			
	Surgical ICU, mean (SD)	Medical ICU, mean (SD)	COVID-19 ICU-A, mean (SD)	COVID-19 ICU-B, mean (SD)
Total alarms	122.84 (38.95)	122.54 (27.85)	142.47 (41.49)	157.40 (42.58)
Red alarms	10.94 (2.43)	10.99 (2.82)	14.99 (5.89)	15.81 (4.01)
Yellow alarms	108.48 (37.86)	107.71 (26.68)	126.23 (37.93)	140.56 (41.08)
Technical alarms	5.10 (2.17)	5.03 (2.97)	10.30 (2.30)	10.84 (2.51)

Table 2. The results from the hypothesis testing for H1 (n=113).

	Alarm load (alarms per bed per day)		P value	Cohen d
	Non–COVID-19 ICUs[a], mean (SD)	COVID-19 ICUs, mean (SD)		
Total alarms	122.21 (22.37)	151.26 (31.02)	<.001	1.04
Red alarms	10.98 (1.90)	15.51 (3.33)	<.001	1.67
Yellow alarms	108.33 (21.73)	134.59 (29.24)	<.001	1.02
Technical alarms	5.06 (1.87)	10.62 (1.86)	<.001	2.98

[a]ICU: intensive care unit.

Hypothesis 2: More Alarms From ECG, NIBP, and IBP Devices Are Issued in COVID-19 ICUs Than in Non–COVID-19 ICUs

Table 3 shows the results of hypothesis testing for alarms issued by the IBP and ECG devices, subdivided by alarm color and ICU type. In both ICU types, the IBP device was responsible for the majority of the alarms, followed by the ECG and NIBP devices. Yellow alarms issued by the IBP and NIBP devices and red alarms issued by the ECG device occurred significantly more often in COVID-19 ICUs (P<.001). However, we did not find significant differences in the occurrence of yellow ECG alarms and red IBP alarms. The results of the alarm load, subdivided by medical device and alarm color, from all ICUs are presented in Table S4 in Multimedia Appendix 1. While certain alarm types exhibited significant differences, the overall impact was not profound enough to affirm the second hypothesis, with the exception of the notable difference in the frequency of red ECG alarms (as indicated by the Cohen d value of 1.13).

Table 3. Results of the hypothesis testing for H2 (n=113).

Devices	Alarm load (alarms per bed per day)		*P* value	Cohen *d*[a]
	Non–COVID-19 ICUs[b], mean (SD)	COVID-19 ICUs, mean (SD)		
ECG[c] (yellow)	39.01 (17.65)	38.79 (18.70)	.99	0.00
ECG (red)	2.51 (1.18)	4.60 (2.34)	<.001	1.13
IBP[d] (yellow)	53.87 (11.38)	66.16 (18.22)	<.001	0.81
IBP (red)	5.16 (1.14)	5.22 (1.34)	.99	0.05
NIBP[e]	0.34 (0.40)	0.56 (0.51)	<.001	0.49

[a]Cohen *d* has been reported in absolute values.

[b]ICU: intensive care unit.

[c]ECG: electrocardiogram.

[d]IBP: invasive blood pressure.

[e]NIBP: noninvasive blood pressure.

Hypothesis 3: The Alarm Duration Is Higher in COVID-19 ICUs Than in Non–COVID-19 ICUs

All ICUs had a median clinical alarm duration of 10 seconds. Yellow alarms issued by the NIBP device had the longest alarm durations (Table 4; all group sizes are reported in Table S5 in Multimedia Appendix 1). The durations of yellow alarms triggered by IBP and ECG devices were shorter than those of alarms from all other medical devices, a pattern consistent across all types of ICUs. The median duration of technical alarms in COVID-19 ICUs was significantly longer. Yellow alarms issued by the NIBP, temperature, and SpO_2 devices had significantly longer durations in COVID-19 ICUs (*P*<.001; Table 5; all group sizes are reported in Table S5 in Multimedia Appendix 1). The results of the median alarm duration, subdivided by medical device and alarm color, are presented in Figure S2 in Multimedia Appendix 1. The results show a mixed picture: <50% of the total alarms had a significant difference in alarm duration between the ICU types, implying that most alarm durations did not differ significantly between the 2 ICU types. As such, our data do not provide sufficient empirical evidence to support H3.

Table 4. Results of the hypothesis testing for H3 across each type of intensive care unit (ICU).

	Alarm duration (s)			
	Surgical ICU, median (IQR)	Medical ICU, median (IQR)	COVID-19 ICU-A, median (IQR)	COVID-19 ICU-B, median (IQR)
Clinical alarms	10 (4-27)	10 (4-25)	10 (3-28)	10 (3-29)
Technical alarms	7 (4-66)	4 (4-62)	13 (4-65)	14 (4-68)

Table 5. Results of the hypothesis testing for H3 across the 2 types of intensive care units (ICUs).

Alarms	Alarm duration (s)		*P* value	Cohen *d*[a] type
	Non–COVID-19 ICUs, median (IQR)	COVID-19 ICUs, median (IQR)		
Technical	5 (3-27)	14 (4-67)	<.001	0.016
NIBP[b]	62 (25.00-179.75)	99 (29-337)	<.001	0.094
Temperature	24 (8-86)	32 (9-192)	<.001	0.009
SpO_2[c] (yellow)	14 (7-32)	23 (11-57)	<.001	0.069
SpO_2 (red)	28 (11-89)	33 (14-75)	.11	0.004
IBP[d] (yellow)	13 (6-32)	11 (4-30)	.99	0.014
IBP (red)	24 (9-84)	20 (9-48)	.99	0.005
ECG[e] (yellow)	4 (3-9)	3 (2-7)	.99	0.014
ECG (red)	23 (9-119)	16 (8-35)	.99	0.005

[a]Cohen *d* has been reported in absolute values.

[b]NIBP: noninvasive blood pressure.

[c]SpO_2: peripheral oxygen saturation.

[d]IBP: invasive blood pressure.

[e]ECG: electrocardiogram.

Hypothesis 4: Staff Alarm Fatigue Is Higher in COVID-19 ICUs Than in Non–COVID-19 ICUs

The questionnaire was completed by 707 participants (n=78, 11% returned blank questionnaires; n=44, 6.2% returned incomplete questionnaires). Of the 585 participants who returned complete questionnaires, we included 144 (24.6%) in the analysis. Of these 144 participants, 88 (61.1%) were from non–COVID-19 ICUs (n=32, 36% from the medical ICU; n=56, 64% from the surgical ICU), and 56 (38.9%) were from COVID-19 ICUs (n=48, 86% from COVID-19 ICU-A; n=8, 14% from COVID-19 ICU-B). The majority of the respondents (92/144, 63.9%) were intensive care nurses (COVID-19 ICUs: 25/56, 45%; non–COVID-19 ICUs: 67/88, 76%). The COVID-19 ICUs had a notable proportion of additional support staff among the respondents, including nursing students and nurses from regular wards (24/56, 43%) compared with non–COVID-19 ICUs (9/88, 10%). The least represented group among the respondents were physicians (COVID-19 ICUs: 7/56, 12%; non–COVID-19 ICUs: 9/88, 10%). The overall alarm fatigue score was higher in COVID-19 ICUs (mean 56.00, SD 15.80) than in non–COVID-19 ICUs (mean 55.27, SD 13.76). Statistical testing of the alarm fatigue score revealed no significant differences between COVID-19 and non–COVID-19 ICUs ($t_{105.41}$=0.2841; *P*=.39; Cohen *d*=0.05). Importantly, when considering only experienced ICU staff—nurses and physicians—the alarm fatigue scores were significantly higher in COVID-19 ICUs than in non–COVID-19 ICUs (t_{109}=1.7332; *P*=.04; Cohen *d*=0.363). Figure S3 in Multimedia Appendix 1 depicts the results of the questionnaire, subdivided by profession. Nurses and physicians reported a higher alarm fatigue score than support staff, who generally reported an overall low alarm fatigue score in both ICU types. Given the results of the hypothesis testing (Table 6), we cannot conclusively validate H4, especially when considering all staff types; however, among ICU staff, there is evidence suggesting higher alarm fatigue in COVID-19 ICUs than in non–COVID-19 ICUs.

Table 6. Results of the hypothesis testing for H4.

	Alarm fatigue questionnaire scores		*P* value	Cohen *d*
	Non–COVID-19 ICUs, mean (SD)	COVID-19 ICUs, mean (SD)		
All participants	55.27 (13.76)	56.00 (15.80)	.39	0.050
Clinicians (nurses and physicians)	59.00 (13.05)	64.28 (14.20)	.04	0.363

Insights From the Exploratory Data Analysis

While the surgical and medical ICUs were approximately fully occupied over the entire period, the COVID-19 ICUs were often only partly occupied depending on the patient load. Average bed occupancy was 97.26% and 92.27% in the surgical and medical ICUs, respectively; and 88.79% and 82.75% in COVID-19 ICU-A and COVID-19 ICU-B, respectively. Figure S4 in Multimedia Appendix 1 displays the unit occupation over the entire period.

The use of the alarm pause function per bed per day was substantially less frequent in COVID-19 ICUs (mean 5.08, SD

1.69) compared with alarm pauses per bed per day in non–COVID-19 ICUs (mean 12.21, SD 2.21). The medical ICU recorded the highest proper pause-to-pause ratio of 0.08, followed by the surgical ICU and COVID-19 ICU-A, both at 0.04. COVID-19 ICU-B had the lowest pause-to-pause ratio: 0.03. Figure S5 in Multimedia Appendix 1 displays the number of threshold changes per bed per day and profile changes per bed per day from all ICUs.

In COVID-19 ICUs, alarm flood conditions per bed per day occurred on average 35% more frequently (COVID-19 ICU-A: mean 2.60, SD 1.29; COVID-19 ICU-B: mean 2.73, SD 1.47) compared with their non–COVID-19 counterparts (surgical ICU: mean 1.71, SD 1.33; medical ICU: mean 1.95, SD 0.92). In addition, COVID-19 ICUs experienced on average 27% more instances of concurrent alarm duration per bed per day (mean 5201.76, SD 1156.00) than non–COVID-19 ICUs (mean 4101.42, SD 965.00).

Discussion

Overview

We compared alarm situations in 2 provisional COVID-19 ICUs with those in 2 non–COVID-19 ICUs. Interviews with nurses and physicians who worked in the COVID-19 ICUs led us to hypothesize that COVID-19 ICUs have a higher alarm load, a higher number of specific alarm signals, and longer-sounding alarms than non–COVID-19 ICUs. We also hypothesized that staff working in COVID-19 ICUs experience more alarm fatigue than staff working in non–COVID-19 ICUs.

There Was a Higher Alarm Load in the Provisional COVID-19 ICUs

COVID-19 ICUs had a significantly higher alarm load from red, yellow, and technical alarms. This higher alarm load led to an increased number of alarm flood conditions and concurrent alarm duration, escalating the nurses' workload and potentially causing sensory overload [38]. While some differences can be attributed to the COVID-19 condition itself—such as the higher number of red critical alarms, possibly due to the high mortality rate and bad conditions of the patients—we suspect that most differences in alarm load were due to the interaction of staff with the alarm system (eg, not using the pause function when turning a patient from prone to supine position or vice versa, not adjusting thresholds specifically to patients' conditions, not installing or not using monitoring profiles specific to patients with COVID-19 infection, or improperly applying sensors).

Patients With COVID-19 Infection Had Similar Alarm Signals as Those Without COVID-19 Infection

We anticipated that some devices—such as ECG, NIBP, and IBP devices—would generate more alarms in COVID-19 ICUs due to the unique physiological manifestations of COVID-19, but this was not the case. Only red ECG alarms were notably more frequent in COVID-19 ICUs than in non–COVID-19 ICUs; other numbers of different clinical alarm signals were similar across all ICUs and seem to be a recurring theme because similar results were reported in previous studies [12,22].

In both ICU types, *ECG lead off* was the most frequent technical alarm signal. Interestingly, it occurred more than twice as often in COVID-19 ICUs. This also might be attributed to practices such as moving patients between prone and supine positioning, which necessitates the removal and subsequent reattachment of all sensors and electrodes each time.

The Alarm Duration Was Equally Long in Both ICU Types

While our initial theory posited that alarm durations would be longer in COVID-19 ICUs—owing to the time required for staff to don protective equipment before entering a patient's room—our findings did not confirm this hypothesis. We found no significant differences in the overall median alarm durations between COVID-19 and non–COVID-19 ICUs, except in the case of a few medical devices.

Devices displaying visual information such as ECG waveforms allow staff to rapidly assess alarm urgency. By contrast, the numerical values displayed by devices such as those measuring temperature or NIBP require more time for staff to interpret, potentially causing longer alarm durations.

We must also acknowledge the impact of other factors, for example, the floor layout of the unit, different unit policies [23], nurse-patient ratio [39], and the individual traits of the staff members [40].

Interestingly, red alarm durations were consistently longer than yellow alarm durations across all medical devices and in both ICU types. This may be due to health care providers promptly turning off yellow alarms first (without checking the patient's condition) [39], placing more emphasis on critical alarms, and only then checking the patient's condition.

In our exploratory analyses, we found that the alarm pause function was used more frequently in non–COVID-19 ICUs than in COVID-19 ICUs. This might be explained by the fact that many nurses in COVID-19 ICUs had limited critical care experience and therefore did not know that this function exists or how to use it; COVID-19 ICU staff also might have entered the patients' rooms less often because they were required to don protective gear, which is a time-consuming process.

ICU Staff Experienced More Alarm Fatigue in COVID-19 ICUs

While overall alarm fatigue questionnaire scores were similar between both types of ICUs, the situation varied among different health care professionals. Nurses and physicians in COVID-19 ICUs had significantly higher alarm fatigue scores, whereas support staff in both types of ICUs reported low alarm fatigue scores. Notably, a substantial portion of questionnaire participants in COVID-19 ICUs consisted of support staff, that is, students or nurses from normal wards (COVID-19 ICUs: 25/56, 45%; non–COVID-19 ICUs: 9/88, 10%), reflecting the recruitment of nurses from other services and students with varying levels of critical care experience due to the exceptional circumstances. Unlike the experienced and well-rehearsed teams typically found in non–COVID-19 ICUs, COVID-19 ICU teams often incorporated diverse teams that lacked experience in critical care medicine, training, and familiarity with the

monitoring system and alarm management. This inexperience extended to their training and understanding of monitoring systems and alarm management, mirroring findings from existing literature [41]. In some hospitals, the patient-nurse ratio was increased, which occasionally resulted in poorer quality of care [42]. However, it is important to note that greater fatigue was identified among those working with patients with COVID-19 infection, rather than directly associating alarm fatigue with the COVID-19 condition itself.

There seemed to have been challenges with the alarm system in the COVID-19 ICUs that might have impeded appropriate monitoring. Due to the absence of central monitors in the corridors, dispensary room, and physicians' room, it was difficult for health care providers to locate or swiftly identify and respond to alarms. Interviewees mentioned the protective gear worn in the COVID-19 ICUs as an additional burden when responding to alarms. Protective equipment can impede swift movement, thus slowing response times to alarms, which might intensify the stress and sensory overload associated with alarm fatigue. This finding aligns with the conclusions drawn by Akturan et al [43] that suggested that personal protective equipment could contribute to increased alarm fatigue.

Recommendations for ICU Alarm Systems in Future Pandemics

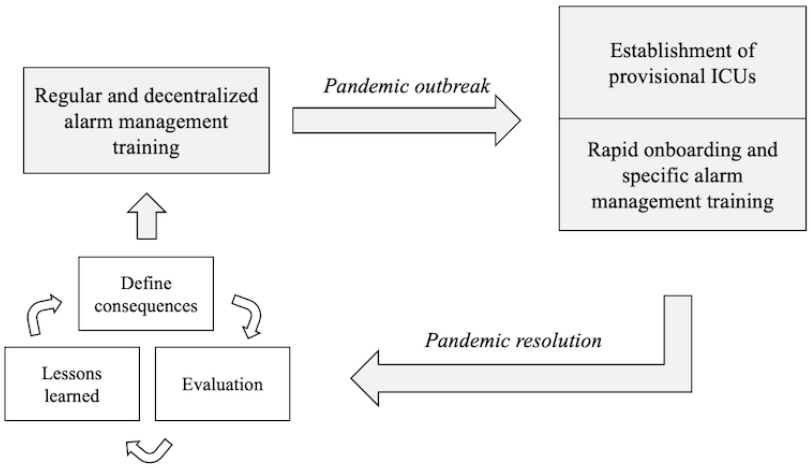
Due to extensive international traffic, the risk of future pandemics remains high in a globalized world [44], and the COVID-19 pandemic will be followed by a new pandemic at some point. Such pandemics will again likely require the setting up of provisional ICUs to cope with rapid patient admissions. When preparing for such events, we recommend also preparing the alarm systems and their human operators. Even outside of pandemics, the shortage of specialist staff is increasing the willingness of authorities to deploy untrained personnel in certain functions in the medical field.

Alarm management aims to reduce the number of unnecessary alarms on the premise that a lower alarm rate decreases alarm fatigue among staff [22]. Provisional ICUs can benefit from effective alarm management, as do regular ICUs, where it helps to significantly decrease the number of false alarms and the overall alarm burden [45-49].

Unnecessary alarms stem from insufficient alarm management knowledge, including using default instead of customized alarm limits, rarely using the pause function during patient manipulation or being unaware of its existence, using insufficient consumables, or attaching electrodes improperly [50]. Previous studies [51] highlight a lack of knowledge about these functions among nurses, pointing to a pressing need for education on physiological monitors [52].

Given the urgency with which new staff members have to be onboarded, the training should be decentralized and easily accessible. To cater to these requirements, we recommend developing educational microcredentials, such as video tutorials that provide guidance on adjusting alarm limits on specific monitoring devices, implementing remote alarm turn-off, and transitioning from individual to smart alarm systems. To visualize and operationalize our approach to pandemic preparedness in ICU alarm systems, we propose a bicyclical strategy, as illustrated in Figure 1. This strategy underscores the significance of continuous learning and adaptability in alarm management. Operating modern bedside monitors requires a blend of cognitive knowledge, psychomotor skills, critical thinking, and an understanding of alarm systems [53]. Hence, this educational program should use the skillmap from the study by Sowan et al [51] as a foundation, expanding it to address the broader technical and practical aspects of monitor use. Modular content that teaches specific skills can be easily adapted to meet unique situational needs; for example, tutorials can be designed to explain how to respond to alarms while wearing protective gear or how to mobilize patients in prone position without causing alarm artifacts. These resources could be beneficial for ICU staff training not only during pandemics but also in routine situations.

Figure 1. Two-phase approach to intensive care unit (ICU) alarm system preparedness and refinement. This figure details continuous training during nonpandemic periods, rapid onboarding, and specialized training at the onset of a pandemic, followed by an evaluation cycle after the pandemic to improve future alarm management strategies.



Limitations

ICUs function as complex sociotechnical systems, making it inherently challenging to compare them. All ICUs were equipped with similar technical equipment (eg, mechanical ventilators and dialysis devices), but our comparison between ICUs was complicated by the variations in the specific devices used across units, leading to differences in alarm signals. Therefore, we only included alarms that occurred in all units. In addition, numerous devices in all ICUs are not connected to the central monitoring system, meaning that the alarms could not be recorded and evaluated. Alarms from mechanical ventilation had to be excluded from the ICU data due to a technical error, preventing us from testing our hypothesis about an increase in such alarms in COVID-19 ICUs.

The monitoring systems in the 2 ICU types varied in their version numbers and settings, further affecting the comparability. In COVID-19 ICUs, the *ECG lead off* alarm was set as a yellow alarm, while in non–COVID-19 ICUs, it was set as a blue alarm. This discrepancy likely stemmed from a lack of awareness or understanding about the settings. We had to exclude this alarm from the yellow alarm load comparison due to this discrepancy.

Our metrics, calculated relative to the number of occupied beds per day, could potentially skew the results. When metrics are calculated this way, it might not accurately reflect the real-life impact of alarm flood conditions on health care workers. In a larger ICU (such as the one with 24 beds), the same rate of alarms per bed would create a larger absolute number of alarms because there are more beds. Consequently, the staff in this ICU would be exposed to a higher number of total alarms than the staff in a smaller ICU (one with, say, 10 beds), although the rate of alarms per bed is the same. While the surgical and medical ICUs were approximately fully occupied over the entire period under study, the occupancy of the COVID-19 ICUs often varied, depending on the patient load. The average bed occupancy was 97.26% and 92.27% in the surgical and medical ICUs, respectively; and 88.79% and 82.75% in COVID-19 ICU-A and COVID-19 ICU-B, respectively. The occupancy of all examined ICUs is reported in Figure S2 in Multimedia Appendix 1. The staff members interviewed from the COVID-19 ICUs were not the same as those from the non–COVID-19 ICUs, which introduces a limitation related to team differences and experience levels.

Regarding the alarm durations, we could not investigate the impact of the isolation process by analyzing alarm response times between COVID-19 ICUs and isolation rooms in non–COVID-19 ICUs. Due to the provisional and time-limited nature of the COVID-19 ICUs, the alarm fatigue in health care providers might reflect not only the conditions in the COVID-19 ICUs but also the individual exposure to alarm load during their former career. We cannot definitively determine whether the increased alarm load was caused by the COVID-19 condition itself or by the differences in ICU settings. Unfortunately, more detailed clinical data for every patient associated with the alarm data were not available at the time of the study. From a clinical perspective, the patient cohorts in the COVID-19 ICUs and general ICUs were comparable in terms of the severity of their conditions. In addition, fatigue among health care professionals working with patients with COVID-19 infection was inherently greater due to the overall stress and workload of managing COVID-19 cases, which cannot be attributed solely to alarm load. We suggest propensity score matching or similar statistical techniques as areas for future research. Similarly, determining isolation rooms from non–COVID-19 ICUs is recommended for further studies to gain a deeper understanding of the impact of the isolation process on alarm response times.

Conclusions

In this study, the COVID-19 ICUs registered significantly more alarms than the non–COVID-19 ICUs. The higher number of alarms led to a higher level of alarm fatigue among the clinicians working in COVID-19 ICUs. We believe that this was caused by the high proportion of untrained staff who were deployed to the temporary ICUs during the pandemic and the provisional setting. The absence of central monitors in individual rooms and corridors further compounded these challenges, making it difficult for health care providers to swiftly identify and respond to alarms. However, it is important to note that our findings are limited by the study design and specific circumstances during the pandemic, which might affect the strength of our conclusions. Further studies are warranted to better understand the broader implications of alarm management in different ICU settings.

To mitigate alarm overload in provisional ICUs during future pandemics, we recommend creating skill-oriented video tutorials on alarm management and monitor use. These tutorials should provide easily accessible training for new staff, who may be rapidly recruited and could have limited or no prior ICU experience. This educational material could equip them with the necessary knowledge to effectively navigate the ICU alarm system.

Acknowledgments

The authors express their gratitude to the intensive care unit staff for their participation in this study. ASP is a participant in the Digital Clinician Scientist Program funded by Charité–Universitätsmedizin Berlin and the Berlin Institute of Health. The cost of publishing open access was covered by Projekt DEAL.

Data Availability

The data sets generated and analyzed during this study are available in the Zenodo open data repository [54]. The code is available from the corresponding author on reasonable request.

Authors' Contributions

ASP and MMW had the idea for the study. The study was conceived by NF, MMW, ASP, and FB. NF and MMW analyzed the data and conducted the interviews, supported by ASP. NF wrote the manuscript, supported by ASP, CH, SA-W, and MMW. FB supervised all parts of the study. All authors critically reviewed and approved the manuscript. The paper was extracted from NF's MSc thesis.

Conflicts of Interest

FB reports grants from the German Federal Ministry of Education and Research, the German Federal Ministry of Health, the Berlin Institute of Health, the Hans Böckler Foundation, the Einstein Foundation, the Berlin University Alliance, and from the Robert Koch Institute, as well as personal fees from Elsevier Publishing, outside the submitted work. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Alarm load and management differences between COVID-19 and non–COVID-19 intensive care units (ICUs). Figures include alarm frequencies by device and color, unit occupancy, alarm durations, and adjustments in settings across ICUs. Table data cover interview questions, alarm log structure, included alarm signals, hypothesis testing group sizes, and statistical analysis of alarm load per bed per day.
[DOCX File , 1421 KB - medinform_v12i1e58347_app1.docx]

References

1. Shen X, Zou X, Zhong X, Yan J, Li L. Psychological stress of ICU nurses in the time of COVID-19. Crit Care 2020 May 06;24(1):200 [FREE Full text] [doi: 10.1186/s13054-020-02926-2] [Medline: 32375848]
2. Ferrer R. COVID-19 pandemic: the greatest challenge in the history of critical care. Med Intensiva (Engl Ed) 2020 Aug;44(6):323-324 [FREE Full text] [doi: 10.1016/j.medin.2020.04.002] [Medline: 32376091]
3. Bello M, Segura V, Camputaro L, Hoyos W, Maza M, Sandoval X, et al. Hospital El Salvador: a novel paradigm of intensive care in response to COVID-19 in Central America. The Lancet Global Health 2021 Mar;9(3):e241-e242. [doi: 10.1016/s2214-109x(20)30513-1]
4. McCabe R, Kont MD, Schmit N, Whittaker C, Løchen A, Baguelin M, et al. Modelling intensive care unit capacity under different epidemiological scenarios of the COVID-19 pandemic in three Western European countries. Int J Epidemiol 2021 Jul 09;50(3):753-767 [FREE Full text] [doi: 10.1093/ije/dyab034] [Medline: 33837401]
5. Cai Y, Wu X, Zhang Y, Xia J, Li M, Feng Y, et al. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) contamination in air and environment in temporary COVID-19 ICU wards. Research Square Preprint posted online April 6, 2020 [FREE Full text] [doi: 10.21203/rs.3.rs-21384/v1]
6. Hittesdorf E, Panzer O, Wang D, Stevens JS, Hastie J, Jordan DA, et al. Mortality and renal outcomes of patients with severe COVID-19 treated in a provisional intensive care unit. J Crit Care 2021 Apr;62:172-175 [FREE Full text] [doi: 10.1016/j.jcrc.2020.12.012] [Medline: 33385774]
7. Peters AW, Chawla KS, Turnbull ZA. Transforming ORs into ICUs. N Engl J Med 2020 May 07;382(19):e52. [doi: 10.1056/nejmc2010853]
8. Singh S, Ambooken GC, Setlur R, Paul SK, Kanitkar M, Singh Bhatia S, et al. Challenges faced in establishing a dedicated 250 bed COVID-19 intensive care unit in a temporary structure. Trends Anaesth Crit Care 2021 Feb;36:9-16 [FREE Full text] [doi: 10.1016/j.tacc.2020.10.006] [Medline: 38620737]
9. Uppal A, Silvestri DM, Siegler M, Natsui S, Boudourakis L, Salway RJ, et al. Critical care and emergency department response at the epicenter of the COVID-19 pandemic. Health Aff (Millwood) 2020 Aug 01;39(8):1443-1449. [doi: 10.1377/hlthaff.2020.00901] [Medline: 32525713]
10. Hravnak M, Pellathy T, Chen L, Dubrawski A, Wertz A, Clermont G, et al. A call to alarms: current state and future directions in the battle against alarm fatigue. J Electrocardiol 2018 Nov;51(6S):S44-S48 [FREE Full text] [doi: 10.1016/j.jelectrocard.2018.07.024] [Medline: 30077422]
11. Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. AACN Adv Crit Care 2013;24(4):378-386. [doi: 10.1097/nci.0b013e3182a903f9]
12. Varisco G, van de Mortel H, Cabrera-Quiros L, Atallah L, Hueske-Kraus D, Long X, et al. Optimisation of clinical workflow and monitor settings safely reduces alarms in the NICU. Acta Paediatr 2021 Apr 22;110(4):1141-1150 [FREE Full text] [doi: 10.1111/apa.15615] [Medline: 33048364]
13. Bach TA, Berglund LM, Turk E. Managing alarm systems for quality and safety in the hospital setting. BMJ Open Qual 2018 Jul 25;7(3):e000202 [FREE Full text] [doi: 10.1136/bmjoq-2017-000202] [Medline: 30094341]
14. Baker K, Rodger J. Assessing causes of alarm fatigue in long-term acute care and its impact on identifying clinical changes in patient conditions. Inform Med Unlocked 2020;18:100300. [doi: 10.1016/j.imu.2020.100300]

15. Poncette AS, Spies C, Mosch L, Schieler M, Weber-Carstens S, Krampe H, et al. Clinical requirements of future patient monitoring in the intensive care unit: qualitative study. JMIR Med Inform 2019 Apr 30;7(2):e13064 [FREE Full text] [doi: 10.2196/13064] [Medline: 31038467]
16. Pugh RJ, Jones C, Griffiths RD. The impact of noise in the intensive care unit. In: Intensive Care Medicine. Yearbook of Intensive Care and Emergency Medicine. Berlin, Germany: Springer; 2007:942-949.
17. Melnyk BM, Hsieh AP, Tan A, Teall AM, Weberg D, Jun J, et al. Associations among nurses' mental/physical health, lifestyle behaviors, shift length, and workplace wellness support during COVID-19: important implications for health care systems. Nurs Adm Q 2022;46(1):5-18 [FREE Full text] [doi: 10.1097/NAQ.0000000000000499] [Medline: 34551423]
18. Sasangohar F, Jones SL, Masud FN, Vahidy FS, Kash BA. Provider burnout and fatigue during the COVID-19 pandemic: lessons learned from a high-volume intensive care unit. Anesth Analg 2020 Jul;131(1):106-111 [FREE Full text] [doi: 10.1213/ANE.0000000000004866] [Medline: 32282389]
19. Sikaras C, Ilias I, Tselebis A, Pachi A, Zyga S, Tsironi M, et al. Nursing staff fatigue and burnout during the COVID-19 pandemic in Greece. AIMS Public Health 2022;9(1):94-105 [FREE Full text] [doi: 10.3934/publichealth.2022008] [Medline: 35071671]
20. Galanis P, Vraka I, Fragkou D, Bilali A, Kaitelidou D. Nurses' burnout and associated risk factors during the COVID-19 pandemic: a systematic review and meta-analysis. J Adv Nurs 2021 Aug 25;77(8):3286-3302 [FREE Full text] [doi: 10.1111/jan.14839] [Medline: 33764561]
21. Allan SH, Doyle PA, Sapirstein A, Cvach M. Data-driven implementation of alarm reduction interventions in a cardiovascular surgical ICU. Jt Comm J Qual Patient Saf 2017 Feb;43(2):62-70. [doi: 10.1016/j.jcjq.2016.11.004] [Medline: 28334564]
22. Poncette AS, Wunderlich MM, Spies C, Heeren P, Vorderwülbecke G, Salgado E, et al. Patient monitoring alarms in an intensive care unit: observational study with do-it-yourself instructions. J Med Internet Res 2021 May 28;23(5):e26494 [FREE Full text] [doi: 10.2196/26494] [Medline: 34047701]
23. Wilken M, Hüske-Kraus D, Röhrig R. Alarm fatigue: using alarm data from a patient data monitoring system on an intensive care unit to improve the alarm management. Stud Health Technol Inform 2019 Sep 03;267:273-281. [doi: 10.3233/SHTI190838] [Medline: 31483282]
24. Wunderlich MM, Amende-Wolf S, Krampe H, Kruppa J, Spies C, Weiß B, et al. A brief questionnaire for measuring alarm fatigue in nurses and physicians in intensive care units. Sci Rep 2023 Aug 24;13(1):13860 [FREE Full text] [doi: 10.1038/s41598-023-40290-7] [Medline: 37620385]
25. Helfferich C. Die Qualität qualitativer Daten: Manual für die Durchführung Qualitativer Interviews. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften; 2009.
26. Mayring P. Qualitative Inhaltsanalyse: Grundlagen und Techniken. Weinheim, Germany: Beltz Verlag; 2022.
27. Microsoft Teams. Microsoft. URL: https://www.microsoft.com/de-de/microsoft-teams/group-chat-software [accessed 2024-08-21]
28. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2019. URL: https://www.R-project.org/ [accessed 2024-08-21]
29. RStudio Team. RStudio: integrated development environment for R. RStudio, PBC. 2020. URL: http://www.rstudio.com/ [accessed 2024-08-21]
30. Grolemund G, Wickham H. Dates and times made easy with lubridate. J Stat Softw 2011 Apr 07;40(3):1-25. [doi: 10.18637/jss.v040.i03]
31. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer; 2009.
32. Wickham H, François R, Henry L, Müller K, Vaughan D, Posit Software, PBC. dplyr: a grammar of data manipulation. The Comprehensive R Archive Network. 2023 Nov 17. URL: https://cran.r-project.org/web/packages/dplyr/index.html [accessed 2024-08-21]
33. Torchiano M. Effsize - a package for efficient effect size computation. Zenodo. 2016 Nov 13. URL: https://zenodo.org/records/196082 [accessed 2024-08-21]
34. Canty A, Ripley B, Brazzale AR. boot: bootstrap functions. The Comprehensive R Archive Network. 2024 Feb 26. URL: https://cran.r-project.org/web/packages/boot/boot.pdf [accessed 2024-08-21]
35. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
36. Wunderlich MM, Amende-Wolf S, Poncette AS, Krampe H, Kruppa J, Spies C, et al. Ein deutschsprachiger Fragebogen zur Messung der Alarmmüdigkeit bei Pflegekräften und Ärzt:innen auf Intensivstationen. Kongress der Deutschen Interdisziplinären Vereinigung für Intensiv- und Notfallmedizin e.V. 2021. URL: https://medinfo.charite.de/forschung/ag_intelligent_patient_monitoring/alarmzen/ [accessed 2024-08-21]
37. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45(3):1-67 [FREE Full text] [doi: 10.18637/jss.v045.i03]
38. Cosper P, Zellinger M, Enebo A, Jacques S, Razzano L, Flack MN. Improving clinical alarm management: guidance and strategies. Biomed Instrum Technol 2017;51(2):109-115. [doi: 10.2345/0899-8205-51.2.109] [Medline: 28296432]

39. Hüske-Kraus D, Wilken M, Röhrig R. Measuring alarm system quality in ICUs (WIP). EasyChair. Preprint posted online 2018 URL: https://easychair.org/publications/preprint_open/6nzL [accessed 2024-08-21]
40. Deb S, Claudio D. Alarm fatigue and its influence on staff performance. IIE Trans Healthc Syst Eng 2015 Jul 29;5(3):183-196. [doi: 10.1080/19488300.2015.1062065]
41. González-Gil MT, González-Blázquez C, Parro-Moreno AI, Pedraz-Marcos A, Palmar-Santos A, Otero-García L, et al. Nurses' perceptions and demands regarding COVID-19 care delivery in critical care units and hospital emergency services. Intensive Crit Care Nurs 2021 Feb;62:102966 [FREE Full text] [doi: 10.1016/j.iccn.2020.102966] [Medline: 33172732]
42. Bergman L, Falk AC, Wolf A, Larsson IM. Registered nurses' experiences of working in the intensive care unit during the COVID-19 pandemic. Nurs Crit Care 2021 Nov;26(6):467-475 [FREE Full text] [doi: 10.1111/nicc.12649] [Medline: 33973304]
43. Akturan S, Güner Y, Tuncel B, Üçüncüoğlu M, Kurt T. Evaluation of alarm fatigue of nurses working in the COVID-19 Intensive Care Service: a mixed methods study. J Clin Nurs 2022 Sep;31(17-18):2654-2662. [doi: 10.1111/jocn.16190] [Medline: 34985160]
44. Shafaati M, Chopra H, Priyanka, Khandia R, Choudhary OP, Rodriguez-Morales AJ. The next pandemic catastrophe: can we avert the inevitable? New Microbes New Infect 2023 Mar;52:101110 [FREE Full text] [doi: 10.1016/j.nmni.2023.101110] [Medline: 36937540]
45. Sahoo T, Joshi M, Madathil S, Verma A, Sankar MJ, Thukral A. Quality improvement initiative for reduction of false alarms from multiparameter monitors in neonatal intensive care unit. J Educ Health Promot 2019 Oct 24;8:203 [FREE Full text] [doi: 10.4103/jehp.jehp_226_19] [Medline: 31807593]
46. Seifert M, Tola DH, Thompson J, McGugan L, Smallheer B. Effect of bundle set interventions on physiologic alarms and alarm fatigue in an intensive care unit: a quality improvement project. Intensive Crit Care Nurs 2021 Dec;67:103098. [doi: 10.1016/j.iccn.2021.103098] [Medline: 34393010]
47. Van de Pol I, Wirds J. St. Antonius hospital reduces non-actionable ICU alarms by 40% to improve patient care and staff satisfaction. Philips Clinical Services. URL: https://www.philips.com/c-dam/b2bhc/master/articles/acute-care/forty-percent-st-antonius-customer-story.pdf [accessed 2024-08-21]
48. Yeh J, Wilson R, Young L, Pahl L, Whitney S, Dellsperger KC, et al. Team-based intervention to reduce the impact of nonactionable alarms in an adult intensive care unit. J Nurs Care Qual 2020;35(2):115-122. [doi: 10.1097/NCQ.0000000000000436] [Medline: 31513051]
49. Yue L, Plummer V, Cross W. The effectiveness of nurse education and training for clinical alarm response and management: a systematic review. J Clin Nurs 2017 Sep;26(17-18):2511-2526. [doi: 10.1111/jocn.13605] [Medline: 27685951]
50. Wilken M, Hüske-Kraus D, Klausen A, Koch C, Schlauch W, Röhrig R. Alarm fatigue: causes and effects. Stud Health Technol Inform 2017;243:107-111. [Medline: 28883181]
51. Sowan AK, Vera AG, Fonseca EI, Reed CC, Tarriela AF, Berndt AE. Nurse competence on physiologic monitors use: toward eliminating alarm fatigue in intensive care units. Open Med Inform J 2017 Apr 14;11:1-11 [FREE Full text] [doi: 10.2174/1874431101711010001] [Medline: 28567167]
52. Cvach M, Kitchens M, Smith K, Harris P, Flack MN. Customizing alarm limits based on specific needs of patients. Biomed Instrum Technol 2017;51(3):227-234. [doi: 10.2345/0899-8205-51.3.227] [Medline: 28530858]
53. Phillips J, Sowan A, Ruppel H, Magness R. Educational program for physiologic monitor use and alarm systems safety: a toolkit. Clin Nurse Spec 2020;34(2):50-62. [doi: 10.1097/NUR.0000000000000507] [Medline: 32068633]
54. Wunderlich MM, Frey N, Amende-Wolf S, Hinrichs C, Balzer F, Poncette AS. Alarm management in provisional COVID-19 intensive care units: a retrospective analysis and recommendations for future pandemics. Zenodo. URL: https://zenodo.org/records/10418595 [accessed 2024-08-21]

Abbreviations

ECG: electrocardiogram
IBP: invasive blood pressure
ICU: intensive care unit
NIBP: noninvasive blood pressure
REDCap: Research Electronic Data Capture
SpO2: peripheral oxygen saturation

Edited by C Lovis; submitted 14.03.24; peer-reviewed by K Bin, JS Jerng; comments to author 24.04.24; revised version received 10.06.24; accepted 21.07.24; published 09.09.24.

Please cite as:
Wunderlich MM, Frey N, Amende-Wolf S, Hinrichs C, Balzer F, Poncette AS
Alarm Management in Provisional COVID-19 Intensive Care Units: Retrospective Analysis and Recommendations for Future Pandemics
JMIR Med Inform 2024;12:e58347
URL: https://medinform.jmir.org/2024/1/e58347
doi:10.2196/58347
PMID:

Original Paper

The Impact of International Classification of Disease–Triggered Prescription Support on Telemedicine: Observational Analysis of Efficiency and Guideline Adherence

Tarso Augusto Duenhas Accorsi[1], MD, PhD; Anderson Aires Eduardo[2], PhD; Carlos Guilherme Baptista[1], MD, MSc; Flavio Tocci Moreira[1], MD, PhD; Renata Albaladejo Morbeck[1], MR; Karen Francine Köhler[1], MSc, PhD; Karine de Amicis Lima[1], MSc, PhD; Carlos Henrique Sartorato Pedrotti[1], MD

[1]Telemedicine Department, Hospital Israelita Albert Einstein, São Paulo, Brazil

[2]Digital Platform, Hospital Israelita Albert Einstein, São Paulo, Brazil

Corresponding Author:
Tarso Augusto Duenhas Accorsi, MD, PhD
Telemedicine Department
Hospital Israelita Albert Einstein
Avenue Albert Einstein, 627 - Bloco B, 2º andar
Secretaria da Unidade de Telemedicina
São Paulo, 05652-900
Brazil
Phone: 55 1121515420
Email: taccorsi@einstein.br

Abstract

Background: Integrating decision support systems into telemedicine may optimize consultation efficiency and adherence to clinical guidelines; however, the extent of such effects remains underexplored.

Objective: This study aims to evaluate the use of *ICD* (*International Classification of Disease*)-coded prescription decision support systems (PDSSs) and the effects of these systems on consultation duration and guideline adherence during telemedicine encounters.

Methods: In this retrospective, single-center, observational study conducted from October 2021 to March 2022, adult patients who sought urgent digital care via direct-to-consumer video consultations were included. Physicians had access to current guidelines and could use an *ICD*-triggered PDSS (which was introduced in January 2022 after a preliminary test in the preceding month) for 26 guideline-based conditions. This study analyzed the impact of implementing automated prescription systems and compared these systems to manual prescription processes in terms of consultation duration and guideline adherence.

Results: This study included 10,485 telemedicine encounters involving 9644 patients, with 12,346 prescriptions issued by 290 physicians. Automated prescriptions were used in 5022 (40.67%) of the consultations following system integration. Before introducing decision support, 4497 (36.42%) prescriptions were issued, which increased to 7849 (63.57%) postimplementation. The physician's average consultation time decreased significantly to 9.5 (SD 5.5) minutes from 11.2 (SD 5.9) minutes after PDSS implementation (P<.001). Of the 12,346 prescriptions, 8683 (70.34%) were aligned with disease-specific international guidelines tailored for telemedicine encounters. Primary medication adherence in accordance with existing guidelines was significantly greater in the decision support group than in the manual group (n=4697, 93.53% vs n=1389, 49.14%; P<.001).

Conclusions: Most of the physicians adopted the PDSS, and the results demonstrated the use of the *ICD*-code system in reducing consultation times and increasing guideline adherence. These systems appear to be valuable for enhancing the efficiency and quality of telemedicine consultations by supporting evidence-based clinical decision-making.

(JMIR Med Inform 2024;12:e56681) doi:10.2196/56681

KEYWORDS

telemedicine; clinical decision support systems; electronic prescriptions; guideline adherence; consultation efficiency; International Classification of Disease–coded prescriptions; teleheath; eHealth

Introduction

Telemedicine increasingly serves as a primary point of entry into the health care system for patients, particularly in urgent care scenarios [1]. Physicians providing digital consultations are tasked with maximizing efficiency in terms of encounter duration while ensuring that prescriptions issued adhere to established guidelines, which is a crucial component of the cost-effectiveness and quality of telemedicine services [2].

The role of prescription decision support systems (PDSSs) is critical in the digital health care environment [3]. Within electronic health records (EHRs), PDSSs can streamline the amount of time that clinicians spend navigating complex medical terminology [4]. Research indicates that standardizing data input can enhance routine documentation in medical records [5]. Apart from time efficiency, EHRs have been linked to improved care quality and increased compliance with clinical guidelines [6]. Although EHR adoption is associated with significant challenges, the development of strategies to facilitate their use is gradually progressing [7]. To date, however, research exploring voluntary PDSS use and its impact on the outcomes of direct-to-consumer telemedicine consultations for acute conditions has been lacking.

We propose that physicians' adoption of a PDSS will demonstrate both a reduction in the time needed to deliver care and increased adherence to clinical guidelines. This study explored the relationships between telemedicine use and *ICD* (*International Classification of Diseases*)-triggered PDSS scores, consultation duration, and guideline adherence.

Methods

Study Design and Participants

A single-center retrospective study is conducted at the Telemedicine Center of Hospital Israelita Albert Einstein in São Paulo, Brazil. The data were collected by physicians at the Telemedicine Center, which ensured secure data storage. All authors contributed to the initial draft of the study and conducted a thorough examination of the complete data set, to which they had full access. The paper was exclusively written by the named authors, without contributions from nonauthors. All data analyses were conducted internally by the supervisory team of the Telemedicine Center. All authors collectively decided to submit this paper for publication and also supported the authenticity and integrity of the reported data.

The study included patients aged 16 years and older who voluntarily accessed digital direct-to-consumer care from October 2021 to March 2022. We considered all patients who were presented with any medical condition for inclusion in this study. The only exclusion criterion was the occurrence of connection issues that precluded the creation of medical records; these patients were excluded because they did not undergo a complete medical evaluation and were consequently not documented in the institution's database.

Ethical Considerations

The protocol for the study, which is known as the "Tele AUTOMATION" trial, and a consent waiver (based on an analysis of anonymized retrospective data from routine care) were approved by the Hospital Israelita Albert Einstein Review Board (registration: CAAE 69981423.6.0000.0071). Institutional digital archives, which were not linked to any external financial support, served as repositories for all of the data related to this study. The collected data were treated confidentially and protected by strict security measures, in accordance with the internal data protection policies of the Hospital Israelita Albert Einstein. All stages of the study involving privacy and personal data protection were conducted in accordance with Brazil's General Data Protection Law (LGPD). No compensation was provided to participants.

Telemedicine Consultations

Telemedicine consultations were conducted over the internet using proprietary videoconferencing software and EHRs. All participating physicians were board-certified and had additional training in both telemedicine and emergency medicine. The Telemedicine Center provided streamlined access to up-to-date clinical guidelines. Medical information was recorded in EHRs, which featured a specific field for the primary *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) diagnosis.

Decision Support and Guideline-Directed Prescribing

A suite of 26 international guidelines adapted for telemedicine was available for immediate reference during the consultations. These guidelines were structured to suggest appropriate prescriptions, as outlined in Multimedia Appendix 1. Physicians received training on how to effectively consult these guidelines, with the aim of aligning prescriptions with scientific evidence, resource efficiency, side effect mitigation, and overall service safety. In 2021, a voluntary feature allowing prescription autopopulation was introduced, providing shortcuts to medications listed in guidelines corresponding to the *ICD-10*, which is central to patient care. With a single keystroke, the autopopulation feature filled prescription fields with medications, dosages, administration routes, and durations as indicated by the relevant guidelines. This new functionality was introduced during the testing period, and no further adaptations were made to the system. By January 2022, this functionality was fully integrated into the medical record system, and physicians were instructed on its use according to clinical findings. Prescription adaptations for comorbidities were tailored based on clinical judgment and evaluated on an individual basis.

Patient and physician acceptance was not directly assessed. The functionality was made available for voluntary use. Although system adoption was not directly measured, a substantial portion of prescriptions were issued using the decision support system, allowing us to infer that a significant number of physicians accepted this strategy.

Data Extraction and Adherence to Institutional Protocols

The data from each TM encounter were extracted from the institution's medical record database. The *ICD-10* code

associated with each telemedicine consultation was used to cross-reference the institutional protocol. Prescriptions issued by the telemedicine physicians were compared to the protocol's medication list using data from the medication prescription field in the telemedicine consultation record.

First, the *ICD-10* code from each telemedicine consultation was used to identify the proper institutional protocol. Then, the set of medications provided in the protocol was compared with the medications prescribed by the telemedicine physicians as documented in the dedicated fields for prescriptions on the telemedicine consultation record. We added a binary feature (1=matching and 0=not matching) to the data set to denote matches between prescribed medication and institutional directives. After applying this procedure to all records in our data set, we were able to compute summary statistics to compare protocol adherence before and after autocomplete system implementation. Table 1 provides the main summary statistics for our data set.

Table 1. Summary statistics for the data set before and after implementation of the clinical decision support autocomplete system.

Data set feature (year)	2021	2022	Total
Consultations, n (%)	3873 (36.94)	6612 (63.06)	10,485
Patients, n (%)	3628 (37.62)	6174 (68.38)	9644
Physicians, n (%)	211 (72.76)	231 (79.65)	290
Medication prescriptions, n (%)	4497 (36.42)	7849 (63.58)	12,346
Mean encounter duration (in minutes), mean (SD)	11.23 (5.92)	9.49 (5.57)	10.13 (5.77)
ICD-10[a] codes documented, n (%)	286 (65.15)	312 (71.07)	439

[a]*ICD-10*: *International Statistical Classification of Disease, Tenth Revision*.

Statistical Analysis

The study analyzed a convenience sample of all patients who were consecutively registered during the defined period. Continuous variables are reported as the mean (SD) or as the median (IQR) ranges for descriptive purposes, whereas categorical variables are summarized as counts and percentages. To test for normality in the distribution of our sample, we used the Kolmogorov-Smirnov test. The Mann-Whitney *U* test was used for continuous variables that were not normally distributed. A *P* value less than .001 indicated statistical significance and 95 % CIs were calculated. All statistical analyses were conducted using IBM SPSS Statistics (version 22.0) for Windows.

Results

Throughout the 6-month study period, we analyzed a total of 10,485 encounters with 9644 patients. Patient demographic data and the most common diagnoses are described in Table 2. These encounters resulted in 12,346 prescriptions being issued by 290 different attending physicians. Notably, some patients had multiple consultations within the study timeframe, and multiple prescriptions were occasionally dispensed during a single consultation; these prescriptions were not individualized. Before the implementation of the self-report prescription system, 4497 (36.42%) prescriptions were issued, which increased to 7849 (63.58%) after its implementation. A preliminary test of the system commenced in 2021; however, the system was not fully operational and made available for voluntary use by all physicians until January 2022. Following its implementation, the automated prescription feature was used in 5022 (40.67%) of the encounters.

During the brief trial period of the PDSS tool in 2021, a total of 261 (5.80%) prescriptions were made with electronic assistance, while 4236 (94.19%) prescriptions were issued without such assistance. Following the full deployment of the PDSS tool in 2022, the figures shifted significantly, with 5022 (63.98%) prescriptions being made with electronic assistance compared to 2827 (36.02%) without assistance.

The data demonstrated a significant reduction in consultation time after self-reporting: 9.5 (SD 5.5) minutes versus 11.2 (SD 5.9) minutes (P<.001; Mann-Whitney U test=12,118,181.5; Figure 1). Figure 2 shows that the decrease in the average consultation duration was correlated with service density, which was defined as the volume of services delivered by the center within a specified timeframe.

Regarding guideline adherence among the 12,346 medications prescribed, a substantial number of these medications (n=8683, 70.34%) were prescribed in accordance with guidelines. Notably, compared with manual prescription entry, the use of the automated filling system significantly increased guideline adherence (n=4697, 93.5% vs n=1389, 49.1%; P<.001; z score=45.24).

Regarding adherence to the institutional protocol for medical prescriptions, when the PDSS tool was not used, 1438 (50.87%) prescriptions deviated from the standard recommendations, while 1389 (49.13%) complied. Our analysis demonstrated that the application of the PDSS tool significantly improved adherence, with only 325 (6.47%) prescriptions failing to comply with the protocol, compared to 4697 (93.53%) that were aligned.

Table 2. Patient demographic data and the most common diagnoses.

Variable	2021-2022
Encounters	10,485
Patients	9644
Sex, n (%)	
Male	6669 (69.15)
Female	2875 (29.81)
Not declared	100 (1.04)
Age (years), mean (SD)	0.0-92.0 (32.2)
Most common diagnoses, n (%)	
J01, acute sinusitis	2627 (25.05)
N30, cystitis	1839 (17.54)
J03, acute tonsillitis	1271 (12.12)
J02, acute pharyngitis	478 (4.56)
J06.9, acute upper respiratory infection, unspecified	324 (3.09)

Figure 1. Comparison of consultation times with and without the use of autocomplete functionality for medical prescriptions revealed a statistically significant decrease in consultation time (*P<.001; 9.5, SD 5.5 minutes vs 11.2, SD 5.9 minutes; Mann Whitney U test=12,118,181.5).

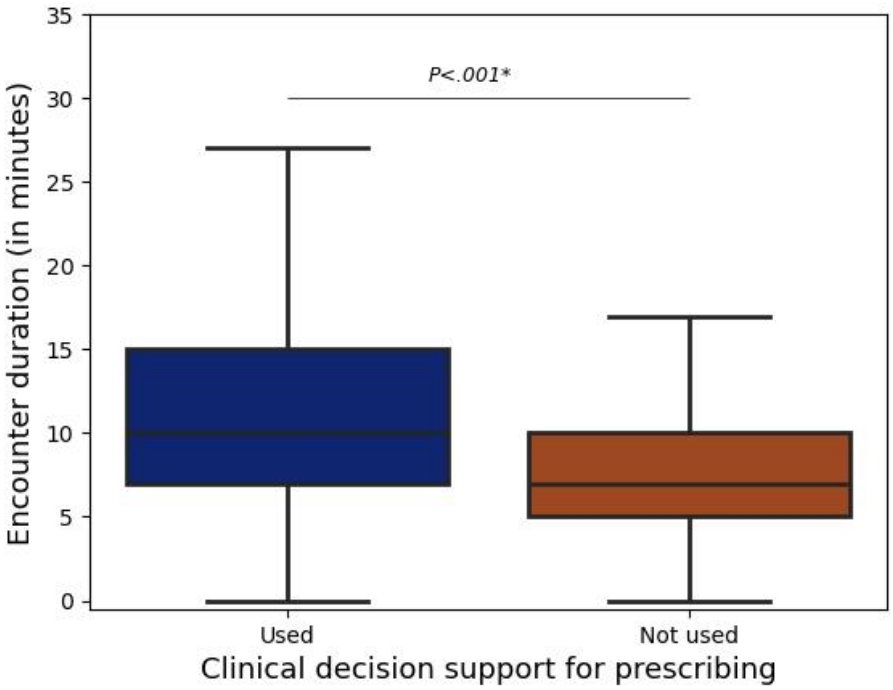
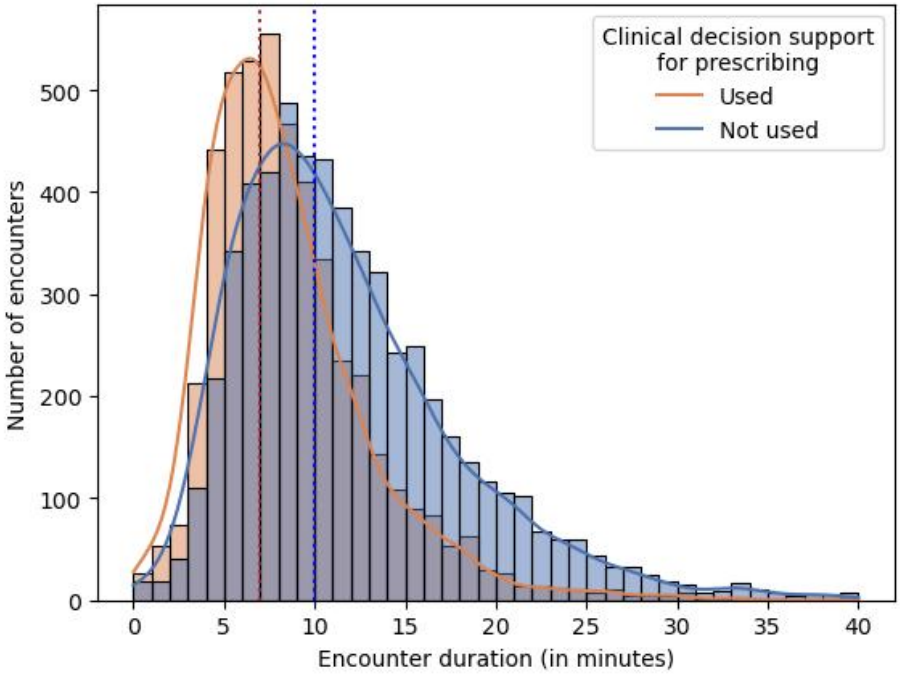


Figure 2. Distributions of consultation times and encounter frequency. The dotted vertical lines indicate the median values for each distribution (8 and 10 minutes, respectively).



Discussion

This study demonstrated that shortly after becoming available, most physicians opted to voluntarily use the decision support autocomplete functionality for medical prescriptions. Innovative approaches that facilitate the use of EHRs tend to be well received by medical practitioners [8]. The integration of new information technologies has transformed numerous aspects of health care and thus revolutionized care delivery within health systems [9], representing a significant advancement in digital health care, although certain challenges still need to be overcome. These challenges include poor interface designs, suboptimal performance, maintenance issues, overreliance, and dependency, all of which could jeopardize patient safety [10].

This study revealed an impressive rate of 94.7% adherence to recommendations, suggesting a potential improvement in the safety of telemedicine consultations with the use of autocomplete prescriptions. Guideline adherence is a measure of how well health care providers follow established clinical guidelines and protocols and is crucial for minimizing risks and ensuring optimal care quality and safety [11]. Notably, given that evidence suggests frequently deficient adherence with respect to in-person consultations, telemedicine shows promise as a method of care associated with better compliance [12-14].

The safety of prescribing practices and reductions in medication errors have been supported by evidence from electronic prescriptions selected through the use of standardized medication lists, codified instructions, and multimodal decision support [15]. Nonetheless, the prevalence of electronic prescribing errors can be high, with rates nearing 60%, particularly when considering incorrect field entries [16]. Previous studies have identified numerous potential error types in electronic prescriptions [17], and the electronic prescription strategy chosen has been noted to be a source of medication errors [18], which underscores the importance of continuing to improve electronic prescription systems.

The option for voluntary prescription autopopulation has several potential benefits for health care delivery. This approach may alleviate bureaucratic burdens associated with current medical practices, thereby enhancing communication with patients [19]. In high-demand settings, an autocomplete option can optimize consultation times as part of a broader management strategy [20]. Additionally, access to checklists for possible medication recommendations may reduce bias [21].

Our analysis revealed a 15.18% decrease in physicians' average consultation time, from 1123 to 949 minutes, when the autocomplete feature was used, although reliance entirely on manual prescription entry was noted in 31% of consultations. Intriguingly, even with voluntary use, most prescriptions issued using the autocomplete function strictly adhered to medication recommendations, indicating high adherence to these recommendations. Prior to our study, no research had specifically investigated telemedicine quality based on guideline-directed prescribing facilitated by an autocomplete function. Accordingly, the establishment of high-quality telemedicine centers that continuously update management guidelines and develop strategies to meet policy requirements and deliver excellent remote medical consultations is imperative to enhance EHR systems. In addition to prescribing practices alone, administrators need to ensure the usability and appropriate implementation of coding for the autocomplete method to prevent misuse and maintain safety.

No specific security assessment was conducted. However, protocol adherence is recognized as a primary indicator of safety in care delivery, including referrals for in-person care when warning signs are identified. This study revealed enhanced

adherence to guideline recommendations through the use of the self-report system, indirectly suggesting that increased safety was observed within this group.

Among the limitations of this study, the observed outcomes from the use of the clinical decision support system may not be solely attributable to changes in care practices. Instead, these outcomes might also reflect changes in the *ICD* codes selected by physicians, possibly for the convenience of prescribing. Moreover, the clinical support decision system may have encouraged excessive prescribing despite many of these conditions being manageable with nonpharmacological interventions, especially for those with mild symptoms. Notably, care quality assessed based solely on whether *ICD* codes matched medication recommendations may not effectively represent the true quality of care and may reflect decreased care quality, as previously mentioned. Another possible limitation was that the group that used decision support may have included patients with simpler diagnoses.

In this study, a PDSS was voluntarily used by physicians for most encounters. Decision support incorporation for prescription selection within EHRs, particularly in scenarios where policies are clearly established, may contribute to reducing consultation times and promoting high rates of adherence to guideline-directed prescriptions. This finding highlights the potential role of these systems in improving both the efficiency and quality of health care delivery, especially within telemedicine environments where prompt and precise decision-making is essential.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Compilation of ICD-10 codes with corresponding guideline-directed prescriptions. ICD-10: International Statistical Classification of Disease, Tenth Revision.
[DOCX File , 26 KB - medinform_v12i1e56681_app1.docx]

References

1. Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]
2. de la Torre-Díez I, López-Coronado M, Vaca C, Aguado JS, de Castro C. Cost-utility and cost-effectiveness studies of telemedicine, electronic, and mobile health systems in the literature: a systematic review. Telemed J E Health 2015 Feb;21(2):81-85 [FREE Full text] [doi: 10.1089/tmj.2014.0053] [Medline: 25474190]
3. Loh LC. Autocomplete: Dr Google's "helpful" assistant? Can Fam Physician 2016 Aug;62(8):622-623 [FREE Full text] [Medline: 27521382]
4. Zaghir J, Goldman J, Bjelogrlic M, Gaudet-Blavignac C, Lovis C. Caregivers interactions with clinical autocomplete tool: a retrospective study. Stud Health Technol Inform 2022 Jun 29;295:132-135. [doi: 10.3233/SHTI220679] [Medline: 35773825]
5. Madandola O, Bjarnadottir R, Yao Y, Ansell M, Dos Santos F, Cho H, et al. The relationship between electronic health records user interface features and data quality of patient clinical information: an integrative review. J Am Med Inform Assoc 2023;31(1):240-255. [doi: 10.1093/jamia/ocad188] [Medline: 37740937]
6. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health 2016;26(1):60-64. [doi: 10.1093/eurpub/ckv122] [Medline: 26136462]
7. Kruse CS, Mileski M, Alaytsev V, Carol E, Williams A. Adoption factors associated with electronic health record among long-term care facilities: a systematic review. BMJ Open 2015;5(1):e006615 [FREE Full text] [doi: 10.1136/bmjopen-2014-006615] [Medline: 25631311]
8. West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: a systematic review. J Am Med Inform Assoc 2015;22(2):330-339 [FREE Full text] [doi: 10.1136/amiajnl-2014-002955] [Medline: 25336597]
9. Yutong T, Yan Z, Qingyun C, Lixue M, Mengke G, Shanshan W. Information and communication technology based integrated care for older adults: a scoping review. Int J Integr Care 2023;23(2):2 [FREE Full text] [doi: 10.5334/ijic.6979] [Medline: 37033366]
10. Añel Rodríguez RM, García Alfaro I, Bravo Toledo R, Carballeira Rodríguez JD. [Electronic medical record and prescription: risks and benefits detected since its implementation. Safe designing, rollout and use]. Aten Primaria 2021;53 Suppl 1(Suppl 1):102220 [FREE Full text] [doi: 10.1016/j.aprim.2021.102220] [Medline: 34961584]
11. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. Lancet 2003;362(9391):1225-1230. [doi: 10.1016/S0140-6736(03)14546-1] [Medline: 14568747]

12. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PC, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999;282(15):1458-1465. [doi: 10.1001/jama.282.15.1458] [Medline: 10535437]
13. Kulchar RJ, Chen K, Moon C, Srinivas S, Gupta A. Telemedicine, safe medication stewardship, and COVID-19: digital transformation during a global pandemic. J Interprof Educ Pract 2022;29:100524 [FREE Full text] [doi: 10.1016/j.xjep.2022.100524] [Medline: 35935734]
14. Pedrotti CHS, Accorsi TAD, De Amicis Lima K, Serpa Neto A, Lira MTDS, Morbeck RA, et al. Antibiotic stewardship in direct-to-consumer telemedicine consultations leads to high adherence to best practice guidelines and a low prescription rate. Int J Infect Dis 2021;105:130-134 [FREE Full text] [doi: 10.1016/j.ijid.2021.02.020] [Medline: 33578013]
15. Schiff G, Mirica MM, Dhavle AA, Galanter WL, Lambert B, Wright A. A prescription for enhancing electronic prescribing safety. Health Aff 2018;37(11):1877-1883. [doi: 10.1377/hlthaff.2018.0725] [Medline: 30395495]
16. Reed-Kane D, Kittell K, Adkins J, Flocks S, Nguyen T. E-prescribing errors identified in a compounding pharmacy: a quality-improvement project. Int J Pharm Compd 2014;18(1):83-86. [Medline: 24881345]
17. Reed-Kane D, Vasquez K, Pavlik A, Peragine J, Sandberg M. E-prescription errors and their resolution in a community compounding pharmacy. Int J Pharm Compd 2014;18(2):159-161. [Medline: 24881120]
18. Kannry J. Effect of e-prescribing systems on patient safety. Mt Sinai J Med 2011;78(6):827-833. [doi: 10.1002/msj.20298] [Medline: 22069206]
19. Lorkowski J, Maciejowska-Wilcock I, Pokorski M. Overload of medical documentation: a disincentive for healthcare professionals. Adv Exp Med Biol 2021;1324:1-10. [doi: 10.1007/5584_2020_587] [Medline: 33034843]
20. Savioli G, Ceresa IF, Gri N, Bavestrello Piccini G, Longhitano Y, Zanza C, et al. Emergency department overcrowding: understanding the factors to find corresponding solutions. J Pers Med 2022;12(2):279 [FREE Full text] [doi: 10.3390/jpm12020279] [Medline: 35207769]
21. Kramer HS, Drews FA. Checking the lists: A systematic review of electronic checklist use in health care. J Biomed Inform 2017 Jul;71S:S6-S12 [FREE Full text] [doi: 10.1016/j.jbi.2016.09.006] [Medline: 27623535]

Abbreviations

EHR: electronic health record
ICD: International Classification of Diseases
ICD-10: International Statistical Classification of Diseases, Tenth Revision
PDSS: prescription decision support system

Edited by C Lovis; submitted 24.01.24; peer-reviewed by J Pevnick, J Knitza; comments to author 23.03.24; revised version received 14.05.24; accepted 25.05.24; published 25.10.24.

Please cite as:
Accorsi TAD, Eduardo AA, Baptista CG, Moreira FT, Morbeck RA, Köhler KF, Lima KDA, Pedrotti CHS
The Impact of International Classification of Disease–Triggered Prescription Support on Telemedicine: Observational Analysis of Efficiency and Guideline Adherence
JMIR Med Inform 2024;12:e56681
URL: https://medinform.jmir.org/2024/1/e56681
doi:10.2196/56681
PMID:39453703

An Electronic Health Record–Integrated Application for Standardizing Care and Monitoring Patients With Autosomal Dominant Polycystic Kidney Disease Enrolled in a Tolvaptan Clinic: Design and Implementation Study

Maroun Chedid[1], MD; Fouad T Chebib[2], MD; Erin Dahlen[3], BSN; Theodore Mueller[3], BSN; Theresa Schnell[3], BSN; Melissa Gay[3], BSN; Musab Hommos[4], MBBS; Sundararaman Swaminathan[4], MBBS; Arvind Garg[5], MBBS; Michael Mao[2], MD; Brigid Amberg[3], BSN; Kirk Balderes[6], BS; Karen F Johnson[6], MA; Alyssa Bishop[6], MBA; Jackqueline Kay Vaughn[6], MA; Marie Hogan[3], MD, PhD; Vicente Torres[3], MD, PhD; Rajeev Chaudhry[7], MBBS, MPH; Ziad Zoghby[3], MBA, MD

1
2
3
4
5
6
7

Corresponding Author:
Ziad Zoghby, MBA, MD

Abstract

Background: Tolvaptan is the only US Food and Drug Administration–approved drug to slow the progression of autosomal dominant polycystic kidney disease (ADPKD), but it requires strict clinical monitoring due to potential serious adverse events.

Objective: We aimed to share our experience in developing and implementing an electronic health record (EHR)–based application to monitor patients with ADPKD who were initiated on tolvaptan.

Methods: The application was developed in collaboration with clinical informatics professionals based on our clinical protocol with frequent laboratory test monitoring to detect early drug-related toxicity. The application streamlined the clinical workflow and enabled our nursing team to take appropriate actions in real time to prevent drug-related serious adverse events. We retrospectively analyzed the characteristics of the enrolled patients.

Results: As of September 2022, a total of 214 patients were enrolled in the tolvaptan program across all Mayo Clinic sites. Of these, 126 were enrolled in the Tolvaptan Monitoring Registry application and 88 in the Past Tolvaptan Patients application. The mean age at enrollment was 43.1 (SD 9.9) years. A total of 20 (9.3%) patients developed liver toxicity, but only 5 (2.3%) had to discontinue the drug. The 2 EHR-based applications allowed consolidation of all necessary patient information and real-time data management at the individual or population level. This approach facilitated efficient staff workflow, monitoring of drug-related adverse events, and timely prescription renewal.

Conclusions: Our study highlights the feasibility of integrating digital applications into the EHR workflow to facilitate efficient and safe care delivery for patients enrolled in a tolvaptan program. This workflow needs further validation but could be extended to other health care systems managing chronic diseases requiring drug monitoring.

(JMIR Med Inform 2024;12:e50164) doi:10.2196/50164

KEYWORDS

ADPKD; autosomal dominant polycystic kidney disease; polycystic kidney disease; tolvaptan; EHR; electronic health record; digital health solutions; monitoring; kidney disease; drug-related toxicity; digital application; management; chronic disease

Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is the leading genetic cause and the fourth overall cause of end-stage kidney failure (ESKF) [1]. Patients with polycystic kidney disease 1 (PKD1) mutations develop ESKF 20 years earlier than those with PKD2 mutations [2,3]. The Mayo imaging classification (MIC) is a validated tool that identifies patients at risk for rapid progression to kidney failure, and disease-modifying therapy is recommended for patients with class 1C, 1D, or 1E, who have higher total kidney volume (TKV) growth rates [4]. In 2018, tolvaptan (brand name Jynarque; Otsuka America Pharmaceutical) was approved by the US Food and Drug Administration (FDA) as the first drug to slow kidney function decline in patients with rapidly progressive ADPKD. Tolvaptan reduces kidney volume growth and estimated glomerular filtration rate (eGFR) decline, delaying the need for kidney replacement therapy [5,6]. Tolvaptan acts by blocking the vasopressin V2 receptors in the distal nephron and collecting duct, inhibiting urinary concentration and sodium reabsorption and reversing the tubuloglomerular feedback inhibition induced by vasopressin, thus acutely and reversibly decreasing eGFR and possibly glomerular hyperfiltration [5]. However, tolvaptan is associated with several side effects, including polyuria, urinary frequency, thirst, and nocturia, which require patient education on adequate hydration. Tolvaptan can also cause significant hepatotoxicity in 5% of patients; thus, periodic liver function tests are mandated by the FDA through the risk evaluation and mitigation strategy (REMS) safety program. Due to the side effects profile and the necessary frequent laboratory test monitoring, the cost associated with staff time to manage the program beyond face-to-face care can limit the ability of health care teams to safely provide this disease-modifying therapy [7,8].

Tools that are directly integrated in the electronic health record (EHR) workflow can increase efficiency, reduce cost, and improve drug monitoring and quality of care [9-12]. For example, a cluster randomized clinical trial in primary care provided access, within the EHR, to a prescription drug monitoring program (PDMP) before the prescription of opioids. The integration increased PDMP-querying rates, suggesting that direct access reduced hassle costs and could improve adherence to guideline-concordant care practices [13]. Another study reported that the design and implementation of an electronic registry with a complementary workflow established an active tracking system that improved monitoring of patients on anticoagulation therapy [14]. However, no prior EHR-integrated workflow has been developed and validated to safely and successfully monitor patients with ADPKD treated with tolvaptan.

This paper describes the design, development, and implementation of an intelligent automated application within the EHR to efficiently manage and monitor ADPKD patients enrolled in the Mayo Clinic tolvaptan program. The goal of this paper is to illustrate how digital applications integrated into the EHR workflow can facilitate efficient and safe care for patients enrolled in a drug monitoring program and how this workflow can be extended to similar programs in chronic disease management and lay the groundwork for quality improvement efforts.

Methods

Ethical Considerations

This work was reviewed by the Mayo Clinic Institutional Review Board (21-005428). The study was exempt from clinical research overight because it was considered to be a quality improvement project. Informed consent was waived and data were deidentified.

Practice Setting

The Mayo Clinic is an integrated, multispecialty, multistate, large academic health system with locations in Minnesota, Florida, and Arizona; there are also other Mayo Clinic health system hospitals across Minnesota and Wisconsin. Since 2018, the Mayo Clinic uses a single, integrated EHR (Epic Systems) across all campuses. The Minnesota practice where the tolvaptan EHR application was initially launched includes 5 experienced nephrologists and 4 nurses directly involved in the ADPKD practice and various other specialists (ie, geneticists, hepatologists, liver surgeons, neurologists, neurosurgeons, pain specialists, interventional radiologists, transplant experts, and research coordinators) who care for these patients as needed. The tolvaptan EHR application was eventually adopted enterprise-wide in 2022, although the workflow may differ slightly by site based on the specificity and resources available in each practice. The 3 main Mayo Clinic campuses in Minnesota, Florida, and Arizona are designated as centers of excellence for ADPKD care by the Polycystic Kidney Disease Foundation.

Clinical Protocol—Indications and Monitoring of Tolvaptan Treatment

Tolvaptan is prescribed in patients aged 18 to 55 years with eGFR ≥25 mL/min/1.73 m^2 and at risk of rapid progression, defined by having an age-indexed height-adjusted TKV within MIC class 1C, 1D, and 1E [5,6,15]. Contraindications to initiate tolvaptan include history of liver injury, uncorrected hypernatremia, hypovolemia, inability to sense thirst, urinary tract obstruction, and concomitant use of strong CYP3A (cytochrome P450, family 3, subfamily A) enzyme inhibitors [16]. Tolvaptan initiation requires a multidisciplinary approach led by the treating nephrologist and a well-trained nursing team. In our program, following a shared decision discussion, eligible patients who agree to start tolvaptan are referred to a specialized nephrology nurse for a detailed educational session. The nurse visit includes a blood pressure check, assessment of alcohol consumption, dietary review, and in-depth education about the side effects of tolvaptan and the need for routine laboratory test monitoring. Patients are instructed to have a drug holiday in certain situations, such when they are unable to maintain adequate fluid intake, are hospitalized, are about to undergo an elective procedure, or are traveling. After confirming their willingness to take the medication, the nurse enrolls the patient in the mandatory REMS program, a drug safety program developed by the FDA for certain medications with serious safety concerns. As part of the tolvaptan REMS program, the

following laboratory tests are performed before the morning dose of tolvaptan: aspartate transaminase (AST), alanine transaminase (ALT), total bilirubin, serum sodium (advised but optional), and creatinine (advised but optional). Results are collected 2 and 4 weeks after tolvaptan initiation, then monthly for 18 months, and every 3 months thereafter [16]. Staff must log in to complete a REMS attestation every 3 months for each patient. Liver enzyme elevation and changes in serum sodium or creatinine are reviewed after each test in a timely fashion by the nursing team and the prescribing nephrologist. This process is designed to detect any laboratory test abnormality or the development of drug complications that could otherwise go unnoticed. For example, one threshold for suspending the medication is elevation in AST or ALT twice above their baseline level, which might not be flagged in the test report. However, this process can be very cumbersome and time consuming for the clinical team. An intelligent, automated, and streamlined real-time EHR-based process of tracking and monitoring is essential for efficient and safe care delivery, especially in specialized centers with a large patient population.

Architecture and Application Development by the Cohort Knowledge Intelligence Solutions Team

At the Mayo Clinic, the Cohort Knowledge Intelligence Solutions (CKIS) team is behind the development of many innovative patient cohort management solutions using the Epic Healthy Planet module. The CKIS team uses a collaborative, agile approach that incorporates feedback from clinical stakeholders and informatics to create care management solutions based on agreed-upon protocols of care that improve and automate processes for clinical staff, all managed within the EHR. All projects are reviewed through a standard intake process that factors in the scope of the project, enterprise impact, patient safety, quality of care, and revenue impact, among other criteria. Once approved and assigned, a business analyst and a builder will work with a group of stakeholders anywhere between several weeks to several months, depending on the scope of the project, to complete a solution build.

After the scope of a project is defined, a registry is used to gather a patient cohort along with a subset of metrics required to support the practice needs. The registry is an internal tool housed within Epic's software and uses a rule-based framework consisting of 2 main components: an inclusion rule and metrics. The inclusion rule is used to define the population and uses a combination of charted data, such as the patient's diagnosis, medication, and surgeries, or general demographics (eg, age and gender). The metrics (ie, rules) define what data will be captured for the population. Once a patient meets the defined inclusion criteria, the underlying metrics are processed, and data is captured. Metrics are designed to support the monitoring workflow in addition to future quality analysis and outcomes. They typically capture dates, laboratory test values, appointment information, patient demographics, and more. Finally, a report is built allowing users to visualize and interact with the registry data. Within the reports, specific patient metrics are displayed pertinent to the practice and may include laboratory test results, appointment information, or customized algorithms to create alerts for care team members to help them prioritize their work. The last phase of the build process includes testing and ensuring that the initial agreed-upon requirements have been met. Several months after the build is complete, the CKIS team meets with the customers to complete a value assessment that measures the impact of the solution provided.

Process of Tolvaptan Application Development

The nephrology ADPKD practice assembled a team of stakeholders to streamline the enrollment and monitoring of patients in the tolvaptan program. After an initial discussion in early 2020, the clinical team determined the content of the application. The stakeholders met on average every 2 weeks over a 3-month period to develop, in an iterative fashion, the initial application and test the efficiency of the system over the subsequent 3 to 4 months. The team determined that 2 applications were required to serve the clinical need. The first and main application, titled Tolvaptan Monitoring Registry, manages all patients actively treated with tolvaptan by consolidating all relevant information in one screen. Patients who discontinue tolvaptan are removed from the first application and automatically added to the second application, titled Past Tolvaptan Patients. The second application allows the care team to maintain a log of all past participants and record the reason for drug discontinuation, such as adverse effects or requiring renal replacement therapy.

Results

In September 2020, 2 EHR-based applications for monitoring patients taking tolvaptan were activated for clinical use. The tolvaptan clinic was established 2 years earlier when the FDA approved tolvaptan for the treatment of ADPKD. All patients enrolled in the program prior to September 2020 (n=32) were retrospectively added to the tolvaptan monitoring application.

Clinical Workflow Using the Tolvaptan Application

The tolvaptan application workflow involves the submission of an electronic prescription order by a nephrology nurse (Figure 1) and completion of an electronic activation form to enroll patients in the EHR-based tolvaptan monitoring application (Figure 2). This form includes the patient's Mayo Clinic site, primary nephrology clinician, and date of treatment initiation. The automated addition of patients into the registry reduces the risk of missing any patient prescribed tolvaptan, ensuring that all treated patients are closely monitored for any adverse events that might occur while on therapy. Quarterly meetings take place between the nursing team and the nephrologists to review workflow issues and assess any new complications that may arise.

Figure 1. Tolvaptan order report. REMS: risk evaluation and mitigation strategy.

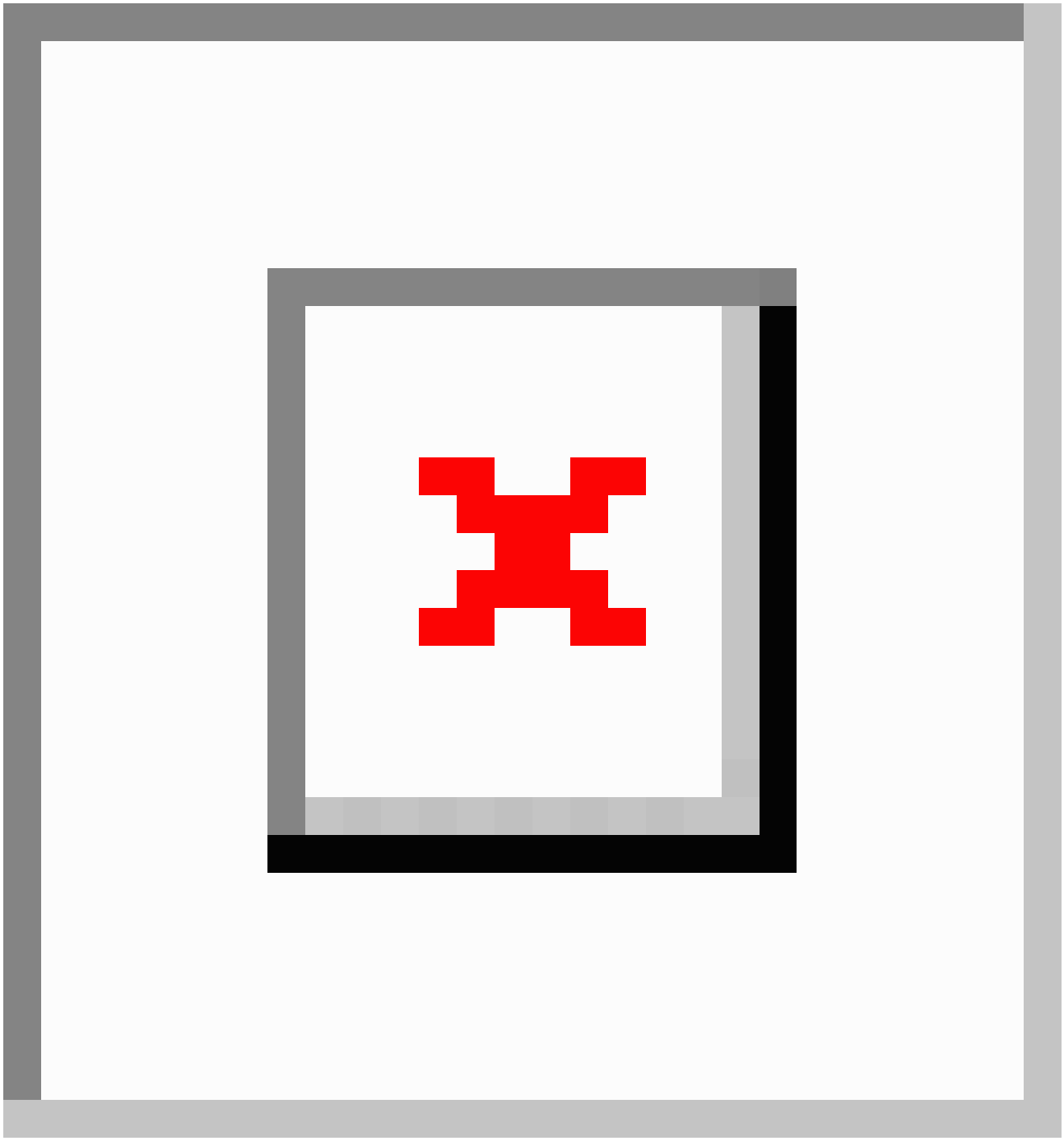
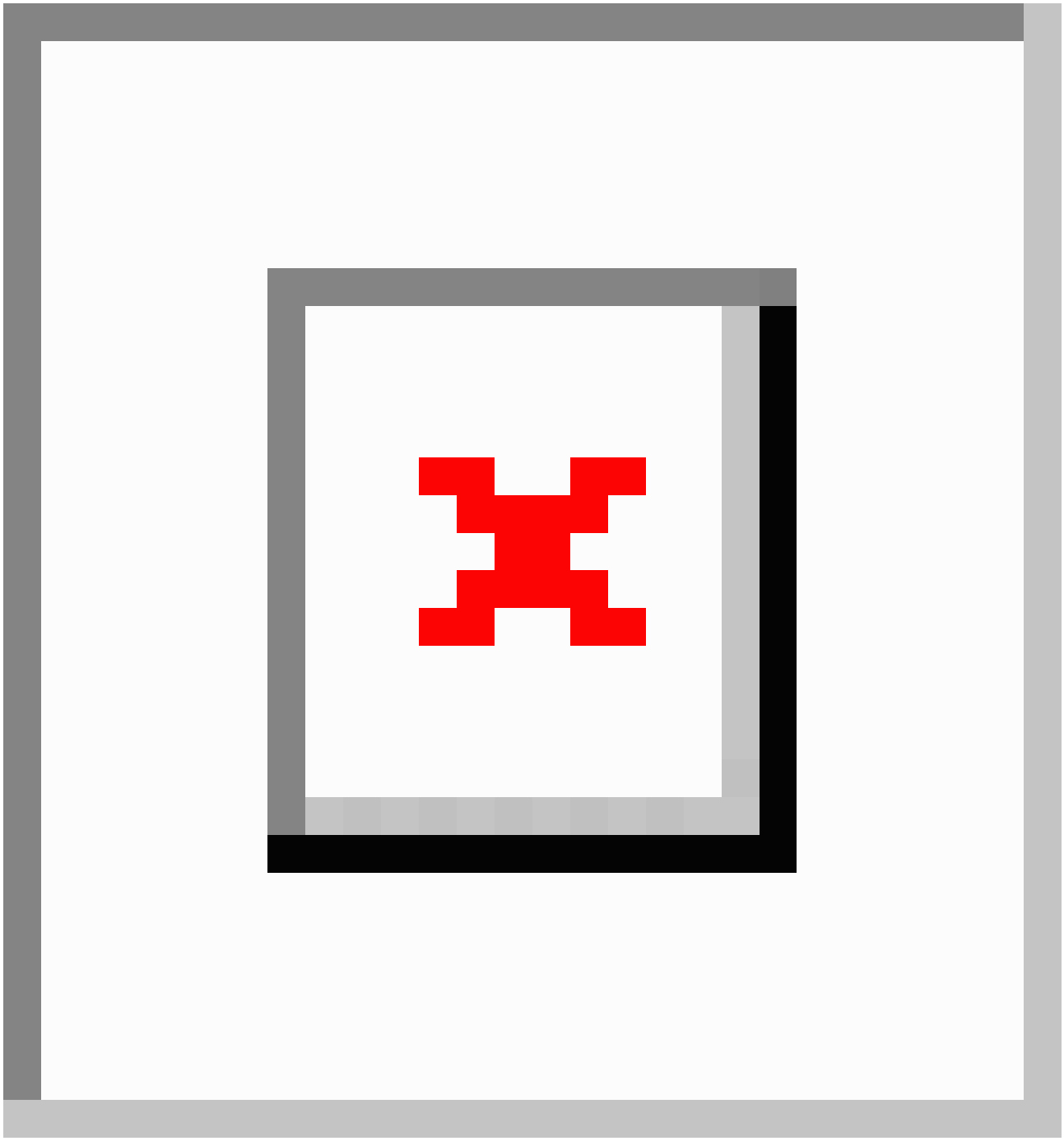


Figure 2. Smart form. MCHS: Mayo Clinic Health System; MN: Minnesota; NE: northeast; NW: northwest; REMS: risk evaluation and mitigation strategy; SW: southwest; WI: Wisconsin.



The main application lists all patients actively taking tolvaptan and includes several columns with relevant information, such as the patient's name, medical record number, date of initiation of tolvaptan, date of last liver function test, abnormal or urgent laboratory test flags, a "needs review" flag (Multimedia Appendix 1 provides details and flag criteria), the recommended laboratory test frequency based on the first dose date (ie, monthly or quarterly), the laboratory test's due date and the date of the next scheduled laboratory test, the last and next (when applicable) nursing outreach dates to the patient, the name of the treating nephrologist, and the last clinic visit date (Figure 3). The application allows filtering based on these variables, such as visualizing only patients who have abnormal laboratory tests or need review based on new laboratory tests since the last outreach date. The application also provides more detailed information for a specific patient based on several reports embedded at the bottom of the screen. These include Tolvaptan Monitoring Summary, Patient Visits, Nephrology Notes/Orders, and Patient Message Review. In our clinical workflow, every week, 1 of 4 dedicated nurses (on a rotation basis) reviews all flagged patients and takes appropriate action based on our clinical protocol. The EHR-based tolvaptan monitoring application provides several reports that allow for a more detailed review of a specific patient without having to open their chart. The Tolvaptan Monitoring Summary displays all monitored laboratory test results, such as AST, ALT,

bilirubin, serum creatinine, eGFR, serum sodium, and urine osmolarity (Figure 4). The Tolvaptan Monitoring Metrics window in the same section allows for quick access to recorded baseline laboratory test measurements and any abnormalities recorded. For example, the report displays a question and response: "Any Abnormal Liver Labs?" (answers are yes or no) (Figure 5). Additionally, all attempted or completed outreach interactions are listed with the name of the nurse conducting the activity and most recent nephrology note (Figure 6).

The Patient Visits report shows future scheduled appointments and surgeries, as well as a record of the patient's last 10 outpatient visits. This report also includes the patient's care team, demographics, and emergency contacts. The Nephrology Notes/Orders report displays pertinent medical history, current medication, immunizations, renal replacement therapy status, allergies, procedures, and the latest nephrology notes and specific ADPKD management–related comments by the nephrologist. Lastly, the Patient Message Review report includes all the patient's communications with personnel, nurses, and clinicians, as well as patient online services.

Figure 3. Tolvaptan monitoring snapshot. Abn: abnormal; Dt: date; REMS: risk evaluation and mitigation strategy.

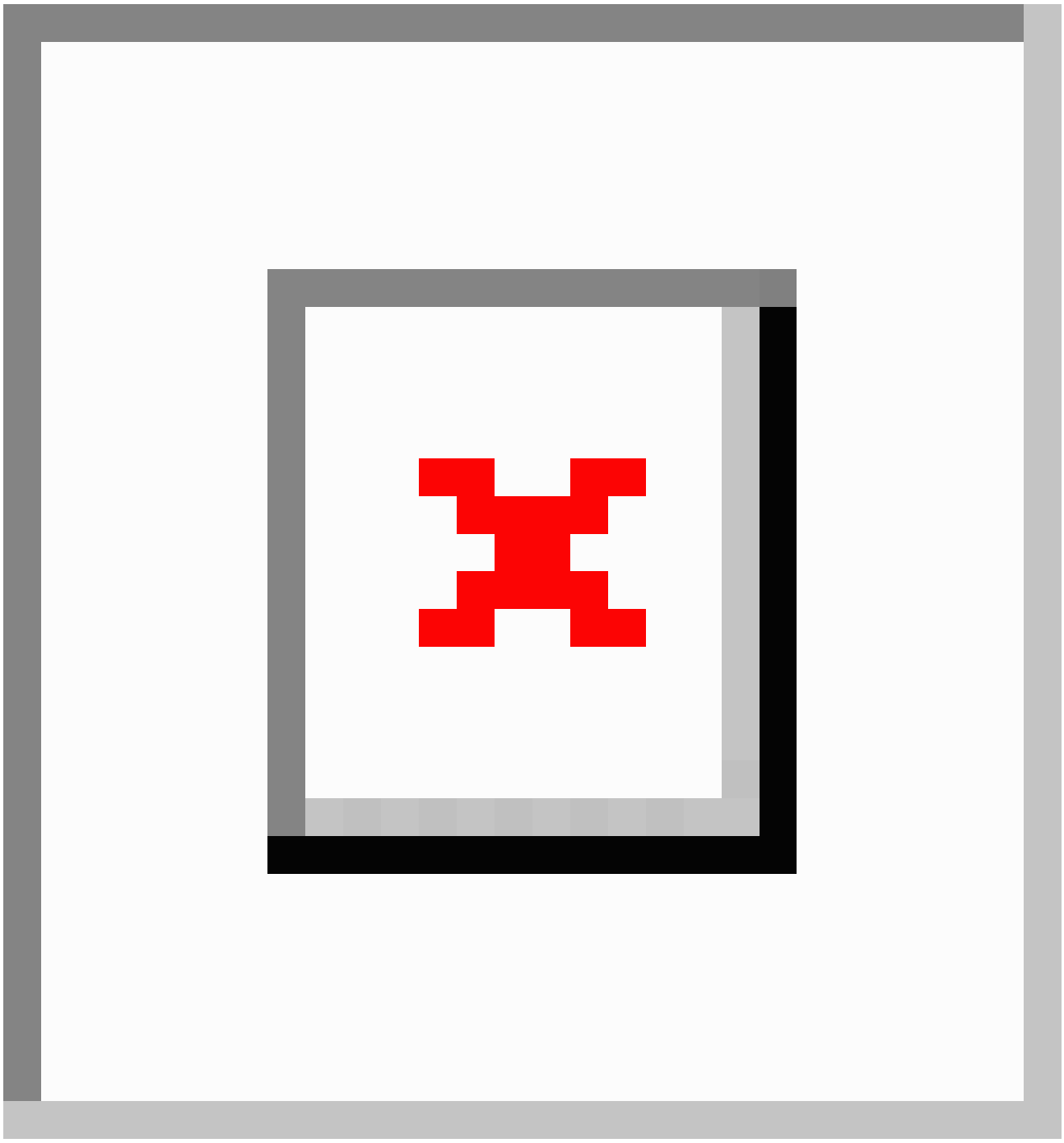


Figure 4. Tolvaptan Monitoring Summary. eGFR: estimated glomerular filtration rate.

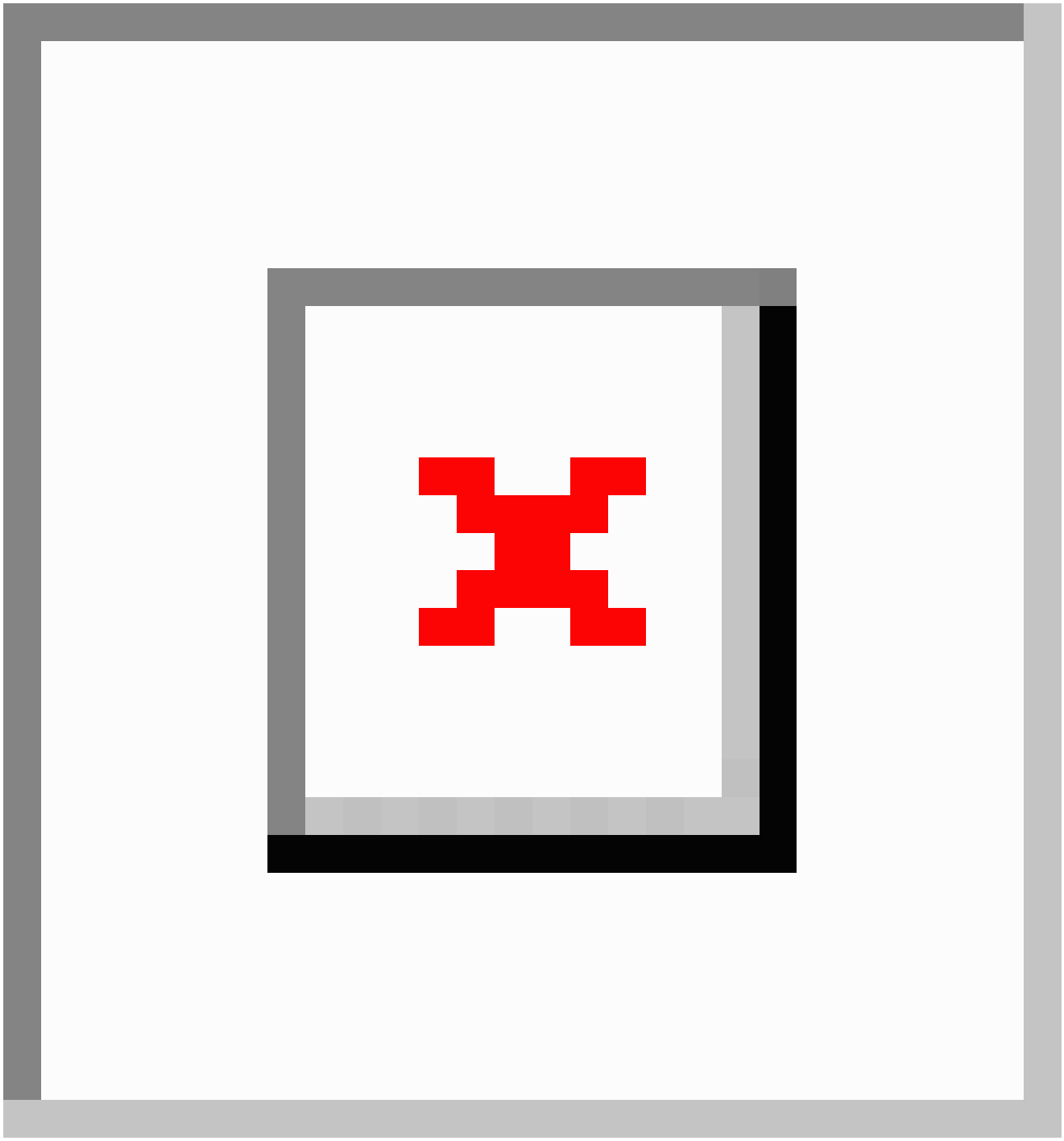


Figure 5. Tolvaptan Monitoring Metrics. ALK Phos: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; eGFR: estimated glomerular filtration rate; Tot Bili: total bilirubin.

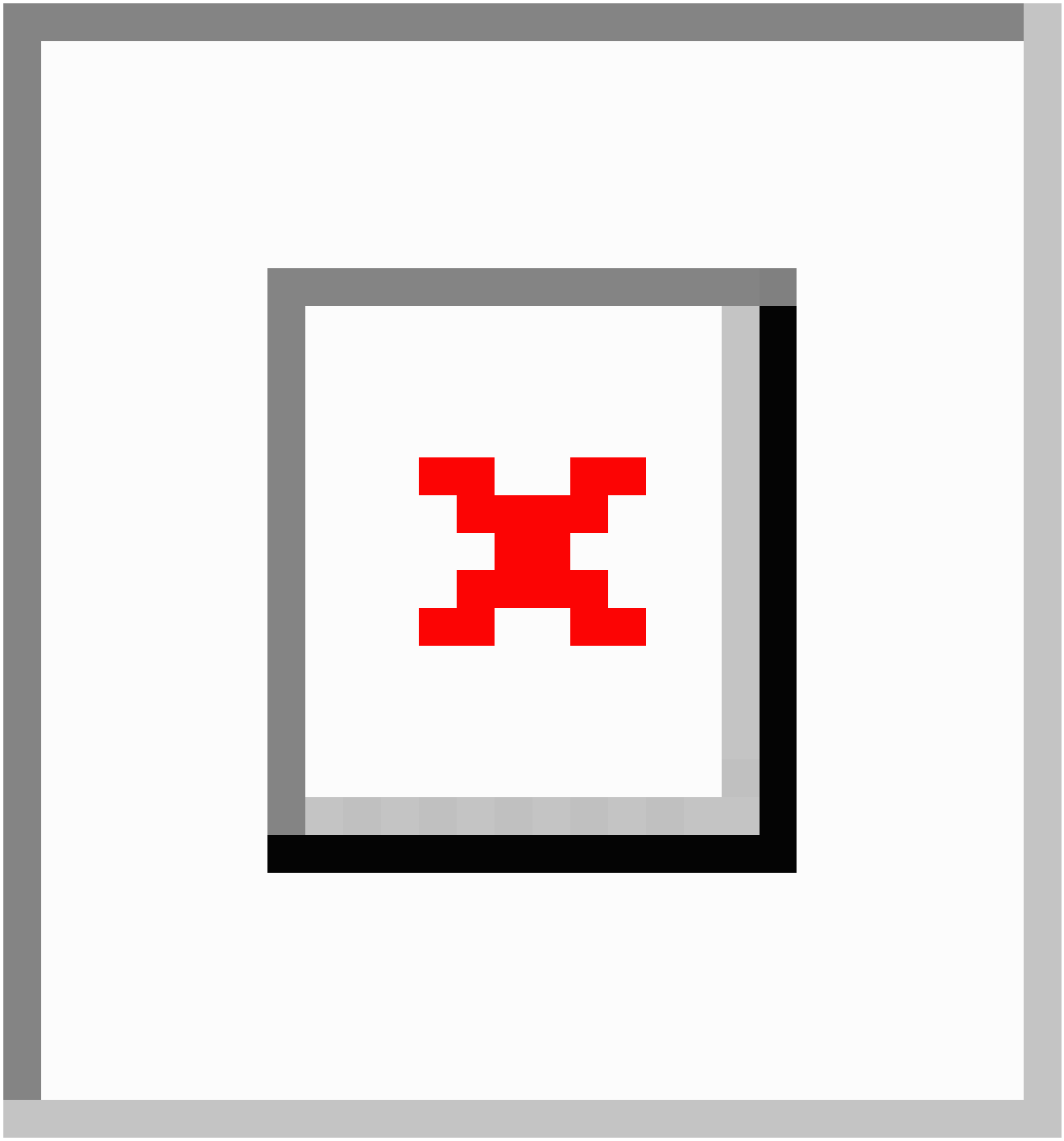
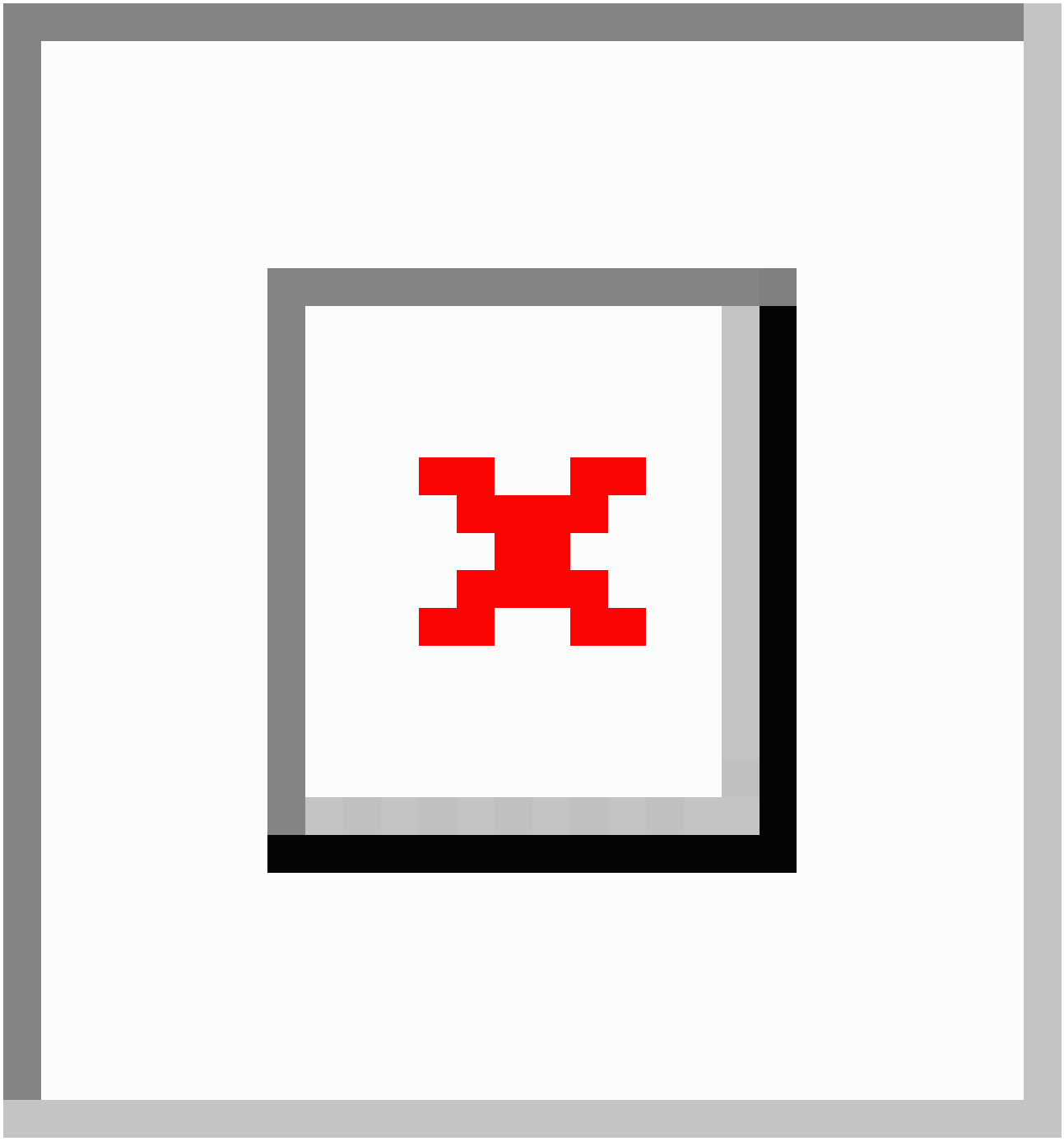


Figure 6. Tolvaptan detailed summary report.



Characteristics of Enrolled Patients

As of September 1, 2022, a total of 214 patients have been enrolled in the tolvaptan program across the Mayo Clinic Health System in Minnesota, Arizona, Florida, and Wisconsin (Table 1).

Of the 214 patients, 126 (59%) were enrolled in the Tolvaptan Monitoring Registry, and the remaining 88 patients were included in Past Tolvaptan Patients. A total of 10 nephrologists were caring for these patients across the 4 locations. Table 2 displays characteristics of the patients in the Tolvaptan Monitoring Registry, including their demographics and MIC status.

The registry included 57.9% (n=124) female, 96.2% (n=206) White, 2.8% (n=6) Hispanic, and 0.9% (n=2) African American individuals. The mean age at enrollment was 43.1 (SD 9.9) years, and 86 patients had a documented MIC. Most patients had an MIC of 1C (n=38, 44.2%), followed by 1D (n=25, 29.1%) and 1E (n=19, 22.1%). Additionally, 3 patients (5.1%) had an MIC of 1B but were prescribed tolvaptan based on a non-MIC criterion. Of note, 33 (15.4%) patients were taking tolvaptan as part of a clinical trial and remained on the drug after trial completion (following FDA approval of the drug). These patients were added retrospectively to the application because they were taking tolvaptan prior to the creation of the EHR-based application.

Table . Distribution of patients with autosomal dominant polycystic kidney disease receiving tolvaptan treatment across the Mayo Clinic system.

Region	Tolvaptan Monitoring Registry (n=126), n (%)	Past Tolvaptan Patients (n=88), n (%)	Total (N=214), n (%)
Minnesota	80 (63.5)	75 (85.2)	155 (72.4)
Arizona	17 (13.5)	7 (8)	24 (11.2)
Florida	22 (17.4)	4 (4.5)	26 (12.1)
MCHS[a], WI[b]	7 (5.6)	2 (2.3)	9 (4.2)

[a]MCHS: Mayo Clinic Health System.

[b]WI: Wisconsin.

Table . Patient demographics and Mayo imaging class.

		Tolvaptan Monitoring Registry	Past Tolvaptan Patients	Total patients
Demographics				
	Patients, n	126	88	214
	Female, n (%)	72 (57.1)	52 (59)	124 (58)
	Age at registry inclusion (years), mean (SD)	43.8 (9.9)	40.4 (9.9)	43.1 (9.9)
	White, n (%)	123 (97.6)	83 (94.3)	206 (96.3)
	Hispanic, n (%)	3 (2.4)	3 (3.4)	6 (2.8)
	African American, n (%)	0 (0)	2 (2.3)	2 (0.9)
	Enrollment through clinical trials, n (%)	29 (22.4)	4 (4.5)	33 (15.4)
Mayo imaging class				
	Patients, n	58	28	86
	1A	0 (0)	0 (0)	0 (0)
	1B	3 (5.2)	1 (3.6)	4 (4.7)
	1C	27 (46.6)	11 (39.3)	38 (44.2)
	1D	19 (32.8)	6 (21.4)	25 (29.1)
	1E	9 (15.5)	10 (35.7)	19 (22.1)

Outcomes of Using the Tolvaptan Application

The implementation of the tolvaptan EHR-based application streamlined the monitoring process of patients treated with tolvaptan in several ways. First, the automated addition of patients into the registry reduced the risk of missing any patients started on tolvaptan, thus assuring that all treated patients were closely monitored for any adverse events that might occur while on therapy. Second, the application allowed for efficient and timely identification of patients who had abnormal laboratory test results and enabled nursing outreach to patients who might need further intervention or education on medication management. Overall, 20 (9.3%) patients had liver function test abnormalities, but only 5 (2.3%) had to discontinue the drug because of hepatotoxicity. The most common reason for drug discontinuation was related to the aquaretic effect, in 10 patients (4.7%), while only 4 (1.8%) could not continue in the program because of medical insurance–related issue. Third, the application provides a comprehensive and up-to-date summary of all pertinent clinical information related to the management of ADPKD, including medications, appointments, laboratory tests, and notes from the care team. Fourth, the application facilitated communication and collaboration among the multidisciplinary team involved in the care of patients with ADPKD. The standardization process and easy data access to all enrolled patients in the registry provided an opportunity for the care team to meet quarterly to review and discuss specific scenarios. These discussions sometimes led team members to share their experiences regarding challenging situations or drug-related adverse events or drug intolerance and at other times to propose enhancements to the EHR application. Finally, the application enhanced the efficiency of the tolvaptan program by reducing the time and effort (informally reported by the care team) required for enrollment, tracking, and monitoring of patients. More specifically, for the physicians, the only required task to enroll a patient in the program was identifying the candidate and making an electronic referral to the nursing team. The nursing team then initiated the education, treatment, and monitoring without any further escalation to the physician, unless there were concerns. For the nurses, all relevant information was consolidated, reducing the need to navigate to various EHR screens and modules.

Discussion

Principal Findings

In this report, we share our experience in developing and implementing an EHR application to manage and monitor patients with ADPKD who were enrolled in the tolvaptan program across several sites at our institution. This application streamlined the clinical workflow and enabled the nephrology nursing team to proactively take appropriate action to mitigate drug-related serious adverse events. Tolvaptan is the first and only FDA-approved drug to slow the progression of ADPKD, but it has multiple adverse effects, most seriously liver toxicity, which can be potentially severe, albeit rare. Therefore, frequent laboratory test monitoring is required to detect early drug-related toxicity. This application is crucial in facilitating the monitoring of patients taking tolvaptan, especially in large centers with high case load or smaller clinics with limited staff and resources. Overall, 20 (9.3%) patients had liver function test abnormalities, but only 5 (2.3%) had to discontinue the drug because of hepatotoxicity. The frequency of these events is very similar to those reported in the REPRISE clinical trial (10.9% hepatotoxicity and 1.6% discontinuation for a liver event) [6]. The most common reason for drug discontinuation was related to the aquaretic effect, which occurred in 4.7% (n=10) of patients. This is higher than the frequency reported in the REPRISE trial (2.1%) but not surprising since participants enrolled in clinical trials may be more motivated to adhere to the treatment protocol. Nonetheless, these clinical outcomes are reassuring.

The logistical requirements of any tolvaptan program may limit the ability of nephrology practices to provide this effective therapy. With the shortage of physicians and their high level of burnout [17-19], well-designed EHR integration that helps review, in a consolidated manner, relevant data for clinical care is important, although it comes with a higher up-front cost [20-23]. This is now more relevant because about 90% of office-based physicians in the United States use an EHR [24], and higher perceived EHR usability is associated with higher levels of perceived positive outcomes (improved patient care) and lower levels of perceived negative outcomes (worse patient interactions and work-life integration) [25]. Whether developing such digital systems is worth the investment is a relevant question for health care systems [26], but they can certainly be scaled in real-world settings. The cost of creating similar EHR-based applications will vary depending on each organization structure and is mostly an up-front cost. This includes the time required by both the clinical care team (nurses, physicians, and other clinical staff) and informatics team (program manager and technical build team) to identify the clinical need and develop and test the product. For our practice, it required at least 1 physician and 1 nurse champion to be present at each meeting (4 staff members were engaged) with the informatics team for 1 hour every 2 weeks on average over a 6-month period (12 hours per staff member involved). Since all our staff are salary based, this work was primarily supported by discretionary efforts and during nonclinical activities (lunch hour or administrative time). Regarding the informatics team (CKIS), our institution has allocated an operational budget to support various EHR-related projects across the enterprise; thus, we did not have to request extra funds to support this effort.

The advantages of these applications and data analytics capabilities within the EHR have been well described for various diseases and conditions, recently including more COVID-19–related activities, to manage the clinical practice safely [27-37]. Besides keeping track of a defined patient population, aggregating data, and identifying care gaps, communication with patients through the patient portal is readily accessible. In addition, bulk messaging (sending the same message to a group of patients in one click) is a convenient feature of the application. For example, staff can easily remind patients to do their monthly or quarterly laboratory tests when these have lapsed and do a synchronous or asynchronous quick health check if needed.

Limitations

The design, development, and deployment in clinical practice of this integrated digital application has limitations. The process is iterative and requires buy-in from various stakeholders, an up-front investment in time, resources, and change management capabilities. Our clinical team was receptive, open to change, and willing to embrace new workflows because of the perceived value of adopting the application (more efficient and safer care delivery). One limitation of our study is that it was conducted in a single health care system. However, the successful implementation of this application in our Minnesota practice, followed by its expansion to all Mayo Clinic practices, highlights the potential for scaling to other health care systems. Another limitation of our study is the lack of objective efficiency outcome measures. The workflow improvement and satisfaction were not evalutated in a formal manner by the physicians and nurses. Ideally, our study would assess the impact of the application using (1) direct observation (time-motion studies), (2) EHR log-based analysis (EHR log data), (3) care team pre- and postimplementation surveys, or a combination of these. However, prior to the implementation of the application, the management of the tolvaptan program was ad hoc, carried out by a care team that performed multiple unrelated clinical activities. This made it impractical to use time-motion studies and impossible to meaningfully use EHR log data. A care team survey was considered, but because the transition to the application was done during the COVID-19 pandemic when our personnel resources were very strained, noncritical activities were paused. Prospective studies are necessary to validate the effectiveness of this application and its potential for improving care processes and ultimately patient outcomes.

Conclusion

In conclusion, our multidisciplinary team developed an EHR-integrated digital monitoring protocol that could facilitate safe, efficient, and high-quality care for patients with ADPKD who were prescribed tolvaptan. The implementation of this application in our health care system can be scaled to other health care systems or smaller clinics after further validation. This can reduce some barriers and help safely provide the best available treatment for eligible patients.

Conflicts of Interest

ZZ serves as a member of the Epic nephrology steering board committee. MH has received consulting fees from Otsuka in the past for work unrelated to this study. All other authors report no conflicts of interest.

Multimedia Appendix 1
Definitions of terms in columns and flags.
[DOCX File, 13 KB - medinform_v12i1e50164_app1.docx]

References

1. Shukoor SS, Vaughan LE, Edwards ME, et al. Characteristics of patients with end-stage kidney disease in ADPKD. Kidney Int Rep 2020 Dec;6(3):755-767. [doi: 10.1016/j.ekir.2020.12.016] [Medline: 33732990]
2. Hateboer N, v Dijk MA, Bogdanova N, et al. Comparison of phenotypes of polycystic kidney disease types 1 and 2. Lancet 1999 Jan;353(9147):103-107. [doi: 10.1016/S0140-6736(98)03495-3] [Medline: 10023895]
3. Chebib FT, Torres VE. Autosomal dominant polycystic kidney disease: core curriculum 2016. Am J Kidney Dis 2016 May;67(5):792-810. [doi: 10.1053/j.ajkd.2015.07.037] [Medline: 26530876]
4. Irazabal MV, Rangel LJ, Bergstralh EJ, et al. Imaging classification of autosomal dominant polycystic kidney disease: a simple model for selecting patients for clinical trials. J Am Soc Nephrol 2015 Jan;26(1):160-172. [doi: 10.1681/ASN.2013101138] [Medline: 24904092]
5. Torres VE, Chapman AB, Devuyst O, et al. Tolvaptan in patients with autosomal dominant polycystic kidney disease. N Engl J Med 2012 Dec 20;367(25):2407-2418. [doi: 10.1056/NEJMoa1205511] [Medline: 23121377]
6. Torres VE, Chapman AB, Devuyst O, et al. Tolvaptan in later-stage autosomal dominant polycystic kidney disease. N Engl J Med 2017 Nov 16;377(20):1930-1942. [doi: 10.1056/NEJMoa1710030] [Medline: 29105594]
7. Tai-Seale M, Olson CW, Li J, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff (Millwood) 2017 Apr 1;36(4):655-662. [doi: 10.1377/hlthaff.2016.0811] [Medline: 28373331]
8. Prasad K, Poplau S, Brown R, et al. Time pressure during primary care office visits: a prospective evaluation of data from the Healthy Work Place study. J Gen Intern Med 2020 Feb;35(2):465-472. [doi: 10.1007/s11606-019-05343-6] [Medline: 31797160]
9. Witry M, Marie BS, Reist J. Provider perspectives and experiences following the integration of the prescription drug monitoring program into the electronic health record. Health Informatics J 2022;28(3):14604582221113435. [doi: 10.1177/14604582221113435] [Medline: 35829729]
10. Seki T, Aki M, Furukawa TA, et al. Electronic health record-nested reminders for serum lithium level monitoring in patients with mood disorder: randomized controlled trial. J Med Internet Res 2023 Mar 22;25:e40595. [doi: 10.2196/40595] [Medline: 36947138]
11. Mishra V, Chouinard M, Keiser J, et al. Automating vancomycin monitoring to improve patient safety. Jt Comm J Qual Patient Saf 2019 Nov;45(11):757-762. [doi: 10.1016/j.jcjq.2019.07.001] [Medline: 31526711]
12. Bundy DG, Marsteller JA, Wu AW, et al. Electronic health record-based monitoring of primary care patients at risk of medication-related toxicity. Jt Comm J Qual Patient Saf 2012 May;38(5):216-223. [doi: 10.1016/s1553-7250(12)38027-6] [Medline: 22649861]
13. Neprash HT, Vock DM, Hanson A, et al. Effect of integrating access to a prescription drug monitoring program within the electronic health record on the frequency of queries by primary care clinicians: a cluster randomized clinical trial. JAMA Health Forum 2022 Jun;3(6):e221852. [doi: 10.1001/jamahealthforum.2022.1852] [Medline: 35977248]
14. Lee SY, Cherian R, Ly I, Horton C, Salley AL, Sarkar U. Designing and implementing an electronic patient registry to improve warfarin monitoring in the ambulatory setting. Jt Comm J Qual Patient Saf 2017 Jul;43(7):353-360. [doi: 10.1016/j.jcjq.2017.03.006] [Medline: 28648221]
15. Chebib FT, Torres VE. Assessing risk of rapid progression in autosomal dominant polycystic kidney disease and special considerations for disease-modifying therapy. Am J Kidney Dis 2021 Aug;78(2):282-292. [doi: 10.1053/j.ajkd.2020.12.020] [Medline: 33705818]
16. Chebib FT, Perrone RD, Chapman AB, et al. A practical guide for treatment of rapidly progressive ADPKD with tolvaptan. J Am Soc Nephrol 2018 Oct;29(10):2458-2470. [doi: 10.1681/ASN.2018060590] [Medline: 30228150]
17. Physician workforce projections: the complexities of physician supply and demand. Association of American Medical Colleges. 2021. URL: https://www.aamc.org/data-reports/workforce/report/physician-workforce-projections [accessed 2024-04-23]
18. Shanafelt TD, West CP, Dyrbye LN, et al. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022 Dec;97(12):2248-2258. [doi: 10.1016/j.mayocp.2022.09.002] [Medline: 36229269]

19. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 1;27(4):531-538. [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
20. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426. [doi: 10.1370/afm.2121] [Medline: 28893811]
21. Kroth PJ, Morioka-Douglas N, Veres S, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 2;2(8):e199609. [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
22. Tawfik DS, Sinha A, Bayati M, et al. Frustration with technology and its relation to emotional exhaustion among health care workers: cross-sectional observational study. J Med Internet Res 2021 Jul 6;23(7):e26817. [doi: 10.2196/26817] [Medline: 34255674]
23. Sinsky CA, Shanafelt TD, Ripp JA. The electronic health record inbox: recommendations for relief. J Gen Intern Med 2022 Nov;37(15):4002-4003. [doi: 10.1007/s11606-022-07766-0] [Medline: 36036837]
24. Electronic medical records/electronic health records (EMRs/EHRs). US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm [accessed 2023-01-31]
25. Melnick ER, Sinsky CA, Dyrbye LN, et al. Association of perceived electronic health record usability with patient interactions and work-life integration among US physicians. JAMA Netw Open 2020 Jun 1;3(6):e207374. [doi: 10.1001/jamanetworkopen.2020.7374] [Medline: 32568397]
26. Shanafelt TD, Larson D, Bohman B, et al. Organization-wide approaches to foster effective unit-level efforts to improve clinician well-being. Mayo Clin Proc 2023 Jan;98(1):163-180. [doi: 10.1016/j.mayocp.2022.10.031] [Medline: 36603944]
27. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752. [doi: 10.7326/0003-4819-144-10-200605160-00125] [Medline: 16702590]
28. Dreyer NA, Garner S. Registries for robust evidence. JAMA 2009 Aug 19;302(7):790-791. [doi: 10.1001/jama.2009.1092] [Medline: 19690313]
29. Hersh W. Electronic health records facilitate development of disease registries and more. Clin J Am Soc Nephrol 2011 Jan;6(1):5-6. [doi: 10.2215/CJN.09901110] [Medline: 21127135]
30. Jaffe MG, Lee GA, Young JD, Sidney S, Go AS. Improved blood pressure control associated with a large-scale hypertension program. JAMA 2013 Aug 21;310(7):699-705. [doi: 10.1001/jama.2013.108769] [Medline: 23989679]
31. Jolly SE, Navaneethan SD, Schold JD, et al. Development of a chronic kidney disease patient navigator program. BMC Nephrol 2015 May 3;16:69. [doi: 10.1186/s12882-015-0060-2] [Medline: 26024966]
32. Mendu ML, Waikar SS, Rao SK. Kidney disease population health management in the era of accountable care: a conceptual framework for optimizing care across the CKD spectrum. Am J Kidney Dis 2017 Jul;70(1):122-131. [doi: 10.1053/j.ajkd.2016.11.013] [Medline: 28132720]
33. Navaneethan SD, Jolly SE, Schold JD, et al. Pragmatic randomized, controlled trial of patient navigators and enhanced personal health records in CKD. Clin J Am Soc Nephrol 2017 Sep 7;12(9):1418-1427. [doi: 10.2215/CJN.02100217] [Medline: 28778854]
34. Rana JS, Karter AJ, Liu JY, Moffet HH, Jaffe MG. Improved cardiovascular risk factors control associated with a large-scale population management program among diabetes patients. Am J Med 2018 Jun;131(6):661-668. [doi: 10.1016/j.amjmed.2018.01.024] [Medline: 29576192]
35. Mendu ML, Ahmed S, Maron JK, et al. Development of an electronic health record-based chronic kidney disease registry to promote population health management. BMC Nephrol 2019 Mar 1;20(1):72. [doi: 10.1186/s12882-019-1260-y] [Medline: 30823871]
36. Peralta CA, Livaudais-Toman J, Stebbins M, et al. Electronic decision support for management of CKD in primary care: a pragmatic randomized trial. Am J Kidney Dis 2020 Nov;76(5):636-644. [doi: 10.1053/j.ajkd.2020.05.013] [Medline: 32682696]
37. Jose T, Warner DO, O'Horo JC, et al. Digital health surveillance strategies for management of coronavirus disease 2019. Mayo Clin Proc Innov Qual Outcomes 2021 Feb;5(1):109-117. [doi: 10.1016/j.mayocpiqo.2020.12.004] [Medline: 33521582]

Abbreviations

ADPKD: autosomal dominant polycystic kidney disease
ALT: alanine transaminase
AST: aspartate transaminase
CKIS: Cohort Knowledge Intelligent Solutions
CYP3A: cytochrome P450, family 3, subfamily A
eGFR: estimated glomerular filtration rate

EHR: electronic health record
ESKF: end-stage kidney failure
FDA: US Food and Drug Administration
MIC: Mayo imaging classification
PDMP: prescription drug monitoring program
PKD: polycystic kidney disease
REMS: risk evaluation and mitigation strategy
TKV: total kidney volume

Edited by C Perrin; submitted 21.06.23; peer-reviewed by C Kwoh, J Walsh, O Amro; revised version received 06.03.24; accepted 25.03.24; published 01.05.24.

Please cite as:
Chedid M, Chebib FT, Dahlen E, Mueller T, Schnell T, Gay M, Hommos M, Swaminathan S, Garg A, Mao M, Amberg B, Balderes K, Johnson KF, Bishop A, Vaughn JK, Hogan M, Torres V, Chaudhry R, Zoghby Z
An Electronic Health Record–Integrated Application for Standardizing Care and Monitoring Patients With Autosomal Dominant Polycystic Kidney Disease Enrolled in a Tolvaptan Clinic: Design and Implementation Study
JMIR Med Inform 2024;12:e50164
URL: https://medinform.jmir.org/2024/1/e50164
doi:10.2196/50164

Enhancing Clinical History Taking Through the Implementation of a Streamlined Electronic Questionnaire System at a Pediatric Headache Clinic: Development and Evaluation Study

Jaeso Cho[1,*], MD, MS; Ji Yeon Han[1,2,*], MD, MS; Anna Cho[1], MD, PhD; Sooyoung Yoo[3], PhD; Ho-Young Lee[4], MD, PhD; Hunmin Kim[1,5], MD, PhD

1

2

3

4

5

[*]these authors contributed equally

Corresponding Author:
Hunmin Kim, MD, PhD

Abstract

Background: Accurate history taking is essential for diagnosis, treatment, and patient care, yet miscommunications and time constraints often lead to incomplete information. Consequently, there has been a pressing need to establish a system whereby the questionnaire is duly completed before the medical appointment, entered into the electronic health record (EHR), and stored in a structured format within a database.

Objective: This study aimed to develop and evaluate a streamlined electronic questionnaire system, BEST-Survey (Bundang Hospital Electronic System for Total Care-Survey), integrated with the EHR, to enhance history taking and data management for patients with pediatric headaches.

Methods: An electronic questionnaire system was developed at Seoul National University Bundang Hospital, allowing patients to complete previsit questionnaires on a tablet PC. The information is automatically integrated into the EHR and stored in a structured database for further analysis. A retrospective analysis compared clinical information acquired from patients aged <18 years visiting the pediatric neurology outpatient clinic for headaches, before and after implementing the BEST-Survey system. The study included 365 patients before and 452 patients after system implementation. Answer rates and positive rates of key headache characteristics were compared between the 2 groups to evaluate the system's clinical utility.

Results: Implementation of the BEST-Survey system significantly increased the mean data acquisition rate from 54.6% to 99.3% (P<.001). Essential clinical features such as onset, location, duration, severity, nature, and frequency were obtained in over 98.7% (>446/452) of patients after implementation, compared to from 53.7% (196/365) to 85.2% (311/365) before. The electronic system facilitated comprehensive data collection, enabling detailed analysis of headache characteristics in the patient population. Most patients (280/452, 61.9%) reported headache onset less than 1 year prior, with the temporal region being the most common pain location (261/703, 37.1%). Over half (232/452, 51.3%) experienced headaches lasting less than 2 hours, with nausea and vomiting as the most commonly associated symptoms (231/1036, 22.3%).

Conclusions: The BEST-Survey system markedly improved the completeness and accuracy of essential history items for patients with pediatric headaches. The system also streamlined data extraction and analysis for clinical and research purposes. While the electronic questionnaire cannot replace physician-led history taking, it serves as a valuable adjunctive tool to enhance patient care.

(JMIR Med Inform 2024;12:e54415) doi:10.2196/54415

KEYWORDS

electronic questionnaire system; electronic questionnaire; history taking; medical history; headache; migraine; neuralgia; pediatric; paediatric; infant; neonatal; toddler; child; youth; adolescent

Introduction

Headache is one of the most common neurological symptoms in children, with a reported prevalence of 54.4% to 58.4% in previous population-based studies [1-4]. Using an appropriate approach in the differential diagnosis of pediatric headache is critical because it can potentially impact children's quality of life in a significant manner [5,6]. A thorough history taking is required to differentiate the symptoms of different primary headaches and to rule out a small but critical number of life-threatening secondary headaches, such as brain tumors or intracranial hypertensions [7].

Despite its importance, many studies have shown that history taking can be the major contributing factor in the misdiagnosis of pediatric headache due to miscommunications and limited time available, leading to missing key information in a real-world setting [8,9]. To overcome such obstacles, various questionnaires for efficient history taking and early detection of specific types of primary headaches have been developed in the past, such as the Diagnostic Headache Diary [10], ID Migraine [11], Migraine Screen Questionnaire [12], and Brief Headache Screen [13]. Even with the development of different history-taking tools, a paper-based, self-administered questionnaire itself poses a significant limitation in data collection and organization since it requires laborious, time-intensive, manual input [14]. Consequently, there has been a pressing need to establish a system whereby the questionnaire is duly completed before the medical appointment, entered into the electronic health record (EHR), and stored in a structured format within a database.

Numerous studies across multiple clinical fields comparing the effectiveness of paper-based and computer-based questionnaires have consistently shown that electronic health history questionnaires have higher usability scores and are more cost-effective than paper-based ones [14-20]. In this study, we developed an electronic questionnaire system named BEST-Survey (Bundang Hospital Electronic System for Total Care-Survey) that enables a streamlined previsit electronic questionnaire, with automatic integration into the EHR to aid clinicians' history taking and the construction of a structured database for further data analysis. This study aimed to develop and evaluate the BEST-Survey system among patients with pediatric headaches.

Methods

Electronic Questionnaire System (BEST-Survey) Development

At Seoul National University Bundang Hospital, we constructed an electronic questionnaire system named BEST-Survey to enhance the clinical utility of patient-provided medical information. Using the BEST-Survey system, patients complete the questionnaire on a tablet PC before their visit, and the information is automatically integrated into our EHR to aid clinicians during their history taking. The patient-provided information is stored in a structured database for further data analysis.

The BEST-Survey system is composed of 2 parts: questionnaire archives and electronic questionnaire system structure. A task force comprising 30 members, including doctors, nurses, and medical information technologists, evaluated the questionnaires based on 5 criteria: predicted demand, target age group, questionnaire availability, clinical utilization within the EHR, and copyright issues. A total of 59 questionnaires were created based on the needs of 12 departments, one of which was the pediatric headache questionnaire. The electronic questionnaire system structure within BEST-Survey consists of the following components: (1) an electronic questionnaire input system, delivered using technology such as a tablet PC or computer, for patients to complete during the preclinic visit; (2) automatic integration of patient response to the EHR to aid clinicians' history taking; and (3) construction of a structured database based on patient input for further data analysis.

Pediatric Headache Questionnaire Development

At our pediatric neurology outpatient clinic, located in Gyeonggi Province, the largest province in South Korea by population, we see 200‐250 new patients with pediatric headaches annually. As a tertiary referral center, we primarily see patients with red-flag symptoms or refractory headaches referred from primary and secondary clinics. To improve the history-taking process and enhance clinical assessments for these patients, we developed a previsit questionnaire specifically for patients with headache attending their initial consultation. The questionnaire was developed with reference to Swaiman's *Pediatric Neurology 5th edition* [21]. To ensure patient comprehension without guidance from medical staff, we used simple language that is easily comprehensible and provided detailed explanations for complex concepts.

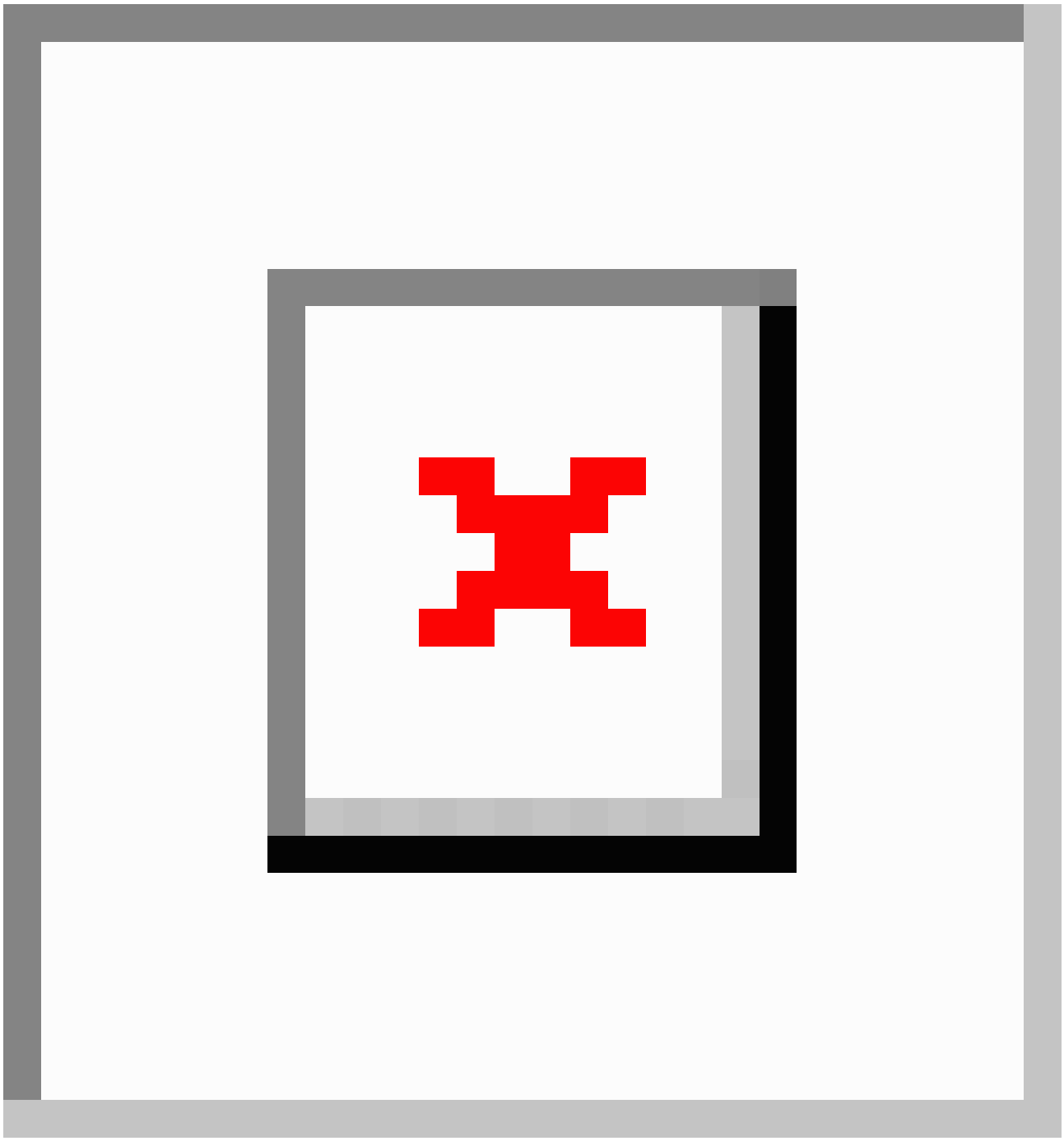
The initial visit questionnaire asks about detailed characteristics of the patient's headaches, utilizing multiple-choice questions with an option for free-text entry. To evaluate detailed headache patterns, patients were asked to provide information about their headache patterns, including location, nature, duration, frequency, recent exacerbation and its nature, frequent headache timing, and aggravating or alleviating factors. The accompanying aura and its types were also inquired to provide detailed information for migraine classification. Red-flag symptoms; family history of headache; limitation in daily activity; and previous headache diagnosis, evaluations, and treatments were also included in the questionnaire to aid clinicians in the detailed evaluation of patient's headache history and characteristics. The questionnaire ends with an open-ended question for patients to freely write down any additional concerns or questions for the initial visit. Headache severity is rated on the Numeric Pain Rating (NRS) scale or as mild, moderate, or severe for infants, as observed by their parents. A total of 35 previsit questions were developed for integration into the BEST-Survey system. The complete questionnaire created in Korean was translated into English. Both versions are shown in Multimedia Appendix 1.

Clinical Utilization of BEST-Survey in the Pediatric Headache Clinic

When the patient visited the pediatric headache clinic, trained medical staff introduced the BEST-Survey system in the waiting

room and provided education on completing the headache questionnaire using a tablet PC. Patients were asked to fill out the survey with the help of their parents when necessary. The completed questionnaire was automatically integrated into the EHR, in both a free format and a structured format for the clinicians to review. The completed questionnaire was further stored into a structured database for further data analysis and tracking of the patient's headache history. Neither the time it took for the patient to complete the survey nor the number of patients who refused to participate in the survey was recorded. Upon request for data collection, data retrieval was conducted automatically in a common data warehouse format, ensuring the exclusion of personal identifiers. Clinical information, including headache characteristics, age, sex, and other patient details, was retrieved in accordance with the preauthorized institutional review board (IRB) approval. An overview of the BEST-Survey system utilized in the pediatric headache clinic is summarized in Figure 1.

Figure 1. Overview of the BEST-Survey system. BEST-Survey: Bundang Hospital Electronic System for Total Care-Survey; EHR: electronic health record; HIS: hospital information system.



Comparison of Acquired Clinical Information Before and After the Implementation of BEST-Survey

We conducted a retrospective analysis on the health records of patients with headaches who visited the pediatric neurology outpatient clinic at Seoul National University Bundang Hospital. The analysis included patients who visited from March 2013 to March 2015—before the implementation of the BEST-Survey system ("Before e-system group")—and patients who visited from September 2015 to September 2017—after the implementation of the system ("After e-system group"). All new patients with pediatric headaches aged <18 years were recruited in the study, and patients were given the option to not to use the electronic questionnaire system. There was no lower age limit to the study. Patients and families who refused to use electronic questionnaire system were not included in the study. We collected information on 18 headache characteristics and associated symptoms, such as severity, onset, frequency, duration, location, characteristics, nausea, vomiting, dizziness, photophobia, phonophobia, visual aura, timing, sleep breakage, relieving factor, aggravating factor, aggravation by increased intracranial pressure, and family history in both groups. We then compared the answer rates and positive rates of 10 specific findings (onset, frequency, location, nature, duration, severity, frequent timing of the headache, factors of aggravation, associated symptoms, and visual aura) between the 2 groups to evaluate the clinical utility of BEST-Survey. The overall headache characteristics, including 10 previously mentioned, were further described to provide clinical overview our pediatric headache population group. Patients and parents were asked to complete the entire survey but were permitted to skip questions if necessary. Data analysis was processed using SPSS Statistics for Windows (version 27.0; IBM Corp). Descriptive statistics were used to express data as means, ranges, and percentages. The Mann-Whitney *U* test was used for nonparametric comparisons between the 2 groups. Statistical significance was set at $P<.05$.

Ethical Considerations

This study was approved by the IRB of Seoul National University Bundang Hospital (B-1805-468-106). All participants were deidentified by assigning them random codes instead of hospital numbers and personal identifiers. No compensation was provided to the participants. The waiver of informed consent was approved by the IRB, as the study utilized deidentified medical records.

Results

Comparison of Information Achieved Before and After the Implementation of BEST-Survey

This study included 365 patients in the before e-system group and 452 patients in the after e-system group. There were no differences in onset age and sex between the 2 groups: the mean age at the visit was 9.7 (95% CI 9.4-10.2) years in the before e-system group and 10.0 (95% CI 9.7-10.4) years in the after e-system group, and 52.9% (193/365) and 49.3% (223/452) of the patients in the before and after e-system groups were female, respectively. The clinical characteristics of headaches obtained from the 2 groups were further analyzed. The mean rate of data acquisition increased significantly ($P<.001$) from 54.6% (range 27/365, 7.4% to 311/365, 85.2%) to 99.3% (range 410/452, 90.7% to 452/452, 100%) after the implementation of the e-system. The data acquisition rate for the 6 cardinal clinical features of headaches (onset, location, duration, severity, nature, and frequency) increased from a range of 53.7% (196/365) to 85.2% (311/365) in the before e-system group to a range of 98.7% (446/452) to 100% (452/452) in the after e-system group. The 3 least obtained clinical features in the before e-system group were aggravating factor (27/365, 7.4%), aggravation by increased intracranial pressure (70/365, 19.2%), and relieving factors (110/365, 30.1%), while onset (311/365, 85.2%), vomiting (298/365, 81.6%), and duration (287/365, 78.6%) were the features with the highest level of data acquisition. All clinical features were obtained in over 90% (range 410/452, 90.7% to 452/452, 100%) of patients in the after e-system group. Timing (410/452, 90.7%), location (446/452, 98.7%), and severity (450/452, 99.6%) were the 3 least obtained clinical features after the implementation of the e-system (Table 1).

Table . Comparison of data acquisition rate between before and after introduction of the electronic questionnaire system (BEST-Survey[a]).

Clinical features	Before e-system (n=365), n (%)	After e-system (n=452), n (%)
Severity (NRS[b] score)	196 (53.7)	450 (99.6)
Onset	311 (85.2)	452 (100)
Frequency	285 (78.1)	452 (100)
Duration	287 (78.6)	451 (99.8)
Location	248 (67.9)	446 (98.7)
Characteristics	221 (60.5)	452 (100)
Nausea	276 (75.6)	452 (100)
Vomiting	298 (81.6)	452 (100)
Dizziness	263 (72.1)	452 (100)
Photophobia	174 (47.7)	452 (100)
Phonophobia	156 (42.7)	452 (100)
Visual aura	130 (35.6)	452 (100)
Timing	228 (50.9)	410 (90.7)
Sleep breakage	219 (60.0)	452 (100)
Relieving factors	110 (30.1)	452 (100)
Aggravating factors	27 (7.4)	452 (100)
Aggravation by IICP[c]	70 (19.2)	452 (100)
Family history	133 (36.4)	452 (100)

[a]BEST-Survey: Bundang Hospital Electronic System for Total Care-Survey.

[b]NRS: Numeral Rating Scale.

[c]IICP: increased intracranial pressure.

Clinical Characteristics of Headaches From Patients Who Visited Our Hospital After the Implementation of BEST-Survey

Detailed clinical headache characteristics of our pediatric patient group was acquired from the BEST-Survey system. In our patient group, most patients (280/452, 61.9%) had their onset of headache less than or equal to 1 year ago, and they most frequently complained of daily headaches (116/452, 25.6%). The temporal region was the most common location of headaches (261/703, 37.1%; unilateral: 111/703, 15.8% and bilateral 150/703, 21.3%), followed by the frontal head region (111/703, 15.8%). Pain characteristics were often pulsatile (188/861, 21.8%), pressing (140/861, 16.3%), or dull (159/861, 18.5%). More than half of patients (232/452, 51.3%) had headaches lasting less than 2 hours, with nausea and vomiting (231/1036, 22.3%) as the most commonly associated symptoms. Red-flag symptoms were reported by 316 (69.9%) out of 452 patients. Of the patients reporting red-flag symptoms, sleep breakage and improvement by vomiting accounted for 137 (26.5%) and 71 (13.7%) out of 517 red-flag symptoms, respectively. The clinical characteristics of the patients with headaches obtained from BEST-Survey are shown in Table 2.

Table . Clinical characteristics of headaches obtained from BEST-Survey[a]. Multiple answers were allowed in some questions.

Clinical characteristics	Value, n (%)
Headache onset (n=452 patients)	
Within 1 month	70 (15.5)
Within 3 months	77 (17)
Within 6 months	50 (11.1)
Within 1 year	83 (18.4)
Within 2 years	70 (15.5)
Within 3 years	44 (9.7)
Within 4 years	21 (4.6)
Within 4 years	13 (2.9)
>5 years	24 (5.3)
Headache frequency (n=452 patients)	
Daily	116 (25.6)
3 - 4 times/week	74 (16.3)
1 - 2 times/week	89 (19.6)
1 - 3 times/month	76 (16.8)
<1 time/month	29 (6.4)
Not definite	68 (15)
Headache nature (n=452 patients; multiple answers allowed: n=861 answers)	
Stabbing	26 (3)
Pressing	140 (16.3)
Tightening	151 (17.5)
Pulsatile	188 (21.8)
Aching	99 (11.5)
Dull	159 (18.5)
Not definite	98 (11.4)
Headache duration (n=452 patients)	
<30 minutes	76 (16.8)
30 minutes to 1 hour	89 (19.7)
1-2 hours	67 (14.8)
2–3 hours	52 (11.5)
3–4 hours	39 (8.6)
4–5 hours	11 (2.4)
5–6 hours	14 (3.1)
6–12 hours	24 (5.3)
12–24 hours	23 (5.1)
>24 hours	12 (2.7)
Unspecified	45 (10)
Headache location (n=452 patients; multiple answers allowed: n=703 answers)	
Whole	80 (11.4)
Bilateral temporal	150 (21.3)
Unilateral temporal	111 (15.8)
Vertex	74 (10.5)

Occipital	90 (12.8)
Frontal	111 (15.8)
Around eyes	59 (8.4)
Neck	21 (3)
Not definite	7 (1)
Associated symptoms (n=452 patients; multiple answers allowed: n=1036 answers)	
Nausea or vomiting	231 (22.3)
Irritability	183 (17.7)
Distraction	177 (17.1)
Transient amnesia	6 (0.6)
Sensitivity to lights or sounds	213 (20.6)
Cramping	25 (2.4)
Tearing or ptosis	56 (5.4)
Yawning	72 (6.9)
Increased urination	12 (1.2)
Diarrhea	14 (1.4)
Depression	47 (4.5)
Aura (n=452 patients; multiple answers allowed: n=405 answers)	
Vertigo	14 (3.5)
Dizziness	84 (20.7)
Ataxia	15 (3.7)
Altered consciousness	5 (1.2)
Motor aura	38 (9.4)
Language aura	9 (2.2)
Sensory aura	4 (1)
Auditory aura	105 (25.9)
Visual aura	113 (27.9)
Not definite	18 (4.4)
Red-flag symptoms (n=316 patients; multiple answers allowed: n=517 answers)	
Headache after hyperventilation	154 (29.8)
Sleep breakage	137 (26.5)
Improved by vomiting	71 (13.7)
Morning headache	155 (30)

[a]BEST-Survey: Bundang Hospital Electronic System for Total Care-Survey.

Discussion

Principal Findings in Comparison to Prior Works

In this study, we developed an electronic questionnaire system named BEST-Survey and utilized it to collect medical history from patients with pediatric headaches. This system, which includes a streamlined previsit electronic headache questionnaire, automatic integration into the EHR, and construction of a structured database for further analysis, was significantly effective in ensuring the completeness of collecting the key clinical features of pediatric headaches. Furthermore, the system allowed easy data extraction and analysis for the comprehensive clinical characterization of pediatric headaches. To the best of our knowledge, this is the first study describing an integrated electronic questionnaire system linked to an EHR for pediatric headache.

Traditional history taking involves interactive conversations with patients to establish rapport, assess communication skills, observe their condition, and collect relevant clinical information [22]. However, comprehensive history taking could be hindered by miscommunication between a patient or parent and their physician, as well as the limited time available. Inquiring all necessary questions and retrieving accurate answers is crucial for accurate diagnosis and appropriate management [23].

Omitting important questions can significantly impact diagnosis and treatment outcomes [24]. Moreover, the time-consuming EHR documentation process further hinders accurate history taking and medical care delivery.

Studies on ambulatory practices have shown that physicians spend 18% to 20% of their time on documentation or writing [25]. Another study revealed that 52.9% of their time is dedicated to direct clinical face time, while 37% is spent on EHR and desk work [26]. Taking medical histories from children poses unique challenges compared to adults. Limited expression of symptoms and reliance on parents or guardians for information [27], as well as the potential for distraction and poor cooperation from the children [28], require detailed explanations and additional time for history taking. To address the limitations to accurate history taking, our system offers several advantages. First, a predetermined questionnaire form ensures that essential data are gathered without omitting important questions. Second, utilizing the waiting time before a consultation gives the patients more time to report their symptoms. Third, automatic integration with the EHR saves efforts required for documentation. Additionally, the computerized and stored data facilitate further processing and analysis.

Strengths of the Streamlined Electronic Questionnaire System

Our system's advantage in capturing essential history items without omission was demonstrated through a comparison of data acquisition rates before and after its implementation. The findings revealed a significant increase in obtaining answers to each item after introducing our system. Prior to using the system, interviews alone resulted in less than 90% of patients providing essential headache diagnosis information such as onset time, frequency, duration, location, and characteristics. Furthermore, inquiries regarding factors worsening the condition, increased intracranial pressure, and relief had the lowest response rates. These results align with previous studies demonstrating the efficacy of digital questionnaires in enhancing data collection. In a systematic review examining the benefits of electronic patient-reported outcome measures, 10 (31%) out of 32 studies reported having missing or incomplete data. Among those 10 studies, 7 (70%) reported lower rates of missing data and more complete information with electronic methods compared to paper-based approaches [19]. Another study comparing histories acquired by physicians and a computer program directly interacting with patients found that the computer-acquired histories revealed 160 problems not documented by physicians. Conversely, there were 13 problems reported by physicians but not by the computer program, indicating the usefulness of computer programs in acquiring more comprehensive and detailed medical histories [29].

We also examined differences in the rate of positive findings for each history item. We observed variations in the rate of positive responses, with interview-based inquiries generally yielding higher positive rates, except for sleep breakage. The high positive rates of various clinical characteristics obtained through physician interviews may be due to discrepancies and inaccuracies in reflecting the actual phenomenon. For instance, in our study, the proportion of patients with photophobia was 41.4% in the before e-system group and 12.2% in the after e-system group. In previous studies of patients with pediatric headaches in South Korea, 43% were diagnosed with migraines, 35% had tension headaches, and 22% had other primary headaches [30]. Considering that approximately half of patients with migraines experience photophobia [31], the finding of 41.4% of patients with photophobia in our study is likely an overestimation even when considering the higher likelihood of patients with debilitating headaches seeking care at tertiary hospitals. Discrepancies between physician interview–based information and patient-reported data have been reported in previous studies. In an analysis of clinical interviews and computer-acquired history data among patients with chest pain, inconsistencies were observed in the collected data [32]. Both our study and previous studies on chest pain have consistently found higher levels of missing data when obtaining medical history through interviews compared to using e-questionnaires or computerized history taking. The primary reason for these discrepancies can be attributed to data incompleteness during the interview-based, history-taking process.

Another advantage of the proposed system in this study is that the computerization of data enables easy data processing, analysis, and storage for future use. This facilitates efficient data analysis and saves processing time. For instance, we conducted an analysis on the clinical characteristics of patients with headaches visiting our hospital, providing an overview of their features. In our study population, predominant headache characteristics included daily occurrence (25.7%), severity rating around an NRS score of 7 - 8 (40.9%), pulsatile nature (21.6%), onset between 6 months to 1 year (18.4%), duration less than 4 hours (71.5%), and involvement of the entire or bilateral temporal area of the head (37.1%). Several studies have evaluated headache characteristics in children attending headache outpatient clinics. In one study with a questionnaire collected from 437 pediatric patients referred to headache outpatient clinics, 5.9% had underlying organic diseases, while 94% had primary headaches [33]. The characteristics of patients with primary headaches analyzed in this study showed similarities to our study, including a significant proportion of patients with headache duration between 2 - 12 hours (68.8%), a high incidence of bilateral pain (63.9%), and a high frequency of severe intensity ratings (63.9%) and pulsatile features (27.1%). The incidence of accompanying symptoms such as nausea (49.1%), vomiting (37%), photophobia (27.8%), phonophobia (24.4%), and aura (13.5%) was similar or slightly higher compared to our study population. Another study of 194 pediatric patients with primary headaches presenting to an outpatient clinic in Jordan also showed similarities to the headache characteristics in our study. The main patient population experienced moderate to severe headaches (80.4%), daily occurrence (34.5%), duration of 0.5 - 4 hours (25.5%), and a higher proportion of bilateral pain (78.9%). The prevalence of accompanying symptoms, including nausea or vomiting (43.1%), dizziness (31%), and photophobia (38.8%), was similar or slightly higher than those in our study [34]. In this study, we included both primary and secondary headaches, whereas previous studies solely focused on primary headaches. However, due to the relatively small proportion of patients with secondary

headaches within the overall group of patients with headache patients, similar results were obtained.

We were also able to identify key considerations for implementing the EHR-integrated, e-questionnaire system in clinical practice. First, using plain language in the questions is crucial as they are asked without additional explanations. However, care must be taken to avoid excessive simplification, which can result in a higher false positive rate. Therefore, continuous updating of the questions is necessary to minimize discrepancies between physician-led history taking and patient-reported questionnaires. For example, the question "Is there weakness in the arms or legs?" is intended to assess symptoms of transient ischemic attacks, a key symptom of Moyamoya disease. However, patients often misinterpret this question as fatigue or general weakness. Thus, more detailed questions differentiating between fatigue and paralysis are necessary. Second, the system should serve as a supplement to, rather than a replacement for, physician-led history taking. Double-checking items marked positively by the patient through the questionnaire is necessary. Third, appropriate question styles should be considered for different age groups to compensate for different levels of understanding. Tailored questionnaires for different age groups are essential for the success of electronic questionnaire system in the clinical field.

Limitations and Future Directions

Our study has several limitations. First, we did not directly compare the history-taking methods used by physicians with the electronic system for the same patient simultaneously. Second, we did not assess the validity of the questionnaire. Third, we did not evaluate whether the introduction of the system resulted in clinically significant changes in diagnosis and management. Based on our limitations, we suggest the following future studies utilizing the system developed in this study. First, a study should be conducted to assess the system's validity by comparing data obtained from physician-led history taking and patient-completed questionnaires for the same individuals. Second, investigations are needed to examine the impact of the system on patient care and outcomes in real-world clinical settings. Third, the development of a clinical decision support system based on the electronic system should be explored. Fourth, the development of our survey was conducted internally and did not undergo validation by external reviewers. Due to this lack of validation, our survey maybe limited in detailed characterization for certain subtypes of headaches, such as tension-type headaches. Fifth, the lack of investigation into the time spent to complete the questionnaire limits us in comparing the total saved clinic time. Sixth, the lack of quantitative feedback from patients, families, and clinicians about the acceptability of the electronic questionnaire limits us in understanding the acceptability of the system. Finally, future studies could investigate the improved diagnostic rate of headache subtypes following system implementation and explore changes in headache burden after administering various medications. A streamlined questionnaire could enhance the completeness of phenotypic data, facilitating the diagnosis of complex, rare diseases and potentially shortening their diagnostic odyssey. Our proposed system could also be adapted for use in other disease contexts [35]. Such studies would help demonstrate the qualitative improvement of the information collected by our electronic questionnaire system.

Conclusion

In this study, we developed an electronic questionnaire system named BEST-Survey and utilized it to implement a patient-reported electronic questionnaire for patients with pediatric headaches. A streamlined, previsit electronic questionnaire that is automatically integrated to the EHR showed several strengths over traditional interview-based history taking. First, it ensures the completeness of essential history items. Second, it enables more accurate history taking, particularly in key clinical features that may be overestimated in traditional methods. Finally, the system facilitates easy data extraction, processing, and analysis, enabling detailed clinical characterization of the specific patient population of interest. While the electronic questionnaire system cannot replace the complex role of physician-led history taking, it can serves as a helpful adjunctive tool to improve patient care.

Data Availability

The datasets used and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

HK contributed to study conception and design. JYH, JC, AC, SY, and HYL contributed to system construction and data collection. JYH, JC, and HK contributed to the interpretation of results and drafting of the manuscript. All authors contributed to manuscript revision and approval.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of questions (in English and Korean).
[DOCX File, 35 KB - medinform_v12i1e54415_app1.docx]

References

1. Maffioletti E, Ferro F, Pucci E, et al. P035. headache prevalence and disability among Italian adolescents aged 11-15 years: a population cross-sectional study. J Headache Pain 2015 Dec;16(Suppl 1):A148. [doi: 10.1186/1129-2377-16-S1-A148] [Medline: 28132240]
2. Abu-Arafeh I, Razak S, Sivaraman B, Graham C. Prevalence of headache and migraine in children and adolescents: a systematic review of population-based studies. Dev Med Child Neurol 2010 Dec;52(12):1088-1097. [doi: 10.1111/j.1469-8749.2010.03793.x] [Medline: 20875042]
3. Wöber-Bingöl C. Epidemiology of migraine and headache in children and adolescents. Curr Pain Headache Rep 2013 Jun;17(6):341. [doi: 10.1007/s11916-013-0341-z] [Medline: 23700075]
4. Wander A, Meena AK, Choudhary PK, Peer S, Singh R. Pediatric headache: a comprehensive review. Ann Child Neurol 2024 Oct;32(4):207-218. [doi: 10.26815/acn.2024.00521]
5. Philipp J, Zeiler M, Wöber C, et al. Prevalence and burden of headache in children and adolescents in Austria - a nationwide study in a representative sample of pupils aged 10-18 years. J Headache Pain 2019 Nov 6;20(1):101. [doi: 10.1186/s10194-019-1050-8] [Medline: 31694547]
6. Son HJ, Jin JO, Lee KH. Impact of the COVID-19 pandemic on behavioral and emotional factors in pediatric patients with headache. Ann Child Neurol 2024 Jul;32(3):161-166. [doi: 10.26815/acn.2024.00486]
7. Conti R, Marta G, Wijers L, Barbi E, Poropat F. Red flags presented in children complaining of headache in paediatric emergency department. Children (Basel) 2023 Feb 13;10(2):366. [doi: 10.3390/children10020366] [Medline: 36832495]
8. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med 2013 Mar 25;173(6):418-425. [doi: 10.1001/jamainternmed.2013.2777] [Medline: 23440149]
9. Albahri AH, Abushibs AS, Abushibs NS. Barriers to effective communication between family physicians and patients in walk-in centre setting in Dubai: a cross-sectional survey. BMC Health Serv Res 2018 Aug 14;18(1):637. [doi: 10.1186/s12913-018-3457-3] [Medline: 30107799]
10. Russell MB, Rasmussen BK, Brennum J, Iversen HK, Jensen RA, Olesen J. Presentation of a new instrument: the Diagnostic Headache Diary. Cephalalgia 1992 Dec;12(6):369-374. [doi: 10.1111/j.1468-2982.1992.00369.x] [Medline: 1473140]
11. Lipton RB, Dodick D, Sadovsky R, et al. A self-administered screener for migraine in primary care: the ID Migraine validation study. Neurology 2003 Aug 12;61(3):375-382. [doi: 10.1212/01.WNL.0000078940.53438.83] [Medline: 12913201]
12. Láinez MJA, Domínguez M, Rejas J, et al. Development and validation of the Migraine Screen Questionnaire (MS-Q). Headache 2005;45(10):1328-1338. [doi: 10.1111/j.1526-4610.2005.00265.x] [Medline: 16324165]
13. Maizels M, Burchette R. Rapid and sensitive paradigm for screening patients with headache in primary care settings. Headache 2003 May;43(5):441-450. [doi: 10.1046/j.1526-4610.2003.03088.x] [Medline: 12752748]
14. Lee SJ, Kavanaugh A, Lenert L. Electronic and computer-generated patient questionnaires in standard care. Best Pract Res Clin Rheumatol 2007 Aug;21(4):637-647. [doi: 10.1016/j.berh.2007.02.001] [Medline: 17678825]
15. Richter JG, Becker A, Koch T, et al. Self-assessments of patients via tablet PC in routine patient care: comparison with standardised paper questionnaires. Ann Rheum Dis 2008 Dec;67(12):1739-1741. [doi: 10.1136/ard.2008.090209] [Medline: 18647853]
16. Abernethy AP, Herndon JEII, Wheeler JL, et al. Improving health care efficiency and quality using tablet personal computers to collect research-quality, patient-reported data. Health Serv Res 2008 Dec;43(6):1975-1991. [doi: 10.1111/j.1475-6773.2008.00887.x] [Medline: 18761678]
17. Fritz F, Balhorn S, Riek M, Breil B, Dugas M. Qualitative and quantitative evaluation of EHR-integrated mobile patient questionnaires regarding usability and cost-efficiency. Int J Med Inform 2012 May;81(5):303-313. [doi: 10.1016/j.ijmedinf.2011.12.008] [Medline: 22236957]
18. VanDenKerkhof EG, Goldstein DH, Blaine WC, Rimmer MJ. A comparison of paper with electronic patient-completed questionnaires in a preoperative clinic. Anesth Analg 2005 Oct;101(4):1075-1080. [doi: 10.1213/01.ane.0000168449.32159.7b] [Medline: 16192524]
19. Meirte J, Hellemans N, Anthonissen M, et al. Benefits and disadvantages of electronic patient-reported outcome measures: systematic review. JMIR Perioper Med 2020 Apr 3;3(1):e15588. [doi: 10.2196/15588] [Medline: 33393920]
20. Wurster F, Beckmann M, Cecon-Stabel N, et al. The implementation of an electronic medical record in a German hospital and the change in completeness of documentation: longitudinal document analysis. JMIR Med Inform 2024 Jan 19;12:e47761. [doi: 10.2196/47761] [Medline: 38241076]
21. Shapiro E, Ziegler R. Pediatric neuropsychology and pediatric neurology: Kenneth Swaiman's legacy. Pediatr Neurol 2021 Sep;122:122-124. [doi: 10.1016/j.pediatrneurol.2021.05.003] [Medline: 34294470]
22. Slack WV, Slack CW. Patient-computer dialogue. N Engl J Med 1972 Jun 15;286(24):1304-1309. [doi: 10.1056/NEJM197206152862408]
23. Termine C, Ozge A, Antonaci F, Natriashvili S, Guidetti V, Wöber-Bingöl C. Overview of diagnosis and management of paediatric headache. part II: therapeutic management. J Headache Pain 2011 Feb;12(1):25-34. [doi: 10.1007/s10194-010-0256-6] [Medline: 21170567]
24. Pappas Y, Anandan C, Liu J, Car J, Sheikh A, Majeed A. Computer-assisted history-taking systems (CAHTS) in health care: benefits, risks and potential for further development. Inform Prim Care 2011;19(3):155-160. [doi: 10.14236/jhi.v19i3.808] [Medline: 22688224]

25. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: a systematic review. Health Policy 2018 Aug;122(8):827-836. [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
26. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016 Dec 6;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
27. Bloom L. Talking, understanding, and thinking: developmental relationship between receptive and expressive language. In: Language Perspectives, Acquisition, Retardation and Intervention: University Park Press; 1974:285-311. [doi: 10.7916/D88S4VMC]
28. Higgins AT, Turnure JE. Distractibility and concentration of attention in children's development. Ch Dev 1984 Oct;55(5):1799-1810. [doi: 10.2307/1129927]
29. Zakim D, Braun N, Fritz P, Alscher MD. Underutilization of information and knowledge in everyday medical practice: evaluation of a computer-based solution. BMC Med Inform Decis Mak 2008 Nov 5;8:50. [doi: 10.1186/1472-6947-8-50] [Medline: 18983684]
30. Arruda MA, Guidetti V, Galli F, Albuquerque RCAP, Bigal ME. Primary headaches in childhood--a population-based study. Cephalalgia 2010 Sep;30(9):1056-1064. [doi: 10.1177/0333102409361214] [Medline: 20713556]
31. Kawatu N, Wa Somwe S, Ciccone O, et al. The prevalence of primary headache disorders in children and adolescents in Zambia: a schools-based study. J Headache Pain 2022 Sep 9;23(1):118. [doi: 10.1186/s10194-022-01477-x] [Medline: 36085007]
32. Zakim D, Brandberg H, El Amrani S, et al. Computerized history-taking improves data quality for clinical decision-making-comparison of EHR and computer-acquired history data in patients with chest pain. PLoS ONE 2021 Sep 27;16(9):e0257677. [doi: 10.1371/journal.pone.0257677] [Medline: 34570811]
33. Wöber-Bingöl C, Wöber C, Karwautz A, et al. Diagnosis of headache in childhood and adolescence: a study in 437 patients. Cephalalgia 1995 Feb;15(1):13-21. [doi: 10.1046/j.1468-2982.1995.1501013.x] [Medline: 7758092]
34. Al Momani M, Almomani BA, Masri AT. The clinical characteristics of primary headache and associated factors in children: a retrospective descriptive study. Ann Med Surg (Lond) 2021 May 2;65:102374. [doi: 10.1016/j.amsu.2021.102374] [Medline: 34026104]
35. Cho J, Joo YS, Yoon JG, et al. Characterizing families of pediatric patients with rare diseases and their diagnostic odysseys: a comprehensive survey analysis from a single tertiary center in Korea. Ann Child Neurol 2024 Jul;32(3):167-175. [doi: 10.26815/acn.2024.00472]

Abbreviation

BEST-Survey: Bundang Hospital Electronic System for Total Care-Survey
EHR: electronic health record
IRB: institutional review board
NRS: Numeric Pain Rating

Edited by C Lovis; submitted 08.11.23; peer-reviewed by A Omoloja, A Raggi; revised version received 11.10.24; accepted 14.10.24; published 08.11.24.

Please cite as:
Cho J, Han JY, Cho A, Yoo S, Lee HY, Kim H
Enhancing Clinical History Taking Through the Implementation of a Streamlined Electronic Questionnaire System at a Pediatric Headache Clinic: Development and Evaluation Study
JMIR Med Inform 2024;12:e54415
URL: https://medinform.jmir.org/2024/1/e54415
doi:10.2196/54415

Original Paper

The Implementation of an Electronic Medical Record in a German Hospital and the Change in Completeness of Documentation: Longitudinal Document Analysis

Florian Wurster[1], MSc; Marina Beckmann[1], DPhil; Natalia Cecon-Stabel[1], MSc; Kerstin Dittmer[1], MA; Till Jes Hansen[1], MSc; Julia Jaschke[2], MSc; Juliane Köberlein-Neu[2], Prof Dr; Mi-Ran Okumu[1], MA; Carsten Rusniok[1], MA; Holger Pfaff[1], Prof Dr; Ute Karbach[1], PD, Dr

[1]Chair of Quality Development and Evaluation in Rehabilitation, Institute of Medical Sociology, Health Services Research, and Rehabilitation Science, Faculty of Human Sciences & Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[2]Center for Health Economics and Health Services Research, University of Wuppertal, Wuppertal, Germany

Corresponding Author:
Florian Wurster, MSc
Chair of Quality Development and Evaluation in Rehabilitation, Institute of Medical Sociology, Health Services Research, and Rehabilitation Science, Faculty of Human Sciences & Faculty of Medicine and University Hospital Cologne, University of Cologne
Eupener Str. 129
Cologne, 50933
Germany
Phone: 49 22147897116
Email: florian.wurster@uni-koeln.de

Abstract

Background: Electronic medical records (EMR) are considered a key component of the health care system's digital transformation. The implementation of an EMR promises various improvements, for example, in the availability of information, coordination of care, or patient safety, and is required for big data analytics. To ensure those possibilities, the included documentation must be of high quality. In this matter, the most frequently described dimension of data quality is the completeness of documentation. In this regard, little is known about how and why the completeness of documentation might change after the implementation of an EMR.

Objective: This study aims to compare the completeness of documentation in paper-based medical records and EMRs and to discuss the possible impact of an EMR on the completeness of documentation.

Methods: A retrospective document analysis was conducted, comparing the completeness of paper-based medical records and EMRs. Data were collected before and after the implementation of an EMR on an orthopaedical ward in a German academic teaching hospital. The anonymized records represent all treated patients for a 3-week period each. Unpaired, 2-tailed *t* tests, chi-square tests, and relative risks were calculated to analyze and compare the mean completeness of the 2 record types in general and of 10 specific items in detail (blood pressure, body temperature, diagnosis, diet, excretions, height, pain, pulse, reanimation status, and weight). For this purpose, each of the 10 items received a dichotomous score of 1 if it was documented on the first day of patient care on the ward; otherwise, it was scored as 0.

Results: The analysis consisted of 180 medical records. The average completeness was 6.25 (SD 2.15) out of 10 in the paper-based medical record, significantly rising to an average of 7.13 (SD 2.01) in the EMR (t_{178}=–2.469; *P*=.01; *d*=–0.428). When looking at the significant changes of the 10 items in detail, the documentation of diet (*P*<.001), height (*P*<.001), and weight (*P*<.001) was more complete in the EMR, while the documentation of diagnosis (*P*<.001), excretions (*P*=.02), and pain (*P*=.008) was less complete in the EMR. The completeness remained unchanged for the documentation of pulse (*P*=.28), blood pressure (*P*=.47), body temperature (*P*=.497), and reanimation status (*P*=.73).

Conclusions: Implementing EMRs can influence the completeness of documentation, with a possible change in both increased and decreased completeness. However, the mechanisms that determine those changes are often neglected. There are mechanisms that might facilitate an improved completeness of documentation and could decrease or increase the staff's burden caused by

documentation tasks. Research is needed to take advantage of these mechanisms and use them for mutual profit in the interests of all stakeholders.

Trial Registration: German Clinical Trials Register DRKS00023343; https://drks.de/search/de/trial/DRKS00023343

(JMIR Med Inform 2024;12:e47761) doi:10.2196/47761

KEYWORDS

clinical documentation; digital transformation; document analysis; electronic medical record; EMR; Germany; health services research; hospital; implementation

Introduction

The digital transformation of the health care system is considered an essential subject to meet current and future societal challenges such as an aging population or rising health care expenditures while at the same time maintaining a high quality of care [1]. An important early step in hospitals' digitalization and a fundamental requirement for expanding digital maturity is the implementation of an electronic medical record (EMR) [2]. This EMR is considered to be an **"**electronic record of health care information of an individual that is created, gathered, managed, and consulted by authorized clinicians and staff within 1 health care organization" [3] and replaces the internal clinical documentation on preprinted paper-based charts. Studies show that the implementation of an EMR can lead to various improvements in the clinical context (eg, in the availability of information [4], coordination of care [5], or patient safety [6]). Moreover, the EMR facilitates the secondary usage of the documented data for research purposes through its digital accessibility [7]. To reach those benefits, it is indispensable that the EMR contain documentation that is of high quality. However, there are varying definitions regarding the quality of documentation. In that matter, the Institute of Medicine defined completeness, legibility, accuracy, and meaning as the main aspects of a medical record's data quality [8]. For those, the completeness of documentation was shown to be the most common dimension of data quality when empirically analyzing the documentation in EMRs [9], and it was highlighted to be especially important for secondary uses such as big data analyses [10].

Our recent systematic review also stated the completeness of documentation as the state of the art for the comparison of paper-based and EMRs [11]. This comparison is important since the implementation of an EMR and the associated transition from handwritten documentation to digital documentation can heavily affect the documentation subject since the transition offers the possibility to adjust which information has to be documented in which way [12]. For example, digitization enables the adoption of certain functionalities that can alter the completeness of documentation, like automatically transferring information from other digital devices to the EMR [13]. Moreover, when working with the EMR, information can be documented remotely, while the paper-based medical record had to be located and physically accessed first. In this matter, several studies conducted in the inpatient setting showed increased completeness in the EMR compared to the paper-based medical record, for example, for the documentation of signs and symptoms [13,14], weight and height, or malnutrition screening [15]. This suggests that the implementation of an EMR might lead to improvements in the completeness of documentation in general. It is therefore the main purpose of this study to evaluate the change in completeness due to the implementation of an EMR in an inpatient setting. Literature already provides proof of a change of completeness in regard to some specific documented information that is analyzed in this work (eg, the documentation of vital signs) [13,14]. Those empirical results might thus be validated for the presented work's specific setting and discipline. In addition, some of the information that is analyzed in this work is not described in literature yet (eg, the documentation of pain). It is examined for the first time with regard to changes in completeness after the implementation of an EMR.

The knowledge gained can not only support the implementation of new EMRs but could also help understand and optimize arising changes in documentation when existing EMRs need to be adapted [16,17]. This is an important aspect, as the implementation of new EMRs is described as one of the most important interventions to improve the quality of documentation [18]. In this process, mechanisms affecting the completeness of documentation in medical records are not completely understood [10]. On the other hand, this knowledge is needed to fulfill reported educational needs regarding how to reach the optimum quality of documentation [19]. In this context, this study contributes to a more comprehensive understanding of the impact of an EMR on the quality of documentation.

Methods

Overview

This study follows the "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE) statement [20] whenever it is applicable. It offers reporting standards to ensure the reporting of any important information in empirical research studies. A checklist with details, where the STROBE information is mentioned in the manuscript, can be found in Multimedia Appendix 1.

Ethical Considerations

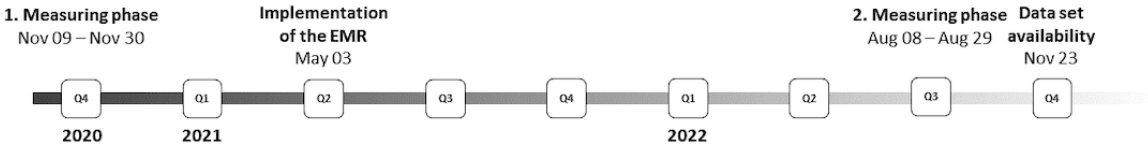
The study has been approved by the ethics committee of the Medical Faculty of the University of Cologne, Germany (20-1349). All data was anonymized at all times during the scientific analysis. No compensation was paid.

Setting and Participants

The study took place as part of the research project eCoCo, which Beckmann et al [21] described in detail. Within the eCoCo project, the researchers collected various types of data

(observations, surveys, interviews, documents, and administrative data) to investigate a possible change in interprofessional collaboration and clinical workflows following the implementation of an inpatient EMR. This study is part of the related work package on documentation content and quality, which took place in a large academic teaching hospital in Germany. The hospital replaces its internal documentation on preprinted paper-based charts with a commercial EMR system (Meona; Mesalvo Freiburg GmbH). The EMR runs on multiple computers that can be moved flexibly over the ward on trolleys. The study follows a pre-post design, retrospectively analyzing the content of the medical records before and after the implementation of the EMR on the hospitals' orthopaedical ward. Within the first measuring phase, the paper-based medical records were provided as a digital copy of the paper sheets. Those paper-based records represent all patients who were treated on the ward during the last 3 weeks in November 2020 (t0). After 6 months, employees received training on how to use the EMR before the implementation of the EMR took place in May 2021. The EMRs were again provided as a digital copy within a second measuring phase, representing all patients who were treated on the same ward during the first 3 weeks of August 2022 (t1). This resulted in a gap of 15 months between the first and second measuring phases. The complete data set was available to the research team in November 2022 (Figure 1). The hospital provided anonymized medical records to the research team after the records were archived and cleared of sensitive personal data (eg, the patient's name or date of birth) in the hospital's internal processes. Any assignment of the patient data or linking of the records' contents to any individual patient was therefore impossible for the research team, which is, thus, in compliance with the European Union General Data Protection Regulation. This also implies the absence of sociodemographic information for describing the compared samples. The hospital's mandatory annual quality report, which is available to the public through a designated database [22], is therefore used to describe the ward's patient sample and the performed treatments in general. This allows an approximation to a description and comparison of the compared samples in terms of their *International Classification of Diseases* (*ICD*)–diagnoses distribution.

Figure 1. Data collection. EMR: electronic medical record.

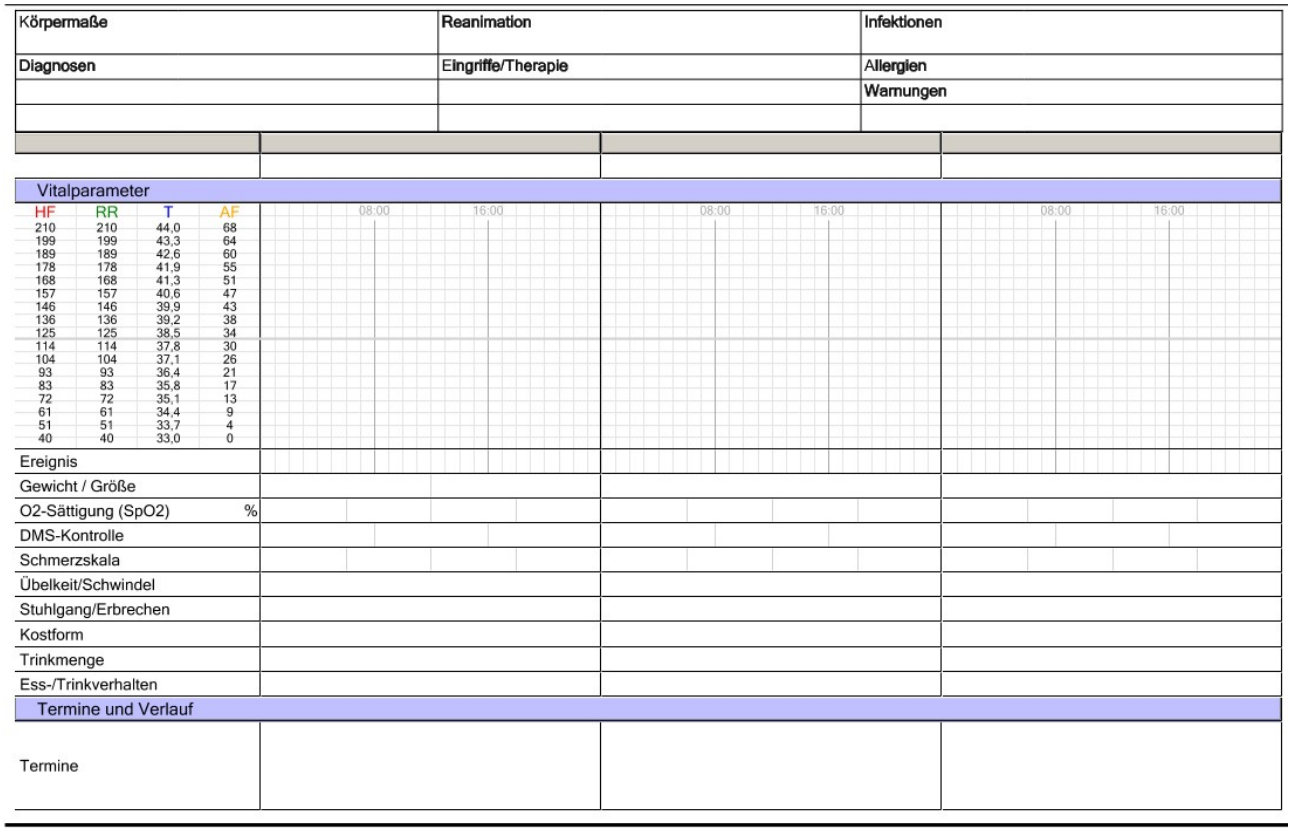


Study Objective

To answer the question of a possible change in completeness, the records were analyzed by content [23]. The change from paper-based documentation to EMRs always offers the opportunity to fundamentally change the structure of the records. This was shown exemplarily by Montagna et al [12] when the documentation as a continuously written text in the paper-based record was changed to a list of events in the EMR. It is therefore important to ensure the comparability between the 2 record types for the purpose of analyzing a possible change in their completeness. To achieve comparability, the medical records progress note was selected as a specific object of interest for this study's analyses since it retained the same structure and format in both record types. Part of this progress note is the fever chart (Figures 2 and 3), which includes basic details about vital signs, personal health data, etc [24].

Figure 2. Paper-based fever chart.

Figure 3. Electronic fever chart.



All information that was commonly documented in both of the 2 record types (paper-based and electronic) became part of this work. Weiskopf and Weng [9] described this selection mechanism for assessing data quality based on the parallels between the EMR and the paper-based record. This procedure resulted in a total of 10 key items that were analyzed for completeness in this work: blood pressure, body temperature, diagnosis, diet, excretions, height, pain, pulse, reanimation

status, and weight. The documentation of this information is equally possible and performed by nurses and physicians. However, there is no information available about who specifically entered the information.

All of those items should be documented immediately when patient care begins on the ward [25]. However, while the documentation of vital signs can take place up to several times a day, the documentation of the patient's diet usually occurs once a day, and the documentation of the reanimation status (patient's preference regarding a possible resuscitation) is probably documented only once per hospital stay. Because of these varying documentation practices and to ensure comparability, the analysis focuses on certain documentation in the progress notes that was entered on the first day of patient care on the ward. With regard to the documentation of a diagnosis, it is therefore the diagnosis with which a patient is admitted to the hospital. This diagnosis is mainly responsible for the allocation to specific medical specialties as well as a certain ward and does not necessarily have to match the final diagnosis at the time of discharge, which is important for reimbursement purposes.

Statistical Analysis of Completeness

For every record, each of the 10 items received a dichotomous score of 1 if it was documented on the first day of patient care on the ward; otherwise, it was scored as 0. This resulted in a percentage of completeness for each item per record type. Chi-square tests for independence were used to assess statistically significant differences in the percentage of completeness per item between the 2 record types. Relative risks were calculated for the association between the electronic record type and a possible increase in completeness. To improve the reliability of the associated confidence intervals, they were calculated with 5000 bootstrap replications since the original sample sizes are unbalanced. Moreover, the overall completeness was assessed as sum of the 10 items, resulting in a mean score of completeness per record type ranging from 0 (no item documented) to 10 (all 10 items documented). Those mean scores of completeness per record type were analyzed for equality of variance and statistical difference using unpaired, 2-tailed *t* tests. Assumptions were checked using several methods (normal distribution: QQ plots and Shapiro-Wilk test; homogeneity of variances: Levene test; and linearity: scatter plot). The level of significance was set to be $P<.05$ for all calculations. The data were stored in Microsoft Excel (Microsoft Corp) and analyzed in December 2022 using SPSS software (version 29; IBM Corp).

Results

Participants

During the first measuring phase (November 2020), a total of 44 patients (paper-based) were treated on the orthopaedical ward. They were encountering a total of 136 treated patients (electronic) during the second measuring phase (August 2022). This resulted in a total of 180 medical records that became part of this analysis. Due to the data protection regulation and the accompanied anonymization of the records data, there is no information regarding the demographics of the specific study population. Therefore, the ward's ICD-diagnosis distribution is given as an approximation of a sample description. In 2020, the 3 most frequently coded diagnoses for the orthopaedical ward were complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84), dorsalgia (ICD-M54), and fracture of shoulder and upper arm (ICD-S42). This report is not yet published for 2022, but the top 3 treated diagnoses in 2019 or 2021 were similar to those in 2020 (Table 1). It can therefore be expected that the treated diagnoses will be similar in 2022, too. Another supporting fact is that the most frequently performed procedure (surgical access to the lumbar spine, the sacrum, or the coccyx [coded as OPS-5-032 in the German adaptation of the International Classification of Procedures in Medicine which is part of the coding system for hospitals reimbursement]) was the same in all 3 years (2019-2021).

Table 1. Most frequently coded diagnoses.

ICD[a] Code	Values, n[b]/N[c] (%)
2019	
Dorsalgia (ICD-M54)	213/3147 (6.77)
Other spondylopathies (ICD-M48)	133/3147 (4.23)
Fracture of forearm (ICD-S52)	131/3147 (4.16)
2020	
Complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84)	166/2912 (5.7)
Dorsalgia (ICD-M54)	148/2912 (5.08)
Fracture of shoulder and upper arm (ICD-S42)	121/2912 (4.16)
2021	
Complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84)	164/3091 (5.3)
Dorsalgia (ICD-M54)	163/3091 (5.27)
Fracture of forearm (ICD-S52)	159/3091 (5.14)

[a]ICD: International Classification of Diseases.

[b]Frequency of coded diagnosis.

[c]Total inpatient cases.

Change of Completeness

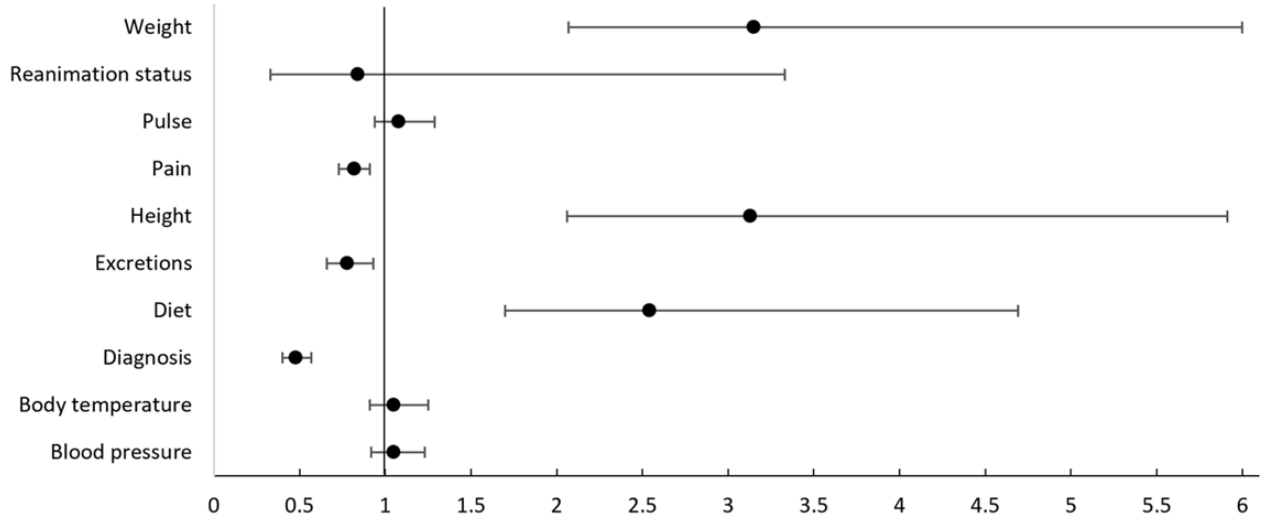
The mean number of documented items was 6.25 (SD 2.15) out of 10 in paper-based medical records and 7.13 (SD 2.01) out of 10 in EMRs. The Levene test confirmed the homogeneity of variances. The Shapiro-Wilk test did not confirm normal distributions, but the QQ plots show an approximation to a normal distribution and a comparable degree of normality (Multimedia Appendix 2). The unpaired *t* test confirmed the EMRs were statistically significantly more complete than the paper-based medical records under equal variances in the 2 record types (t_{178}=–2.469; *P*=.01; *d*=–0.428). When looking at the 10 items separately, data from chi-square tests showed that the documentation of diet increased from being present in 30% (13/44) of the paper-based medical record to 75% (102/136; *P*<.001) in the EMR, height from 27% (12/44) to 85.3% (116/136; *P*<.001), and weight from 27% (12/44) to 86% (117/136; *P*<.001). At the same time, documentation of diagnosis decreased from being present in 100% (44/44) of the paper-based medical records to 49% (66/136; *P*<.001) in the EMR, excretions from 86% (38/44) to 68% (92/136; *P*=.02), and pain from 95% (42/44) to 78% (106/136; *P*=.008). The documentation of vital signs such as blood pressure (*P*=.47), body temperature (*P*=.497), and pulse (*P*=.28) remained unchanged on a high level of completeness, while the documentation of reanimation status (*P*=.73) remained unchanged on a low level of completeness (Table 2). Positive relative risks (Figure 4) illustrate the association of the electronic record type (exposure) with complete documentation (outcome). The confidence intervals represent 5000 bootstrap replications.

Table 2. Change of completeness.

Variable	Type of record		Chi-square (*df*)	*P* value	RR[a] (95% CI)
	Paper (n=44), n (%)	Electronic (n=136), n (%)			
Blood pressure	37 (84)	120 (88.2)	0.5 (1)	.47	1.05 (0.92-1.23)
Body temperature	36 (81.8)	117 (86)	0.5 (1)	.497	1.05 (0.91-1.25)
Diagnosis	44 (100)	66 (48.5)	37.1 (1)	<.001	0.48 (0.40-0.57)
Diet	13 (29.6)	102 (75)	29.8 (1)	<.001	2.54 (1.70-4.69)
Excretions	38 (86.4)	92 (67.7)	5.8 (1)	.02	0.78 (0.66-0.93)
Height	12 (27.3)	116 (85.3)	54.5 (1)	<.001	3.13 (2.06-5.91)
Pain	42 (95.4)	106 (77.9)	7.0 (1)	.008	0.82 (0.73-0.91)
Pulse	36 (81.8)	120 (88.2)	1.2 (1)	.28	1.08 (0.94-1.29)
Reanimation status	5 (11.4)	13 (9.6)	0.1 (1)	.73	0.84 (0.33-3.33)
Weight	12 (27.3)	117 (86)	56.5 (1)	<.001	3.15 (2.07-6.00)

[a]RR: relative risk.

Figure 4. Forest plot of relative risks.



Discussion

Principal Findings and Comparison to Previous Work

The main findings of this study confirm an improved completeness of the analyzed information in the EMR on average. This provides further evidence for the suggestion that the general completeness of documentation can improve after the implementation of an EMR. The findings align with the results of similar studies, showing improvements in other data quality dimensions like the accuracy [26] or legibility [27] of documentation. However, when looking at the completeness of the analyzed 10 items in detail, the improvements can only be seen in 3 out of 10 items (diet, height, and weight), while 3 different items exhibited a deterioration in completeness (diagnosis, excretions, and pain). This links to the results of Coffey et al [28], who found 5 of their 11 analyzed items to be more complete while also proving 1 of their elements to be less complete. The reason for the variation in the change in completeness may lie in the mechanism of how information reaches the record. In the paper-based medical records, all information was documented by hand by the various professional groups. EMRs, on the other hand, offer technical features, for example, automatically obtaining information from other digital sources, like patients' health insurance data [29]. This was manifested as a possible mechanism by Jang et al [30], who showed improved completeness in the EMR for the automatically filled information but not for the manually documented ones.

The analysis shows that roughly every second EMR was missing the documentation of a diagnosis. This is a remarkable change, as it was present in every paper-based record (44/44, 100% vs 66/136, 48.5%). In the first place, it must be clarified that the diagnosis is determined by a physician who enters it into an independently run hospital information system (HIS). This documented diagnosis can also be a preliminary diagnosis, which is used for distribution to the clinical disciplines and is present for every admitted patient. The HIS was already in operation when medical staff was still using the paper-based preprints for documentation purposes. After the EMR's implementation, the HIS was still in operation along with the EMR. That being said, it is undisputed that during the paper-based period as well as the electronic period, a diagnosis was indeed present for the patients. In the paper-based period, the diagnosis was transferred manually from the HIS into the paper-based preprints, when a record for a recently admitted patient was prepared by a nurse. Since the HIS and the EMR are produced by different software developers, the diagnosis cannot be transferred automatically from the HIS into the EMR. Due to this noninteroperability of the 2 independent digital systems, the manual transfer is still necessary in the electronic period. With the drop of completeness in mind, this double documentation was accepted and carried out in the period of the paper-based record. In the electronic period, the described double documentation has decreased. One possible explanation is that the HIS was not automatically accessible, when an employee had the paper-based record at hand. With the introduction of the EMR, the availability of the EMR became synonymous with the availability of the HIS, since both are accessible from a computer. Therefore, the transfer of the diagnosis from the HIS to the EMR may no longer have been considered necessary. Nevertheless, the reason for this difference remaining unclear illustrates that the sole analyzation of completeness of the documentation alone does not provide sufficient information about the actual quality of the provided treatment. In that matter, it must also be highlighted that the record can contain additional qualitative data entries, like free texts, which might complement the analyzed quantitative information. This underlines that an insufficient quality of documentation does not necessarily allow conclusions to be drawn about the quality of care, and vice versa.

Brown [31] emphasizes this by cautioning people to always consider the circumstances under which people put information into the record before drawing conclusions. This is a major issue because the completeness of documentation might be biased due to aspects that do not directly derive from clinical care. On the one hand, the hospital's reimbursement for the delivered care depends on what is documented and might cause a possible strengthened thorough filling of certain fields [32]. On the other hand, the burden caused by documentation tasks is critically

heavy. It is responsible for a high prevalence of burnout among physicians and nurses [33]. Therefore, clinically or legally unnecessary documentation might be evaded [34]. However, even though complete documentation might neither necessarily arise from nor be essential for the delivery of excellent clinical care, it is likewise of concern under the aspect of big data analytics. In this regard, it would be desirable for the discussed diagnosis to indeed be present in the EMR, even if it already exists in the HIS. An automatic transfer of this information could help to prevent the burden on staff resulting from manual transmission and ensure a complete data set. This is an important point, as the insights gained from analyzing big data offer numerous opportunities, like data-based personalized care in diagnostics and therapy or the support of scientific activities, both with the chance of saving lives and reducing health care costs at the same time [7,35]. It is therefore indispensable to recognize the possibility of changes in documentation due to the implementation or adaptation of EMRs. Only with this attention will it become possible to optimize the documentation process with a focus on the various benefits for all stakeholders, like patients [6], practitioners [36], organizations [5], and society [7].

Strengths and Limitations

The German health care system, in which the study was conducted, was heavily strained by the high number of COVID-19 cases and the associated use of intensive care units during the study period. Especially the first measuring phase (November 2020) fell into the first pandemic year when many planned procedures were suspended to increase hospital capacities. For the first lockdown period in Germany (March 2020), a decrease in orthopedic surgeries is described by approximately 80% [37]. A lockdown-like situation was again declared during the first measuring phase [38], which probably explains the difference in treated patients over the 2 measuring phases (n_{Paper}=44 vs $n_{Electronic}$=136). However, the similarity between the coded ICD diagnoses over different years (Table 1) suggests that the proven changes in completeness of documentation are not due to significant changes in the studied patient sample, but a detailed sample description based on socioeconomical data is missing due to data protection regulations. On the other hand, there is a study assuming a positive influence of the pandemic on the completeness of documentation since an incomplete documentation might have led to repetitive contacts with the patient, which could have been avoided if the documentation would have been complete in the first place [39]. However, this cannot be verified in this paper due to the lack of further measuring phases. Within this given context, the generalizability of the presented results remains limited.

Further, limitations regarding the analyzed data set have to be stated. The chosen unpaired *t* test is theoretically based on the assumption of normal distributions. This could not be confirmed statistically for the mean completeness scores by the Shapiro-Wilk test. Although *t* test has been shown to be robust to a missing normal distribution [40] and the QQ plots (Multimedia Appendix 2) indicate an approximation to a normal distribution, the results could still be biased by the broken assumption.

Moreover, the analyzed data set is missing any information on which person was entering the documentation regarding which patient. On the one hand, it might be arguable that the same physicians or nurses were documenting during the first and also the second measuring phases. This circumstance would make the 2 compared measuring phases dependent samples, having an impact on the chosen statistical model. Since the analyzed data set is missing this information, the results might be biased regarding a possible dependent or independent sample. However, the time passed between the 2 measuring phases might have led to a change of the employees since the teaching status of the hospital results in many young physicians or nurses who do not necessarily stay on the same ward for a long time. Moreover, the hospital in which the study was conducted has a rotation system in which clinicians rotate hospital-wide across different wards of the same discipline. Those 2 facts let us assume that the 2 compared samples are indeed independent. However, the lack of information regarding the documenting individual is preventing the use of advanced tests like mixed effect models. These could equally consider the record type on the one hand and the possible documenting individuals on the other hand, potentially advancing the results' reliability. However, the 15-month interval from the implementation date of the EMR to the second data collection signifies that there is only little risk of any possible changes in documentation due to a bias from the described effects of preimplementational documentation training [41] since the employees indeed underwent software training before they were allowed to use the EMR. Therefore, the shown changes in completeness are, in fact, most likely due to the implementation of the EMR.

Conclusions

The results show that implementing EMRs can influence the completeness of documentation. A demonstrated improved completeness might also facilitate an improvement of the described outcomes that depend on documentation that is of high quality, like the availability [4] and analyzability of information [7,35], the coordination of care [5], or patient safety [6]. However, at the same time, the results show that a deterioration of completeness is also conceivable with the accompanied risks. This highlights the importance of understanding the underlying mechanisms that determine these changes. The knowledge may help stakeholders manage the implementation of new EMRs or the optimization of existing EMRs. Future research should address mechanisms that can improve documentation while simultaneously reducing the burden on practitioners caused by documentation tasks.

Acknowledgments

This work is funded by the German Federal Ministry of Education and Research (grant 01GP1906B). The sponsor had no influence on study design, data collection, analysis, or the writing process.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

FW and UK conceptualized the article. Data collection and analysis were performed by FW, supervised by UK. The original draft of the manuscript was written by FW, and all authors reviewed and edited previous versions of the manuscript and contributed to the interpretation of the data. All authors read and approved the final manuscript.

Conflicts of Interest

None Declared.

Multimedia Appendix 1
STROBE Checklist.
[PDF File (Adobe PDF File), 150 KB - medinform_v12i1e47761_app1.pdf]

Multimedia Appendix 2
Q-Q-Plots.
[PDF File (Adobe PDF File), 86 KB - medinform_v12i1e47761_app2.pdf]

References

1. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. Clin Chem Lab Med 2019;57(3):328-335 [FREE Full text] [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]
2. Mangiapane M, Bender M. EMR Adoption Model (EMRAM). In: Mangiapane M, Bender M, editors. Patientenorientierte Digitalisierung im Krankenhaus. Wiesbaden: Springer Vieweg; 2020:33-39.
3. Jacob PD. Chapter 3 - Management of patient healthcare information: healthcare-related information flow, access, and availability. In: Gogia S, Novaes M, Basu A, Gogia K, Gogia S, editors. Fundamentals of Telemedicine and Telehealth. London: Academic Press; 2020:35-57.
4. Embi PJ, Weir C, Efthimiadis EN, Thielke SM, Hedeen AN, Hammond KW. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. J Am Med Inform Assoc 2013;20(4):718-726 [FREE Full text] [doi: 10.1136/amiajnl-2012-000946] [Medline: 23355462]
5. Vos JFJ, Boonstra A, Kooistra A, Seelen M, van Offenbeek M. The influence of electronic health record use on collaboration among medical specialties. BMC Health Serv Res 2020;20(1):676 [FREE Full text] [doi: 10.1186/s12913-020-05542-6] [Medline: 32698807]
6. Yanamadala S, Morrison D, Curtin C, McDonald K, Hernandez-Boussard T. Electronic health records and quality of care: an observational study modeling impact on mortality, readmissions, and complications. Medicine (Baltimore) 2016;95(19):e3332 [FREE Full text] [doi: 10.1097/MD.0000000000003332] [Medline: 27175631]
7. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 [FREE Full text] [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
8. Dick RS, Steen EB, Detmer DE. The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition. Washington: National Academies Press; 1997.
9. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
10. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
11. Wurster F, Fütterer G, Beckmann M, Dittmer K, Jaschke J, Köberlein-Neu J, et al. The analyzation of change in documentation due to the introduction of electronic patient records in hospitals: a systematic review. J Med Syst 2022;46(8):54 [FREE Full text] [doi: 10.1007/s10916-022-01840-0] [Medline: 35781136]

12. Montagna S, Croatti A, Ricci A, Agnoletti V, Albarello V, Gamberini E. Real-time tracking and documentation in trauma management. Health Informatics J 2020;26(1):328-341 [FREE Full text] [doi: 10.1177/1460458219825507] [Medline: 30726161]
13. Zargaran E, Spence R, Adolph L, Nicol A, Schuurman N, Navsaria P, et al. Association between real-time electronic injury surveillance applications and clinical documentation and data acquisition in a South African trauma center. JAMA Surg 2018;153(5):e180087 [FREE Full text] [doi: 10.1001/jamasurg.2018.0087] [Medline: 29541765]
14. Thoroddsen A, Ehnfors M, Ehrenberg A. Content and completeness of care plans after implementation of standardized nursing terminologies and computerized records. Comput Inform Nurs 2011;29(10):599-607 [FREE Full text] [doi: 10.1097/NCN.0b013e3182148c31] [Medline: 22041791]
15. McCamley J, Vivanti A, Edirippulige S. Dietetics in the digital age: The impact of an electronic medical record on a tertiary hospital dietetic department. Nutr Diet 2019;76(4):480-485 [FREE Full text] [doi: 10.1111/1747-0080.12552] [Medline: 31199071]
16. Karp EL, Freeman R, Simpson KN, Simpson AN. Changes in efficiency and quality of nursing electronic health record documentation after implementation of an admission patient history essential data set. Comput Inform Nurs 2019;37(5):260-265 [FREE Full text] [doi: 10.1097/CIN.0000000000000516] [Medline: 31094915]
17. Meier-Diedrich E, Davidge G, Hägglund M, Kharko A, Lyckblad C, McMillan B, et al. Changes in documentation due to patient access to electronic health records: protocol for a scoping review. JMIR Res Protoc 2023;12:e46722 [FREE Full text] [doi: 10.2196/46722] [Medline: 37639298]
18. Wiebe N, Varela LO, Niven DJ, Ronksley PE, Iragorri N, Quan H. Evaluation of interventions to improve inpatient hospital documentation within electronic health records: a systematic review. J Am Med Inform Assoc 2019;26(11):1389-1400 [FREE Full text] [doi: 10.1093/jamia/ocz081] [Medline: 31365092]
19. Emekli E, Coscun Ö, Budakoglu I, Kiyak YS. Clinical record keeping education needs in a medical school and the quality of clinical documentations. Konuralp Med J 2023;15(2):257-265 [FREE Full text] [doi: 10.18521/ktd.1259969]
20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
21. Beckmann M, Dittmer K, Jaschke J, Karbach U, Köberlein-Neu J, Nocon M, et al. Electronic patient record and its effects on social aspects of interprofessional collaboration and clinical workflows in hospitals (eCoCo): a mixed methods study protocol. BMC Health Serv Res 2021;21(1):377 [FREE Full text] [doi: 10.1186/s12913-021-06377-5] [Medline: 33892703]
22. Referenzdatenbank der Qualitätsberichte der Krankenhäuser. Gemeinsamer Bundesausschuss. 2023. URL: https://qb-referenzdatenbank.g-ba.de/ [accessed 2023-12-22]
23. Prior L. Repositioning documents in social research. Sociology 2008;42(5):821-836 [FREE Full text] [doi: 10.1177/0038038508094564]
24. Ranegger R, Hackl WO, Ammenwerth E. Implementation of the Austrian Nursing Minimum Data Set (NMDS-AT): a feasibility study. BMC Med Inform Decis Mak 2015;15:75 [FREE Full text] [doi: 10.1186/s12911-015-0198-7] [Medline: 26384111]
25. Toney-Butler TJ, Unison-Pace WJ. Nursing Admission Assessment and Examination. Treasure Island (FL): StatPearls Publishing; 2018.
26. Yadav S, Kazanji N, Narayan KC, Paudel S, Falatko J, Shoichet S, et al. Comparison of accuracy of physical examination findings in initial progress notes between paper charts and a newly implemented electronic health record. J Am Med Inform Assoc 2017;24(1):140-144 [FREE Full text] [doi: 10.1093/jamia/ocw067] [Medline: 27357831]
27. Muallem YA, Dogether MA, Househ M, Saddik B. Auditing the completeness and legibility of computerized radiological request forms. J Med Syst 2017;41(12):199 [FREE Full text] [doi: 10.1007/s10916-017-0826-0] [Medline: 29101478]
28. Coffey C, Wurster LA, Groner J, Hoffman J, Hendren V, Nuss K, et al. A comparison of paper documentation to electronic documentation for trauma resuscitations at a level i pediatric trauma center. J Emerg Nurs 2015;41(1):52-56 [FREE Full text] [doi: 10.1016/j.jen.2014.04.010] [Medline: 24996509]
29. Seroussi B, Bouaud J. The (Re)-Relaunching of the DMP, the French shared medical record: new features to improve uptake and use. In: Ugon A, Karlsson D, Klein GO, editors. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created EHealth. Amsterdam: IOS Press; 2018:256-260.
30. Jang J, Yu SH, Kim CB, Moon Y, Kim S. The effects of an electronic medical record on the completeness of documentation in the anesthesia record. Int J Med Inform 2013;82(8):702-707 [FREE Full text] [doi: 10.1016/j.ijmedinf.2013.04.004] [Medline: 23731825]
31. Brown ML. Can't you just pull the data? The limitations of using of the electronic medical record for research. Paediatr Anaesth 2016;26(11):1034-1035 [FREE Full text] [doi: 10.1111/pan.12951] [Medline: 27747978]
32. Pruitt Z, Pracht E. Upcoding emergency admissions for non-life-threatening injuries to children. Am J Manag Care 2013;19(11):917-924 [FREE Full text] [Medline: 24511988]
33. Gesner E, Gazarian P, Dykes P. The burden and burnout in documenting patient care: an integrative literature review. Stud Health Technol Inform 2019;264:1194-1198. [doi: 10.3233/SHTI190415] [Medline: 31438114]

34. Saravi BM, Asgari Z, Siamian H, Farahabadi EB, Gorji AH, Motamed N, et al. Documentation of medical records in Hospitals of Mazandaran University of medical sciences in 2014: a quantitative study. Acta Inform Med 2016;24(3):202-206 [FREE Full text] [doi: 10.5455/aim.2016.24.202-206] [Medline: 27482136]
35. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data 2022;9(1):3 [FREE Full text] [doi: 10.1186/s40537-021-00553-4] [Medline: 35013701]
36. Ommaya AK, Cipriano PF, Hoyt DB, Horvath KA, Tang P, Paz HL, et al. Care-Centered clinical documentation in the digital environment: solutions to alleviate burnout. National Academy of Medicine. 2018. URL: https://nam.edu/care-centered-clinical-documentation-digital-environment-solutions-alleviate-burnout/ [accessed 2023-12-22]
37. Kapsner LA, Kampf MO, Seuchter SA, Gruendner J, Gulden C, Mate S, et al. Reduced rate of inpatient hospital admissions in 18 German University Hospitals during the COVID-19 lockdown. Front Public Health 2020;8:594117 [FREE Full text] [doi: 10.3389/fpubh.2020.594117] [Medline: 33520914]
38. Videokonferenz der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder am 28. Oktober 2020. Presse- und Informationsamt der Bundesregierung. 2020. URL: https://www.bundesregierung.de/resource/blob/997532/1805024/5353edede6c0125ebe5b5166504dfd79/2020-10-28-mpk-beschluss-corona-data.pdf?download=1 [accessed 2023-12-22]
39. Curtis CA, Nguyen MU, Rathnasekara GK, Manderson RJ, Chong MY, Malawaraarachchi JK, et al. Impact of electronic medical records and COVID-19 on adult goals-of-care document completion and revision in hospitalised general medicine patients. Intern Med J 2022;52(5):755-762 [FREE Full text] [doi: 10.1111/imj.15543] [Medline: 34580964]
40. Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. Amsterdam: Academic Press; 2011.
41. Prokosch HU, Ganslandt T. Perspectives for medical informatics. reusing the electronic medical record for clinical research. Methods Inf Med 2009;48(1):38-44. [Medline: 19151882]

Abbreviations

EMR: electronic medical record
HIS: hospital information system
ICD: International Classification of Diseases
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by J Hefner; submitted 31.03.23; peer-reviewed by S Veeranki, N Nekliudov, C Hudak; comments to author 23.07.23; revised version received 10.08.23; accepted 23.10.23; published 19.01.24.

Please cite as:
Wurster F, Beckmann M, Cecon-Stabel N, Dittmer K, Hansen TJ, Jaschke J, Köberlein-Neu J, Okumu MR, Rusniok C, Pfaff H, Karbach U
The Implementation of an Electronic Medical Record in a German Hospital and the Change in Completeness of Documentation: Longitudinal Document Analysis
JMIR Med Inform 2024;12:e47761
URL: https://medinform.jmir.org/2024/1/e47761
doi:10.2196/47761
PMID:38241076

Original Paper

Application of Failure Mode and Effects Analysis to Improve the Quality of the Front Page of Electronic Medical Records in China: Cross-Sectional Data Mapping Analysis

Siyi Zhan[1*], MMed; Liping Ding[1*], BSc; Hui Li[1], MMed; Aonan Su[1], MMed

Zhejiang Provincial People's Hospital, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Aonan Su, MMed
Zhejiang Provincial People's Hospital
No. 158, Shangtang Rd
Hangzhou, 310000
China
Phone: 86 18814885258
Email: suaonan_512917@126.com

Abstract

Background: The completeness and accuracy of the front pages of electronic medical records (EMRs) are crucial for evaluating hospital performance and for health insurance payments to inpatients. However, the quality of the first page of EMRs in China's medical system is not satisfactory, which can be partly attributed to deficiencies in the EMR system. Failure mode and effects analysis (FMEA) is a proactive risk management tool that can be used to investigate the potential failure modes in an EMR system and analyze the possible consequences.

Objective: The purpose of this study was to preemptively identify the potential failures of the EMR system in China and their causes and effects in order to prevent such failures from recurring. Further, we aimed to implement corresponding improvements to minimize system failure modes.

Methods: From January 1, 2020, to May 31, 2022, 10 experts, including clinicians, engineers, administrators, and medical record coders, in Zhejiang People's Hospital conducted FMEA to improve the quality of the front page of the EMR. The completeness and accuracy of the front page and the risk priority numbers were compared before and after the implementation of specific improvement measures.

Results: We identified 2 main processes and 6 subprocesses for improving the EMR system. We found that there were 13 potential failure modes, including data messaging errors, data completion errors, incomplete quality control, and coding errors. A questionnaire survey administered to random physicians and coders showed 7 major causes for these failure modes. Therefore, we established quality control rules for medical records and embedded them in the system. We also integrated the medical insurance system and the front page of the EMR on the same interface and established a set of intelligent front pages in the EMR management system. Further, we revamped the quality management systems such as communicating with physicians regularly and conducting special training seminars. The overall accuracy and integrity rate of the front page (P<.001) of the EMR increased significantly after implementation of the improvement measures, while the risk priority number decreased.

Conclusions: In this study, we were able to identify the potential failure modes in the front page of the EMR system by using the FMEA method and implement corresponding improvement measures in order to minimize recurring errors in the health care services in China.

(JMIR Med Inform 2024;12:e53002) doi:10.2196/53002

KEYWORDS

front page; EMR system; electronic medical record; failure mode and effects analysis; FMEA; measures

Introduction

The electronic medical record (EMR) system is the main carrier of medical information that has details about the whole process of a physician's treatment for a patient [1]. The information on the front page of the EMR is condensed, which includes a patient's basic information, disease diagnosis, information on surgical or invasive operations, and medical expenses [2]. Since January 1, 2013, almost all tertiary hospitals in China have submitted the front pages of the EMRs of inpatients to the Hospital Quality Monitoring System led by the Bureau of Medical Administration and Medical Service Supervision and National Health and Family Planning Commission of the People's Republic of China [3]. The quality and management of the front pages of EMRs are critical for their application in medical services [4], research [2,5], education [6], and hospital management [7]. For example, some indicators for assessing the capacity of hospital medical services, such as the services for surgery and disease diagnosis, often utilize the information through the front page of the EMR for statistical purposes. However, there are many difficulties in the management of the front page of EMR. A survey conducted by the National Medical Record Management Quality Control Center of China [8] showed that more than 230 million front pages of EMRs in 2020 are established in China. Each of them contain over 100 fields. However, there are only 2.5 full-time coders on average in each hospital among 5439 medical institutions, and only 67.9% of them perform special quality control, while 24.2% of them use information technology to control the quality of the front page of the EMR system.

For reforming the medical insurance payment methods in China, the Chinese State Council's version of health insurance issued a notice in 2019 on the issuance of technical specifications and grouping schemes for the national pilot of diagnosis-related grouping payments for diseases [9]. Therefore, the front page of an EMR needs to be uploaded on the websites of the Health and Wellness Committee and the Health Insurance Authority, which means coders need to edit a front page twice to meet the different needs of both the sectors. The former is for hospital performance evaluation and the latter is for patient health insurance payment. The introduction of this policy in 2019 increased the difficulty of medical record management.

Failure mode and effects analysis (FMEA) is a proactive risk management tool that originated in the US military in the 1940s. It is widely applicable to human, equipment, and system failure modes, as well as hardware and software programs. FMEA finds out all the potential failure modes in a system and analyzes their possible consequences by mapping the subsystems and each subprocess that makes up the process one by one in the product design stage and process design stage [10]. Thus, the advantage of FMEA is that problems can be identified and improved during the system development phase to avoid possible problems. Moreover, the costs incurred to address software defects and failures at an early stage are lower compared to those incurred to address defects at a later stage. Initially, FMEA was widely used in engineering [11], food safety management [12], financial management [13], and so on. Thereafter, with the rising demands in health care services, FMEA was used for proactive health care risk analysis. Doctors often use the EMR system to record patients' visits. Any issue in the EMR system can affect the patient's visit process and visit records. According to a systematic review [14], 158 studies published from 1998 to 2018 and classified under 4 categories, namely, health care process, hospital management, hospital informatization, and medical equipment and production, reported the use of FMEA in health care systems for proactive health care risk evaluation. In FMEA, the risk priority number (RPN) is calculated by giving a numerical value (scoring) for the severity, frequency, and detectability of the risks or failures, which enable risk assessment of the system [10]. An EMR system named Heren (Zhejiang Heren Technology Corporation), which is installed in many hospitals in China, is used by physicians and medical record management coders and quality controllers for filling out the front page. The purpose of this study was to identify the possible failures in the front page data of the EMR and their causes and effects and to propose specific improvement measures to minimize errors. Moreover, we aimed to compare the EMRs before and after introducing the measures to verify the efficacy of the improvement measures. For this, we reviewed previous relevant literature through PubMed, Embase, Web of Science, and Cochrane Library. During this review, we found that although FMEA has been used in some studies for improvement of some facets of EMRs, no study has used FMEA for improving the efficiency the front page of the EMR [15,16]. Thus, to the best of our knowledge, ours is the first study to apply FMEA to identify the potential failures on the front page of the EMR in China and the causes and effects of these failures and to perform a before-and-after comparison of the revised front page of the EMR.

Methods

Study Design

We conducted a cross-sectional study from January 1, 2020, to May 31, 2022, in Zhejiang People's Hospital, which is one of the largest public hospitals in Zhejiang province with more than 100,000 hospital discharges per year. During the period of our research, the number of hospital discharges reached 250,774, which means the same number of front pages of EMRs needed to be filled and coded.

Steps of FMEA

Assembling a Panel for FMEA

Ten experts, including clinicians, medical record coders, and hospital administrators, were invited to assess the potential risks of the EMR system in China. Since coders and quality controllers were necessary to ensure the accuracy of the front page of the EMR, only those who had been working full-time on this task for more than 5 years and who had achieved a coding accuracy rate of more than 95% and who had checked more than thousands of medical records for quality were included. Before we began our study, the organizer introduced the theme of our study to ensure that every expert knew the process of FMEA and the importance of a front page of an EMR. Then, the time and place for each discussion was planned to ensure that the process ran smoothly.

Mapping the Process and Subprocesses

Each expert mapped the process and subprocess of completing a front page of an EMR alone initially to avoid interference from others. For example, there are 2 data sources for the content on the front page of the EMR: information automatically imported from the hospital information system that is mainly used by physicians and information that is filled in manually by the physician. Thus, different experts could map their own process according to their work experience. Thereafter, all experts were gathered to draw the final process and subprocess to achieve the completeness of the whole system.

Brainstorming to Identify Potential Failure Modes in Each Subprocess and Their Causes and Effects

The implementation process of this step is consistent with the mapping process. At first, each expert could think about every potential failure mode individually. Then, all the experts summarized all the modes and discussed many more potential failure modes by brainstorming once again. In addition, the views on effects and reasons for failure modes were exchanged by experts. Since there were so many issues that could result in potential failure modes, our team summarized the main causes and created a questionnaire for randomly selected physicians and coders to answer.

Calculating the RPN

A scoring criterion was used to evaluate the severity, frequency, and detectability of the failures, and each dimension was divided into 10 points. Then, the RPN was calculated by using the score of the 3 dimensions (RPN = Severity × Frequency × Detectability) to evaluate the final score of each failure mode, which ranges from 0 to 1000. To improve the consistency and accuracy of scoring, the rating weight of each expert was based on their professional title grade, work experience, and familiarity with FMEA. In addition, a risk assessment criterion was established to avoid any dispute about the scores given by the experts.

Proposing Improvement Measures for Each Failure Mode

Since a low RPN could result from severity, frequency, or detectability and a low score for each dimension could be caused by many different reasons, it is necessary to find out the main issues. According to the Pareto principle, 80% of the consequences are due to only 20% of the potential causes [17]. Our team used the Pareto principle to identify the pressing causes that need to be addressed. Then, the experts proposed one or more corresponding improvement measures for each failure mode. Further, the feasibility and effectiveness of improvement measures were also discussed.

Comparing the Quality Before and After the Improvements

The experts evaluated the quality of the front page of the EMR before and after the application of the improvement measures. The RPN score was bound to improve if these improvement measures were effective.

Ethical Considerations

This study did not involve any patient data or ethical data, and the ethics approval committee of Zhejiang Provincial People's Hospital specified that no ethics approval was required.

Statistical Analysis

We performed statistical analyses using SPSS (version 20.0; IBM Corp). Two-sided *t* tests were performed to compare the RPNs of the front page of the EMR before and after applying the improvement measures. *P* values <.05 were considered statistically significant.

Results

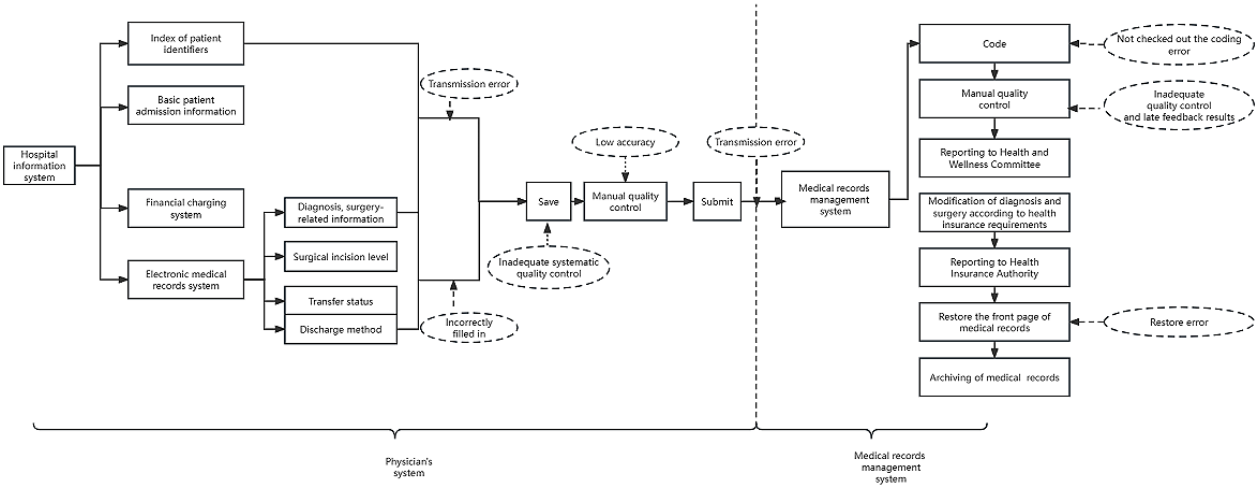
Assembling a Panel for FMEA

Our expert panel consisted of 10 experts in 5 different fields. There were 4 physicians, 2 coders, 2 hospital administrators, 1 quality control staff, and 1 information engineer who expressed different views and opinions on the front page of the EMR.

Mapping the Process and Subprocesses of the Front Page of EMR

The expert panel identified 2 main process steps and 6 subprocesses after discussion (Figure 1). The 2 main process steps were management of the physician’s system and management of the medical record system. The 6 subprocesses were information import from physicians’ hospital information system, front page filled by physicians, front page quality control by physicians, transmission of information in the EMR system, coders’ proofreading and coding, and front page quality control by EMR management.

Figure 1. Process map of the front page of the electronic medical record system.



Brainstorming to Identify Potential Failure Modes in Each Subprocess and Their Causes and Effects

The front page of the EMR that was evaluated in this study is shown in Figure 2. According to the process map of the front page of the EMR, the expert panel found that there were 13 potential failure modes, which can be mainly divided into 2 categories. One category is the low accuracy in a variety of information, including basic patient information, treatment information, and cost information. The other category is the low detection in a variety of information, including incorrect case header coding, incomplete quality control, and transmission errors. Regarding the causes for the failures, 115 physicians and coders filled out a questionnaire and summarized 15 main causes. The main causes were incompleteness of the front page, error in information, or incorrect diagnosis-related group, which is risky for hospital medical quality management, academic research, and medical insurance payment (Table 1).

Figure 2. Front page of the electronic medical record.

The Front Page of the Medicla Record of Inpatients

Medical Payment Options:			Medical institution code:			
Health Card Number:	No. of Hospitalizations:		Medical Card Number:			
Name:	Sex:	Birthday:	Age:	Nationality:		
Age(Less Than One Year Old):	Newborn Birth Weight:	Newborn Admission Weight:				
Birthplace:	Registered Birthplace:	Ethnological:				
I.D. Number:	Occupation:	Matrimonial:				
Current Address:	Telephones:	Postcode:				
Residential Address:		Postcode:				
Workplace And Address:	Unit Phone Number:	Postcode:				
Contact Name:	Relationship:	Contact Person'S Address:	Telephones:			
Pathway to Hospitalization:						
Date of Admission:	Admission Department:	Admission Ward:	Referral Unit:			
Date of Discharge:	Discharge Department:	Discharge Ward:	Days of Hospitalization:			
Outpatient Diagnosis:			Disease Codes:			
Discharge Diagnosis:	Disease Codes::	Admission Condition:	Discharge Diagnosis:	Disease Codes:	Admission Condition:	
Admission Condition: 1. Yes; 2. Clinically Undetermined; 3. Unknown; 4. None						
External Causes of Injuries, Poisoning:			Disease Codes:			
Pathological Examinations:			Pathological Number:			
Drug Allergy:						
Blood Type:	Rh:					
Chairman of Section:	Chief (Deputy) Physician:	Attending Doctor:	Resident Doctor:			
Student Nurse:	Refresher Doctor:	Intern:	Coder:			
Quality of Medical Record:	Quality Control Doctor:	Quality Control Nurse:	Quality Control Date:			
Surgery:						
Surgical Coding:	Date of Surgery:	Surgical Level:	Name of Surgery: Operator, First Assistant, Second	Incision Healing Grade:	Anesthesia:	Anesthesiologist
Methods of Discharge:	Whether there Is A Plan to be Admitted to the Hospital After 31 Days of Discharge			Coma Time In Patients With Craniocerebral Injuries:		
Hospitalization Expenses (Yuan):	Total Cost (Yuan):	(Out-of-Pocket Expenses) :				
1. Comprehensive Medical Service	(1) General Medical Service Expenses:	(2) General Treatment Operation Expenses:		(3) Nursing Care Expenses	(4) Other Expenses:	
2. Diagnostic Category:	(5) Diagnostic Pathology Expenses:	(6) Laboratory Diagnostic Expenses:		(7) Diagnostic Imaging Expenses:	(8) Clinical Diagnostic Program Expenses:	
3. Therapeutic Category:	(9) Non-Surgical Item Treatment Expenses:	(Clinical Physiotherapy Expenses:)				
	(10) Surgical Treatment Expenses:	(Anesthesia Expenses; Surgical Expenses:)				
4. Rehabilitation:	(11) Rehabilitation Expenses:					
5. Traditional Chinese Medicine:	(12) Chinese Medicine Treatment Expenses:					
6. Western Medicines:	(13) Western Medicine Expenses	(Antimicrobial Drug Expenses:)				
7. Traditional Chinese Herb:	(14)Chinese Patent Medicines Expenses:	(15)Chinese Medicinal Herb				
8. Blood Products:	(16) Blood Expenses:	(17) Albumin-Based Products Expenses:		(18) Expenses for Globulin-Based Products:	(19) Coagulation Factor-Based Products Expenses:	(20) Cytokine-Based Products Expenses:
9. Consumables :	(21) Charges for Disposable Medical Materials for Examinations :	(22) Charges for Disposable Medical Materials for Therapeutic Use :		(23) Charges for Surgical Disposable Medical Materials		
10. Other Categories :	(24) Other Expenses:					
Diagnostic Compliance:	Outpatient & Discharge:	Admission & Discharge:	Pre-Operative And Post-Operative:	Radiation And Pathology:	Clinical And	
0. Not Done; 1. Conform; 2. Not Conform; 3. Not Sure						
Single-Case Management:	Clinical Pathway Management:	Resuscitation:	Outcome Situation:			
Description:						
(A) Medical Payment Methods 1. Basic Medical Insurance for Urban Workers 2. Basic Medical Insurance for Urban Residents 3. New Rural Cooperative Medical Care 4. Poverty Relief 5. Commercial Medical Insurance 6. Full Public Expense 7. Full Self-Funding 8. Other						
(B) Where the Hospital Information System Can Provide A List of Inpatient Expenses, the Front Page of the Inpatient Medical Record May Not be Filled In With "Inpatient Expenses"						

Table 1. Potential failure modes with their causes and effects.

Process	Failure modes	Reasons	Effects
Transmission in the hospital information system	• Basic information transmission error • Inpatient information transmission error • Expenses information transmission error	• Data interface errors	• The original data on the front page are erroneous • The DRG[a] is erroneous • Affects patients' medical reimbursement
Front page filled by physicians	• Incorrectly filled-in medical information • Incorrectly filled in other information	• Do not understand the filling criteria • Do not fill in carefully • Incomplete quality control reminders	• The original data on the front page are erroneous • The DRG is erroneous • Affects patients' medical reimbursement
Front page quality control by physicians	• Inadequate quality control • Inaccurate quality control	• No emphasis on quality control • Unfamiliar with quality control rules • Complexity of quality control rules • Lack of information assistance	• The original data on the front page are erroneous • The DRG is wrong • Affects patients' medical reimbursement
Transmission in the physicians' EMR[b] system	• Inconsistency between the received data in the EMR system and original data	• Data interface errors • Encoding conversion error	• The original data on the front page are wrong • The DRG is wrong
Coders' proofreading and coding	• No data errors were found • Wrong code for diagnosis, surgery, or operation • Restoration error • Diagnostic and surgical operation codes do not meet the requirements of patients' insurance	• Formal quality control rules are too simple • Lack of internal quality control reminders • Insufficient professional capacity of coders • Few training opportunities for coders • Inadequate communication between coders and doctors • The criteria are different between the requirements of patients' insurance and front page	• Erroneous data persist • The DRG is wrong
Front page quality control by EMR management	• Inadequate quality control • Late feedback for the results of quality control	• Using a sampling model to conduct quality control • Insufficient professional capacity of quality control staff • Complexity of quality control rules • Lack of information assistance	• Unable to find all errors on the front page • Erroneous data persist • The DRG is wrong

[a]DRG: diagnosis-related group.

[b]EMR: electronic medical record.

Calculating the RPN

Before calculating the RPN, a risk assessment criterion was established to evaluate the quality of the front page of the EMR (Table 2).

The rating weight of each expert was calculated to reduce the influence caused by the individual subjective factors of the experts (Table 3).

Table 2. Risk assessment criteria for the quality management of the front page of the electronic medical record system.

Grade	Severity	Criteria for risk severity	Frequency	Criteria for risk frequency	Detectability	Criteria for risk detectability
10	Very high	Make the score of the front page of the EMR[a] below 20	Extremely high	Every time	Very low	Cannot be detected
9	Very high	Make the score of the front page of the EMR between 20 and 30	Very high	Almost every time	Very low	Hard to detect
8	High	Make the score of the front page of the EMR between 30 and 40	Very high	One time every half day	Low	Seldom detected
7	High	Make the score of the front page of the EMR between 40 and 50	High	More than one time every day	Low	Seldom detected
6	Middle	Make the score of the front page of the EMR between 50 and 60	High	More than one time every week	Middle	Easy to be detected
5	Middle	Make the score of the front page of the EMR between 60 and 70	Middle	More than one time every month	Middle	Easy to be detected
4	Middle	Make the score of the front page of the EMR between 70 and 80	Middle	More than one time every year	High	Very easy to be detected
3	Low	Make the score of the front page of the EMR between 80 and 90	Low	One time every year	High	Very easy to be detected
2	Low	Make the score of the front page of the EMR between 90 and 100	Very low	Less than one time every year	High	Very easy to be detected
1	Very low	Does not affect the score of the front page of the EMR	Extremely low	Never	Very high	No failure modes

[a]EMR: electronic medical record.

Table 3. Details of the expert panel.

Position	Rating weight	Working experience (years)	Familiarity with FMEA[a]	Rating weight
Physician	High	>20	General	9/10
Physician	Middle	10-20	General	7/10
Physician	Middle	1-5	Familiar	6/10
Physician	Primary	6-10	General	5/10
Coder	High	10-20	Familiar	9/10
Coder	Middle	6-10	Familiar	7/10
Quality control staff	High	10-20	Not very familiar	7/10
Administrator	High	>20	Not very familiar	8/10
Administrator	Middle	10-20	Familiar	8/10
Information engineer	Primary	1-5	Familiar	5/10

[a]FMEA: failure mode and effects analysis.

Proposing Improvement Measures for Each Failure Mode

According to the principle of Pareto, there were 7 causes in our study that contributed to 80% of the consequences (Figure 3), which can be addressed by revamping the information and quality management. For example, we integrated the medical insurance system with the front page of the EMR on the same interface and established a set of intelligent front pages for the EMR management system. In addition, we revamped the management of quality, such as communicating with physicians regularly and conducting special training seminars (Table 4).

Figure 3. The 7 causes that contributed to 80% of the failure modes in the electronic medical record system, according to the principle of Pareto. QC: quality control.

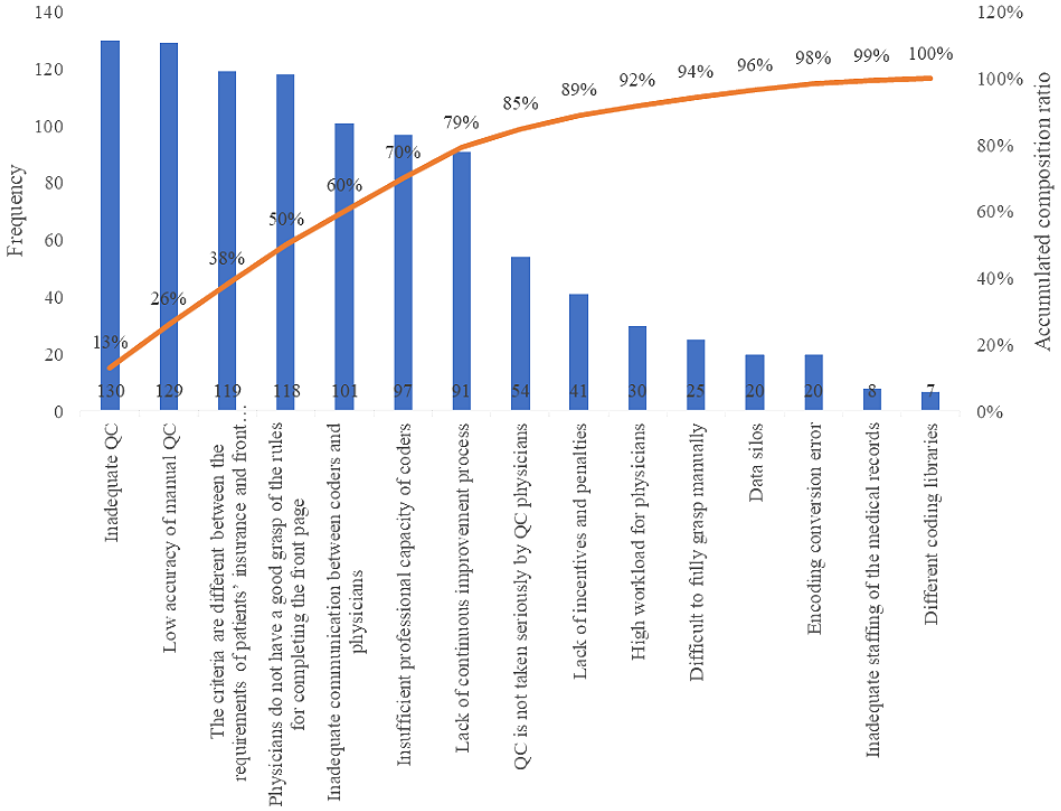


Table 4. Improvement measures for the causes of potential failure modes.

	Why	What	How and When	Where	Who
Revamp the information	• The rules for quality control are inadequate • Low accuracy in manual quality control • The criteria are different between the requirements of patients' insurance and front page details	• Establish a set of intelligent front pages in the EMR[a] management system	• Added quality control rules from May to September 2021 • Embedded quality control systems into the physician's system and front page system from July to October 2021 • Integrated medical insurance system and the front page of the EMR from July to December 2021	• EMR center and information center	• Staffs of EMR center and information engineers
Revamp quality management	• Formal quality control rules are simple • Insufficient professional capacity of coders • Inadequate communication between coders and physician • Lack of a continuous improvement process	• Improve the correctness of the front page by physician • Improve the professional capacity of coders • Establish effective communication channels • Form a continuous improvement process	From January to May 2021 • Conducted special training seminars • Prepared a quality and information management manual • Implemented professional training regularly • Invited experts for guidance • Communicated with physicians regularly • Provided timely feedback to physicians on quality control results	• EMR center and clinical departments	• Staffs of EMR center, information engineers, and physicians

[a]EMR: electronic medical record.

Comparing the Front Page of EMR Before and After Improvement

Before carrying out improvement measures, the highest RPN given by the experts was 296.3 for the failure mode "wrong code for diagnosis, surgery, or operation," which was due to the quality rules being too simple, while the lowest was 50.6 for "basic information transmission error," which was caused by wrong data interface or conversion error. On average, the final RPN was 181.2. The highest score for severity was for "wrong code for diagnosis, surgery, or operation" and the lowest was for "expense information transmission error." The highest score for frequency was for "incorrectly filled-in medical information" and the lowest was for "expense information transmission error." As for detectability, the highest score was for "wrong diagnosis-related group code of diagnosis, surgery, or operation," "restoration error," and "late feedback for quality control results." The lowest score for detectability was for "basic information transmission error." Our team calculated the RPN of the revised front page of the EMR after implementing the improvement measures mentioned in Table 5, and the final RPN was 95.0, which was lower than that of the original front page of the EMR (RPN=181.2).

The RPN of every failure mode decreased after implementing the improvements, and the mode for the late feedback for quality control decreased the most remarkably (Table 5). In addition, the accuracy rate of the basic information (χ^2_1=269.6; P<.001); inpatient information (χ^2_1=175.9; P<.001); diagnosis, surgery, and operation code (χ^2_1=32.9; P<.001); and the overall accuracy rate of the front page (χ^2_1=239.3; P<.001) and the integrity rate of the front page (χ^2_1=110.4; P<.001) increased significantly (Table 6).

Table 5. Comparison of the risk analysis before and after failure mode and effects analysis model improvement.

Process, failure modes	Before FMEA[a]				After FMEA			
	Severity	Frequency	Detectability	Risk priority number	Severity	Frequency	Detectability	Risk priority number
Information import from HIS[b]								
Basic information transmission error	3.1	3.4	4.8	50.6	3.1	3.1	3.7	35.6
Inpatient information transmission error	3.9	2.3	6.6	59.2	3.6	2.1	6.2	46.9
Expenses information transmission error	2.4	2.1	6.6	33.3	2.2	2.0	6.0	26.4
Front page filled by physicians								
Incorrectly filled-in medical information	6.1	6.6	6.4	257.7	5.8	6.2	6.0	215.8
Incorrectly filled-in other information	2.7	5.8	6.1	95.5	2.3	5.1	5.5	64.5
Front page quality control by physicians								
Inadequate quality control	6.5	6.4	6.4	266.2	6.0	5.1	5.2	159.1
Inaccurate quality control	6.1	6.5	6.5	257.7	4.9	5.0	4.7	115.2
Information import from the physician's EMR[c] system								
Inconsistency between the received data in EMR system and the original data	3.6	3.3	5.3	63.0	3.1	3.1	4.7	45.2
Coders' proofreading and coding								
No data errors were found	5.2	6.3	6.3	206.4	4.5	5.3	4.8	114.5
Wrong code for diagnosis, surgery, or operation	6.7	6.6	6.7	296.3	4.6	4.5	5.0	103.5
Restoration error	6.6	6.1	6.7	269.7	4.4	4.6	4.7	95.1
Front page quality control by EMR management								
Inadequate quality control	6.1	6.5	6.0	237.9	4.6	5.1	4.6	107.9
Late feedback of quality control results	6.1	6.4	6.7	261.6	4.7	4.9	4.6	105.9
Average	N/A[d]	N/A	N/A	181.2	N/A	N/A	N/A	95.0

[a]FMEA: failure mode and effects analysis.

[b]HIS: hospital information system.

[c]EMR: electronic medical record.

[d]N/A: not applicable.

Table 6. Comparison of the accuracy and integrity of the front page of the electronic medical records before and after failure mode and effects analysis model improvement.

Items	Front pages (n)	Accuracy rate of basic information	Accuracy rate of inpatient information	Accuracy rate of diagnosis, surgery, and operation code	Overall accuracy rate of front page	Integrity rate of front page
Before	48,509	94.09	95.28	97.29	93.44	96.15
After	78,890	96.09	96.74	97.81	95.48	97.26
Chi-square *(df)*	N/A[a]	269.6 (1)	175.9 (1)	32.9 (1)	239.3 (1)	110.4 (1)
P value	N/A	<.001	<.001	<.001	<.001	<.001

[a]N/A: not applicable.

Discussion

The quality of the front page of an EMR is quite important not only for hospital performance management [2] but also for insurance payments to patients [15]. Thus, it is necessary to improve the effectiveness of the front page of the EMR. There are many risk management tools for investigating the potential problems in an EMR system, such as Expert Delphi [18], scenario analysis method [19], and SWOT (strengths, weaknesses, opportunities, and threats) analysis method [20]. The advantage of Expert Delphi is that everyone's opinions are collected and that of scenario analysis is that it identifies risks by designing multiple possible future scenarios. The advantage of SWOT is that it identifies the strengths, weaknesses, opportunities, and costs of the project, thus qualitatively identifying the project risks from multiple perspectives. FMEA is a risk management tool that has most of the advantages of the above tools. FMEA can not only change the occurrence of risk from postprocessing to preemptive prevention but is also a simple and a practical risk quantification method [10]. In recent years, FMEA has been widely used in various fields, including the medical field. Studies on medical services [21], medicine distribution [22], infection control [23], and medical equipment operation and maintenance [24] have used FMEA to date.

In this study, we found potential failures existing in the EMR system of China and proposed improvement measures to solve the problems by using FMEA. Our results showed that there were 2 main processes and 6 subprocesses in the EMR system that showed 13 potential failure modes. The 2 main process steps were management of the physician's system and management of the medical record system. The 6 subprocesses were information transmission in the hospital information system, front page filled by physicians, front page quality control by physicians, information transmission in the EMR system, coders' proofreading and coding, and front page quality control by EMR management. This finding is similar to that reported in a study performed in Indonesia [25], wherein potential failure modes included incomplete or missing medical record files, mistakes caused by coders, and excessive code writing or upcoding [25].

According to the principle of Pareto and from questionnaire responses, we found that there were 7 causes in our study that contributed to 80% of the consequences, which can be divided into 2 aspects for the resolution of errors. One aspect was to revamp the information by establishing a set of intelligent front pages in the EMR management system to solve the problems of inadequate information and inaccurate quality control and to implement different codes of management or payment. In this study, we established quality control rules for medical records and embedded them in the system first. Accurate quality control rules are important for maintaining data quality. For example, Carlson et al [26] used quality control rules to identify common logical problems, including incomplete data, invalid values, and inconsistent data, in a clinical data set of an intensive care unit. Hart et al [27] reported >50% decrease in rejected records across patient information, service information, and financial information in 6 months by using quality rules. In addition, we integrated the medical insurance system with the front page of the EMR on the same interface. The other aspect was to revamp the management of quality by conducting multichannel trainings for doctors and coders, creating a quality and information management manual, and communicating with physicians and coders regularly. Previous studies [28-30] have shown a high rate of errors in physician coding for professional services, which can be risky in medical care services. One study [31] showed that clinicians and coders differ in their understanding of disease coding and need to communicate in a timely manner. Some of our measures are also consistent with those previously reported [25] that a hospital needs to update coding training for coders and provide guidance and validation of coding for physicians as well.

After implementing improvement measures, we found that the RPN of every failure mode decreased. The most significant decline in RPN was for the mode on the late feedback for quality control results. Many studies have proved the benefits of artificial intelligence. For example, machine learning could improve the content of medical records by identifying patients' medical information [32] or by predicting the onset of disease [33]. Therefore, we applied artificial intelligence to establish an intelligent front page for the EMR management system and then embedded it in the doctor's medical record writing interface and medical record quality control interface, which made it possible for real-time quality control of the front page. The second indicator of decline was inaccurate quality control of the front page by physicians. The original data on the front page, such as basic patient information, expenses, and surgery, are filled by physicians who decide the quality of the front page mostly [34]. After the amendments, the accuracy and integrity

of the front page were both improved for those measures, which helped the diagnosis-related group to be more specific and the evaluation of the hospital performance more precise. In addition, the quality of the front page of EMR is quite important for patients. A complete front page of the medical record enables doctors to grasp important information about the patient in a short period, such as family history, allergy history, and important test results and facilitates doctors to quickly and accurately judge the patient's condition and formulate diagnosis and treatment plans, thereby reducing overmedication and even erroneous medical treatment.

Human factors engineering and user-centered designs are indispensable components of mobile health technology design and implementation [35]. Human factors emphasize human needs and capabilities as the core of the design technology system, making people the most important consideration in the design process and aiming to achieve the goal of "making machines fit people." Regarding EMR system update, physicians, medical record coders, and quality controllers are the target users, and they will resist the technology when they believe it does not meet their expectations and needs [36]. For this reason, this study was conducted through brainstorming and questionnaires to inform the needs of physicians, coders, and others regarding the front page of the EMR system. For example, incorrectly filled-in medical information and quality control proposed by physicians, coders, and other users prompted engineers to establish a set of intelligent front pages in the EMR management system. The usability of the EMR system is evaluated by its effectiveness, efficiency, and suitability for target users. Although we did not use questionnaires to analyze the satisfaction of doctors, coders, and others with the improved EMR system, the results of FMEA showed that RPN was greatly reduced after the system was improved; thus, it can be hypothesized that the user's satisfaction with the system has been enhanced. Moreover, the overall accuracy rate of the front page (P<.001) and the integrity rate of the front page (P<.001) were significantly enhanced after implementing the improvement measures, thereby demonstrating the increase in the effectiveness of the system. The number of front pages of EMR increased from 48,509 to 78,890 with the same amount of time and labor, which proves that the efficiency of the system was also improved.

Our study has several strengths. First, medical research FMEA has mostly been performed for health care processes, hospital management, etc. For example, a study performed in Sri Lanka used FMEA to improve medication safety in the dispensing process [22], while another study aimed to increase the efficiency and success rate of patients with acute ischemic stroke [25]. No study has used FMEA for improving the front page of EMR in China before. Therefore, this is the first study performed in China, which can provide the base for future studies. Second, most studies only used FMEA to find potential failure modes and propose improvement measures, but the system was not evaluated after the application of those measures. However, our study used FMEA to compare the RPN of the front page of the EMR before and after applying the improvement measures to verify the efficiency of the system. Our study also has some limitations. The first limitation was that the method we used is not advanced since there are many better methods such as data envelopment analysis [37] and fuzzy RPN method [38]. The second limitation was that the process of scoring the system by the experts was subjective although we had set weights for the experts' scores. The third limitation was that we did not use additional methods to validate the results, which we aim to improve in the future. Lastly, although the EMR system called Heren has been used in many hospitals, different hospitals may use different types of Heren. Consequently, the generalizability of this study and the findings should be considered cautiously. In conclusion, we improved the front pages of the EMRs in China based on the potential failure modes found by the FMEA method.

Acknowledgments

This project was supported by Project of 2023 Zhejiang Province Archives Science and Technology Project Research (2023-34).

Data Availability

Data sets are available from the corresponding author on reasonable request.

Authors' Contributions

SZ wrote the main manuscript. AS, HL, and SZ prepared the figures and tables. AS, HL, SZ, and LD designed the study. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;Suppl 1:S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
2. Wu C, Zhang D, Bai X, Zhou T, Wang Y, Lin Z, China PEACE Collaborative Group. Are medical record front page data suitable for risk adjustment in hospital performance measurement? Development and validation of a risk model of in-hospital

mortality after acute myocardial infarction. BMJ Open 2021 Apr 09;11(4):e045053 [FREE Full text] [doi: 10.1136/bmjopen-2020-045053] [Medline: 33837102]
3. Tan Y, Yu F, Long J, Gan L, Wang H, Zhang L, et al. Frequency of systemic lupus erythematosus was decreasing among hospitalized patients from 2013 to 2017 in a national database in China. Front Med (Lausanne) 2021;8:648727 [FREE Full text] [doi: 10.3389/fmed.2021.648727] [Medline: 33889586]
4. Sutherland SM. Electronic health record-enabled big-data approaches to nephrotoxin-associated acute kidney injury risk prediction. Pharmacotherapy 2018 Aug;38(8):804-812. [doi: 10.1002/phar.2150] [Medline: 29885015]
5. Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. The Lancet HIV 2019 Oct;6(10):e688-e695. [doi: 10.1016/s2352-3018(19)30137-7]
6. Chung J, Cho I. The need for academic electronic health record systems in nurse education. Nurse Educ Today 2017 Jul;54:83-88. [doi: 10.1016/j.nedt.2017.04.018] [Medline: 28500984]
7. Shin J, Ko H, Kim JA, Song Y, An JS, Nam SJ, et al. Hospital cancer pain management by electronic health record-based automatic screening. Am J Manag Care 2018 Nov 01;24(11):e338-e343 [FREE Full text] [Medline: 30452201]
8. National Health Commission. 2020 National Medical Services and Quality Safety Report. Beijing: Science and technology literature press; Oct 1, 2021.
9. Notice on printing and distributing the national pilot technical specifications and grouping plans for the payment of disease diagnosis related groups (DRGs). The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/zhengce/zhengceku/2019-11/18/content_5562261.htm [accessed 2023-12-20]
10. Schneider H. Failure mode and effect analysis: FMEA from theory to execution. Technometrics 1996 Feb;38(1):80-80. [doi: 10.1080/00401706.1996.10484424]
11. Fu Y, Qin Y, Wang W, Liu X, Jia L. An extended FMEA model based on cumulative prospect theory and type-2 intuitionistic fuzzy VIKOR for the railway train risk prioritization. Entropy 2020 Dec 15;22(12):1418. [doi: 10.3390/e22121418]
12. Lee JC, Daraba A, Voidarou C, Rozos G, Enshasy HAE, Varzakas T. Implementation of food safety management systems along with other management tools (HAZOP, FMEA, Ishikawa, Pareto). The case study of Listeria monocytogenes and correlation with microbiological criteria. Foods 2021 Sep 13;10(9):2169. [doi: 10.3390/foods10092169]
13. Edu AS, Agoyi M, Agozie D. Digital security vulnerabilities and threats implications for financial institutions deploying digital technology platforms and application: FMEA and FTOPSIS analysis. Peer J Comput Sci 2021;7:e658 [FREE Full text] [doi: 10.7717/peerj-cs.658] [Medline: 34435101]
14. Liu H, Zhang L, Ping Y, Wang L. Failure mode and effects analysis for proactive healthcare risk evaluation: A systematic literature review. J Eval Clin Pract 2020 Aug;26(4):1320-1337. [doi: 10.1111/jep.13317] [Medline: 31849153]
15. Notice of the Office of the National Medical Security Administration on issuing the detailed grouping plan for diagnosis related groups of medical security diseases (CHS-DRG) (version 1.0). The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/zhengce/zhengceku/2019-11/18/content_5562261.htm [accessed 2023-12-20]
16. Asgari Dastjerdi H, Khorasani E, Yarmohammadian MH, Ahmadzade MS. Evaluating the application of failure mode and effects analysis technique in hospital wards: a systematic review. J Inj Violence Res 2017 Jan;9(1):51-60 [FREE Full text] [doi: 10.5249/jivr.v9i1.794] [Medline: 28039688]
17. Harvey HB, Sotardi ST. The Pareto principle. J Am Coll Radiol 2018 Jun;15(6):931. [doi: 10.1016/j.jacr.2018.02.026] [Medline: 29706287]
18. Vázquez L, Salavert M, Gayoso J, Lizasoaín M, Ruiz Camps I, Di Benedetto N, Study Group of Risk Factors for IFI using the Delphi Method. Delphi-based study and analysis of key risk factors for invasive fungal infection in haematological patients. Rev Esp Quimioter 2017 Apr;30(2):103-117 [FREE Full text] [Medline: 28198173]
19. Jones CH, Wylie V, Ford H, Fawell J, Holmer M, Bell K. A robust scenario analysis approach to water recycling quantitative microbial risk assessment. J Appl Microbiol 2023 Mar 01;134(3):029-029. [doi: 10.1093/jambio/lxad029] [Medline: 36796790]
20. Dominguez JA, Pacheco LA, Moratalla E, Carugno JA, Carrera M, Perez-Milan F, et al. Diagnosis and management of isthmocele (Cesarean scar defect): a SWOT analysis. Ultrasound Obstet Gynecol 2023 Sep;62(3):336-344. [doi: 10.1002/uog.26171] [Medline: 36730180]
21. Maughan NM, Garcia-Ramirez JL, Huang FS, Willis DN, Iravani A, Amurao M, et al. Failure modes and effects analysis of pediatric I-131 MIBG therapy: Program design and potential pitfalls. Pediatr Blood Cancer 2022 Dec;69(12):e29996. [doi: 10.1002/pbc.29996] [Medline: 36102748]
22. Anjalee JAL, Rutter V, Samaranayake NR. Application of failure mode and effects analysis (FMEA) to improve medication safety in the dispensing process - a study at a teaching hospital, Sri Lanka. BMC Public Health 2021 Jul 20;21(1):1430 [FREE Full text] [doi: 10.1186/s12889-021-11369-5] [Medline: 34284737]
23. Lin L, Wang R, Chen T, Deng J, Niu Y, Wang M. Failure mode and effects analysis on the control effect of multi-drug-resistant bacteria in ICU patients. Am J Transl Res 2021;13(9):10777-10784 [FREE Full text] [Medline: 34650755]
24. Frosini F, Miniati R, Grillone S, Dori F, Gentili GB, Belardinelli A. Integrated HTA-FMEA/FMECA methodology for the evaluation of robotic system in urology and general surgery. THC 2016 Nov 14;24(6):873-887. [doi: 10.3233/thc-161236]

25. Yang Y, Chang Q, Chen J, Zou X, Xue Q, Song A. Application of integrated emergency care model based on failure modes and effects analysis in patients with ischemic stroke. Front Surg 2022;9:874577 [FREE Full text] [doi: 10.3389/fsurg.2022.874577] [Medline: 35449548]
26. Carlson D, Wallace CJ, East TD, Morris AH. Proc Annu Symp Comput Appl Med Care 1995:188-192 [FREE Full text] [Medline: 8563264]
27. Hart R, Kuo MH. Better data quality for better healthcare research results - a case study. Stud Health Technol Inform 2017;234:161-166. [Medline: 28186034]
28. Andreae MC, Dunham K, Freed GL. Inadequate training in billing and coding as perceived by recent pediatric graduates. Clin Pediatr (Phila) 2009 Nov;48(9):939-944. [doi: 10.1177/0009922809337622] [Medline: 19483135]
29. Balla F, Garwe T, Motghare P, Stamile T, Kim J, Mahnken H, et al. Evaluating coding accuracy in General Surgery Residents' Accreditation Council for Graduate Medical Education procedural case logs. J Surg Educ 2016;73(6):e59-e63. [doi: 10.1016/j.jsurg.2016.07.017] [Medline: 27886974]
30. Greenky MR, Winters BS, Bishop ME, McDonald EL, Rogero RG, Shakked RJ, et al. Coding education in residency and in practice improves accuracy of coding in orthopedic surgery. Orthopedics 2020 Nov 01;43(6):380-383. [doi: 10.3928/01477447-20200827-10] [Medline: 32882048]
31. Glauser G, Sharma N, Beatson N, Dimentberg R, Savarese F, Gagliardi M, et al. Surgical CPT coding discrepancies: analysis of surgeons and employed coders. Am J Med Qual 2021;36(4):263-269. [doi: 10.1177/1062860620959440] [Medline: 32959674]
32. Willyard C. Can AI fix medical records? Nature 2019 Dec;576(7787):S59-S62. [doi: 10.1038/d41586-019-03848-y] [Medline: 31853075]
33. Kilic A. Artificial intelligence and machine learning in cardiovascular health care. Ann Thorac Surg 2020 May;109(5):1323-1329. [doi: 10.1016/j.athoracsur.2019.09.042] [Medline: 31706869]
34. Reinus JF. The electronic medical record, Lawrence Weed, and the quality of clinical documentation. Am J Med 2021 Mar;134(3):e143-e144. [doi: 10.1016/j.amjmed.2020.09.055] [Medline: 33171102]
35. Human factors engineering and user-centered design for mobile health technology: enhancing effectiveness, efficiency, and satisfaction. In: Human-Automation Interaction. New York: Springer, Cham; Dec 15, 2022.
36. Or C, Dohan M, Tan J. Understanding critical barriers to implementing a clinical information system in a nursing home through the lens of a socio-technical perspective. J Med Syst 2014 Sep;38(9):99. [doi: 10.1007/s10916-014-0099-9] [Medline: 25047519]
37. Lamovšek N, Klun M. Evaluation of biomedical laboratory performance optimisation using the DEA method. Zdr Varst 2020 Sep;59(3):172-179 [FREE Full text] [doi: 10.2478/sjph-2020-0022] [Medline: 32952718]
38. Alizadeh SS, Solimanzadeh Y, Mousavi S, Safari GH. Risk assessment of physical unit operations of wastewater treatment plant using fuzzy FMEA method: a case study in the northwest of Iran. Environ Monit Assess 2022 Jul 23;194(9):609. [doi: 10.1007/s10661-022-10248-9] [Medline: 35870035]

Abbreviations

EMR: electronic medical record
FMEA: failure mode and effects analysis
RPN: risk priority number
SWOT: strengths, weaknesses, opportunities, and threats

Edited by J Hefner; submitted 21.09.23; peer-reviewed by C Or, Y Zhang, PH Liao; comments to author 24.10.23; revised version received 24.11.23; accepted 05.12.23; published 19.01.24.

Please cite as:
Zhan S, Ding L, Li H, Su A
Application of Failure Mode and Effects Analysis to Improve the Quality of the Front Page of Electronic Medical Records in China: Cross-Sectional Data Mapping Analysis
JMIR Med Inform 2024;12:e53002
URL: https://medinform.jmir.org/2024/1/e53002
doi:10.2196/53002
PMID:38241064

Viewpoint

Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review

Elham Asgari[1,2], MD, MSc, PhD; Japsimar Kaur[3], MBBS; Gani Nuredini[4], MBBS, BMedSci; Jasmine Balloch[2], BSc, MSc; Andrew M Taylor[5], MBBS, Prof Dr; Neil Sebire[5], MBBS, MD, Prof Dr; Robert Robinson[5], MBBS, MA, PhD; Catherine Peters[5], MBChB, MD; Shankar Sridharan[5], BSc, MBBS; Dominic Pimenta[2], MBBS, BSc

[1]Guy's and St Thomas' NHS Trust, London, United Kingdom

[2]Tortus AI, London, United Kingdom

[3]Manchester University NHS Foundation Trust, Manchester, United Kingdom

[4]Barts Health NHS Trust, London, United Kingdom

[5]Great Ormond Street Hospital, London, United Kingdom

Corresponding Author:
Elham Asgari, MD, MSc, PhD
Tortus AI
193-197 High Holborn
London, WC1V 7BD
United Kingdom
Phone: 44 7763891802
Email: asgelham@gmail.com

Abstract

The cognitive load theory suggests that completing a task relies on the interplay between sensory input, working memory, and long-term memory. Cognitive overload occurs when the working memory's limited capacity is exceeded due to excessive information processing. In health care, clinicians face increasing cognitive load as the complexity of patient care has risen, leading to potential burnout. Electronic health records (EHRs) have become a common feature in modern health care, offering improved access to data and the ability to provide better patient care. They have been added to the electronic ecosystem alongside emails and other resources, such as guidelines and literature searches. Concerns have arisen in recent years that despite many benefits, the use of EHRs may lead to cognitive overload, which can impact the performance and well-being of clinicians. We aimed to review the impact of EHR use on cognitive load and how it correlates with physician burnout. Additionally, we wanted to identify potential strategies recommended in the literature that could be implemented to decrease the cognitive burden associated with the use of EHRs, with the goal of reducing clinician burnout. Using a comprehensive literature review on the topic, we have explored the link between EHR use, cognitive load, and burnout among health care professionals. We have also noted key factors that can help reduce EHR-related cognitive load, which may help reduce clinician burnout. The research findings suggest that inadequate efforts to present large amounts of clinical data to users in a manner that allows the user to control the cognitive burden in the EHR and the complexity of the user interfaces, thus adding more "work" to tasks, can lead to cognitive overload and burnout; this calls for strategies to mitigate these effects. Several factors, such as the presentation of information in the EHR, the specialty, the health care setting, and the time spent completing documentation and navigating systems, can contribute to this excess cognitive load and result in burnout. Potential strategies to mitigate this might include improving user interfaces, streamlining information, and reducing documentation burden requirements for clinicians. New technologies may facilitate these strategies. The review highlights the importance of addressing cognitive overload as one of the unintended consequences of EHR adoption and potential strategies for mitigation, identifying gaps in the current literature that require further exploration.

(JMIR Med Inform 2024;12:e55499) doi:10.2196/55499

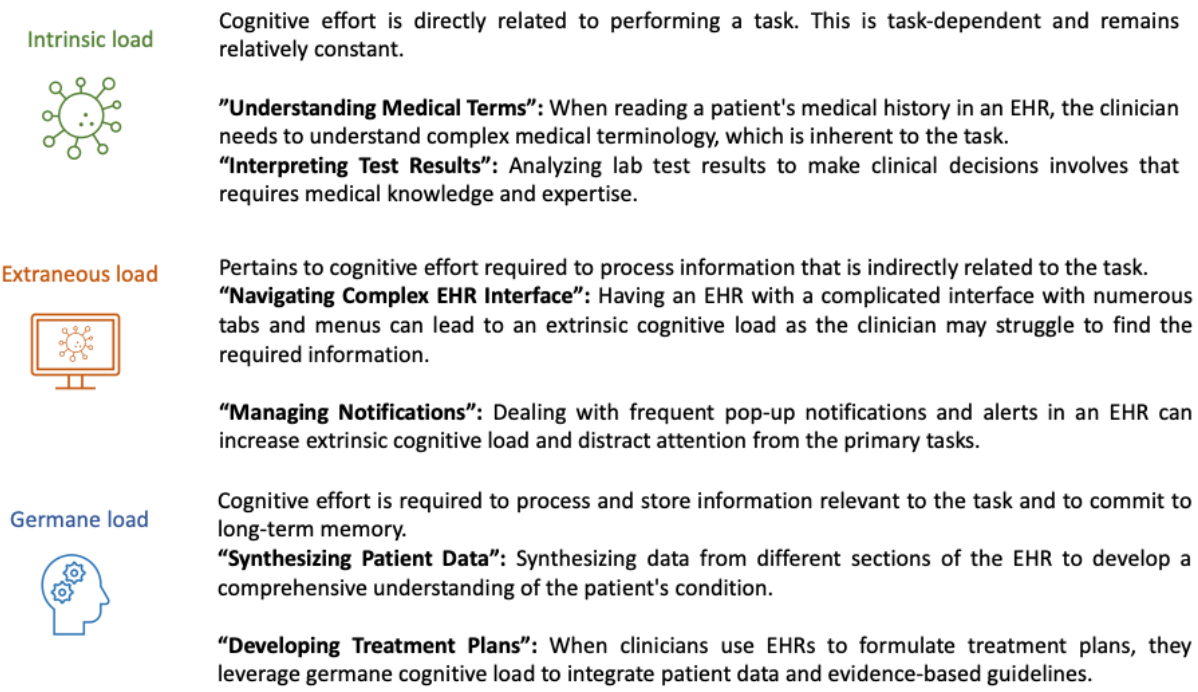
KEYWORDS

electronic health record; cognitive load; burnout; technology; clinician

Introduction

Sweller [1] defined cognitive load as "the amount of mental effort required to process and store information in working memory." The cognitive load theory argues that completing a task requires a complex interplay between sensory inputs, working memory, and long-term memory [1]. The working memory helps interpret the sensory input and then commits processed information into long-term memory. This psychological theory stipulates that while both sensory and long-term memories can handle large volumes of input data, working memory has a comparatively limited capacity and can keep 3 to 5 items in mind at a time [2]. When the amount of information exceeds this given capacity, it leads to cognitive overload. For any given task, there are 3 main factors that contribute to its perceived cognitive load: intrinsic, extraneous, and germane [1]. Figure 1 provides a description of each cognitive load with examples of how they are affected when clinicians use electronic health records (EHRs).

Figure 1. Factors contributing to cognitive load and examples of how they are affected when clinicians use EHRs. EHR: electronic health record.



Physician burnout refers to a state of chronic physical and emotional exhaustion experienced by clinicians, often due to prolonged stress and heavy workloads. It is a growing concern in the medical sector, both in the United Kingdom and globally, and it affects individual health care professionals as well as the health care system as a whole. Key aspects include emotional exhaustion, demoralization, and a reduced sense of accomplishment [3].

Contributing factors to burnout include unremitting workloads, administrative burdens, emotional stress, work-life imbalance, the lack of autonomy, and insufficient support systems. Physicians who work long hours and have high patient loads can experience physical and mental fatigue. Additionally, administrative tasks, paperwork, and EHR requirements add to the workload, causing frustration [4].

According to a report from the World Health Organization, the average life expectancy globally has risen by 6 years, from 66.8 years in 2000 to 73.4 years in 2019 [5]. It is estimated that 1 in 3 adults have multiple long-term conditions [6]. With advancements in science and technology, there are now more diagnostic and treatment options available, which require additional monitoring and follow-up. Consequently, physicians face an increased cognitive load due to the larger volume of information they need to process from complex patients to deliver high-quality care.

Simultaneously, the extraneous load experienced by clinicians is influenced by the presentation of this information. When information is disorganized, unnecessary, or incomplete, clinicians are exposed to a higher extraneous load, placing a greater demand on their working memory [7], which in turn has a knock-on effect on the germane load. This effect is most pronounced among early-career clinicians with a significant amount of new material to learn [7].

Working memory is also attenuated in the presence of physiological or emotional stress [8], which in recent times has become increasingly more common among clinicians following the COVID-19 pandemic, leading to widespread burnout.

Although EHR systems are widely used in health care settings worldwide, there are not enough comprehensive studies evaluating their advantages and disadvantages or examining how they can be improved. In a systematic review, Moy and colleagues [9] aimed to identify the studies on physician and nursing burnout related to using EHR systems. They found that 40% of the 35 studies meeting their inclusion criteria had mentioned clinical burnout. However, they noted a lack of standardized and validated measures to assess the documentation

burden related to EHR use. There is also a lack of objective measures to assess the cognitive load associated with using EHRs.

In summary, the increasing cognitive load experienced by physicians is a sum of the increasing complexity of the information presented, the quality and clarity of that presentation, as well as the emotional and psychological context in which the information is received [1]. The usability of EHRs, including their design, interoperability, and various regulatory requirements, impacts this cognitive burden. By working constantly above the cognitive load threshold, clinicians may exhibit symptoms of cognitive overload, which is considered an immediate precursor to burnout [10]. This narrative review examines the existing literature on cognitive overload experienced by health care professionals, specifically in relation to their use of EHRs and the associated burnout. The review also explores potential solutions that could help reduce the EHR-related cognitive load and improve the well-being of clinicians.

Use of EHRs in Health Care

The digitization of health care has been a growing trend over the past decades, which has seen most patient data transferred from paper records to EHRs. The use of EHR dates back to the 1960s but was only limited to government use [11]. Since the 1970s, EHR systems have been developed with hierarchical or relational databases for various indications, such as to help with hospital billing and scheduling systems, to help improve medical care, and for use in medical research. As computers have become more accessible, larger health care organizations have begun using them to gather patient information [12]. This shift has gradually expanded to encompass nonclinical tasks such as administration, medicolegal work, research, and education [13], which have also increased in demand and complexity over time.

Poor physician acceptance and the lack of incentives, in addition to high costs and errors associated with data entry, hindered EHR implementation uptake in the 1990s, and thus, digital records were not widespread [14]. By the 2000s, countries such as the United States and the United Kingdom began implementing national projects to digitize paper records [12]. The UK government attempted to create the National Programme for Information Technology [15] in 2002 to create a universal EHR system for the entire United Kingdom. The nationwide initiative was centered around 3 main objectives: lifelong EHRs, 24/7 web-based access by public health care professionals, and seamless information sharing throughout all sectors of the National Health Service [16]. The project failed to meet its objectives, and digitization became fragmented and slow once again [17]. However, by 2022, a total of 86% of the UK hospital trusts had successfully transitioned from paper notes to a digital system (although only a minority have enterprise-wide EHR capability), with the figure expected to reach 90% by December 2023 [18]. In the United States, the *Health Information Technology for Economic and Clinical Health Act* was signed into law in February 2009 to promote the adoption of and meaningful use of health information technology (HIT) as part of the American Recovery and Reinvestment Act. Financial incentives that were allocated as part of the *Health Information Technology for Economic and Clinical Health Act* led to a significant increase in the adoption of EHRs in the United States [14].

Factors Impacting the Usability of EHR Systems

The usability of EHR systems continues to be a major concern, whereby clinicians are subjected to too much or too little information, preprogrammed workflows, and multiple alerts [19]. There have been problems with chaotic, nonintuitive visual displays and numerous default settings that might not be relevant for a given task or patient [20]. Navigating through the same information includes unnecessary steps, for example, multiple clicks and duplicated information [21]. Users experience higher fatigue, leading to potential room for errors and decreased efficiency [22,23]. In addition to documentation and chart review, managing inbox tasks has been noted as one of the significant burdens for clinicians [24]. Receiving excessive notification has been shown to cause alert fatigue, leading to missing important information and poor patient outcomes [25]. Although clinical decision support systems have been introduced to enhance patient care, excessive use of interruptive clinical decision support systems in the EHRs can lead to alert fatigue and reduced effectiveness. Chaparro and colleagues [26] have described how interruptive alerts can increase cognitive burden and lead to reduced acceptance of the alert and an increase in the number of errors.

Several factors have been highlighted as contributing to the use of EHR and physician well-being. Nguyen and colleagues [27] studied this in a systematic review, where they found that EHR-related physician well-being is determined by multiple factors, including EHR usability, EHR system features, and physician-level characteristics and beliefs.

The sheer volume of data that a physician can access during a specific clinical encounter proves challenging [28]. As an example, Hill and colleagues [29] found that emergency health care physicians see an average of 2.4 patients per hour and use 4000 mouse clicks in a 10-hour shift. This can result from a combination of poor EHR design and information overload and adds to physician stress [30,31]. Information overload is a part of the 5 main hazards of “information chaos” alongside information underload, information scatter, information conflict, and erroneous information [32]. Clinicians are then required to spend more effort to filter through the information, clarify conflicting documentation, or reassess potentially erroneous information, leading to excess workload and adverse outcomes on not only patient care and health systems but, more importantly, clinician well-being [33].

Gal and colleagues [34] studied this in a pediatric intensive care unit where they calculated that each patient generated an average of 1460 new data points in a 24-hour period. Pediatric intensive care unit attending physicians cared for an average of 11 patients during the day and 22 patients overnight, resulting in exposure to 16,060 (range 11,680-18,980) and 32,120 (range

23,360-37,960) individual data points during the day and night, respectively.

Wanderer and colleagues [35] have described how optimal data visualization in various specialties can lead to improved decision-making for clinicians and more efficient use of their time. Many EHR vendors use visual analytic systems to improve physician workflow and reduce medical errors [36].

Blink rate, measured using eye-tracking technology, has been associated with cognitive workload. Visual tasks that require more focused attention and working memory load have been shown to reduce blink rate [37]. A decreased blink rate has been found to occur in EHR-based tasks that require more cognitive workload [38].

The NASA Task-Load Index (NASA-TLX) is a widely used questionnaire to assess perceived workload (available in Multimedia Appendix 1) [39]. It consists of 6 questions, which can be rated from 1 to 10. Nurses rated their perceived workload from 0 (very low) to 10 (very high).

Using blink rate in addition to the NASA-TLX, Mazur and colleagues [40] tested the implications of the EHR usability interface in a study where they assigned tasks, including the review of medical test results for 20 and 18 individuals using baseline and enhanced EHR versions, respectively, that provided policy-based decision support instructions for next steps. Interestingly, they found that the baseline group had poorer performance and higher cognitive load compared with those who used the enhanced version, suggesting the importance of improving the usability of EHRs to address issues such as clinician burnout and patient safety events.

Harry and colleagues [7] studied the direct relationship between cognitive load with physician burnout in a national sample of US physicians. Using the NASA-TLX, they had responses from 4517 (85.6%) of the 5276 physicians included in the survey. The median age of the physicians was 53 years; 61.8% were male, 37.9% were female, and 0.3% were other gender; and 24 specialties were identified. They identified a dose-response relationship between physician task load and the risk of burnout independent of age, gender, practice setting, and hours worked per week.

To demonstrate a more accurate association between EHR use and stress, Yen and colleagues [41] used blood pulse wave monitoring (previously used as a surrogate for chronic stress) in addition to NASA-TLX on 7 nurses during 132 hours of work. They found that the nursing staff spent 45.54 minutes using EHR during a 4-hour shift, which was much more than the time spent on any other communication or hands-on activities. In addition, the nurses' stress when using EHR was associated with higher perceived physical demand and frustration.

The level of EHR-related burnout has also been shown to be in part influenced by physician specialty. In a large study that used assessing questionnaires among physicians in various specialties with over 25,000 respondents, the investigators found the level of burnout ranged from 22% to 34% by specialty [42]. The specialties with the highest levels of burnout were family medicine (34%) and hematology or oncology (33%). The specialties with the lowest levels of burnout were psychiatry (22%) and anesthesiology (24%). After adjusting for confounding variables, physicians with 5 or fewer hours of weekly out-of-hours charting were twice as likely to report lower levels of burnout than those with 6 or more hours. Those who agree that their organization has performed well with EHR implementation, training, and support were also twice as likely to report lower levels of burnout than those who disagreed. This highlights the importance of training and support following the implementation of EHR for their optimal use.

In a scoping review, Muhiyaddin and colleagues [43] studied the causes and consequences of physician burnout related to the use of EHRs. Reviewing 30 eligible studies out of 500, they identified 6 main causes that are related to physician burnout, including EHR documentation and related tasks, poor design of EHR systems, workload leading to overtime work, inbox alerts, and alert fatigue. Not surprisingly, physician burnout was associated with a low quality of care, behavioral issues, and mental health complications, as well as career dissatisfaction and a reduction in patient safety and satisfaction.

In a survey of 640 clinicians from 3 institutions, with 282 (44.1%) responses to 105 questions, Kroth and colleagues [30] identified 7 EHR design and use factors associated with high stress and burnout. These were information overload, slow system response times, excessive data entry, inability to navigate the system quickly, note bloat, interference with the patient-clinician relationship, fear of missing something, and notes geared toward billing.

Another study [44] aiming to quantify burnout due to the use of HIT used a survey sent to 4197 physicians, where 1792 responded (response rate: 42.7%). They found that HIT-related stress was measurable, prevalent, and specialty related. About 70% of physicians with EHRs experienced HIT-related stress in their sample, and the presence of any of the 3 HIT-related stress measures independently predicted burnout symptoms among respondents. In particular, those with time pressures for documentation or those doing excessive "work after work" on their EHR at home had approximately twice the odds of burnout compared to physicians without these challenges. Time spent after hours on the EHR and the volume of inbox messages have been found to relate to physician exhaustion [45].

Using live observational design and NASA-TLX surveys, Khairat and colleagues [46] assessed the effect of EHRs on emergency department attending and resident physicians' perceived workload, satisfaction, and productivity through completing 6 EHR patient scenarios. They found that EHR frustration levels are significantly higher among more senior attending physicians compared with more junior resident physicians. Among the factors causing high EHR frustrations are (1) remembering menu and button names and commands use; (2) performing tasks that are not straightforward; (3) system speed; and (4) system reliability.

Advantages and Disadvantages of Using EHRs

Overview

As highlighted in the previous section, despite their potential benefits, there have been growing concerns that EHRs also have detrimental effects. Here, we summarize some of the advantages and disadvantages of using EHRs.

Information overload is a significant concern when using EHRs [47-49]. Various studies also suggest a correlation between the usability of the EHR and cognitive load and burnout among clinicians [34,50-52]. Clinicians feel that work-life balance, satisfaction rates, attrition, and burnout are all affected due to the continuous daily interaction with EHR systems [22,53-55].

Advantages

The transition from paper-based medical records to EHRs has been perceived as a positive development in several areas [9,56]. In addition to being easily accessible, EHR systems have been shown to improve communication between clinicians and enhance the continuity of care [57,58]. They can lead to better-informed decisions due to the availability of data and avoid the duplication of diagnostic testing [59]. However, a review of the impact of EHR use on enhancing medication safety, one of the biggest risks to patient care, has shown only modest improvements [60].

EHRs also contain a high volume of clinical data, providing us with multiple opportunities to conduct research and audit [13,59]. A good example of this in the United Kingdom is OpenSAFELY, a secure, transparent, and open-source software platform for the analysis of EHR data [61]. During the COVID-19 pandemic, scientists and statisticians could use the data available on this platform to provide insights into population demographics most at risk of death following COVID-19 infection, which aided with the national policy strategy for prioritizing care [62-65].

Disadvantages

Over the past decade, there has been a reported increase in burnout levels among clinicians, with one potential factor being the introduction of EHR systems [23,34]. The introduction of EHRs has resulted in changes in workflow, with frontline clinicians taking on administrative tasks such as ordering tests, correcting notes, and placing referrals. This has led to increased cognitive load, which is often overlooked [66,67].

On a day-to-day basis, clinicians face time constraints; administrative load; and consequently, elongated workdays. The current documentation methods used in EHRs are under scrutiny by clinicians due to the perceived poor quality of user interfaces, ultimately leading to burnout [52]. Factors such as increased structured documentation requirements, physician order entry, inbox management, and patient portals contribute to more work that is not direct face time with patients [19,68].

Inflated documentation also extends to the excessive use of templates and copy-and-paste workflows in EHR systems that introduce data that are neither required nor accurate [48,49,69]. EHRs allow information to be copied from almost anywhere in the record to another section. This can save time and allow clinicians to focus on clinical tasks rather than documentation; however, it comes with its own challenges. Erroneous information can be copied and pasted without editing, leading to data integrity issues and diagnostic errors [70]. This also creates room for false assumptions and inferred incorrect information between different health care professions, perpetuating previous inaccuracies. It might also sanction junior clinicians to rely solely on readily available information rather than conducting a thorough history and examination for themselves and constructing their own differential thought processes [71].

Although thorough documentation is key for clinical care, there has been a rise in complex and lengthy documentation of content that is required for billing purposes, quality improvement measures, avoiding malpractice, and signs of compliance [72]. In countries such as the United States, the regulatory requirements for data entry beyond what is required for direct patient care can contribute to an increasing workload [73]. Examples of these include collecting data for claim submission, prior authorization, billing, and quality reporting. In addition, a lack of interoperability between EHR systems can result in clinicians not having access to adequate patient information and fragmented care [74]. Often the clinical needs to spend a significant amount of time to obtain this information from various medical records between different health care organizations and sometimes even within one facility. Textbox 1 summarizes some of the advantages and disadvantages of using EHRs.

Textbox 1. Advantages and disadvantages of using electronic health records (EHRs).

Advantages of using EHRs

- Improved communication between clinicians
- Remote access to clinical records enhances care delivery
- Convenient access to patient information for clinicians
- Facilitates research and audit through a high volume of clinical data storage

Disadvantages of using EHRs

- Information overload leading to cognitive overload
- Increased cognitive load due to EHRs can contribute to feelings of exhaustion and burnout
- Continuous interaction with EHRs affects work-life balance and may lead to burnout
- Complex and lengthy documentation required for billing and quality reporting can be cumbersome
- Poor quality of user interfaces in EHRs leads to clinician burnout
- Excessive use of templates and copy-and-paste workflows can lead to data integrity issues
- The lack of interoperability between EHR systems can lead to missed information and duplication of investigations

Overcoming EHR-Related Burnout

Health care organizations and policy makers worldwide are increasingly recognizing the importance of addressing clinician burnout [75]. Here, we have summarized some of the interventions recommended in the literature that can reduce various types of cognitive load and potentially clinician burnout related to EHR use.

Dymek and colleagues [24] have made a case for producing an evidence base to reduce EHR-related clinician burden. Describing documentation, chart review, and inbox tasks as some of the key contributing factors causing burnout, they have made suggestions to help overcome these challenges. Some of these approaches include using speech recognition software and natural language processing to help with documentation and the generation of progress notes; the use of natural language processing and machine learning to process, filter, and rank patient information so that the attention can be paid to where it is most needed; and the use of better inbox design by involving clinicians in their development. Understanding the workflow of the clinicians and involving them in the design of the EHR have been shown to positively impact its usability and user satisfaction [76].

Several studies have reviewed alert burden in EHRs and described potential solutions on how to manage them effectively. One of the very interesting and useful recommended suggestions is developing an Interruptive Alert Stewardship by implementing metrics to assess the alert burden and their effectiveness in improving outcomes [26]. McGreevey and colleagues [77] have comprehensively described reasons for alert fatigue and suggest that organizations develop an alert governance specific to their needs. They propose that key stakeholders including clinicians, informatics, information technology, and administration groups need to participate in developing the alert governance and oversee the design and purpose of alerts. They also recommend using a checklist to assess the purpose and justification of alerts and suggest using metrics to assess their effectiveness and efficiency. Organizations such as Geisinger Health System and Penn Medicine have successfully improved their EHR alert to help with clinician well-being [77].

Clinicians have specific and feasible suggestions for reducing EHR-related burdens, such as providing high-quality EHR training; having an on-site EHR support team; involving support staff or scribes in the documentation process; and, importantly, obtaining physician input and feedback in improving EHRs [27]. Future efforts should focus on implementing the strategies and upgrades requested by these frontline users.

In a recent systematic review, Kang and Sarkar [78] looked at interventions that have been used to reduce EHR-related burnout. The study identified 3 primary interventions, including the use of scribes, EHR training, EHR modifications, and a combination of training and modifications. The use of scribes has been overall well received by clinicians and patients and, in some cases, led to increased productivity, but there were downsides, in particular, the cost, which would be difficult to overcome in smaller centers. EHR training had varying outcomes, with some studies showing a reduction in documentation time, whereas others did not demonstrate this benefit. Nevertheless, subjective EHR proficiency increased, which could help improve clinicians' perception of EHR.

The study [78] has also examined several EHR modification techniques, such as data entry automation technology, improving EHR workflows, reducing unnecessary inbox alerts, and providing support teams to resolve EHR issues promptly. These interventions resulted in positive outcomes, such as a reduction in documentation time ranging from 18.5 to 60%, improved documentation quality and completion rate, decrease in data errors, and subjective EHR usability and satisfaction. However, these positive effects did not lead to a significant reduction in physician burnout. The authors suggest that this could be due to the fact that although EHR enhancements can improve some aspects of the clinician's workflow, they probably do not address the defects in the EHR usability, which contribute to burnout. In addition, there are other factors contributing to burnout, such

as overall workload, organizational culture, and work-life balance that extend beyond EHR systems.

Improving interoperability and health information exchange through understanding the barriers, appropriate incentives, and legislation can facilitate the clinicians' workflow, reduce workload, and enhance patient safety [79].

In 2020, The Office of the National Coordinator for Health Information Technology published a report outlining strategies to reduce regulatory and administrative burden related to the use of HIT and EHR systems [80]. The report focuses on the challenges of EHR and HIT-related burden, which hinder the achievement of interoperability. They recognize that these burdens increase the time and expense clinicians must invest in interacting with EHRs, reducing the value of information, and diverting resources from patient care. They propose a framework for trusted exchange among health information networks to reduce clinician burden while benefiting patients and the health care system.

The National Academy of Medicine published a potential roadmap for EHR optimization and clinician well-being [81]. They have described several strategies currently available that can improve the usability of EHRs, such as EHR optimization, in-basket management techniques, documentation strategies, team-based workflow, and EHR training, as well as the use of artificial intelligence and add-on applications that can help with interoperability, automation, and decision support tools in the future. Gandhi and colleagues [82] have described how the use of artificial intelligence can reduce the cognitive workload by helping with data gathering, documentation, and decision support. They also suggest useful methods to assess the impact of these technologies.

Gaps in the Current Literature

Given the increasingly interconnected digital ecosystem and the complexity of health care systems, which are influenced by physical, emotional, and human factors, it can be challenging to attribute specific outcomes to any particular technology.

Therefore, to maximize the benefits of new tools, it is essential to create frameworks for scientifically assessing the impact of any technology used.

By using user-friendly interfaces, customization options, and context-sensitive information presentation, EHRs can streamline data management [83]. Incorporating decision support tools, data visualization techniques, and smart documentation practices further enhances health care staff's ability to focus on patient care, reducing the risk of errors and burnout [36]. EHR optimization to support clinical workflow and real-life working is a key to uplifting the well-being of health care professionals.

Addressing physician burnout requires systemic changes, including improving work environments with a renewed focus on teamwork, reducing administrative burdens, providing support, and promoting work-life balance within health care organizations.

Individual strategies, such as self-care, stress management, and professional pastoral help, are crucial for clinicians to mitigate and recover from burnout [84]. This will support the well-being of the health care workforce and ensure ongoing high-quality care delivery for all.

As our patients and work environments become more complex and more health technology products become available, it is crucial that we assess their impact through studies and engaging with our health care staff and patients throughout all stages of their development and use [85].

Conclusion

The use of EHR systems may provide benefit for centralizing patient data and simplifying the process of reviewing records, requesting laboratory and imaging tests, and reviewing results, as well as conducting clinical audits, research, and quality improvement projects.

However, there is a noticeable difference in the quality of various EHR systems health care organizations use. Many of these EHR systems do not communicate with each other, keeping data isolated in silos. Our review highlights the cognitive load that their use places on clinical staff, which is not always considered. Improving the design, user interface, and data visualization or retrieval of EHR systems can help to reduce cognitive load, support working memory, and potentially reduce physician workload while enhancing patient care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
NASA Task-Load Index questionnaire.
[DOCX File , 14 KB - medinform_v12i1e55499_app1.docx]

References

1. Sweller J. Cognitive load during problem solvingffects on learning. Cogn Sci 1988 Apr;12(2):257-285. [doi: 10.1016/0364-0213(88)90023-7]
2. Cowan N. The magical mystery four: how is working memory capacity limited, and why? Curr Dir Psychol Sci 2010 Feb 01;19(1):51-57 [FREE Full text] [doi: 10.1177/0963721409359277] [Medline: 20445769]

3. McKinley N, McCain RS, Convie L, Clarke M, Dempster M, Campbell WJ, et al. Resilience, burnout and coping mechanisms in UK doctors: a cross-sectional study. BMJ Open 2020 Jan 27;10(1):e031765 [FREE Full text] [doi: 10.1136/bmjopen-2019-031765] [Medline: 31988223]
4. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med 2018 Jun;283(6):516-529 [FREE Full text] [doi: 10.1111/joim.12752] [Medline: 29505159]
5. Department of Data and Analytics: Division of Data, Analytics and Delivery for Impact. GHE: life expectancy and healthy life expectancy. World Health Organization. 2020. URL: https://tinyurl.com/4zc9csw4 [accessed 2024-04-05]
6. Hajat C, Stein E. The global burden of multiple chronic conditions: narrative review. Prev Med Rep 2018 Dec;12:284-293 [FREE Full text] [doi: 10.1016/j.pmedr.2018.10.008] [Medline: 30406006]
7. Harry E, Sinsky C, Dyrbye LN, Makowski MS, Trockel M, Tutty M, et al. Physician task load and the risk of burnout among US physicians in a national survey. Jt Comm J Qual Patient Saf 2021 Feb;47(2):76-85 [FREE Full text] [doi: 10.1016/j.jcjq.2020.09.011] [Medline: 33168367]
8. Chajut E, Algom D. Selective attention improves under stress: implications for theories of social cognition. J Pers Soc Psychol 2003 Aug;85(2):231-248 [FREE Full text] [doi: 10.1037/0022-3514.85.2.231] [Medline: 12916567]
9. Moy AJ, Schwartz JM, Chen R, Sadri S, Lucas E, Cato KD, et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. J Am Med Inform Assoc 2021 Apr 23;28(5):998-1008 [FREE Full text] [doi: 10.1093/jamia/ocaa325] [Medline: 33434273]
10. Iskander M. Burnout, cognitive overload, and metacognition in medicine. Med Sci Educ 2019 Mar;29(1):325-328 [FREE Full text] [doi: 10.1007/s40670-018-00654-5] [Medline: 34457483]
11. Stone CP. A glimpse at EHR implementation around the world: the lessons the US can learn. Dokumen. 2014 May. URL: https://dokumen.tips/download/link/a-glimpse-at-ehr-implementation-around-the-world-the-glimpse-at-ehr-implementation.html [accessed 2024-04-05]
12. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;Suppl 1(Suppl 1):S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
13. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013 Dec 18;274(6):547-560 [FREE Full text] [doi: 10.1111/joim.12119] [Medline: 23952476]
14. Jha AK, DesRoches CM, Kralovec PD, Joshi MS. A progress report on electronic health records in U.S. hospitals. Health Aff (Millwood) 2010 Oct;29(10):1951-1957 [FREE Full text] [doi: 10.1377/hlthaff.2010.0502] [Medline: 20798168]
15. Crompton P. The National Programme for Information Technology--an overview. J Vis Commun Med 2007 Jun;30(2):72-77 [FREE Full text] [doi: 10.1080/17453050701496334] [Medline: 17671907]
16. The electronic health records system in the UK. Centre for Public Impact. URL: https://www.centreforpublicimpact.org/case-study/electronic-health-records-system-uk [accessed 2024-04-05]
17. Justinia T. The UK's National Programme for IT: why was it dismantled? Health Serv Manage Res 2017 Feb;30(1):2-9 [FREE Full text] [doi: 10.1177/0951484816662492] [Medline: 28166675]
18. A plan for digital health and social care. NHS England. URL: https://tinyurl.com/cp2p2w4m [accessed 2022-06-29]
19. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
20. Williams MS. Misdiagnosis: burnout, moral injury, and implications for the electronic health record. J Am Med Inform Assoc 2021 Apr 23;28(5):1047-1050 [FREE Full text] [doi: 10.1093/jamia/ocaa244] [Medline: 33164089]
21. Bouamrane MM, Mair FS. A study of general practitioners' perspectives on electronic medical records systems in NHSScotland. BMC Med Inform Decis Mak 2013 May 21;13:58 [FREE Full text] [doi: 10.1186/1472-6947-13-58] [Medline: 23688255]
22. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, et al. Electronic medical records and physician stress in primary care: results from the MEMO Study. J Am Med Inform Assoc 2014 Feb;21(e1):e100-e106 [FREE Full text] [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]
23. Khairat S, Coleman C, Ottmar P, Bice T, Carson SS. Evaluation of physicians' electronic health records experience using actual and perceived measures. Perspect Health Inf Manag 2022 Jan 1;19(1):1k [FREE Full text] [Medline: 35440931]
24. Dymek C, Kim B, Melton GB, Payne TH, Singh H, Hsiao CJ. Building the evidence-base to reduce electronic health record-related clinician burden. J Am Med Inform Assoc 2021 Apr 23;28(5):1057-1061 [FREE Full text] [doi: 10.1093/jamia/ocaa238] [Medline: 33340326]
25. Powell L, Sittig DF, Chrouser K, Singh H. Assessment of health information technology-related outpatient diagnostic delays in the US Veterans Affairs health care system: a qualitative study of aggregated root cause analysis data. JAMA Netw Open 2020 Jun 01;3(6):e206752 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.6752] [Medline: 32584406]
26. Chaparro JD, Beus JM, Dziorny AC, Hagedorn PA, Hernandez S, Kandaswamy S, et al. Clinical decision support stewardship: best practices and techniques to monitor and improve interruptive alerts. Appl Clin Inform 2022 May;13(3):560-568 [FREE Full text] [doi: 10.1055/s-0042-1748856] [Medline: 35613913]

27. Nguyen OT, Jenkins NJ, Khanna N, Shah S, Gartland AJ, Turner K, et al. A systematic review of contributing factors of and solutions to electronic health record-related impacts on physician well-being. J Am Med Inform Assoc 2021 Apr 23;28(5):974-984 [FREE Full text] [doi: 10.1093/jamia/ocaa339] [Medline: 33517382]
28. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. J Biomed Inform 2011 Aug;44(4):688-699 [FREE Full text] [doi: 10.1016/j.jbi.2011.03.008] [Medline: 21440086]
29. Hill RG, Sears LM, Melanson SW. 4000 Clicks: a productivity analysis of electronic medical records in a community hospital ED. Am J Emerg Med 2013 Nov;31(11):1591-1594 [FREE Full text] [doi: 10.1016/j.ajem.2013.06.028] [Medline: 24060331]
30. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 02;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
31. Shoolin J, Ozeran L, Hamann C, Bria W. Association of Medical Directors of Information Systems consensus on inpatient electronic health record documentation. Appl Clin Inform 2013 Jun 26;4(2):293-303 [FREE Full text] [doi: 10.4338/ACI-2013-02-R-0012] [Medline: 23874365]
32. Beasley JW, Wetterneck TB, Temte J, Lapin JA, Smith P, Rivera-Rodriguez AJ, et al. Information chaos in primary care: implications for physician performance and patient safety. J Am Board Fam Med 2011;24(6):745-751 [FREE Full text] [doi: 10.3122/jabfm.2011.06.100255] [Medline: 22086819]
33. Patel RS, Bachu R, Adikey A, Malik M, Shah M. Factors related to physician burnout and its consequences: a review. Behav Sci (Basel) 2018 Oct 25;8(11):98 [FREE Full text] [doi: 10.3390/bs8110098] [Medline: 30366419]
34. Gal DB, Han B, Longhurst C, Scheinker D, Shin AY. Quantifying electronic health record data: a potential risk for cognitive overload. Hosp Pediatr 2021 Feb;11(2):175-178 [FREE Full text] [doi: 10.1542/hpeds.2020-002402] [Medline: 33500357]
35. Wanderer JP, Nelson SE, Ehrenfeld JM, Monahan S, Park S. Clinical data visualization: the current state and future needs. J Med Syst 2016 Dec;40(12):275 [FREE Full text] [doi: 10.1007/s10916-016-0643-x] [Medline: 27787779]
36. Rostamzadeh N, Abdullah SS, Sedig K. Visual analytics for electronic health records: a review. Informatics 2021 Feb 23;8(1):12 [FREE Full text] [doi: 10.3390/informatics8010012]
37. Chen S, Epps J. Using task-induced pupil diameter and blink rate to infer cognitive load. Hum Comput Interact 2014 Apr 29;29(4):390-413 [FREE Full text] [doi: 10.1080/07370024.2014.892428]
38. Mosaly PR, Mazur LM, Yu F, Guo H, Derek M, Laidlaw DH, et al. Relating task demand, mental effort and task difficulty with physicians' performance during interactions with electronic health records (EHRs). Int J Hum Comput Interact 2017 Sep 25;34(5):467-475 [FREE Full text] [doi: 10.1080/10447318.2017.1365459]
39. Hart SG. NASA-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2016 Nov 05;50(9):904-908. [doi: 10.1177/154193120605000909]
40. Mazur LM, Mosaly PR, Moore C, Marks L. Association of the usability of electronic health records with cognitive workload and performance levels among physicians. JAMA Netw Open 2019 Apr 05;2(4):e191709 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.1709] [Medline: 30951160]
41. Yen PY, Pearl N, Jethro C, Cooney E, McNeil B, Chen L, et al. Nurses' stress associated with nursing activities and electronic health records: data triangulation from continuous stress monitoring, perceived workload, and a time motion study. AMIA Annu Symp Proc 2019 Mar 4;2019:952-961 [FREE Full text] [Medline: 32308892]
42. Eschenroeder HC, Manzione LC, Adler-Milstein J, Bice C, Cash R, Duda C, et al. Associations of physician burnout with organizational electronic health record support and after-hours charting. J Am Med Inform Assoc 2021 Apr 23;28(5):960-966 [FREE Full text] [doi: 10.1093/jamia/ocab053] [Medline: 33880534]
43. Muhiyaddin R, Elfadl A, Mohamed E, Shah Z, Alam T, Abd-Alrazaq A, et al. Electronic health records and physician burnout: a scoping review. Stud Health Technol Inform 2022 Jan 14;289:481-484. [doi: 10.3233/SHTI210962] [Medline: 35062195]
44. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
45. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 01;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
46. Khairat S, Burke G, Archambault H, Schwartz T, Larson J, Ratwani RM. Perceived burden of EHRs on physicians at different stages of their career. Appl Clin Inform 2018 Apr;9(2):336-347 [FREE Full text] [doi: 10.1055/s-0038-1648222] [Medline: 29768634]
47. What is "cognitive load"—and how can we help clinicians manage it? Nuance. 2022 Aug 11. URL: https://whatsnext.nuance.com/healthcare-ai/cognitive-load-and-impact-on-clinician-burnout/ [accessed 2024-04-05]
48. Downing N, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: are we ignoring the real cause? Ann Intern Med 2018 Jul 03;169(1):50-51. [doi: 10.7326/M18-0139] [Medline: 29801050]

49. Pickering BW, Gajic O, Ahmed A, Herasevich V, Keegan MT. Data utilization for medical decision making at the time of patient admission to ICU. Crit Care Med 2013 Jun;41(6):1502-1510. [doi: 10.1097/CCM.0b013e318287f0c0] [Medline: 23528804]
50. Gawande A. Why doctors hate their computers. New Yorker. 2018 Nov 5. URL: https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers [accessed 2024-04-05]
51. Ash JS, Sittig DF, Dykstra RH, Guappone K, Carpenter JD, Seshadri V. Categorizing the unintended sociotechnical consequences of computerized provider order entry. Int J Med Inform 2007 Jun;76 Suppl 1:S21-S27 [FREE Full text] [doi: 10.1016/j.ijmedinf.2006.05.017] [Medline: 16793330]
52. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc 2004;11(2):104-112 [FREE Full text] [doi: 10.1197/jamia.M1471] [Medline: 14633936]
53. Friedberg MW, Chen PG, Van Busum KR, Aunon F, Pham C, Caloyeras J, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand Health Q 2014 Dec 1;3(4):1 [FREE Full text] [Medline: 28083306]
54. Shanafelt TD, Hasan O, Dyrbye LN, Sinsky C, Satele D, Sloan J, et al. Changes in burnout and satisfaction with work-life balance in physicians and the general US working population between 2011 and 2014. Mayo Clin Proc 2015 Dec;90(12):1600-2613 [FREE Full text] [doi: 10.1016/j.mayocp.2015.08.023] [Medline: 26653297]
55. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012 Oct 08;172(18):1377-1385 [FREE Full text] [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
56. Aziz F, Talhelm L, Keefer J, Krawiec C. Vascular surgery residents spend one fifth of their time on electronic health records after duty hours. J Vasc Surg 2019 May;69(5):1574-1579 [FREE Full text] [doi: 10.1016/j.jvs.2018.08.173] [Medline: 31010521]
57. Ball C, McBeth PB. The impact of documentation burden on patient care and surgeon satisfaction. Can J Surg 2021 Aug 10;64(4):E457-E458 [FREE Full text] [doi: 10.1503/cjs.013921] [Medline: 34388108]
58. Alobo IG, Soyannwo T, Ukponwan G, Akogu S, Akpa AM, Ayankola K. Implementing electronic health system in Nigeria: perspective assessment in a specialist hospital. Afr Health Sci 2020 Jun;20(2):948-954 [FREE Full text] [doi: 10.4314/ahs.v20i2.50] [Medline: 33163063]
59. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
60. Ratwani RM. Modest progress on the path to electronic health record medication safety. JAMA Netw Open 2020 May 01;3(5):e206665 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.6665] [Medline: 32469409]
61. Secure analytics platform for NHS electronic health records. OpenSAFELY. URL: https://www.opensafely.org/ [accessed 2024-04-05]
62. Bhaskaran K, Bacon S, Evans SJW, Bates CJ, Rentsch CT, MacKenna B, et al. Factors associated with deaths due to COVID-19 versus other causes: population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. Lancet Reg Health Eur 2021 Jul;6:100109 [FREE Full text] [doi: 10.1016/j.lanepe.2021.100109] [Medline: 33997835]
63. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020 Aug;584(7821):430-436 [FREE Full text] [doi: 10.1038/s41586-020-2521-4] [Medline: 32640463]
64. Issitt RW, Booth J, Bryant WA, Spiridou A, Taylor AM, du Pré P, et al. Children with COVID-19 at a specialist centre: initial experience and outcome. The Lancet Child & Adolescent Health 2020 Aug;4(8):e30-e31 [FREE Full text] [doi: 10.1016/s2352-4642(20)30204-2]
65. Bourgeois FT, Gutiérrez-Sacristán A, Keller MS, Liu M, Hong C, Bonzel CL, et al. International analysis of electronic health records of children and youth hospitalized with COVID-19 Infection in 6 countries. JAMA Netw Open 2021 Jun 01;4(6):e2112596 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.12596] [Medline: 34115127]
66. Colligan L, Potts HWW, Finn CT, Sinkin RA. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. Int J Med Inform 2015 Jul;84(7):469-476 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.03.003] [Medline: 25868807]
67. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. A model of burnout and life satisfaction amongst nurses. J Adv Nurs 2000 Aug;32(2):454-464 [FREE Full text] [doi: 10.1046/j.1365-2648.2000.01496.x] [Medline: 10964195]
68. Baumann L, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: a systematic review. Health Policy 2018 Aug;122(8):827-836 [FREE Full text] [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
69. Tsou A, Lehmann C, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Appl Clin Inform 2017 Dec 20;26(01):12-34 [FREE Full text] [doi: 10.4338/aci-2016-09-r-0150]
70. Vogel L. Cut-and-paste clinical notes confuse care, say US internists. CMAJ 2013 Dec 10;185(18):E826 [FREE Full text] [doi: 10.1503/cmaj.109-4656] [Medline: 24218539]

71. Cheng CG, Wu DC, Lu JC, Yu CP, Lin HL, Wang MC, et al. Restricted use of copy and paste in electronic health records potentially improves healthcare quality. Medicine (Baltimore) 2022 Jan 28;101(4):e28644 [FREE Full text] [doi: 10.1097/MD.0000000000028644] [Medline: 35089204]
72. Koopman RJ, Steege LMB, Moore JL, Clarke MA, Canfield SM, Kim MS, et al. Physician information needs and electronic health records (EHRs): time to reengineer the clinic note. J Am Board Fam Med 2015;28(3):316-323 [FREE Full text] [doi: 10.3122/jabfm.2015.03.140244] [Medline: 25957364]
73. Tutty M, Carlasare LE, Lloyd S, Sinsky CA. The complex case of EHRs: examining the factors impacting the EHR user experience. J Am Med Inform Assoc 2019 Jul 01;26(7):673-677 [FREE Full text] [doi: 10.1093/jamia/ocz021] [Medline: 30938754]
74. Jacob JA. On the road to interoperability, public and private organizations work to connect health care data. JAMA 2015 Sep;314(12):1213-1215 [FREE Full text] [doi: 10.1001/jama.2015.5930] [Medline: 26393833]
75. National Academies of Sciences, Engineering, and Medicine, National Academy of Medicine, Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. Washington, DC: National Academies Press (US); 2019.
76. Honavar S. Electronic medical records - the good, the bad and the ugly. Indian J Ophthalmol 2020 Mar;68(3):417-418 [FREE Full text] [doi: 10.4103/ijo.IJO_278_20] [Medline: 32056991]
77. McGreevey J, Mallozzi CP, Perkins RM, Shelov E, Schreiber R. Reducing alert burden in electronic health records: state of the art recommendations from four health systems. Appl Clin Inform 2020 Jan;11(1):1-12 [FREE Full text] [doi: 10.1055/s-0039-3402715] [Medline: 31893559]
78. Kang C, Sarkar IN. Interventions to reduce electronic health record-related burnout: a systematic review. Appl Clin Inform 2024 Jan;15(1):10-25 [FREE Full text] [doi: 10.1055/a-2203-3787] [Medline: 37923381]
79. Turbow S, Hollberg JR, Ali MK. Electronic health record interoperability: how did we get here and how do we move forward? JAMA Health Forum 2021 Mar 01;2(3):e210253 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.0253] [Medline: 36218452]
80. U.S. Department of Health and Human Services. Strategy on reducing burden relating to the use of health IT and EHRsng Burden Relating to the Use of Health IT and EHRs. The Office of the National Coordinator for Health Information Technology. 2020. URL: https://tinyurl.com/4z9fv83y [accessed 2024-04-05]
81. Shah T, Kitts AB, Gold JA, Horvath K, Ommaya A, Frank O, et al. Electronic health record optimization and clinician well-being: a potential roadmap toward action. NAM Perspect 2020 Aug 3;2020:10.31478/202008a [FREE Full text] [doi: 10.31478/202008a] [Medline: 35291737]
82. Gandhi TK, Classen D, Sinsky CA, Rhew DC, Vande Garde N, Roberts A, et al. How can artificial intelligence decrease cognitive and work burden for front line practitioners? JAMIA Open 2023 Oct;6(3):ooad079 [FREE Full text] [doi: 10.1093/jamiaopen/ooad079] [Medline: 37655124]
83. Guo U, Chen L, Mehta PH. Electronic health record innovations: helping physicians - one less click at a time. Health Inf Manag 2017 Sep;46(3):140-144 [FREE Full text] [doi: 10.1177/1833358316689481] [Medline: 28671038]
84. Cohen C, Pignata S, Bezak E, Tie M, Childs J. Workplace interventions to improve well-being and reduce burnout for nurses, physicians and allied healthcare professionals: a systematic review. BMJ Open 2023 Jun 29;13(6):e071203 [FREE Full text] [doi: 10.1136/bmjopen-2022-071203] [Medline: 37385740]
85. Examining clinician burnout in healthcare organizations – why it's also an IT concern. Wolters Kluwer. 2023 Apr 12. URL: https://tinyurl.com/3xa4ynef [accessed 2024-04-05]

Abbreviations

EHR: electronic health record
HIT: health information technology
NASA-TLX: NASA Task-Load Index

Edited by C Lovis; submitted 14.12.23; peer-reviewed by R Schreiber, L Ozeran; comments to author 02.01.24; revised version received 15.02.24; accepted 11.03.24; published 12.04.24.

Please cite as:
Asgari E, Kaur J, Nuredini G, Balloch J, Taylor AM, Sebire N, Robinson R, Peters C, Sridharan S, Pimenta D
Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review
JMIR Med Inform 2024;12:e55499
URL: https://medinform.jmir.org/2024/1/e55499
doi:10.2196/55499
PMID:38607672

Health Care Worker Usage of Large-Scale Health Information Exchanges in Japan: User-Level Audit Log Analysis Study

Jun Suzumoto[1,*], Dr med; Yukiko Mori[1,2,3,*], PhD; Tomohiro Kuroda[1,2,3,*], PhD

[1]

[2]

[3]

[*]all authors contributed equally

Corresponding Author:
Jun Suzumoto, Dr med

Abstract

Background: Over 200 health information exchanges (HIEs) are currently operational in Japan. The most common feature of HIEs is remote on-demand viewing or searching of aggregated patient health data from multiple institutions. However, the usage of this feature by individual users and institutions remains unknown.

Objective: This study aims to understand usage of the on-demand patient data viewing feature of large-scale HIEs by individual health care workers and institutions in Japan.

Methods: We conducted audit log analyses of large-scale HIEs. The research subjects were HIEs connected to over 100 institutions and with over 10,000 patients. Each health care worker's profile and audit log data for HIEs were collected. We conducted four types of analyses on the extracted audit log. First, we calculated the ratio of the number of days of active HIE use for each hospital-affiliated doctor account. Second, we calculated cumulative monthly usage days of HIEs by each institution in financial year (FY) 2021/22. Third, we calculated each facility type's monthly active institution ratio in FY2021/22. Fourth, we compared the monthly active institution ratio by medical institution for each HIE and the proportion of cumulative usage days by user type for each HIE.

Results: We identified 24 HIEs as candidates for data collection and we analyzed data from 7 HIEs. Among hospital doctors, 93.5% (7326/7833) had never used HIEs during the available period in FY2021/22, while 19 doctors used them at least 30% of days. The median (IQR) monthly active institution ratios were 0.482 (0.470 - 0.487) for hospitals, 0.243 (0.230 - 0.247) for medical clinics, and 0.030 (0.024 - 0.048) for dental clinics. In 51.9% (1781/3434) of hospitals, the cumulative monthly usage days of HIEs was 0, while in 26.8% (921/3434) of hospitals, it was between 1 and 10, and in 3% (103/3434) of hospitals, it was 100 or more. The median (IQR) monthly active institution ratio in medical institutions was 0.511 (0.487 - 0.529) for the most used HIE and 0.109 (0.0927 - 0.117) for the least used. The proportion of cumulative usage days of HIE by user type was complex for each HIE, and no consistent trends could be discerned.

Conclusions: In the large-scale HIEs surveyed in this study, the overall usage of the on-demand patient data viewing feature was low, consistent with past official reports. User-level analyses of audit logs revealed large disparities in the number of days of HIE use among health care workers and institutions. There were also large disparities in HIE use by facility type or HIE; the percentage of cumulative HIE usage days by user type also differed by HIE. This study indicates the need for further research into why there are large disparities in demand for HIEs in Japan as well as the need to design comprehensive audit logs that can be matched with other official datasets.

(JMIR Med Inform 2024;12:e56263) doi:10.2196/56263

KEYWORDS

health information exchange; audit log; Japan; HIE; audit; logs; usage; medical informatics; rate; hospitals; electronic health record

Introduction

A health information exchange (HIE) is an electronic mobilization system of clinical data across entities such as institutions or organizations, or an organization that controls such systems [1-3]. The appropriate sharing of medical information using HIEs enables fewer duplicated procedures, less duplicated imaging, and fewer total orders. HIE usage is also associated with improved medication reconciliation and immunization [1,4,5]. HIEs have several major features. The first feature enables sharing (ie, sending and receiving) of secure information electronically between care providers to support

coordinated care, known as directed exchange [1,6]. The second enables remote, on-demand viewing or searching of aggregated patient health data from multiple health care institutions. This feature is known as query-based exchange or query-based HIE [1,6,7]. In addition to these two features, consumer-mediated exchange allows patients to aggregate and control the use of their health data among providers [1].

In Japan, HIEs are not widely used. Instead, systems called "Chiiki iryo joho renkei nettowa-ku" have provided features equivalent to HIEs [8-11]. In some literature, these are called "regional health care networks" (RHNs) in English [10]. Past surveys show that over 200 RHNs of various sizes operate in Japan [9]. Most of these RHNs are sponsored by governments or local authorities. According to a Ministry of Health, Labour and Welfare (MHLW) report, 27 RHNs cover entire prefectures, 104 RHNs are within the secondary medical area, 32 are the size of a municipality, and 15 are smaller than a municipality [9]. Item 2.10.2 of the Japan Medical Association Research Institute's 2021 survey [8] investigates the services provided by 229 RHNs, revealing that "sharing of medical data" was the most common service, provided by 190 (83%) RHNs, followed by "sharing of medical images," provided by 187 (81.6%). These features are equivalent to a query-based exchange. In addition, 66 (28.8%) RHNs provide email services and 45 (19.7%) provide electronic patient referral documents, which is equivalent to a directed exchange. Only 9 (3.9%) RHNs provide self-management systems for patients, which is equivalent to a consumer-mediated exchange. In other words, query-based exchange is the most common feature of Japanese HIEs. Since HIEs and RHNs essentially refer to systems with the same features, we will refer to RHNs as HIEs in the subsequent paragraphs. To avoid confusion between computer systems and organizations, we refer to the organization that promotes HIEs as a regional health information organization (RHIO), a term adopted in most literature [7,12].

It is crucial to evaluate the benefits of HIEs for individual health care workers and institutions. Review papers on HIEs have highlighted the importance of understanding whether the system is used [13]. To study the actual use of HIEs, audit logs have often been analyzed [14-25]. The Japanese Association of Healthcare Information Systems Industry published a technical document called "JAHIS's Guide Ver.1.0 on Evaluation Indicators for Regional Medical Collaboration" [26], which emphasizes the importance of evaluating HIE systems using audit log analysis. However, most previous reports [8,9,27] or peer-reviewed journal articles [28-30] about HIE in Japan either did not include audit log analysis or were limited to simple analyses. The MHLW conducted a survey [9] on access to HIEs in 2019. The average monthly active institution ratio based on the MHLW report was 0.381 (SD 0.199) for HIEs connected to 100 or more institutions (Table S1 in Multimedia Appendix 1), meaning that more than half of the connected institutions did not access HIEs. Although the MHLW report suggested low utilization of large-scale HIEs, it did not include a user-level analysis. As there are no studies analyzing the audit logs of multiple HIEs in Japan at the user level, the usage of HIE by individual users or medical institutions remains unknown.

The primary objective of this study was to clarify the extent to which query-based exchange is used by individual health care workers and institutions in large-scale HIEs in Japan by analyzing audit logs at the user level. One reason for investigating only query-based exchange is that, as already mentioned, it is the most common feature of such systems in Japan. The other reason is that, while directed exchange has alternatives such as patient referral letters on paper and consumer-mediated exchange has alternatives such as prescription records on paper, query-based exchange can only be achieved through HIEs. Therefore, analyzing the usage of this feature indicates the significance of HIEs. There are two reasons for investigating only large-scale HIEs. First, it is not realistic to investigate the audit logs of all 200 or more HIEs. Second, large-scale HIEs appear to have spread, as they are accepted by many medical institutions and many patients in the region. When conceptualizing this study, we thought that by investigating large-scale HIEs, we would be able to make suggestions for the increased use of small-scale HIEs.

Methods

Study Design and Data Collection

In this study, we collected data on HIEs that met the following inclusion criteria: (1) they must be included in the list of the survey report, "About the current situation of regional healthcare network" [9] published by the MHLW, and (2) each HIE must be connected to more than 100 institutions and have more than 10,000 patients according to the report above. We asked all RHIOs operating HIEs that met the inclusion criteria to cooperate in this study. When requesting data from each RHIO, we promised to conceal the identity of the RHIO that provided the data in this study. We also agreed to present our published analysis and results in such a way that the RHIO that provided the data would be concealed. This was done to avoid any potential effects that public disclosure of the usage status of each HIE would have on its operation. We obtained data from RHIOs that provided informed consent.

The profile and audit log data of health care workers enrolled in the HIE were collected. In this study, we did not collect patient data from the HIE. Textbox 1 displays the data requested for each HIE; we only received the available data each RHIO could provide. Consequently, datasets and data representation formats differ among HIEs. We also obtained data on the number of connected institutions per month or year for each RHIO. The maximum period of the audit log data was 5 years. We aimed to acquire audit log data from April 1, 2017, to March 31, 2022, but if there were no accumulated data for that period, we asked the RHIO to provide data for the period that could be extracted.

Textbox 1. User data analyzed in this study.

1. Occupation
2. Institution
3. Anonymous identifier
4. Date of account registration and account deletion in health information exchange
5. Date and time of access to query-based exchange
6. Type of data accessed by the user
7. Type of device used for access

Ethical Considerations

This study was approved by the ethics committee of Kyoto University Graduate School and the Faculty of Medicine. The accession number was R3266-7. The disclosure document regarding the research plan and the data to be extracted were published on the Kyoto University Hospital website [31], ensuring that research subjects had the opportunity to opt out. Personal data obtained in the study were pseudonymized by the HIEs that provided the data. Research subjects did not receive compensation.

Measures and Data Analyses

Overview

We refer to "viewing patient medical data using query-based exchange" as "HIE use" in this study. Patient medical data viewed by health care workers can be obtained from multiple storage locations such as hospital electronic medical records (EMRs). If a patient agrees to the disclosure of their medical data stored by the institution, the institution or the RHIO office will take steps to release the stored medical data to HIEs. The types of medical data disclosed by institutions and RHIOs to HIEs vary by institution. Patients can also choose the types of facilities to which their medical data can be disclosed. Several models have been proposed by the MHLW for viewing the medical data disclosed in this way [11]. For example, doctors at a clinic can see which medical tests a patient has previously had when they visit a clinic for the first time. Patients referred from a clinic to a hospital can later check at the clinic the kind of treatment they will receive at the hospital to which they were referred. We investigated how often these types of use cases presented by the MHLW occur by analyzing audit logs. Audit logs for logging into the HIE system are not subject to analysis. Furthermore, the sending and receiving of documents between medical workers using the directed exchange was not included in the analysis.

As a unit for measuring access, 1 man-day was defined as HIE use on 1 day with 1 user account. The cumulative man-days for an institution are the cumulative number of days of HIE use by each user belonging to the institution. For example, consider use by a virtual clinic within a given month. At that clinic, one doctor used HIE on 3 days, and one nurse used HIE on 2 days, which was the clinic's total use of HIE for that month. In this case, the clinic's usage for that month was 5 man-days. If multiple users share a common account, this aggregation methodology may underestimate HIE usage. However, sharing accounts is generally not recommended when using HIEs.

We classified the institutions enrolled in HIEs as hospitals, medical clinics, dental clinics, pharmacies, visiting nursing stations, or nursing facilities. The institutions that could not be classified into these categories were excluded from the analysis; for example, this study did not analyze public institutions such as fire departments, public health centers, local medical associations, or vendors that developed HIEs.

In this study, the financial year (FY) is from April 1 of one year to March 31 of the next year. For example, FY2021/22 started on April 1, 2021, and ended on March 31, 2022. We used R (version 4.3.1; R Foundation for Statistical Computing) to perform the analysis.

Percentage of Days of HIE Use by Each Hospital Doctor in FY2021/22

For each user account of doctors affiliated with hospitals, we calculated the ratio of the number of days of active HIE use. We analyzed the data of HIEs that met the following two criteria: (1) audit log data for all periods in FY2021/22 were available, and (2) each user's date of account registration and date of account deletion in the HIE were available (Textbox 1). For all HIEs that met the criteria, the annual number of days of HIE use by each hospital doctor's account in FY2021/22 was counted. Next, we calculated the number of days that the doctor could use HIEs in FY2021/22. For each doctor's account that was registered with the HIE during FY2021/22, we subtracted from 365 the number of days from April 1, 2021, to the day before the account registration date. For accounts that were removed from the HIE during FY2021/22, we subtracted the number of days from the day after the account deletion date to March 31, 2022, from the remaining number of days. The number of days remaining after these subtractions is the number of days that the doctor could use the HIE in FY2021/22. For accounts that were able to use the HIE on all days in FY2021/22, we used 365 as the number of days that the account could use the HIE. We then calculated the ratio of days of HIE use by each hospital doctor. The ratio of days of HIE use was defined as follows:

(1) Data number of days of HIE use in FY2021/22 / Number of days doctor could use HIE in FY2021/22

Man-Days for Monthly HIE Use by Each Institution in FY2021/22

We calculated the man-days for monthly HIE use by each institution for each month and aggregated them by facility type. We analyzed the data of HIEs that met the following two criteria: (1) the audit log data for all periods in FY2021/22 were available, and (2) data were available on the number of participating institutions by facility type, matching our facility classification. For each institution belonging to any HIEs that meet the criteria, man-days for monthly HIE use in FY2021/22 were aggregated. Next, by each facility type, we tallied the number of months for each man-day group divided into 5 or 10 increments. Finally, for each facility type, the percentage of each man-day group was calculated.

Monthly Active Institution Ratio in FY2021/22

We calculated the monthly active institution ratio for each facility type, defining it as follows:

$$\frac{\text{Number of institutions that used HIE on the last day of each month}}{\text{Number of institutions participating in HIE on the last day of each month}}$$

We analyzed the data of HIEs that met the following two criteria: (1) audit log data for all periods in FY2021/22 were available, and (2) data on the number of participating institutions by facility type, which matched our facility classification, were available. For each facility type, the total number of participating institutions for the last day of each month in FY2021/22 was aggregated in all HIEs that met the criteria. This corresponds to the denominator of Equation 2. Next, all months in FY2021/22 and all institutions were flagged as to whether they used HIE. If at least one account within an institution used HIE for at least one day in a month, the institution was deemed to have used the HIE that month. Subsequently, for each facility type, we calculated the sum of the institutions that used HIE for each month. This corresponds to the numerator in Equation 2. Finally, for each facility type, we calculated the active institution ratio for each month using Equation 2.

Within facility types, hospitals were further classified based on the number of beds. Hospitals were divided into three categories: those with ≥200 beds, those with 100-199 beds, and those with ≤99 beds, and Equation 2 was calculated for each classification.

Monthly Active Institution Ratio of Medical Institutions and Man-Days of HIE Use for Each User Type

We compared the monthly active institution ratios at medical institutions for each HIE and the proportion of man-days of HIE use by user type. Equation 2 defines the monthly active institution ratio. “Medical institutions” include the facility types “hospital,” “medical clinic,” and “dental clinic.” We analyzed all HIEs during all periods for which data could be obtained.

The total number of medical institutions participating in each HIE on the last day of each month was calculated. Next, for each HIE in each month, each medical institution of each HIE was flagged as to whether it used the HIE. If at least one account from within the institution used the HIE for at least one day in a month, the institution was deemed to have used HIE that month. Subsequently, for each HIE in each month, we calculated the sum of the institutions that used it. We then calculated the active institution ratio of medical institutions for each HIE for each month, according to Equation 2.

Next, we classified all user occupation data into 8 user types: “doctor,” “nurse,” “rehabilitation staff,” “pharmacist,” “dental profession,” “nursing care staff,” “other medical professions,” and “type unknown.” We aggregated the total number of man-days of HIE use by user type and calculated the proportion of man-days by user type.

Results

Data Collection

The MHLW report listed 218 HIEs. However, numbers 57-60 in the report are federated, and the aggregated statistics are shown as number 61. After removing these duplicates, there were 214 HIEs listed [9]. Of the 214 HIEs, 36 HIEs were connected to more than 100 institutions and 45 were connected with more than 10,000 patients. Overall, 21 HIEs met both inclusion criteria. Considering number 61 to be 4 HIEs, 24 HIEs were considered candidates for data collection.

Initially, we requested research cooperation from each RHIO administrator via email. The data request document is shown in Multimedia Appendix 2. Thereafter, we requested each RHIO to provide data through a web conference once we were able to have detailed discussions. We obtained research data from 8 HIEs. Although these tasks were sometimes performed free of charge, the extraction of access logs from 2 HIEs was performed for a fee. One of the 8 HIEs was unable to extract comprehensive audit log data of query-based exchange and was removed from the final analysis. The flow diagram is depicted in Figure 1. Among the 24 HIEs that met the inclusion criteria for this study, we calculated the average and standard deviation of the monthly active institution ratio based on the MHLW report separately for the HIEs that were included in the final analysis and those that were not. The results are shown in Table 1.

To avoid identification, all RHIOs and HIEs were assigned a pseudonym using letters from A to G. We obtained the audit log data from RHIO A, RHIO C, and RHIO E from April 2017 to March 2022. We obtained audit log data from April 2018 to March 2022 from RHIO B. From RHIO F and RHIO G, we obtained audit log data from April 2021 to March 2022. We obtained audit log data from October 2021 to March 2022 from RHIO D. HumanBridge [32] (developed by Fujitsu Limited) has an audit log extraction feature added by default. Therefore, audit log extraction in RHIOs that used HumanBridge was performed using this feature. The extraction of audit logs from HIEs other than HumanBridge was conducted by requesting the system vendor to extract audit logs from the RHIO secretariat.

Figure 1. Flow diagram of HIE selection. HIE: health information exchange. *HIE number 61 was considered to be 4 HIEs.

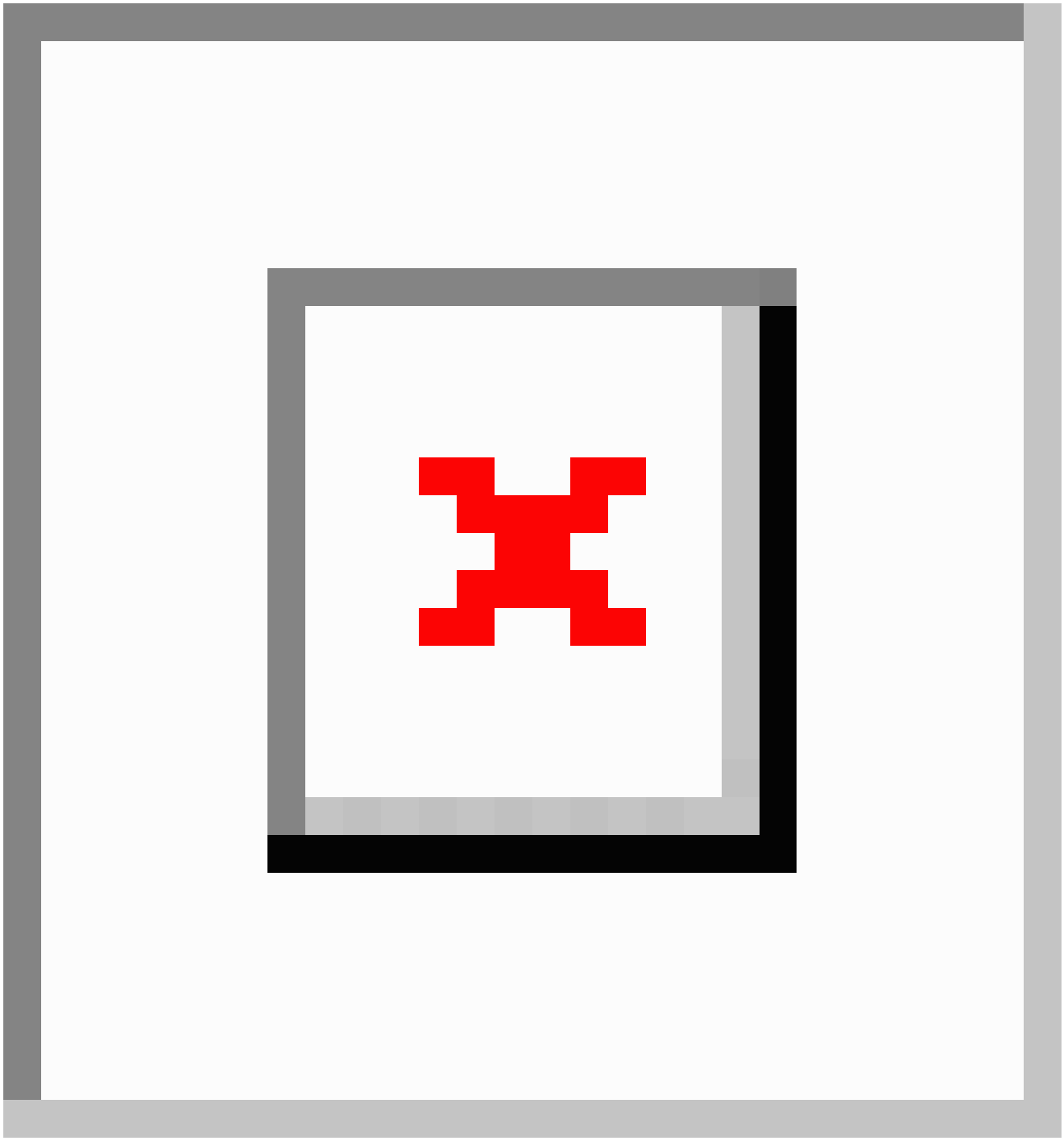


Table . Average monthly active institution ratio based on the MHLW[a] report.

	Health information exchange[b], n	Monthly active institution ratio based on the MHLW report (%)[c], mean (SD)
Included in the final analysis	7	46.1 (24.4)
Not included in the final analysis	17	46.4 (19.1)

[a]MHLW: Ministry of Health, Labour and Welfare.

[b]Among the health information exchanges listed in the MHLW report, those connected to more than 100 institutions and with more than 10,000 patients were included in this analysis.

[c]For each health information exchange, the MHLW report lists the number of participating medical institutions and the number of medical institutions that accessed health information exchange (ie, the number of institutions that used health information exchange during the month covered by the survey). We divided the number of medical institutions that accessed health information exchange by the number of participating medical institutions for each health information exchange and called this number “monthly active institution ratio based on the MHLW reports.”

Of the data items we attempted to obtain in advance (Textbox 1), the type of device used for access was not collected by any HIEs. Consequently, we could not analyze the type of device with which the HIE was accessed. We obtained users' date of account registration and account deletion in the HIE from only 3 RHIOs. For the remaining 4 HIEs, we could not determine the exact number of registrants in each period.

Characteristics of HIEs Included in the Final Analysis

All 7 HIEs began operations between 2010 and 2015. Of the 24 HIEs that met the inclusion criteria, 9 did not have membership fees, and of the 7 HIEs included in the final analysis, 3 did not have participation fees for connected institutions. Overall, 5 RHIOs adopted ID-Link for the HIEs they operate, 3 RHIOs adopted HumanBridge, and 2 RHIOs employ products other than these. The reason that the sum of these products is more than 7 is that some RHIOs operate multiple products in parallel.

For HIE A to G, the patient consent rates for HIE in 2022 were 21.2%, 5.6%, 69.5%, 1.9%, 2%, 4%, and 6.5%. Patient consent rate was obtained by dividing the number of patients connected with the HIE by the population of the area. All 6 HIEs except HIE C required patients to complete a paper consent form for their medical data to be viewed by health care workers using the HIE. HIE C uses paper consent forms as well as the "patient demographic data synchronization feature" provided by ID-Link. For institutions that disclose patient data to HIE C, basic patient profiles such as name, date of visit, and public insurance data are automatically accumulated in the HIE. This feature allows health care workers to use query-based exchange to obtain patient data in the event of emergency treatment, even if the patient cannot explicitly consent to the use of the HIE in advance. Patient enrollment in an HIE using this feature is opt-out, that is, individuals are considered to implicitly consent to participate in the HIE unless participation is explicitly declined. Therefore, the apparent consent rate of HIE C is extremely high.

Percentage of Days of HIE Use by Each Hospital Doctor in FY2021/22

A total of 3 HIEs met the criteria: HIE A, HIE B, and HIE C. The number of hospital doctor accounts registered in HIEs operated by these 3 HIEs in FY2021/22 was 7833. For each of these doctor accounts, we calculated the percentage of days of HIE use according to Equation 1. The overall results are shown in Table 2.

Table . Percentage of days of HIE[a] use by doctors affiliated with the hospital.

Days of HIE use, %	Hospital doctors, n
0	7326
>0 and ≤5	412
>5 and ≤10	39
>10 and ≤15	11
>15 and ≤20	17
>20 and ≤25	5
>25 and ≤30	4
>30 and ≤35	5
>35 and ≤40	5
>40 and ≤45	1
>45 and ≤50	1
>50	7

[a]HIE: health information exchange.

Man-Days for Monthly HIE Use by Each Institution in FY2021/22

Different HIEs met the criteria for each facility type as shown in Table S2 in Multimedia Appendix 1. The cumulative number of months of hospital participation in HIEs in FY2021/22 was 3434. The distribution of man-days for monthly HIE use by hospitals is shown in Table 3. Table 4 shows the analysis results for facility types other than hospitals. The number of institutions connected to the HIEs included in the analysis for each month in FY2021/22 is shown in Table S3 in Multimedia Appendix 1.

Table . Distribution of man-days for monthly HIE[a] use for hospitals.

Man-days for monthly HIE use	Cumulative number of months, n
0	1781
1-10	921
11-20	287
21 - 30	129
31 - 40	67
41 - 50	59
51 - 60	28
61 - 70	14
71 - 80	11
81 - 90	21
91 - 100	13
≥101	103

[a]HIE: health information exchange.

Table . Distribution of man-days for monthly HIE[a] use by medical care institutions other than hospitals.

Man-days for monthly HIE use per institution	Months per institution, n (%)				
	Medical clinic (7791 months)	Dental clinic (1006 months)	Pharmacy (5140 months)	Visiting nursing station (983 months)	Nursing facility (3629 months)
0	5914 (75.9)	971 (96.5)	4081 (79.4)	672 (68.4)	2781 (76.6)
1 - 5	1063 (13.6)	31 (3.1)	733 (14.3)	134 (13.6)	452 (12.5)
6 - 10	261 (3.4)	1 (0.1)	116 (2.3)	42 (4.3)	126 (3.5)
11 - 15	177 (2.3)	1 (0.1)	57 (1.1)	40 (4.1)	73 (2)
16 - 20	137 (1.8)	2 (0.2)	80 (1.6)	19 (1.9)	48 (1.3)
21 - 25	138 (1.8)	0 (0)	16 (0.3)	3 (0.3)	52 (1.4)
≥26	101 (1.3)	0 (0)	57 (1.1)	73 (7.4)	97 (2.7)

[a]HIE: health information exchange.

Monthly Active Institution Ratio in FY2021/22

Table S2 in Multimedia Appendix 1 lists the HIEs that met the criteria in this section. The median (IQR) monthly active institution ratios were 0.482 (0.470 - 0.487) in hospitals, 0.244 (0.231 - 0.247) in medical clinics, 0.030 (0.024 - 0.048) in dental clinics, 0.202 (0.188 - 0.216) in pharmacies, 0.307 (0.301 - 0.325) at visiting nursing stations, and 0.197 (0.185 - 0.204) in nursing facilities. We illustrated the monthly active institution ratios using box plots in Figure 2.

Table 5 shows the monthly active hospital rate for each category subdivided by the number of hospital beds. HIE F data could not be combined with hospital bed data and was therefore excluded from the analysis.

Figure 2. Monthly active institution ratio of health information exchange categorized by facility type.

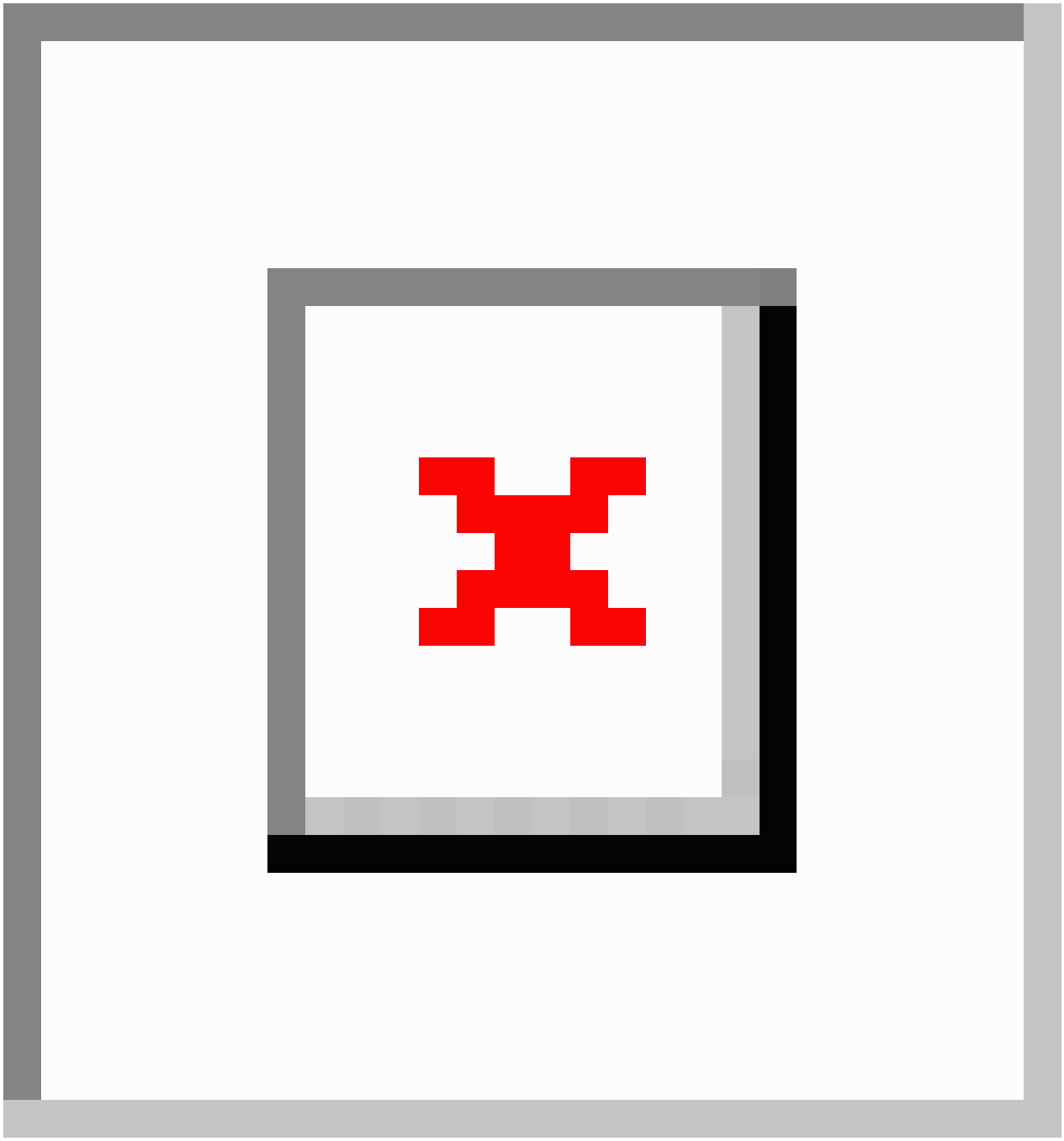


Table . Monthly active hospital ratio subcategorized by the number of hospital beds per institution.

Hospital beds, n	Monthly active institution ratio (%), median (IQR)
≤99	34.2 (30.9 - 39.5)
100 - 199	65.2 (65.2 - 69.6)
≥200	75.0 (72.9 - 77.1)

Monthly Active Institution Ratio of Medical Institutions and Man-Days of HIE Use for Each User Type

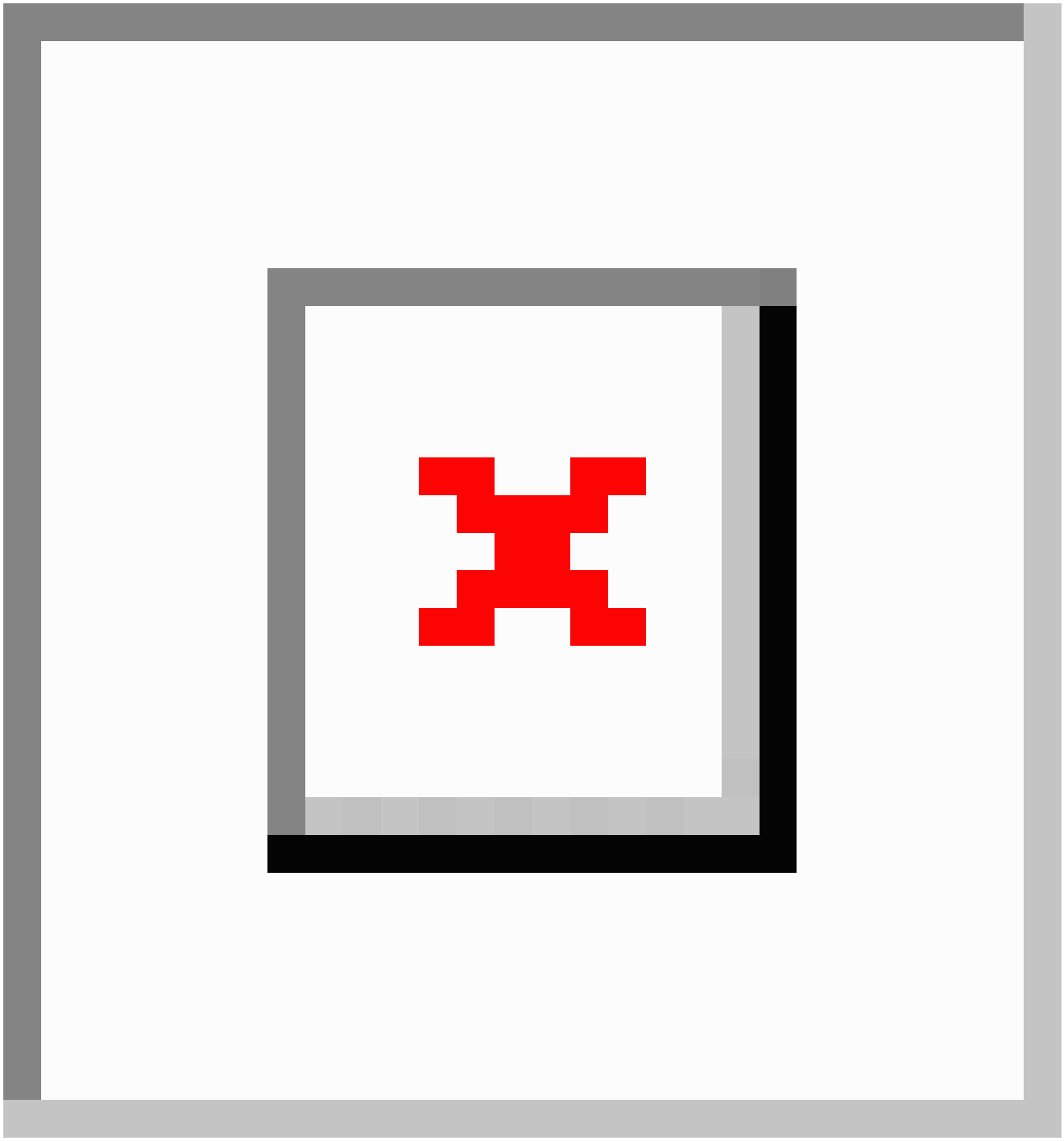
We analyzed all 7 HIEs. As the period of audit log data obtained differs for each HIE, the analysis period also differs. A total of 5 HIEs included all types of medical institutions; however, no dental clinics participated in HIE D and HIE E. Regarding user type data, precise data were not available for HIE D. Occupational data for HIE F could only be obtained for "doctors" and "other medical professions."

Some HIEs have restrictions on the types of users that can use the HIE. In 2 HIEs, users affiliated with medical clinics could only use the HIEs if they were doctors. In one HIE, users who were affiliated with hospitals that did not disclose patient data to the HIE could only use the HIE if they were doctors. In addition, 6 HIEs had no restrictions on the type of data that authorized health care workers could view. One HIE was set up so that people other than doctors affiliated with the hospital could not view outpatient treatment data.

We illustrated the monthly active institution ratio of HIE and the proportion of man-days of HIE use by user type in medical institutions using box plots and bar graphs (Figure 3).

Monthly active institution ratios of HIE in medical institutions are also shown in Table S4 in Multimedia Appendix 1. The proportions of man-days of HIE use by user type in medical institutions are also shown in Table S5 in Multimedia Appendix 1.

Figure 3. (A) Monthly active institution ratio and (B) proportion of man-days of HIE use by user type for each HIE. HIE: health information exchange.



Discussion

Overview of HIE Use

As already mentioned, the low utilization of large-scale HIEs has previously been suggested by MHLW reports (Table S1 in Multimedia Appendix 1). Our analysis included data from only 18% (7/32) of such HIEs. However, among HIEs that met the inclusion criteria, monthly active institution ratios based on the MHLW report were similar for HIEs included in the final analysis and those not included (Table 1). This suggests that the results of this study apply to some extent to large-scale HIEs in general. However, HIEs with fewer participating institutions have a higher monthly active institution ratio based on the MHLW report than large-scale HIEs (Table S1 in Multimedia Appendix 1). Therefore, it is possible that small-scale HIEs are used more actively than the results of this study suggest. However, the MHLW report includes 13 HIEs where the number of participating medical institutions is 1 or 2, and the number of participating institutions is equal to the number of institutions that have accessed HIEs. The active institution ratio based on the MHLW report should be interpreted with caution because extremely small-scale HIEs are driving up the ratio.

Across the large-scale HIEs analyzed in this study, many health care workers and institutions did not use query-based exchange. These results are consistent with MHLW reports (Table S1 in Multimedia Appendix 1). We found that 93.5% (7326/7833) of the doctors at hospitals registered with HIEs do not use them even once per year (Table 2). In addition, 51.9% (1781/3434) of hospitals did not use query-based exchange even once per month (Table 3). This is lower than the values reported by previous studies conducted in other countries [20,33]. The monthly active institution ratio is higher for hospitals with more beds. The median (IQR) monthly active institution ratio for hospitals with ≤99 beds is 34.2% (30.9%-39.5%), but it is 75% (72.9%-77.1%) for hospitals with ≥200 beds (Table 5). Among facilities other than hospitals, the monthly active institution ratio is even lower than for hospitals with ≤99 beds. The median (IQR) monthly active institution ratio for visiting nursing stations reached 30.7% (30.1% - 32.5%), but it was only approximately 20% for medical clinics, pharmacies, and nursing care facilities (Figure 2). As for dental clinics, the median (IQR) monthly active institution ratio was only 3% (2.4%-4.8%). Previous studies outside Japan have also revealed that HIEs are not often used by dental practices [34]. This is the first study to reveal the active institution ratio of HIE by facility type in Japan.

Where query-based exchange was used, most people and institutions only used it for a limited number of days. Previous reports have not provided user-level analysis; therefore, this study is the first to reveal the total number of days of HIE use by health care workers and institutions. Of the 507 hospital-affiliated doctor accounts that actively used query-based exchange in FY2021/22, we found that 81.3% (412/507) used it for 5% or fewer days (Table 2). In other words, assuming the average doctor works 20 days per month, most of these doctors use query-based exchange less than once per month. As the percentage of days of HIE use increases, the number of corresponding hospital doctor accounts tends to decrease. This trend is reflected in man-days for monthly HIE use of hospitals. We found that 90.8% (3118/3434) of all hospitals use HIE for 30 or fewer man-days per month (Table 3). In these hospitals, query-based exchange is used by less than one user daily. The number of man-days of HIE use in hospitals also shows a tendency for the number of applicable months to decrease as the number of man-days increases. This trend remains true for man-days for monthly HIE use at facilities other than hospitals (Table 4). However, some institutions and users use query-based exchange for many days. Of the hospital-affiliated physician accounts, 19 users used query-based exchange for 30% or more days. Of the cumulative months of hospital participation in HIEs in FY2021/22, we found that 3% (103/3434) had over 101 man-days for HIE use (Table 3). This exceeds the number of months when man-days of monthly HIE use are in the range of 91-100. For visiting nursing stations and nursing care facilities, the number of months in which the number of man-days for HIE use exceeds 26 is greater than the number of months in which the number of man-days is 21-25 (Table 4), showing there are significant disparities in HIE use across institutions and users.

Monthly active institution ratios for medical institutions vary widely by HIE. HIE A, which has the highest monthly active institution ratio, has a median rate of 50.9% (IQR 48.7%-52.9%), but some HIEs have a rate of over 10% (Figure 3, Table S4 in Multimedia Appendix 1). The proportion of man-days of HIE use by each user type is not constant for each HIE. However, regarding the HIEs that could be confirmed, the number of man-days of HIE use was low for dental professionals, pharmacists, and rehabilitation workers.

Possible Factors Influencing HIE Use

Many other factors could have influenced monthly active institution ratios and man-days of monthly HIE use, for example, the system used and whether there are membership fees and usage restrictions based on use type. The data viewed by health care professionals when using HIEs is also extremely important. None of these can be shown as individual HIE data due to privacy considerations; therefore, they cannot be discussed in relation to the results shown in Figure 3. This is a significant limitation of this study and indicates the need for further research into why there are such large disparities in demand for HIE in Japan.

Although a detailed elucidation must be reserved for future research, two factors may have influenced the monthly active institution ratio of facilities in the HIEs in our analysis. One is the consent rate of patients to participate in HIEs. HIE A, which had the highest monthly active institution ratio among medical institutions in this analysis, had a relatively high patient consent rate of over 21%. HIE C, which had the second highest monthly active institution ratio, has a partial opt-out policy and a very high consent rate. However, HIEs G, D, and B, where the monthly active institution ratio was 20% or less, had patient consent rates of 7% or less. The patient consent rate in HIEs E and F, where the monthly active institution ratio was in the 20% range, was less than 5%, and therefore lower than in G and B. High consent rates may have contributed to the high active

institution ratios for HIEs A and C, but this study cannot determine whether these factors are causally related.

Another factor that may have influenced the monthly active institution ratio is the number of staff at each participating institution. As already shown, the active institution ratio was higher in hospitals with ≤99 beds than in medical clinics, and it was higher in hospitals with ≥200 beds than in hospitals with ≤99 beds. It is natural to assume that this difference is caused by the absolute number of staff working at each institution. Therefore, when considering the active institution ratio for a given HIE, the value is likely to be high if a large proportion of the institutions participating in the HIE are large hospitals.

Audit Log for Further Research Analysis

To analyze HIE use in detail, proper design of the audit log is extremely important. When analyzing the audit logs in this study, two characteristics of some audit logs posed obstacles. One was that the extracted audit logs are not comprehensive. In HIEs that were configured using products from multiple vendors [35], each product generally had its own unique audit log design and storage. To perform detailed analysis of the usage of such HIEs, it was necessary to extract the logs from each product. This was difficult when each product was controlled by a different institution; specifically, the HIE platform system is managed by the RHIO, but the EMR data viewing system may be managed by each hospital. In this case, we need to obtain consent for research collaboration from each institution to perform overall log extraction. If user access to individual systems via the platform system is recorded in the audit log, analyzing the general usage status may be possible by extracting only the platform system's audit log. However, in practice, access to individual systems is not necessarily recorded in the platform system's audit log. In HIEs configured using products from multiple vendors, careful attention must be paid to facilitating comprehensive log extraction.

The second obstacle was the incompatibility of the institution IDs and user IDs used in HIEs. To clarify the factors that create the disparities in HIE usage across users and across institutions, more detailed data on users and medical institutions were required, such as medical specialty and whether institutions provide acute or chronic care. However, such data are not generally included in the audit log itself; therefore, log data needs to be cross-referenced with institution IDs compiled in the master dataset by the government or the detailed user data of each institution. As the institution IDs used in the audit log data extracted in this study did not necessarily correspond to the institution IDs assigned by the MHLW, it was difficult to cross-reference log data with other datasets. To investigate HIE usage in greater depth than this study, it is recommended to use master data that can be matched with EMR and official datasets when designing audit logs.

Limitations

This study had several limitations. The most significant limitation was the need to maintain the anonymity of the HIEs included in the analysis. Therefore, important data, such as the systems employed by individual HIEs and the types of medical data disclosed, were either kept private or disclosed anonymously. Consequently, it was almost impossible to analyze the causes of differences in active institution ratios for individual HIEs from the data. We also attempted to evaluate the viewing situation for all data types, such as images and prescriptions. However, it was difficult to perform a comprehensive analysis because the data storage format was not standardized for each HIE. The list of HIEs included in the analysis differed for each analysis because the data items that could be obtained differed for each HIE. Data regarding the type of device used to access the HIE could not be obtained from any HIE.

Some HIEs have features other than viewing patient medical data, such as sending and receiving documents or messages [8]. Previous reports indicated that some HIEs actively used these additional features when treating COVID-19 patients [27]. As this study focused on query-based exchange, we did not perform a quantitative analysis of the usage of other features. Another study is required on the actual usage of features other than query-based exchange.

This study revealed that most users do not use query-based exchange or use it infrequently, but it is impossible to prove whether this is due to a lack of patients' medical data in the HIE repository or a lack of need to view data. As mentioned above, HIEs vary widely in both patient consent rates and the types of data that health care professionals can view. Therefore, it is difficult to provide a single answer to this remaining question. To answer this question, a deeper investigation of each HIE is required, using more detailed audit log analysis, system descriptions, and qualitative research.

Conclusions

In the large-scale HIEs surveyed in this study, the overall usage of the on-demand patient data viewing feature was low, consistent with past MHLW reports. User-level analysis of audit logs revealed large disparities in the number of days of HIE use among health care workers and institutions. There were also large disparities in HIE use by facility type or HIE, and the percentage of cumulative HIE usage days by user type also differed by HIE. This study indicates the need for further research into why there are large disparities in demand for HIEs in Japan, as well as the need to design comprehensive audit logs that can be matched with other official datasets.

Acknowledgments

This work was supported by JST, CREST grant number JPMJCR21M1, Japan. We would like to thank Tetsuya Otsubo, Naoki Ishizuka, Taichi Hatta, and Hiroshi Yamazaki for the useful discussion. We also would like to thank Editage for English language editing.

Data Availability

The datasets analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed tables of data analysis results.
[DOCX File, 30 KB - medinform_v12i1e56263_app1.docx]

Multimedia Appendix 2
Request letters to health information exchange operators [document in Japanese].
[PDF File, 370 KB - medinform_v12i1e56263_app2.pdf]

References

1. What is HIE? HealthIT.gov. URL: https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/what-hie [accessed 2023-07-28]
2. Hersh W, Totten A, Eden K, et al. Health information exchange. Evid Rep Technol Assess (Full Rep) 2015 Dec(220):1-465. [doi: 10.23970/AHRQEPCERTA220] [Medline: 30307736]
3. Akhlaq A, Sheikh A, Pagliari C. Health information exchange as a complex and adaptive construct: scoping review. J Innov Health Inform 2017 Jan 25;23(4):889. [doi: 10.14236/jhi.v23i4.889] [Medline: 28346129]
4. Menachemi N, Rahurkar S, Harle CA, Vest JR. The benefits of health information exchange: an updated systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1259-1265. [doi: 10.1093/jamia/ocy035] [Medline: 29718258]
5. Sadoughi F, Nasiri S, Ahmadi H. The impact of health information exchange on healthcare quality and cost-effectiveness: a systematic literature review. Comput Methods Programs Biomed 2018 Jul;161:209-232. [doi: 10.1016/j.cmpb.2018.04.023] [Medline: 29852963]
6. Vest JR, Unruh MA, Casalino LP, Shapiro JS. The complementary nature of query-based and directed health information exchange in primary care practice. J Am Med Inform Assoc 2020 Jan 1;27(1):73-80. [doi: 10.1093/jamia/ocz134] [Medline: 31592529]
7. Vest JR, Hilts KE, Ancker JS, Unruh MA, Jung HY. Usage of query-based health information exchange after event notifications. JAMIA Open 2019 Oct;2(3):291-295. [doi: 10.1093/jamiaopen/ooz028] [Medline: 31984363]
8. Overview of the nationwide regional healthcare network using ICT (2021 Edition) and urgent survey on the continuation of the regional healthcare network. ICT wo riyou sita zenkoku chiiki iryo joho nettowa-ku no gaikyou (2021 nendo ban) oyobi chiki iryo renkei nettowa-ku sonzoku ni kannsuru kinkyu chosa [Website in Japanese]. Japan Medical Association Research Institute. 2021. URL: https://www.jmari.med.or.jp/result/working/post-3560/ [accessed 2023-07-28]
9. About the current situation of a regional healthcare network. Chiki iryo joho renkei nettowaku no genzyo ni tuite [Website in Japanese]. Ministry of Health Labour and Welfare. 2019. URL: https://www.mhlw.go.jp/content/10800000/000683765.pdf [accessed 2023-07-24]
10. Ito A, Tanno T, Okumura T. Strategies for overcoming regional healthcare network stagnation issues. Chiiki iryo nettowa-ku no teitai mondai no kokuhuku ni muketa senryaku [Article in Japanese]. Oukan 2022;16(2):34-45. [doi: 10.11487/trafst.16.2_34]
11. What is a healthcare network? Iryo joho renkei nettowa-ku toha [Website in Japanese]. Ministry of Health, Labour and Welfare. URL: https://www.mhlw.go.jp/content/10808000/000644575.pdf [accessed 2024-03-28]
12. Adler-Milstein J, Landefeld J, Jha AK. Characteristics associated with regional health information organization viability. J Am Med Inform Assoc 2010;17(1):61-65. [doi: 10.1197/jamia.M3284] [Medline: 20064803]
13. Vest JR, Jasperson J. What should we measure? Conceptualizing usage in health information exchange. J Am Med Inform Assoc 2010;17(3):302-307. [doi: 10.1136/jamia.2009.000471] [Medline: 20442148]
14. Mullins AK, Skouteris H, Rankin D, Morris H, Hatzikiriakidis K, Enticott J. Predictors of clinician use of Australia's national health information exchange in the emergency department: an analysis of log data. Int J Med Inform 2022 May;161:104725. [doi: 10.1016/j.ijmedinf.2022.104725] [Medline: 35231719]
15. Vest JR, Zhao H, Jasperson J, Gamm LD, Ohsfeldt RL. Factors motivating and affecting health information exchange usage. J Am Med Inform Assoc 2011;18(2):143-149. [doi: 10.1136/jamia.2010.004812] [Medline: 21262919]
16. Vest JR. Health information exchange and healthcare utilization. J Med Syst 2009 Jun;33(3):223-231. [doi: 10.1007/s10916-008-9183-3] [Medline: 19408456]
17. Bailey JE, Pope RA, Elliott EC, Wan JY, Waters TM, Frisse ME. Health information exchange reduces repeated diagnostic imaging for back pain. Ann Emerg Med 2013 Jul;62(1):16-24. [doi: 10.1016/j.annemergmed.2013.01.006] [Medline: 23465552]

18. Campion TRJ, Edwards AM, Johnson SB, Kaushal R, HITEC investigators. Health information exchange system usage patterns in three communities: practice sites, users, patients, and data. Int J Med Inform 2013 Sep;82(9):810-820. [doi: 10.1016/j.ijmedinf.2013.05.001] [Medline: 23743323]
19. Campion TR, Vest JR, Ancker JS, Kaushal R, HITEC Investigators. Patient encounters and care transitions in one community supported by automated query-based health information exchange. AMIA Annu Symp Proc 2013;2013:175-184. [Medline: 24551330]
20. Johnson KB, Gadd CS, Aronsky D, et al. The MidSouth eHealth Alliance: use and impact in the first year. AMIA Annu Symp Proc 2008 Nov 6;2008:333-337. [Medline: 18999184]
21. Kern LM, Ancker JS, Abramson E, Patel V, Dhopeshwarkar RV, Kaushal R. Evaluating health information technology in community-based settings: lessons learned. J Am Med Inform Assoc 2011;18(6):749-753. [doi: 10.1136/amiajnl-2011-000249] [Medline: 21807649]
22. Lobach DF, Kawamoto K, Anstrom KJ, et al. Proactive population health management in the context of a regional health information exchange using standards-based decision support. AMIA Annu Symp Proc 2007 Oct 11;2007:473-477. [Medline: 18693881]
23. Vest JR. How are health professionals using health information exchange systems? Measuring usage for evaluation and system improvement. J Med Syst 2012 Oct;36(5):3195-3204. [doi: 10.1007/s10916-011-9810-2] [Medline: 22127521]
24. Johnson KB, Unertl KM, Chen Q, et al. Health information exchange usage in emergency departments and clinics: the who, what, and why. J Am Med Inform Assoc 2011;18(5):690-697. [doi: 10.1136/amiajnl-2011-000308] [Medline: 21846788]
25. Devine EB, Totten AM, Gorman P, et al. Health information exchange use (1990-2015): a systematic review. EGEMS (Wash DC) 2017 Dec 7;5(1):27. [doi: 10.5334/egems.249] [Medline: 29881743]
26. Guide for evaluation indicators of regional medical cooperation ver.1.0. Chiki iryo renkei no hyokasihyo ni kansuru gaido ver.1.0 [Website in Japanese]. Japan Association of Healthcare Information Systems Industry. URL: https://www.jahis.jp/standard/detail/id=850 [accessed 2023-07-26]
27. Utilizing regional healthcare networks during the coronavirus pandemic. Korona ka ni okeru chiki iryo zyouhorenkei nettowa-ku no katsuyo [Website in Japanese]. Japan Medical Association Research Institute. URL: https://www.jmari.med.or.jp/result/working/post-3486/ [accessed 2023-07-24]
28. Matsumoto T, Taura N, Kawasaki K, Masuzaki H, Honda M. The impact of the health information exchange system for the hospital management in Japan. Acta Med Nagasaki 2020;64:39-44. [doi: 10.11343/amn.64.39]
29. Ido K, Nakamura N, Nakayama M. Miyagi Medical and Welfare Information Network: a backup system for patient clinical information after the Great East Japan Earthquake and tsunami. Tohoku J Exp Med 2019 May;248(1):19-25. [doi: 10.1620/tjem.248.19] [Medline: 31080195]
30. Nakayama M, Inoue R, Miyata S, Shimizu H. Health information exchange between specialists and general practitioners benefits rural patients. Appl Clin Inform 2021 May;12(3):564-572. [doi: 10.1055/s-0041-1731287] [Medline: 34107543]
31. Disclosure of information regarding medical research involving human subjects. Hito wo taisho to suru igakukei kenkyu ni kansuru joho kokai. Kyoto University Hospital. URL: https://kyoto.bvits.com/rinri/publish.aspx [accessed 2024-09-19]
32. Regional Medical Network HumanBridge EHR Solutions. Fujitsu. URL: https://www.fujitsu.com/jp/solutions/industry/healthcare/products/humanbridge/ [accessed 2024-04-19]
33. Motulsky A, Weir DL, Couture I, et al. Usage and accuracy of medication data from nationwide health information exchange in Quebec, Canada. J Am Med Inform Assoc 2018 Jun 1;25(6):722-729. [doi: 10.1093/jamia/ocy015] [Medline: 29590350]
34. Taylor HL, Apathy NC, Vest JR. Health information exchange use during dental visits. AMIA Annu Symp Proc 2020;2020:1210-1219. [Medline: 33936497]
35. Cooperation between different systems in the regional medical information sharing platform (EHR) - present and future. Chiki iryou joho renkei kiban (EHR) ni okeru kotonaru sisutemu kan no renkei - genjo to shorai [Website in Japanese]. Japan Medical Association Research Institute. 2020. URL: https://www.jmari.med.or.jp/result/working/post-249/ [accessed 2024-07-01]

Abbreviations

EMR: electronic medical record
FY: financial year
HIE: health information exchange
MHLW: Ministry of Health, Labour and Welfare
RHIO: regional health information organization
RHN: regional health care network

Edited by C Lovis; submitted 12.01.24; peer-reviewed by A Ito, T Okumura; revised version received 02.08.24; accepted 17.08.24; published 09.10.24.

Please cite as:
Suzumoto J, Mori Y, Kuroda T
Health Care Worker Usage of Large-Scale Health Information Exchanges in Japan: User-Level Audit Log Analysis Study
JMIR Med Inform 2024;12:e56263
URL: https://medinform.jmir.org/2024/1/e56263
doi:10.2196/56263

A Generic Transformation Approach for Complex Laboratory Data Using the Fast Healthcare Interoperability Resources Mapping Language: Method Development and Implementation

Jesse Kruse[1], MSc; Joshua Wiedekopf[2,3], MSc; Ann-Kristin Kock-Schoppenhauer[2], Dr rer hum biol; Andrea Essenwanger[4], MSc; Josef Ingenerf[2,3], Dr rer nat; Hannes Ulrich[2,5], Dr rer nat

1

2

3

4

5

Corresponding Author:
Hannes Ulrich, Dr rer nat

Abstract

Background: Reaching meaningful interoperability between proprietary health care systems is a ubiquitous task in medical informatics, where communication servers are traditionally used for referring and transforming data from the source to target systems. The Mirth Connect Server, an open-source communication server, offers, in addition to the exchange functionality, functions for simultaneous manipulation of data. The standard Fast Healthcare Interoperability Resources (FHIR) has recently become increasingly prevalent in national health care systems. FHIR specifies its own standardized mechanisms for transforming data structures using StructureMaps and the FHIR mapping language (FML).

Objective: In this study, a generic approach is developed, which allows for the application of declarative mapping rules defined using FML in an exchangeable manner. A transformation engine is required to execute the mapping rules.

Methods: FHIR natively defines resources to support the conversion of instance data, such as an FHIR StructureMap. This resource encodes all information required to transform data from a source system to a target system. In our approach, this information is defined in an implementation-independent manner using FML. Once the mapping has been defined, executable Mirth channels are automatically generated from the resources containing the mapping in JavaScript format. These channels can then be deployed to the Mirth Connect Server.

Results: The resulting tool is called FML2Mirth, a Java-based transformer that derives Mirth channels from detailed declarative mapping rules based on the underlying StructureMaps. Implementation of the *translate* functionality is provided by the integration of a terminology server, and to achieve conformity with existing profiles, validation via the FHIR validator is built in. The system was evaluated for its practical use by transforming Labordatenträger version 2 (LDTv.2) laboratory results into Medical Information Object (*Medizinisches Informationsobjekt*) laboratory reports in accordance with the National Association of Statutory Health Insurance Physicians' specifications and into the HL7 (Health Level Seven) Europe Laboratory Report. The system could generate complex structures, but LDTv.2 lacks some information to fully comply with the specification.

Conclusions: The tool for the auto-generation of Mirth channels was successfully presented. Our tests reveal the feasibility of using the complex structures of the mapping language in combination with a terminology server to transform instance data. Although the Mirth Server and the FHIR are well established in medical informatics, the combination offers space for more research, especially with regard to FML. Simultaneously, it can be stated that the mapping language still has implementation-related shortcomings that can be compensated by Mirth Connect as a base technology.

(JMIR Med Inform 2024;12:e57569) doi:10.2196/57569

KEYWORDS

FHIR; StructureMaps; FHIR mapping language; laboratory data; mapping; standardization; data science; healthcare system; HIS; information system; electronic healthcare record; health care system; electronic health record; health information system

Introduction

Digitalization is progressively transforming health care systems, and its rapid pace is creating new clinical data sources that need to be integrated. This is of enormous importance for patient care, as data integration and the fusion of all sources allow a comprehensive, holistic overview and ensure the best treatment possible. However, the lack of interoperability of health care systems is a significant, long-lasting problem [1]. Nonetheless, interoperability is required to ensure seamless and effortless access to essential health care information. Existing medical data are presented in a multitude of proprietary formats and in line with different standards [2]. Therefore, it is inevitable to transform data into a harmonized structure to enable collective use. Current national and international initiatives are integrating large volumes of data in clinical environments to enable their use [3-5]. The established data integration is close to the origin of the data, which is highly desirable since late mapping harbors disadvantages and issues can be addressed accordingly [6].

Aligning the multitude of standards and thus making data usable is a major task. The mapping from source to target data structures is mainly defined by data stewards and is followed by a qualitive evaluation involving medical professionals. The involvement of these key professional groups is much easier if concise, declarative mapping rules are used instead of less comprehensible program instructions in various scripting languages. Basically, the creation of mapping rules is time-consuming and demands a significant number of resources. It occurs at every site involved, which commits a substantial additional number of stewards and medical professionals across the entire health care system [6,7]. Furthermore, sharing and reusing of declarative mapping rules in a standardized and adaptable manner is highly desirable and can release the needed resources and accelerate data enabling. Hence, this study aims to provide a software solution called "FML2Mirth," which is based on the separation of declarative mapping rules from a transformation pipeline leveraging the functionalities of Mirth.

Methods

The transformation of health care–related information needs two major components to enable generalization: a highly adaptable exchange format and a modular transformation engine. Our approach is based on the HL7 (Health Level Seven) Fast Healthcare Interoperability Resources (FHIR) standard and the Mirth communication server.

Ethical Considerations

No ethics committee approval was required for this study, as the data used comprised 20 anonymized laboratory data carriers (LDT) that were provided by the laboratory company LADR Laboratory Group as part of its quality control and represent a complete blood count. According to institutional and local policies, ethics approval is not necessary for studies using anonymized laboratory testing data, and this poses no risk to individual participants.

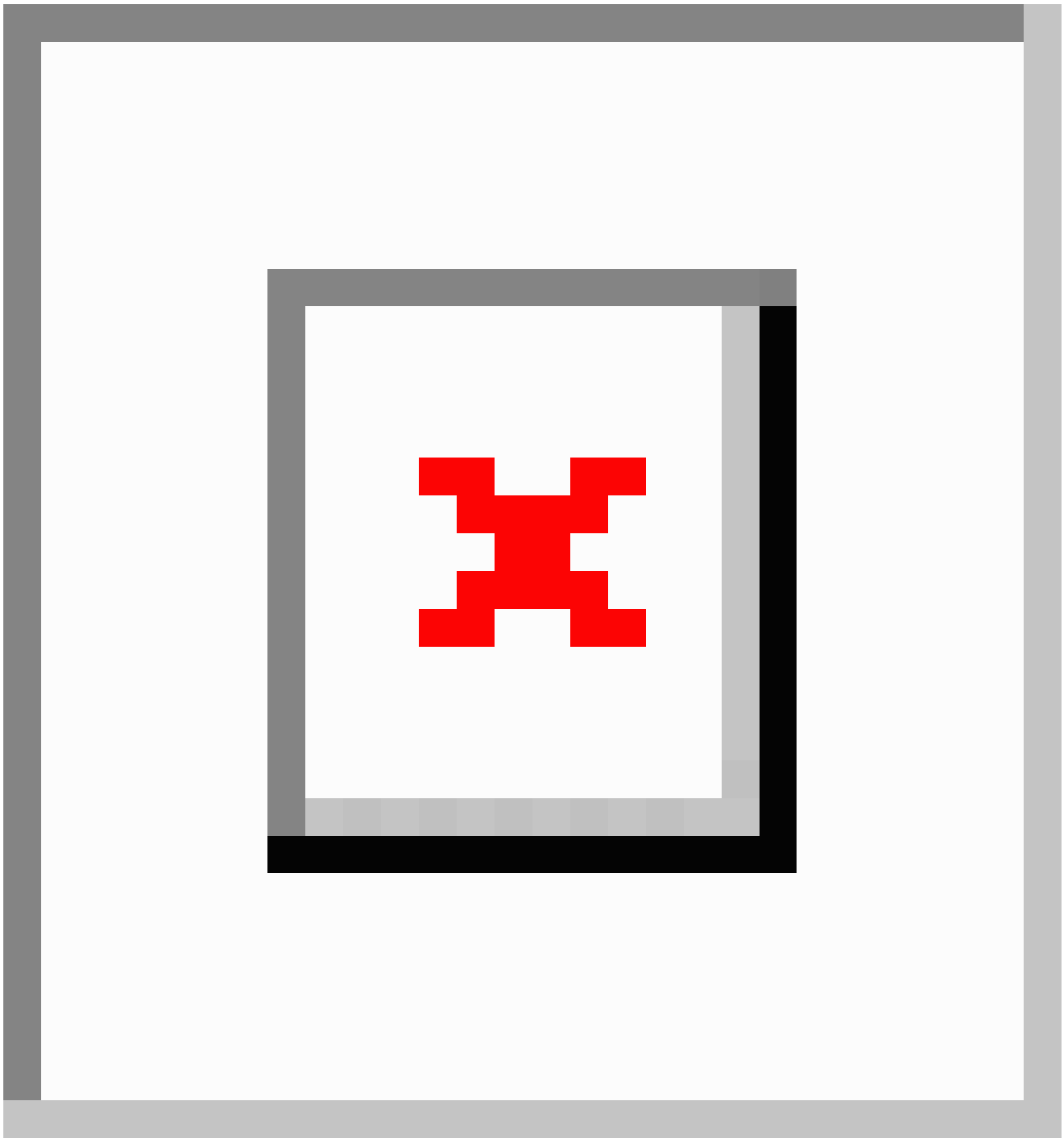
HL7 FHIR and the FHIR Mapping Language

HL7 FHIR is the emerging standard for health care–specific data exchange and has been broadly adapted worldwide [8]. FHIR provides state-of-the-art technologies to modernize the current health care landscape using extensible resources as harmonized and semantically annotatable information units [9]. However, FHIR also provides mechanisms to natively define transformations on its structures. The essential FHIR resource for automated conversion of instance data is the StructureMap, which defines a mapping from a source structure to a target structure [10] and provides all necessary information for an automatic transformation. To aid the definition of StructureMap resources, FHIR has specified a domain-specific, declarative language, the FHIR mapping language (FML) [11], which was introduced as part of FHIR release 3. By specifying the mappings in an implementation-independent manner, the mapping between structures can be easily shared within the medical informatics community.

One major benefit is the use of generic data structures as input formats, since FML can define mappings on non-FHIR structures. Consequently, FML can also be used to map older health care–related formats such as HL7 (version 2) or LDT to FHIR.

The declarative syntax of FML enables users to define mappings concisely. The mapping itself is structured in so-called *groups* within the mapping, being formalized by mapping rules. These rules consist of two parts separated by an arrow. On the left side of the arrow is the source part. Here, fields of source structures can be accessed, and the values can be written in variables. On the right side, the fields of a target structure can be accessed and built-in transformation functions can be invoked (see Figure 1). Those provided functions cover a broad functionality, ranging from a simple *copy* to a *translate* function to resolve given concept codes using an external FHIR ConceptMap.

Figure 1. The code illustrates a snippet of the FHIR (Fast Healthcare Interoperability Resources) mapping language, showing how a conventional laboratory test in Labordatenträger format can be mapped to an FHIR observation resource. The mapping is structured in groups, containing specific rules.



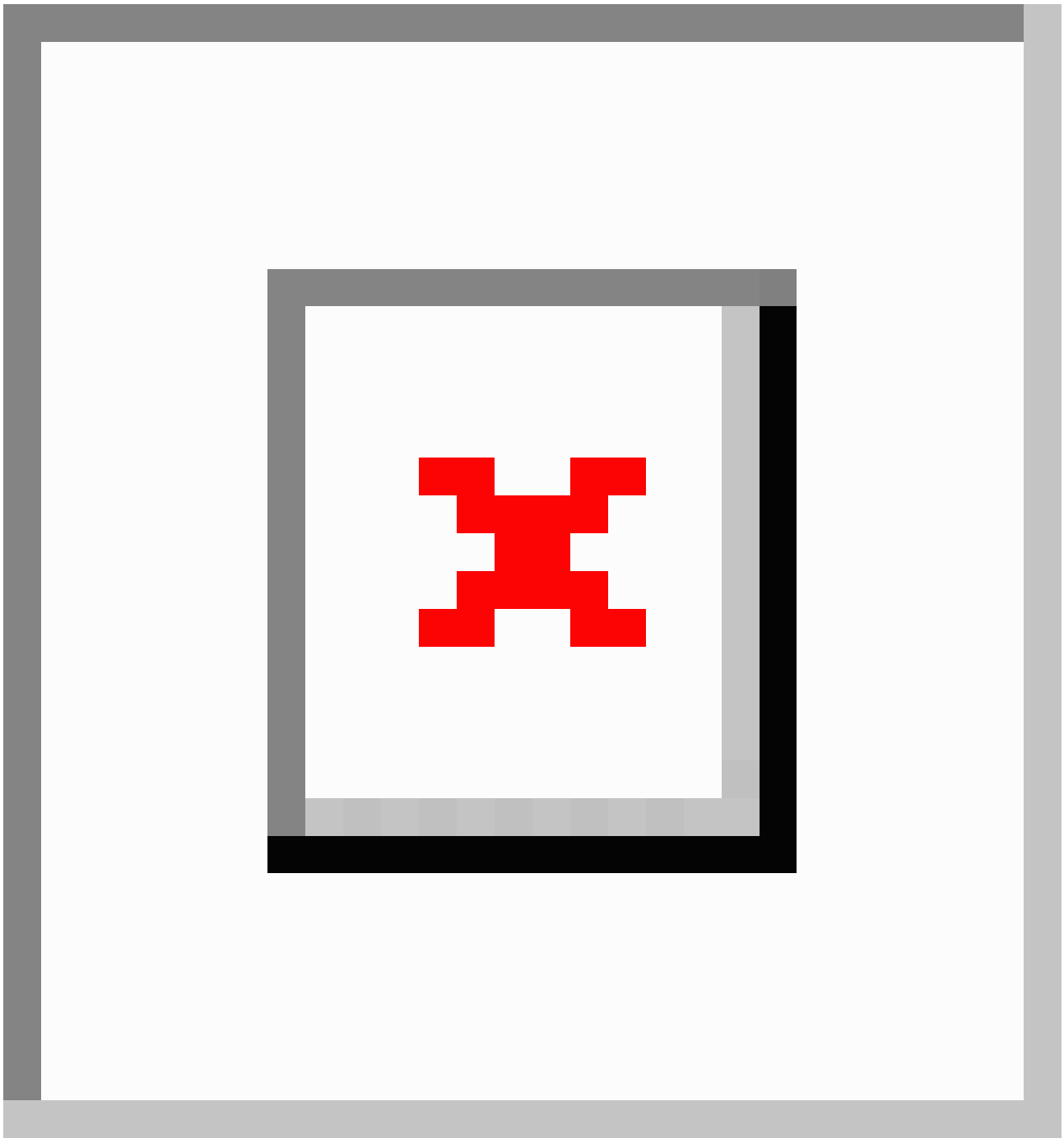
Mirth Connect Communication Server

Mirth Connect is an open-source communication server designed specifically for the health care sector. It features support for domain-specific formats, such as HL7 (version 2), as well as more general formats such as XML and JSON. Mirth is widely used as a communication server and is applied in various areas for data integration [12,13]. As a communication server, Mirth receives messages from one system and forwards them to downstream systems. Mirth introduces the concept of *channels* to establish a connection from a source system to various target systems. Within the channels, rule-based transformers can be defined to enable message manipulation. The received message can be automatically modified in terms of content or structure. The transformer rules are implemented as JavaScript code snippets, which Mirth applies to the messages. The channels including its transformers are serialized as XML files and can easily be shared and redeployed at other sites.

Architecture for Generic Data Transformation

The two-fold process shown in Figure 2 enables light and adaptable data transformation: the first step is transformation design and the second one is staging.

Figure 2. Conceptual overview of the 2-step process. In the first phase, the mapping is formalized as an FHIR (Fast Healthcare Interoperability Resources) mapping language script and corresponding semantic CodeSystems are aligned with each other in a ConceptMap. In the second phase, prior defined rules are used to deploy the mapping to Mirth. If semantic translation is needed, the ConceptMaps can be queried during run time.



During the transformation design phase, the source and target systems are selected, and the mapping is declared in FML. The mapping needs to be created manually by a data steward. This process can be supported by previously proposed algorithms and tools [14]. In addition, semantic concepts and codes can be mapped from source to target based on prepared mapping relationships provided by the FHIR ConceptMap resource, if necessary. It is recommended that this process is carried out by a health care professional to ensure data validity and integrity.

In the staging phase, the Mirth channel, which was built on the basis of prior mapping efforts, is injected in the productive communication server at run time. Mapping within the transformer uses the rules derived from FML for format manipulation, and dynamically translates the semantic codes using the ConceptMap.

Results

Overview

The proposed system implements the transition between the two process phases: transformation design and staging. This transition is achieved by the in Java-implemented generator FML2Mirth [15], which derives Mirth channels from a given mapping. There are two types of files used in the process, which

need to be created upfront. The first type is ConceptMaps used by a terminology server to provide a translation service for concepts that need to be translated during data transformation in the production phase. These ConceptMaps are referenced by the *translate* operations in the FML to specify how a given concept in the source structure should be mapped to a concept used in the target structure.

In the second file, the FML script, the mapping between source and target structure is defined. This script is then used as an input for FML2Mirth. As a first step, the FML script must be parsed, for which we use the official FHIR validator, since it provides functionality to generate a StructureMap from a given FML-Script. An additional advantage of this approach is, that by using a standard tool, it is guaranteed that the produced artifact is a standard-compliant StructureMap. Using these files as the input, FML2Mirth should also be compatible with every other StructureMap that adheres to the standard.

Besides the FML script, FHIR2Mirth needs the API (application programming interface) address of the Mirth server to which the generated channel should be deployed, as well as the address of the terminology server that is queried for the translation service during a transformation in the production phase. The tool then parses the StructureMap and generates the JavaScript code needed for every transformation specified. Once the generation is finished, the resulting code is embedded within a Mirth channel definition and deployed on the specified server via the Mirth REST-API. Subsequently, the incoming messages are automatically transformed into the desired structure. On encountering the FML *translate* function, Mirth queries the given terminology server to carry out translation between the different CodeSystems of the source and the target structure based on the previously created ConceptMaps. In our setting shown in Figure 3, an HAPI FHIR JPA server instance is used as an external terminology server, but any other FHIR terminology server implementation could be used instead. The transformed message is then forwarded to the defined target system, for example, an FHIR repository.

Evaluation of the transformation within Mirth is implemented using the JUnit testing framework. While JUnit is most commonly used for unit testing, it can also be used for end-to-end integration testing. The JUnit runner loads LDT laboratory messages from files and sends them to the Mirth server via the REST API. The destination connector of the test channel is a JavaScript connector that returns the transformed message as a response to the requesting system. The JUnit runner then receives the transformed message and uses HAPI FHIR to parse the incoming bundles as FHIR instances. Thus, the incoming resources are automatically tested for their conformity to the specification and validated against the corresponding profiles using the HAPI instance validator (Figure 4).

Figure 3. Conceptual overview of the proposed architecture for a generic data transformation process.

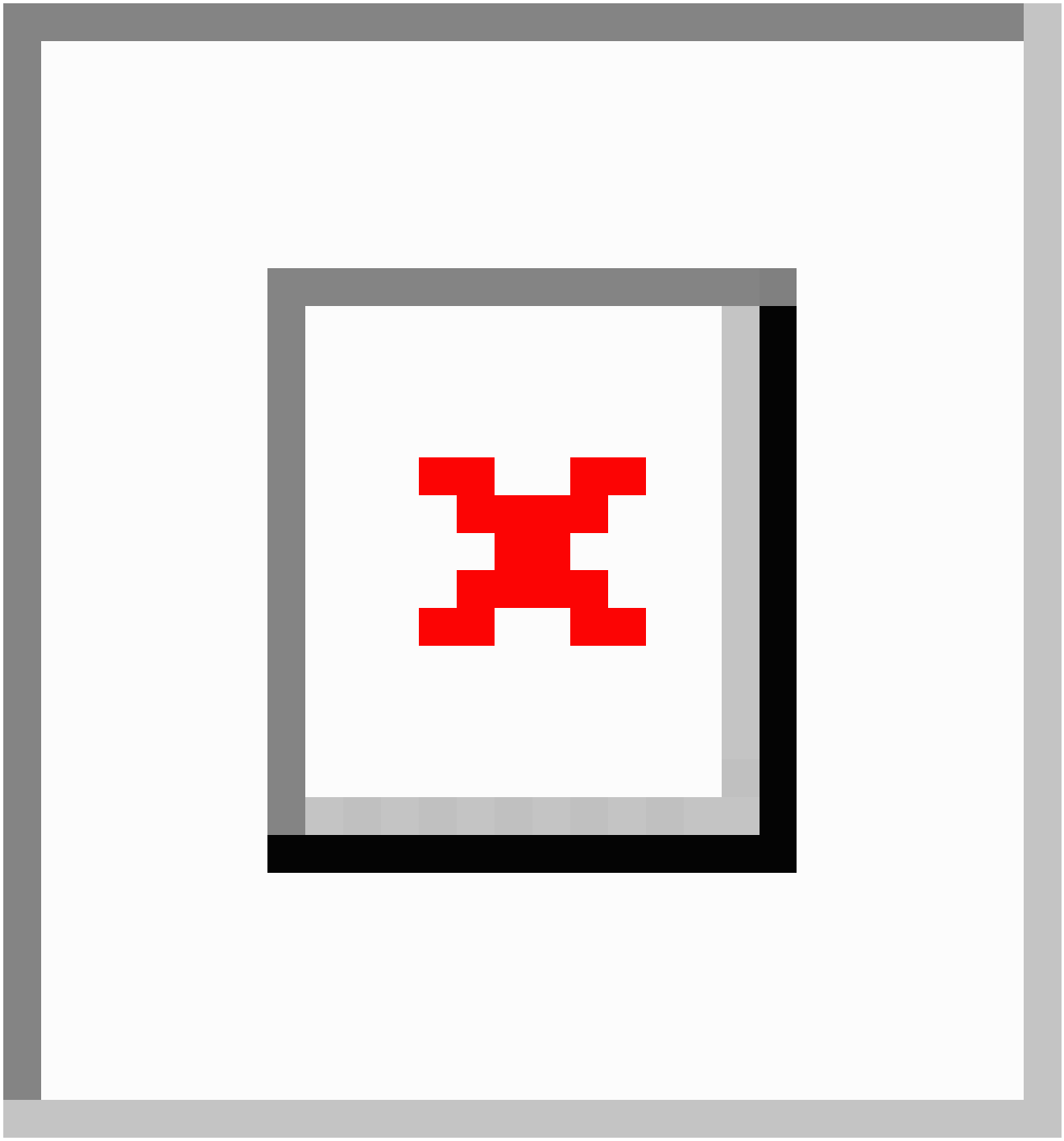
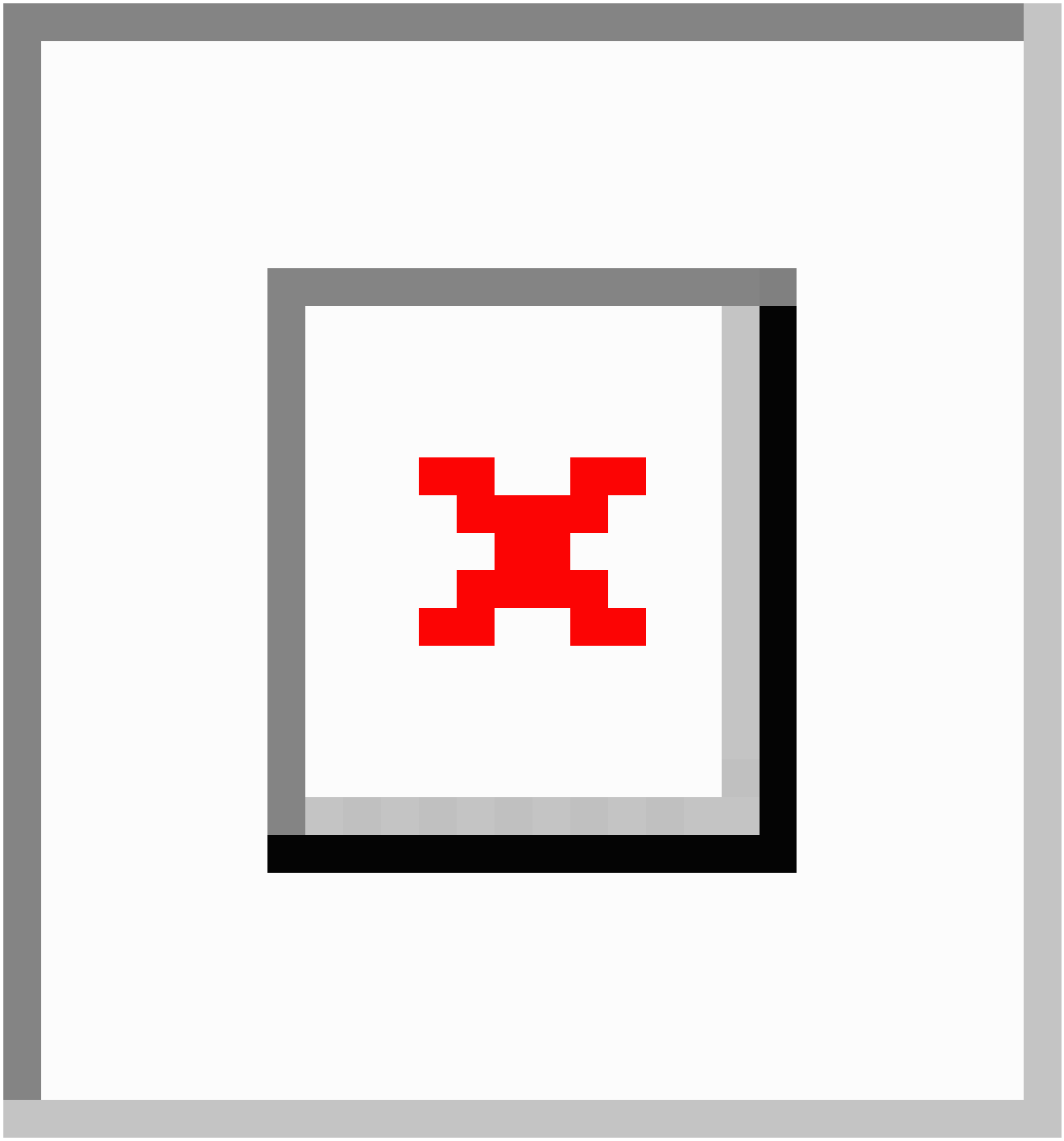


Figure 4. This example shows the different transformation steps, including the FHIR (Fast Healthcare Interoperability Resources) StructureMap and JavaScript result.



Evaluation

Assessment of our proposed transformation process is carried out in cooperation with the LADR laboratory network [16], an association of 19 specialist laboratories throughout Germany serving more than 400 clinics. As an evaluation scenario, laboratory findings standardized in the German LDT [17] format are transformed into FHIR. The LDT format is a perennial standard developed and published by the National Association of Statutory Health Insurance Physicians (NASHIP; German: Kassenärztliche Bundesvereinigung [KBV]) and is used daily for communicating up to 1 million messages in Germany. In our evaluation scenario, LADR provided 20 anonymized LDT laboratory messages representing complete blood counts in LDT (version 2) according to the data format description in FHIR (version 5.12) for the evaluation. Two different FHIR implementation guides were used as target formats: the *German Laboratory finding of the NASHIP* and the *HL7 Europe Laboratory Report* [18].

The LDT format is mapped to both target formats using FML. Using the FML2Mirth generator, the FML scripts are translated into StructureMaps, and two separate channels are generated and deployed on a Mirth instance. Each format has a distinct set of semantic codes, which are permissible values for certain elements (binding). The sending system in our evaluation scenario uses proprietary codes that must be mapped to the specified CodeSystem in the target profiles. A terminology

server is used to provide a translate service based on ConceptMaps, which is accessed by the Mirth channel via a REST API. The initial FML script for the transformation to the target format *MIO Laborbefund* with its 9 profiles [19] of the NASHIP consists of over 400 lines of formalized mapping rules, resulting in StructureMaps with over 3500 lines and more than 2300 elements. The 20 transformed LDT laboratory findings resulted in an average of 62 unique FHIR each. Only one report failed the closing instance validation due to a missing first name within the corresponding patient resource.

The second Mirth channel transforming LDT to the 8 profiles of the Europe Laboratory Report is based on the initial FML script. The same 20 LDT laboratory results sum up a total of over 694 FHIR resources, resulting in an average of 35 resources each. The subsequent validation process revealed for all test cases a structural issue that occurred due to a required extension of the DiagnosticReport profile. The extension should be a Composition reference to the report, but the extension itself is missing in the provided FHIR package by EU (European Union) laboratory report, and the referral link is nonfunctional. Yet, a manual check and a comparison with the NASHIP MIO resources proved their validity.

Discussion

This study presents a concept and the implementation of an adaptable transformation process for various data structures into a standardized format. Our approach is based on FHIR as a standardized data exchange format, FML as declarative transformation rules, and Mirth Connect as the transformation executor.

Principal Results

The approach focusses on separating the transformation rules from the actual transformation to extract the rules out of the process. Due to the detachment, the rules, which are created in a labor-intensive process, can be shared in a declarative and standardized manner. The use of FML for defining the mappings can cover technical operations such as network calls to the terminology server and are triggered by FML2Mirth dynamically during JavaScript generation. In addition, the separation allows a degree of modularization and, thus, the reuse of the created rules. For example, it was possible to modify the mapping for the MIO reports to generate laboratory reports, which conform to the profiles of the EU laboratory report within a few hours. This emphasizes the flexibility of the proposed method. A significant advantage of this approach is that it allows mappings of concepts to be exchanged at run time of the Mirth channel and be developed and updated independently of the structural mapping. Compared to common ETL jobs, this approach needs much less manual intervention.

The mapping is defined using FML to render the StructureMap using the official FHIR Validator. The declarative mapping is enclosed in the generated StructureMap instances as a parse tree. Referring and combining various FML scripts is possible per the specification but has not been implemented yet. Therefore, the mapping shall be edited incrementally during the transformation design phase in the FML scripts rather than in the StructureMaps directly. Nevertheless, initial creation of the FML scripts is less time-consuming but still a labor-intensive task, and currently there is a lack of suitable tool support. An FML script is always a 1-way mapping, and reverse transformation is not automatically assumed.

The functionality provided by Mirth highly constrains the design of our approach, as external dependencies should be minimized to ensure ease of use. With this in mind, the implemented single command line tool in Java injects the transformations into a specified Mirth channel. In addition, only the Mirth server and an FHIR server, which provides terminology services, are required. Simultaneously, the system should be as flexible as possible and support arbitrary StructureMaps. To fulfill these requirements, it was necessary to work with the circumstances provided by Mirth. Since Mirth itself works on XML objects, it is prudent to use this approach as well. Hence, the transformations defined in the StructureMap are translated into transformations on XML objects. Mirth itself first translates incoming data such as HL7v2 to XML as a first processing step, so all formats supported by Mirth can automatically also be transformed using the generated transformer from FML2Mirth. If a format is not supported by Mirth, such as LDT, another transformation step can simply be added to transform the incoming data into an XML structure.

Limitations

During the evaluation, an issue was discovered with the FHIR implementation within the different FHIR servers. In particular, the issue concerned FHIRPath implementation, which is used by the FHIR Validator to evaluate the validation rules. In our test setting, one rule is interpreted differently depending on the implementation used. On the tested laboratory reports, the .NET and JavaScript implementations returned true, which is correct per our understanding. In contrast, both Java implementations tested (HAPI and IBM) returned false.

Alongside the complex structural mapping, semantic integrity must also be ensured. In our approach, the semantic mappings are created and validated by health care professionals and stored in an FHIR server as an external service, which is standard procedure. While existing terminology systems generally provide transition rules to newer versions, the maintenance of FHIR resources is not standardized and is an ongoing topic of investigation in medical informatics [20].

Furthermore, a mechanism was missing to process further StructureMaps, which are referenced from a given StructureMap. FML uses FHIRPath as an embedded language in several instances. The associated specification is extensive and was implemented in this study only in parts.

Comparison With Prior Work

The use of formalized mapping rules in data integration is a well-studied topic, and Mirth as a transformation engine for clinical data integration is also a renowned tool. However, the combination of both topics with the goal of a generic and easily adaptive mapper is missing in the literature, to our knowledge. Alongside declarative rules, the work of Ong et al [21] must be mentioned; they present a dynamic ETL approach that uses a custom mapping language to transform health care–related data

into the OMOP (Observational Medical Outcomes Partnership) common data model. The formalized rules are rich in details but are proprietary and rather database-oriented due to their use case. An innovative feasibility study for FML has been presented by Dimitrov et al [22]; they used FML for transforming the HL7 CDA (Clinical Document Architecture)–based national Austrian electronic patient record. Using FML, a transformation from CDA documents to the International Patient Summary based on FHIR could be accomplished. The usage of Mirth as a transformation engine is an intuitive choice due to its functionality and product scope. Hence, various studies are focusing on the transformation from HL7 (version 2) messages into further target formats, including structural reformation into JSON [23] or standardization into HL7 FHIR [24].

Our approach is based on that of a prior study that evaluated the transformation of StructureMaps into Mirth channels [25]. The results were promising, but it also emerged that the manual definition of StructureMaps became cumbersome and rapidly very complex. The presented approach overcomes this shortage using FML and its intuitive nature.

Conclusions

We successfully implemented a functional system based on Mirth, which automatically generates transformations from StructureMaps and integrates them into a Mirth channel. This system supports large parts of the FML specification. Initial tests revealed that complex structures based on FML maps, generated in a channel in conjunction with a terminology server and ConceptMaps are feasible. If a new format is to be supported, it may only be necessary to insert another Mirth transformation step before the transformation from the StructureMap, to transform the input structure into XML. One advantage of the current solution is its flexibility and robustness in that only one resource is used for translation. The topic of FML is still an ongoing discussion and has not been extensively investigated in the FHIR community so far. Accordingly, the tooling in this area is currently rather rudimentary.

Acknowledgments

This study was funded by the German Federal Ministry of Education and Research (grants 01ZZ2312A and 01ZZ2011). We acknowledge financial support provided by Land Schleswig-Holstein within the funding program Open Access Publikationsfonds.

Conflicts of Interest

None declared.

References

1. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79. [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
2. Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022 Jan 11;24(1). [doi: 10.2196/25440]
3. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018 Jul;57(S 01):e50-e56. [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
4. Koch M, Richter J, Hauswaldt J, Krefting D. How to make outpatient healthcare data in Germany available for research in the dynamic course of digital transformation. Stud Health Technol Inform 2023 Sep 12;307:12-21. [doi: 10.3233/SHTI230688] [Medline: 37697833]
5. Auffray C, Balling R, Barroso I, et al. Making sense of big data in health research: towards an EU action plan. Genome Med 2016 Jun 23;8(1):71. [doi: 10.1186/s13073-016-0323-y] [Medline: 27338147]
6. Kock-Schoppenhauer AK, Schreiweis B, Ulrich H, et al. Medical data engineering – theory and practice. In: Bellatreche L, Chernishev G, Corral A, Ouchani S, Vain J, editors. Advances in Model and Data Engineering in the Digitalization Era: Springer; 2021, Vol. 1481:269-284. [doi: 10.1007/978-3-030-87657-9_21]
7. Bönisch C, Kesztyüs D, Kesztyüs T. Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. Sci Data 2022 Oct 28;9(1):659. [doi: 10.1038/s41597-022-01792-7] [Medline: 36307424]
8. Vorisek CN, Lehne M, Klopfenstein SAI, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724. [doi: 10.2196/35724] [Medline: 35852842]
9. Benson T, Grieve G, editors. Principles of Health Interoperability: FHIR, HL7 and SNOMED CT, 4th edition: Springer International Publishing; 2021. [doi: 10.1007/978-3-030-56883-2]
10. Recource StructureMap - FHIR v4.0.1. HL7 International. 2023. URL: http://build.fhir.org/structuremap.html [accessed 2023-08-07]
11. HL7 International. FHIR Mapping Language - FHIR v400. URL: https://www.hl7.org/fhir/mapping-language.html [accessed 2023-10-23]
12. Camacho Rodriguez JC, Stäubert S, Löbe M. Automated import of clinical data from HL7 messages into OpenClinica and tranSMART using Mirth Connect. Stud Health Technol Inform 2016;228:317-321. [Medline: 27577395]
13. Lin J, Ranslam K, Shi F, Figurski M, Liu Z. Data migration from operating EMRs to OpenEMR with Mirth Connect. Stud Health Technol Inform 2019;257:288-292. [Medline: 30741211]

14. Deppenwiese N, Duhm-Harbeck P, Ingenerf J, Ulrich H. MDRCupid: a configurable metadata matching toolbox. Stud Health Technol Inform 2019 Aug 21;264:88-92. [doi: 10.3233/SHTI190189] [Medline: 31437891]
15. Kruse J. FML2Mirth. Zenodo. 2024. URL: https://doi.org/10.5281/zenodo.10678100 [accessed 2024-02-20]
16. LADR laboratory network. URL: https://www.ladr.de/ [accessed 2024-02-12]
17. Bundesvereinigung KK. LDT Labordatenträger Datensatzbeschreibung, LDT1001. 01 und Elektronisches Leistungsverzeichnis. Köln. 2004. URL: https://update.kbv.de/ita-update/Labor/Labordatenkommunikation/EXT_ITA_VGEX_LDT%203_2_16_Gesamtdokument.pdf [accessed 2024-10-08]
18. HL7 Europe laboratory report. HL7 Europe. URL: https://hl7.eu/fhir/laboratory/0.1.0-ballot/ [accessed 2024-02-12]
19. Laboratory report 1.0.0 [Article in German]. Laborbefund 1.0.0. URL: https://mio.kbv.de/display/LAB1X0X0 [accessed 2024-02-20]
20. Wiedekopf J, Drenkhahn C, Rosenau L, Ulrich H, Kock-Schoppenhauer AK, Ingenerf J. TerminoDiff - detecting semantic differences in HL7 FHIR CodeSystems. Stud Health Technol Inform 2022 May 25;294:362-366. [doi: 10.3233/SHTI220475] [Medline: 35612097]
21. Ong TC, Kahn MG, Kwan BM, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Med Inform Decis Mak 2017 Sep 13;17(1):134. [doi: 10.1186/s12911-017-0532-3] [Medline: 28903729]
22. Dimitrov A, Duftschmid G. Generation of FHIR-based international patient summaries from ELGA data. Stud Health Technol Inform 2022 May 16;293:1-8. [doi: 10.3233/SHTI220339] [Medline: 35592952]
23. Cruz R, Guimarães T, Peixoto H, Santos MF. Architecture for intensive care data processing and visualization in real-time. Proc Comput Sci 2021;184:923-928. [doi: 10.1016/j.procs.2021.03.115]
24. Alkarkoukly S, MdM K, Beyan O. Breaking barriers for interoperability: a reference implementation of CSV-FHIR transformation using open-source tools. In: Hägglund M, Blusi M, Bonacina S, Nilsson L, Cort Madsen I, Pelayo S, et al, editors. Studies in Health Technology and Informatics: IOS Press; 2023. [doi: 10.3233/SHTI230061]
25. Vogl KM, Ulrich H, Ingenerf J. Generation of Message Transformers Based on HL7 FHIR StructureMaps within the Interface Engine Mirth Connect: Infinite Science Publishing; 2020:131-134.

Abbreviations

CDA: Clinical Document Architecture
ETL: Extract, Transform, Load Process
EU: European Union
FHIR: Fast Healthcare Interoperability Resources
FML: FHIR Mapping Language
IPS: International Patient Summaries
LDT: Labordatenträger
MIO: Medical Information Object (*Medizinisches Informationsobjekt*)
NASHIP: National Association of Statutory Health Insurance Physicians
OMOP: Observational Medical Outcomes Partnership

Edited by M Focsa; submitted 20.02.24; peer-reviewed by C Lien, R Saripalle, Z Hou; revised version received 09.07.24; accepted 25.07.24; published 18.10.24.

Please cite as:
Kruse J, Wiedekopf J, Kock-Schoppenhauer AK, Essenwanger A, Ingenerf J, Ulrich H
A Generic Transformation Approach for Complex Laboratory Data Using the Fast Healthcare Interoperability Resources Mapping Language: Method Development and Implementation
JMIR Med Inform 2024;12:e57569
URL: https://medinform.jmir.org/2024/1/e57569
doi:10.2196/57569

The Information and Communication Technology Maturity Assessment at Primary Health Care Services Across 9 Provinces in Indonesia: Evaluation Study

Dewi Nur Aisyah[1,2,3], PhD; Agus Heri Setiawan[2], MSc; Alfiano Fawwaz Lokopessy[2], BSc; Nadia Faradiba[2], BDent; Setiaji Setiaji[2], MSc; Logan Manikam[1,3], PhD; Zisis Kozlakidis[4], PhD

1

2

3

4

Corresponding Author:
Logan Manikam, PhD

Abstract

Background: Indonesia has rapidly embraced digital health, particularly during the COVID-19 pandemic, with over 15 million daily health application users. To advance its digital health vision, the government is prioritizing the development of health data and application systems into an integrated health care technology ecosystem. This initiative involves all levels of health care, from primary to tertiary, across all provinces. In particular, it aims to enhance primary health care services (as the main interface with the general population) and contribute to Indonesia's digital health transformation.

Objective: This study assesses the information and communication technology (ICT) maturity in Indonesian health care services to advance digital health initiatives. ICT maturity assessment tools, specifically designed for middle-income countries, were used to evaluate digital health capabilities in 9 provinces across 5 Indonesian islands.

Methods: A cross-sectional survey was conducted from February to March 2022, in 9 provinces across Indonesia, representing the country's diverse conditions on its major islands. Respondents included staff from public health centers (*Puskesmas*), primary care clinics (*Klinik Pratama*), and district health offices (*Dinas Kesehatan Kabupaten*/*Kota*). The survey used adapted ICT maturity assessment questionnaires, covering human resources, software and system, hardware, and infrastructure. It was administered electronically and involved 121 public health centers, 49 primary care clinics, and 67 IT staff from district health offices. Focus group discussions were held to delve deeper into the assessment results and gain more descriptive insights.

Results: In this study, 237 participants represented 3 distinct categories: 121 public health centers, 67 district health offices, and 49 primary clinics. These instances were selected from a sample of 9 of the 34 provinces in Indonesia. Collected data from interviews and focus group discussions were transformed into scores on a scale of 1 to 5, with 1 indicating low ICT readiness and 5 indicating high ICT readiness. On average, the breakdown of ICT maturity scores was as follows: 2.71 for human resources' capability in ICT use and system management, 2.83 for software and information systems, 2.59 for hardware, and 2.84 for infrastructure, resulting in an overall average score of 2.74. According to the ICT maturity level pyramid, the ICT maturity of health care providers in Indonesia fell between the basic and good levels. The need to pursue best practices also emerged strongly. Further analysis of the ICT maturity scores, when examined by province, revealed regional variations.

Conclusions: The maturity of ICT use is influenced by several critical components. Enhancing human resources, ensuring infrastructure, the availability of supportive hardware, and optimizing information systems are imperative to attain ICT maturity in health care services. In the context of ICT maturity assessment, significant score variations were observed across health care levels in the 9 provinces, underscoring the diversity in ICT readiness and the need for regionally customized follow-up actions.

(JMIR Med Inform 2024;12:e55959) doi:10.2196/55959

KEYWORDS

public health centers; Puskesmas; digital maturity; infrastructure; primary health care; district health office; primary care clinics; Asia; Asian; Indonesia; ICT; information and communication technologies; information and communication technology; maturity; adoption; readiness; implementation; eHealth; telehealth; telemedicine; cross sectional; survey; surveys; questionnaire; questionnaires; primary care

Introduction

Digital health plays a significant role in enhancing health care services for people and has developed beyond providing electronic health records to supporting health care services provision, health surveillance, health literature, health research, and data-driven health policies [1-4]. The massive growth of technology use in health propelled to the top of the global agenda, as summarized in the World Health Organization's (WHO) Global Strategy on Digital Health 2020 - 2025 [1]. According to this report, each country is expected to adopt the strategies best suited to its conditions, culture, and values to reach its digital health sustainability.

In 2022, the Ministry of Health (MoH) of the Republic of Indonesia launched its health system transformation strategy. The transformation strategy had 6 pillars; one of them was primary health care transformation, which focused on strengthening promotive and preventive activities in its implementation to create more healthy people, improve health screening, and increase primary service capacity. Health care services in Indonesia are primarily provided at a public health care facility called *Puskesmas* (short for *Pusat Kesehatan Masyarakat* or public health centers) and community-based public health services (called *Posyandu*) since the 1980s.

Puskesmas in Indonesia are spread across all types of characteristic regions, such as urban, rural, remote, and very remote regions in Indonesia within all 38 provinces and 514 districts or cities. The number of *Puskesmas* has increased since 2017, from 9825 units to 10,374 in 2022. The ratio of *Puskesmas* in Indonesia to subdistricts was 1.4 in 2022. This illustrates that the ideal ratio of *Puskesmas* to subdistricts, namely a minimum of 1 *Puskesmas* in 1 subdistrict, has been fulfilled nationally. In terms of an average ratio, each *Puskesmas* serve 27,000 - 30,000 residents in 2023.

Puskesmas serve two main functions: providing integrated individual health care services and essential public health services. Other tasks of *Puskesmas* are to provide continuous and comprehensive care, refer patients to specialists and hospital services, coordinate health services, and guide patients within the network of public health services. *Puskesmas* play roles in promotive, preventive, and curative services for the population, while primary clinics and other primary health services, such as private general practitioner (DPM) and private midwife (BPM), are more focused on curative and specific health services approaches. Other than that, *Puskesmas* also have a responsibility to provide technical guidance to primary clinics, DPMs, and BPMs as the networking partner institutions in the area. Besides *Puskesmas* and *Posyandu*, essential health services are provided by other primary care centers, such as clinics, DPMs, and BPMs. In 2023, the primary care clinics also increasingly grew and reached more than 11,000 units, while DPM and BPMs reached 5800 units across all 38 provinces in Indonesia.

In terms of organizational governance, *Puskesmas* are considered the technical implementation unit of district health offices. In this case, the district health office is responsible for providing guidance, monitoring, and evaluation for *Puskesmas* in its region. However, *Puskesmas* have autonomy to synchronize and harmonize the health development goals in their working area. All health service activities in the area are conducted by *Puskesmas* and their networking partners (ie, primary clinics, DPMs, and BPMs). *Puskesmas* coordinate and report these activities to the district health office on an annual, monthly, and weekly basis using either electronic or nonelectronic systems. The reports include (1) activity report, (2) financial report, (3) field survey report, (4) related cross-sector reports, and (5) the health services network (clinics, DPM, BPM) reports in the area. This reporting scheme continues by the district health office to the province health office and then to the MoH as the national authority.

The health services and public health programs that were routinely reported using an electronic information system included maternal and child health, nutrition, surveillance, as well as disease prevention and control. These programs use multiple, separate applications that currently lack interoperability data standards. Besides that, there are multiple electronic medical record systems widely developed by hundreds of private vendors, designed to provide patient medical records for health care providers. These systems include *Sistem Informasi Puskesmas* (SIMPUS) for *Puskesmas* and *Sistem Informasi Klinik* (SIMKLINIK; clinic information systems) for primary clinics, as well as telemedicine and tele consultation platforms. These syetms are still not integrated and interoperable with other public health information systems developed by the MoH.

Indonesia has pursued continuous improvements in the national digital health implementation. To this end, the COVID-19 pandemic accelerated the introduction of national digital health capacity and raised the number of daily health applications users, currently exceeding 15 million people [5,6]. Through the launch of the Blueprint of Digital Health Transformation Strategy 2024, the government committed to using digital technology and data for public health to support the realization of a healthy Indonesia. The priority of Indonesia's digital health transformation activities include the integration and development of health data, the integration and development of application systems, and the developmet of a health technology ecosystem. These transformations aimed to collect standardized health data into a centralized platform named *Satu Sehat* (One Health Data). This process began with designing the health data architecture and interoperability as well as assessing the current infrastructure and security levels [7].

Implementing such a large-scale digital transformation requires technical maturity in the health sectors' information and communication technology (ICT). The critical factors of ICT maturity include infrastructure, policies, human capital development, change management, strategy, leadership, partnership, and collaboration [8]. In 2017, the MoH used the Health Metrics Network of the World Health Organization as the basis for a national digital health strategic framework, incorporating a holistic approach in planning, developing, implementing, and evaluating the use of ICT in health services. The strategic framework contained seven components: (1) governance and leadership; (2) strategy and investment; (3) services and application; (4) standards and interoperability; (5)

infrastructure; (6) legislation, policy, and compliance; as well as (7) workforce [2,9].

Chanyagorn and Kungwannarongkun [10] developed an ICT maturity assessment tool to explore ICT readiness in small- and medium-sized organizations in middle-income countries, covering both public and private sectors. This assessment tool was used to evaluate the ability of consumers, businesses, and governments to use ICT to their advantage; however, this study assessed ICT maturity in sample governmental public health services' facilities and institutions. The most significant aspect of this tool was the assessment of maturity for various aspects, as it could be adjusted to reflect best the different mixes of conditions in middle-income countries like Indonesia. The tool assessed four main ICT factors: (1) infrastructure, (2) hardware, (3) software and system information, as well as (4) people and human resources [10].

Therefore, as part of the digital health development in Indonesia, it is important to understand the ICT maturity level in services immediately after the pandemic, as a benchmark for future ICT initiatives. This study focused its assessment on ICT maturity across primary health care services and district health offices in Indonesia, applying the ICT maturity assessment tool for middle-income countries. To maintain a high level of inclusivity of services across different geographical regions, the evaluation of digital health capability was investigated in 9 provinces, spread over the 5 biggest islands in Indonesia. This study is the first study to assess the ICT maturity through middle-income country approaches for public health centers, primary clinics, and district health offices.

Methods

Data Collection

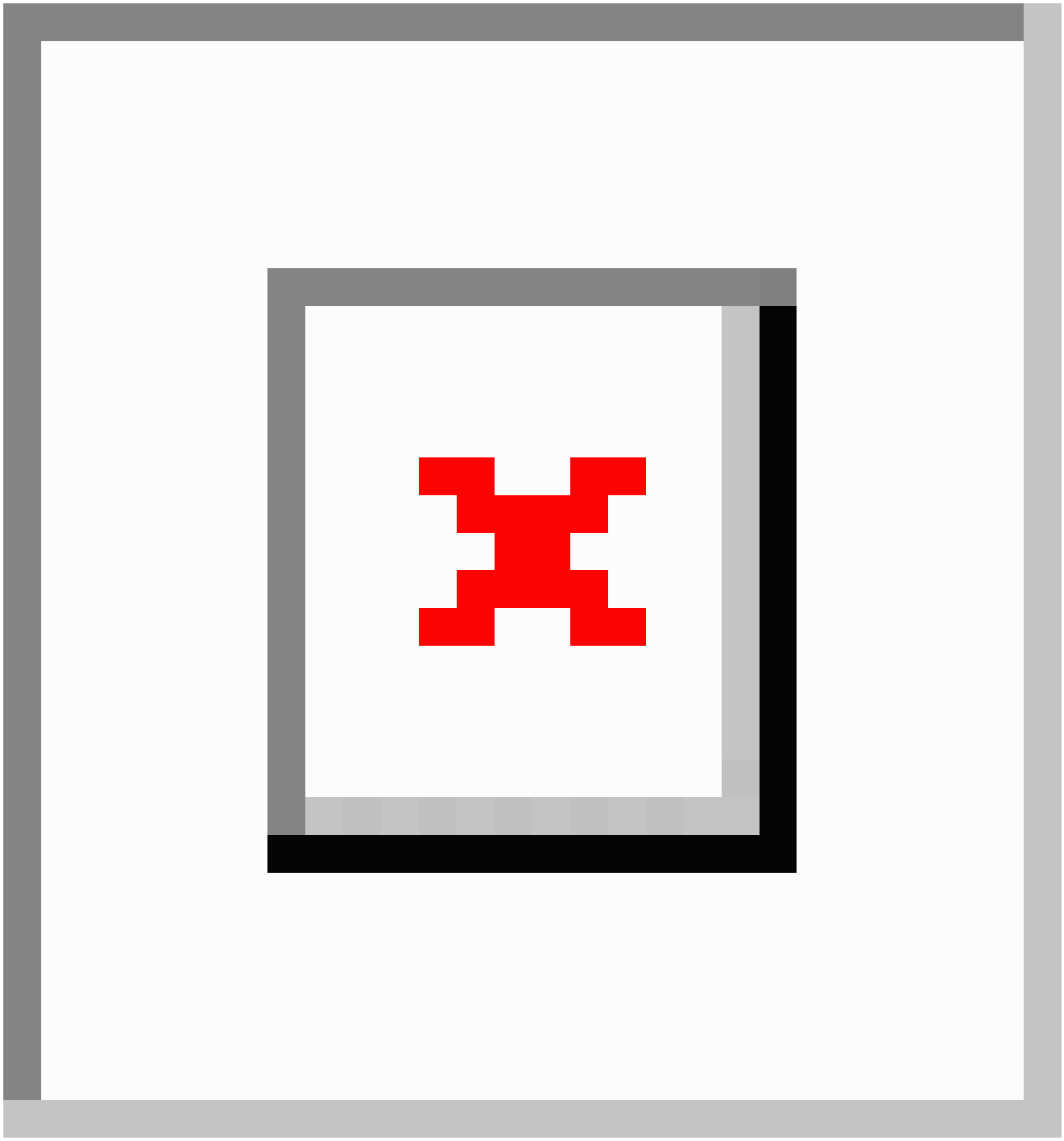
A cross-sectional survey design involving the ICT maturity assessment tool for middle-income countries was used in 9 provinces in Indonesia. Targeted stratified sampling was applied to choose the participants from provinces on the 5 largest islands in the country. DKI Jakarta, West Java, Banten, and East Java provinces on Java Island; Aceh province on Sumatra Island; East Kalimantan province on Kalimantan Island; Central Sulawesi province on Sulawesi Island; West Nusa Tenggara on the Nusa Tenggara Islands; and Maluku province on Maluku and Papua Islands. Provinces chosen on the 5 largest islands were based on region representativeness and health facility characteristics, which included urban, rural, and remote areas.

The targeted respondents were representatives of 3 main health sites in the province, including public health centers (*Puskesmas*), primary care clinics (*Klinik Pratama*), and district health offices (*Dinas Kesehatan Kabupaten*/*Kota*). A letter of invitation was sent to the targeted province district health office as the official invitation for an electronic survey and a forum group discussion (FGD) session. Each participant involved in the survey and FGD session was selected based on top management's assignment to attend the electronic survey and FGD meeting. To address participant selection and response biases, all assigned participants from all regions needed to fulfill the following inclusion criteria: (1) personnel who handled and oversaw the health information system implementation in each site, (2) with knowledge of the health information system used at the health site, and (3) with the ability to communicate with local staff to complete the survey accurately.

An ICT maturity assessment questionnaire was adopted and modified, adjusting to the local country's situation (Multimedia Appendix 1). The modification was made to the questions of each ICT subcomponent, especially the organization knowledge component, software security and document component, and network security component (Figure 1). A pilot survey test was conducted with 20 participants to gather feedback from participants, ensuring the questions were easy to understand, determining the time required to complete the questionnaire, and capturing suggestions to improve the assessment form. The questionnaire (Multimedia Appendix 2) consisted of four sections: (1)human resources, covering the availability and capacity of the personnel in using ICT; (2) software and systems, covering the number of health information systems used, data reporting in the health information system, troubleshooting, and maintenance performance; (3) hardware, including the availability of PCs/laptops, servers, storage, and manual entry data; and (4) infrastructure, including access to the internet access, electricity, and physical facility/building (Figure 1).

The survey was conducted by gathering the respondents through web-based meetings. The process was completed between February and March 2022. There were 237 respondents divided into 7 FGDs, with each FGD facilitated by 5 facilitators. The data were collected through 2 sessions in the meetings, which took 2.5 hours to complete. The 2 sessions were as follows: (1) filling out the questionnaires for the quantitative score and (2) an FGD session to discuss the survey findings and further explore the assessment results obtained. Each individual component of the ICT maturity questionnaire was discussed in more detail to obtain more comprehensive qualitative information. Discussions were continued until thematic saturation was reached.

Figure 1. The modified components evaluated in the information and communication technology (ICT) maturity assessment.

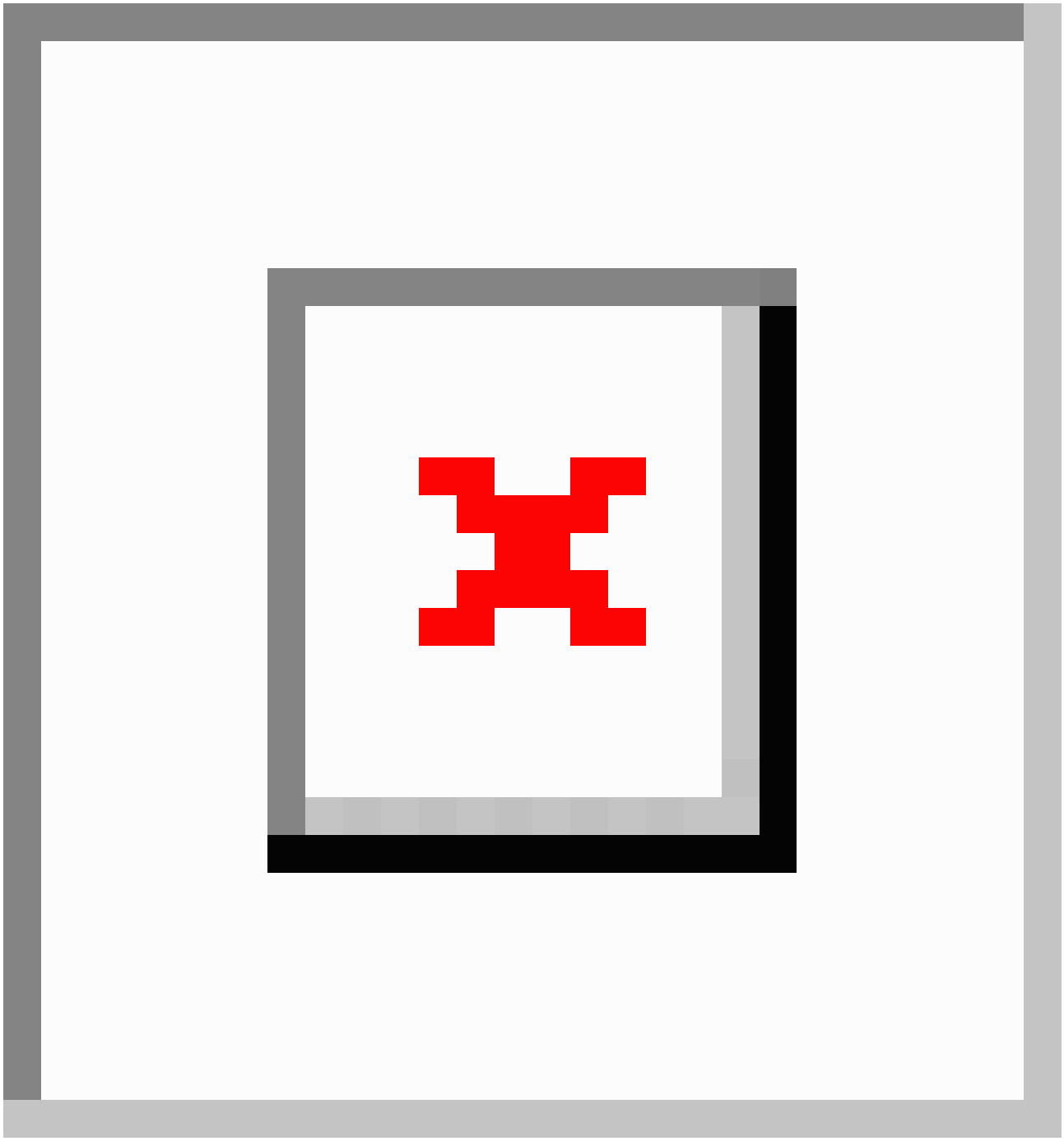


Data Analysis

The ICT maturity assessment tool consisted of 14 subsections containing 30 questions (Multimedia Appendix 2). The participants' responses to these questions were categorized into 5 levels of digital maturity: (1) initial, (2) basic, (3) good, (4) best practice, and (5) excellent (Figure 2), with a score of 1 indicating the lowest level of digital maturity (initial) and 5 representing the highest (excellent). The electronic survey responses were recorded by our team using the Mentimeter forms database tools, and the FGD results were input into a table using Microsoft Excel. The scores of each participating organization, including public health centers, primary care clinics, and district health offices, were used to determine their ICT maturity levels. The organization's score was calculated and averaged for each province.

The analysis was performed in Microsoft Excel. The scores from the 4 components were used to produce the mean score. A comparison analysis at the district and province levels was also generated. The FGD, held via Zoom, was recorded, and qualitative information obtained from participants was categorized based on the ICT component to be summarized as the FGD results for all provinces.

Figure 2. Information and communication technology maturity level.



Ethical Considerations

According to the "Exemption" section of the Indonesian MoH's National Guidelines and Standards for Ethical Research and Development in Health (2017), studies are exempted from review process if there is no/little potential risk/harm arising from the conduct of the research or when the information collected is available from the public domain; if they involve the use of educational tests (cognitive, diagnostic, attitude, and achievement); and if they involve survey or interview procedures, or public behavior observations. Research and demonstrations conducted by or subject to the approval of a department or agency and designed to study, evaluate, or assess the benefits of public programs or services, as well as other goods and services identified in the regulations, are also exempted from obtaining ethics approval. The original guidelines are available in Indonesian language of the MoH's guidelines (Chapter IIIB, point 2) [11]. The participants' data were also anonymized, following the Indonesian MoH's National Guidelines and Standards for Ethical Research and Development in Health (2017), Chapter IIIB, point 2.

Results

The study respondents were 237 in total and included 121 representatives of public health centers, 67 IT staff from district health offices, and 49 primary care clinic staff (Table 1). The provinces of East Java, Aceh, East Kalimantan, and West Nusa

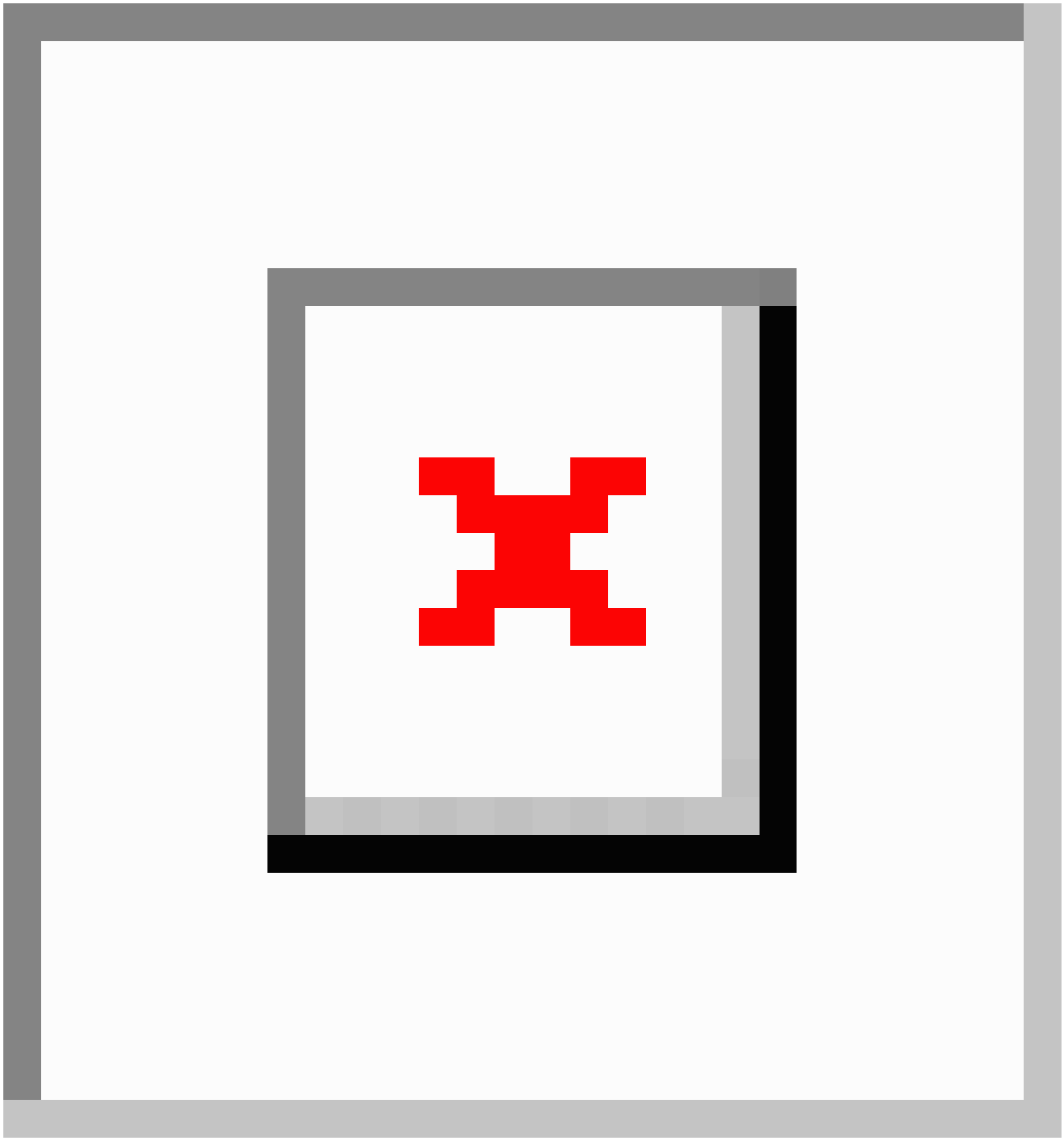
Tenggara had the largest representation among the participating public health centers, with 27, 17, 15, and 15 public health centers from each province, respectively. As for the district health offices, Aceh, West Nusa Tenggara, and East Kalimantan each had 11, 9, and 9 district health offices, respectively, while East Java had the highest representation with 13 district health offices. Notably, representatives from West Nusa Tenggara led the primary care clinic segment, accounting for 12 primary care clinics, followed by East Java with 9 primary care clinics and Aceh with 8 primary care clinics.

On average, the ICT maturity scores were broken down as follows: 2.71 for human resource, 2.83 for software and systems, 2.59 for hardware, and 2.84 for infrastructure, resulting in an overall average score of 2.74 (Figure 3). Based on the analysis, health care providers in Indonesia had an ICT maturity between basic and good according to the ICT maturity level pyramid.

Table . The study participants.

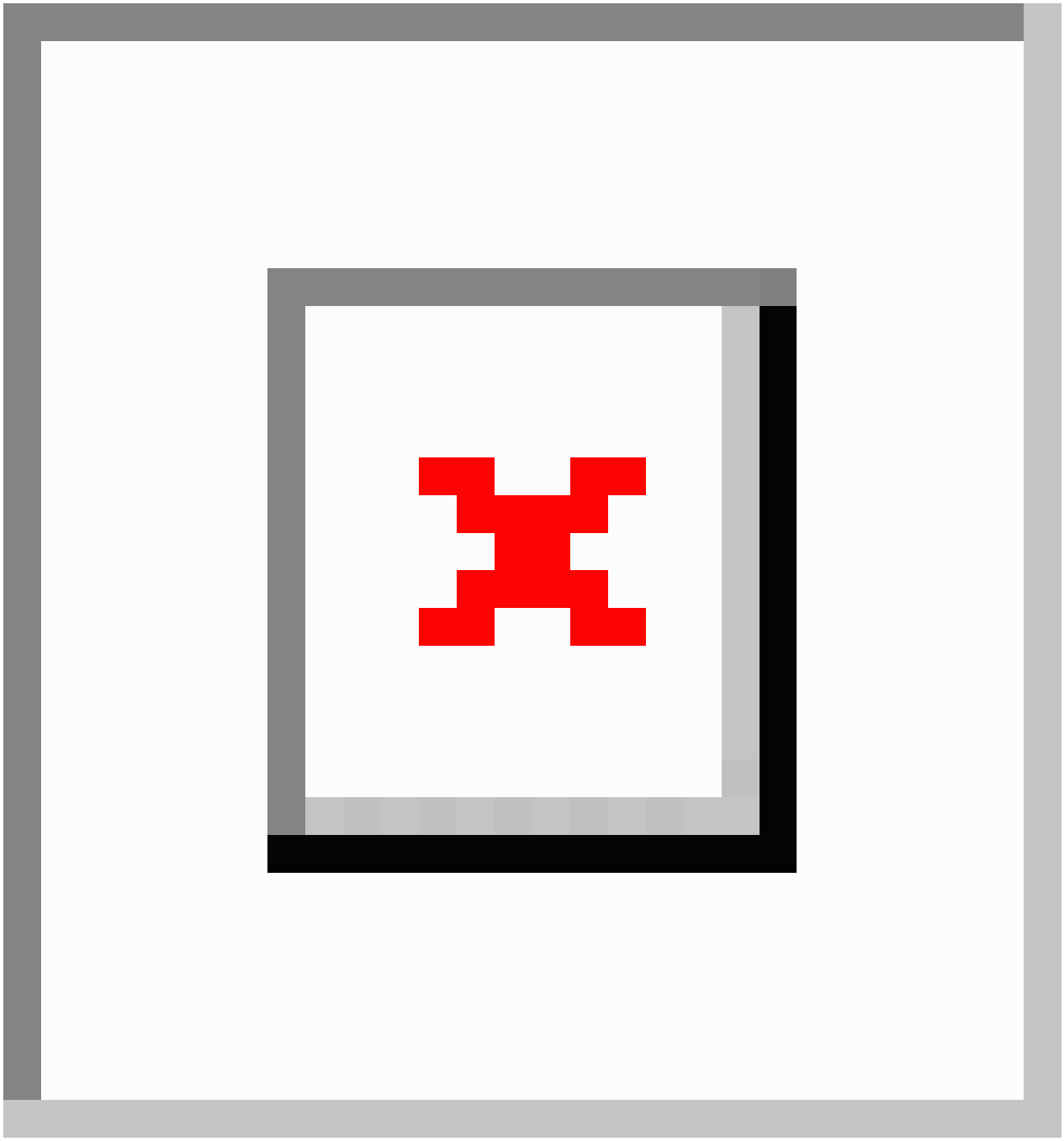
Island and province	Public health center, n	District health office, n	Primary care clinic, n	Subtotal, n
Java, DKI Jakarta	6	5	1	12
Java, West Java	13	6	5	24
Java, Banten	5	2	1	8
Sumatra, Aceh	17	11	8	36
Java, East Java	27	13	9	49
Nusa Tenggara, West Nusa Tenggara	15	9	12	36
Kalimantan, East Kalimantan	15	9	6	30
Sulawesi, Central Sulawesi	11	8	2	21
Maluku, Maluku	12	4	5	21
Total	121	67	49	237

Figure 3. Information and communication technology maturity scores across four main components at each site.



There were variations in the ICT maturity scores by province (Figure 4). On average, the ICT maturity score at the district health offices level was 2.60. Banten Province had the best overall score (3.00), followed by East Java (2.85), DKI Jakarta (2.75), and West Nusa Tenggara (2.75). At the public health center level, the average score was 2.69, with the highest average score belonging to Jakarta (3.18), followed by West Nusa Tenggara (2.93) and East Java (2.83). The average ICT maturity scores at the primary care clinics level were 2.87 for human resource capability and system management, 2.87 for software and information systems, 2.90 for hardware, and 3.11 for infrastructure. The DKI Jakarta province had the highest average ICT maturity score at the primary care clinic level (3.30), followed by Banten (3.10) and Aceh (3.05).

Figure 4. Average scores for the information and communication technology maturity level at (A) district health offices (DHOs); (B) public health centers (PHCs); and (C) primary care clinics.



Our analysis revealed that the public health centers, primary care clinics, and district health offices exhibited varying degrees of maturity across provinces (Figure 5). Important findings included the human resource scores, where Aceh, Banten, and Sulawesi had better developed human resource capacities, with ratings exceeding 3. A total of 9 provinces had scores ranging from 2 to 3 when it came to software and system maturity; the only exceptions were the public health centers and primary care clinics in DKI Jakarta, with scores of 3.54 and 3.38, respectively. In terms of hardware maturity, the majority of regions scored between 2 and 3, although some had significantly lower scores than others, such as Maluku's district health office (1.67) and the *Puskesmas* in Central Sulawesi (1.83). Primary care clinics typically showed greater maturity in terms of infrastructure, with an average score of 3 or higher. However, Maluku's public health center received a score of 2.06, whereas the district health offices in Maluku and Central Sulawesi received a lower score, ranging around 2.

When examining the differences in ICT maturity based on the location of public health centers, the distribution is evident. Figure 6 shows a detailed scatter plot graph created with the ICT maturity assessment scores, indicating differential maturity levels across urban and remote locations. In terms of human resources, it is apparent that isolated locations have lower maturity scores compared to their equivalent urban sites, although urban and rural areas have different variances. In

addition, there are significant differences in the software maturity scores between rural and urban areas. The software maturity varies in rural areas compared to urban areas, where the scores are constantly higher. On the other hand, software maturity varies in remote places.

Figure 5. Information and communication technology maturity assessment score ranges across various provinces and health care levels.

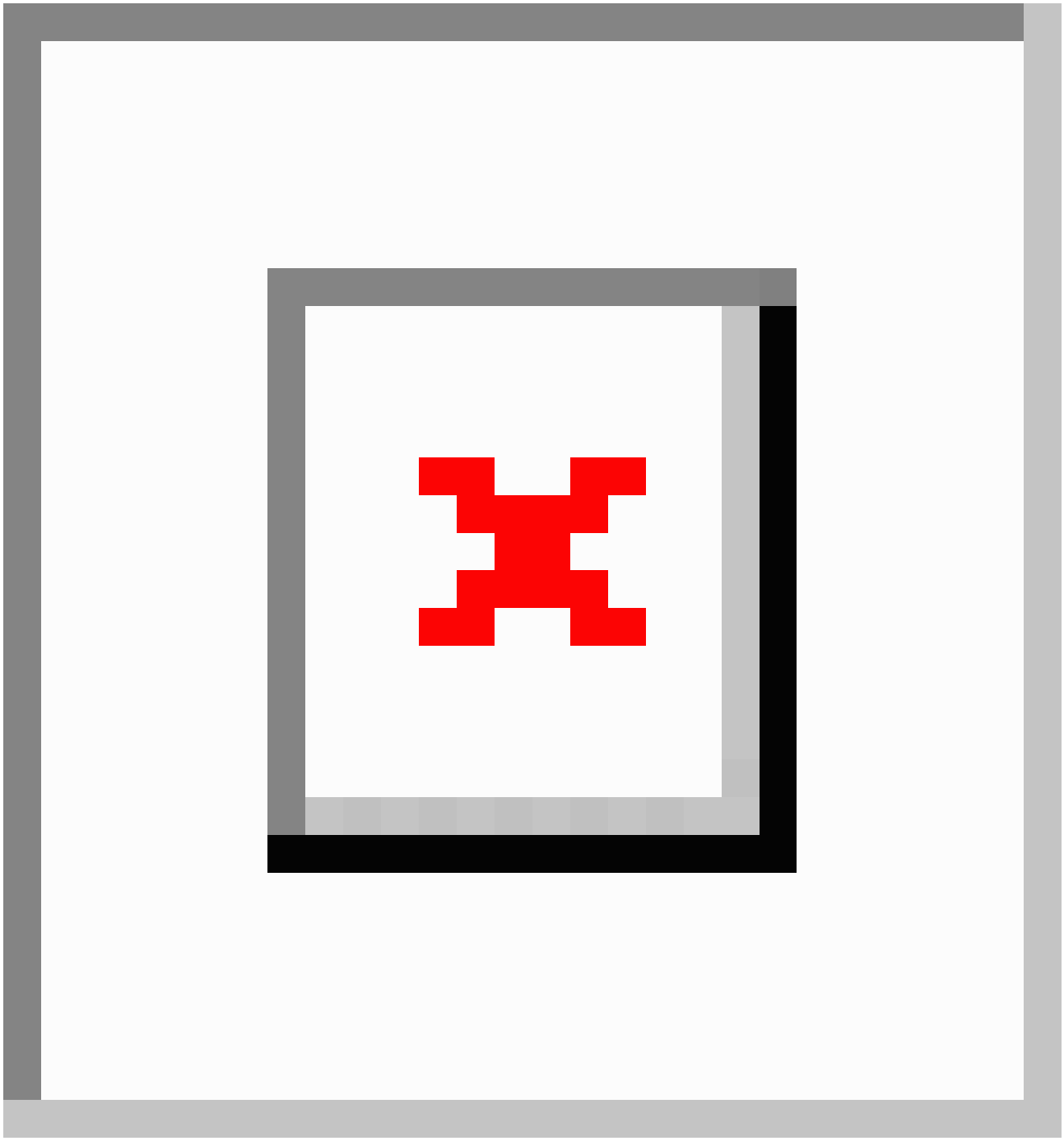
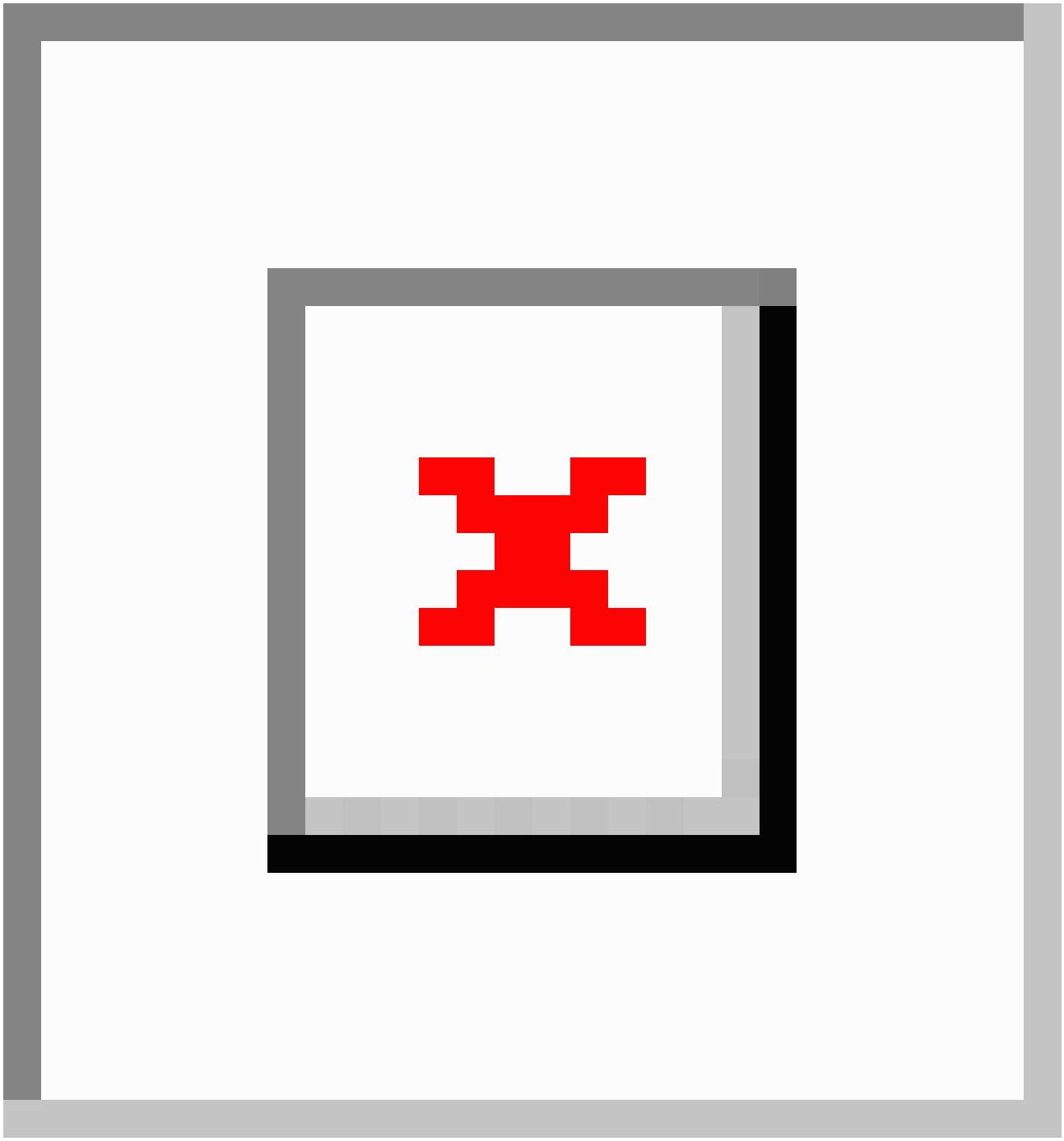


Figure 6. Information and communication technology maturity scores scatter plot by main components and site locations in the 9 targeted provinces.



The lower scores for the software system maturity indicated that the facility's questionnaire responses reflected qualities that hindered digital transformation or digital maturity. Key issues identified included the absence of trained or specialized personnel for data input and IT. Additionally, there was an excessive number of applications or information systems mandated by central or local governments. Despite the abundance of applications, not all were used effectively, and many lacked interoperability, resulting in data duplication and increased workloads for health workers. Hardware components were scored based on several factors, including the reliance on manual, paper-based methods or computer-based Excel tables for data input; the insufficient number of personal computers available for health care services and data input; the type of data storage used, such as computer-based or cloud-based to minimize data loss; and the quality and stability of the data server. For infrastructure, the scores for some regions were lower due to the poor internet quality in health facilities and the inconsistent power availability, which was not guaranteed 24 hours a day. In some cases, power outages occured even during operational hours.

Regional differences in hardware quality were also evident, with varying degrees of quality found in rural, urban, and remote locations. While remote areas had challenges due to lower quality or limits, rural and urban areas generally had higher standards for hardware quality. The trends seen in hardware and infrastructure were strikingly comparable, highlighting a shared tendency between these elements. These findings offer

important insights into the particular opportunities and problems in various situations and they portrait an intricate picture of ICT maturity across various geographies. Additionally, these findings highlight specific area characteristics, which may aid the government in deciding appropriate approaches for ensuring the effective implementation of digital health services in these areas.

In addition to assessing digital maturity through scoring, we conducted FGDs, the findings of which highlighted a number of staffing- and human resources–related issues. First, there were people in the health care facilities and district health offices with a variety of backgrounds, but not all of them were IT literate. Second, data input was typically done by health officials or health workers; these staff members lacked professional IT certification or training. Third, data analysis was usually performed by nurses and staff at health centers due to a shortage of professionals with necessary capabilities. Fourth, a large number of different unintegrated health information systems added to the burden of data reporting. Lastly, the lack of hiring civil servants to fill the paucity of IT professionals with computer knowledge was a further obstacle.

In terms of software and information systems, issues included frequent system disruptions, poor system performance, and operational difficulties. The largest issue with software was the sheer number of health information systems that required data input but were not interoperable. For instance, a question about immunization status appeared in two existing systems: the system for tracking toddlers' nutritional status and the immunization system. As there were no data integration in place, health care personnel were forced to continually enter the same data into multiple systems as a result of this redundancy.

The complexity was further increased by the ongoing addition of new applications from numerous sources, such as the MoH, other governmental organizations, or municipal governments. These applications frequently demanded a lot of data entry, which made data management quite difficult. Furthermore, the prevalence of paper-based reporting techniques persisted even in the face of digital alternatives. These difficulties highlight the need for more efficient and integrated solutions in the health care industry by impeding the overall simplification of data entry and system operations.

Numerous hardware-related issues emerged during the FGDs. Program managers frequently had to use their own laptops because their public health centers did not have enough resources. Despite the fact that the office had multiple computers, they were for shared use and in some subdistricts, the computers were limited and used only for the registration and administrative sections. In others, there were only 2-5 computers available to fulfill the management requirements of the entire public health center. Staff members occasionally used their personal laptops for data reporting. Furthermore, there were service recording delays due to PC storage limits. In addition, operational difficulties were exacerbated when individuals' PCs experienced slowness or malfunctions and when data storage was not centralized but rather restricted to individual PCs without cloud backup.

Regarding infrastructure, the intermittently weak signal on Wi-Fi and internet-connected devices in several places was mentioned. The employees routinely used Wi-Fi, but they frequently ran into network issues. Consequently, all reporting needed to be completed manually in certain locations or by using personal mobile data or paper-based data collection. Moreover, server outages occasionally caused delays or even entirely stopped the data recording process. Additionally, there was no dedicated data storage, thus data were at risk during sporadic power outages, which were particularly problematic in situations such as wildfires.

Discussion

Principal Findings

The analysis of ICT maturity in Indonesia's care providers reveals an overall average score of 2.74, indicating maturity between basic and good levels, aligned with the ICT maturity level pyramid. Variations exist across provinces and health care levels, with Banten Province exhibiting the highest overall score. The DKI Jakarta Province stands out with superior scores at the public health center and primary care clinic levels. Disparities in human resource, software, hardware, and infrastructure maturity exist between provinces and health care levels, with rural areas generally lagging behind urban areas. These findings underscore the importance of tailored strategies to address regional disparities and enhance digital health service implementation effectively.

The ICT maturity level has the potential to affect the quality of national health services, with the highest impact felt in middle-income countries. Excellent ICT maturity can lead to a good digital health implementation and potentially contribute to the improvement of health care, such as quality of care, supplies and logistics, training and communication, community engagement and participation in health services, as well as the availability and use of routine data by decision-makers [12-14]. The current research provides a snapshot of the ICT maturity in several health care services across Indonesia and explores factors that would require further attention.

The results of the questionnaires show the variety of ICT maturity among provinces, districts, and health facilities, with variations in critical components, such as human resource capabilities, software and information systems, hardware, and infrastructure. These disparities not only indicate the different provinces' ICT readiness but also highlight potential areas for improvement and development to enhance health care service delivery and information management in each respective region. Understanding these variations is crucial for devising targeted strategies to address specific ICT-related challenges and opportunities within the health care system across diverse geographical locations.

The development of ICT in Indonesia is facing tough challenges due to the country's geography, consisting of thousands of islands and many cultures, affecting educational, social, and economical aspects in the country. Moreover, Indonesia as an archipelago has many remote areas where telecommunication and internet service providers cannot develop the sufficient

infrastructure; as such, these areas are unavoidably underrepresented [14-16]. Indonesia ranked 111 on the 2017 ICT Development Index, falling behind other countries in Southeast Asia like Singapore (ranking 18), Brunei Darussalam (ranking 53), Malaysia (ranking 63), Thailand (ranking 78), and Philippines (ranking 101) [17].

Regarding human resources, often staff members do not have adequate digital literacy, and yet are required to report using IT processes, as it was also indicated in our work [18,19]. Moreover, there is a shortage of staff with computer expertise and insufficient data management training opportunities to mitigate this issue. The same is described in Malaysia, where challenges related to human resources include workload, readiness, skills, and user dependency. Additionally, tasks often require health workers to focus on mining data, instead of improving service provision [20-23].

Indonesia also faces a mind shift challenge for staff, as the plurality of overlapping systems comes against a context of well-established manual data collection, using paper or status books. In contrast, frequent input of overlapping variables across multiple software, adds to the work burden of individual health care staff and makes the digital health process inefficient. Countries in sub-Saharan Africa have also faced a similar challenge: they documented 738 distinct digital health interventions at different levels of functioning in the sub-Saharan African region over the past 10 years. One in 5 of those did not have a link to any health service outcomes, and only half could be classified as “established” at the end of the study period. Two of every 3 were focused solely on solutions for a single health care activity, limiting integration [20]. This aspect has not been researched as yet for the countries neighboring Indonesia.

The existence of infrastructure remains the main challenge in developing and implementing e-governments in middle-income countries [24,25]. In the health care industry, the continuity of using ICT in daily and routine operations depends on the availability of a robust IT infrastructure. Unfortunately, many middle-income countries lack the necessary infrastructure to support digital health, specifically telecommunication and electricity networks coverage [25-27]. This is also supported by other studies that found that IT systems used in the health care industry would only be optimal, effective, and efficient when adequate facilities and infrastructure supported them [28-31].

The discrepancies in infrastructure availability across the provinces in Indonesia directly affect a stable internet connection. Frequent internet signal downtime, network problems, poor connection, and insufficient computer availability, including the lack of electricity supplies in some remote areas, were found in this study, showing the need for further improvements in technology infrastructure and facilities. This lack and instability of IT infrastructure, such as the limitation of internet access, electricity supply, and availability of computers, is common across low- and middle-income countries, as similar findings were described for Brazil, sub-Saharan Africa as a whole, Sierra Leone, and Tanzania [32-36].

These findings are crucial for gaining a better understanding to support the Indonesian health technology transformation mentioned in the Blueprint of Health Transformation Strategy 2024. The foundation to transform the health systems lies in having a solid platform architecture design and infrastructure to implement integrated and interoperable health systems. The blueprint highlighted the plan to integrate all electronic medical record systems from public health centers, hospitals, primary care clinics, and laboratories into the *Satu Sehat* platform and to adopt the *Satu Sehat* standard. This study showed that to support the implementation of digital health transformation, the government should identify the gap; map the area based on capacity; and provide assistance to improve the software, hardware, infrastructure, and human resource capacity, especially in areas with lower ICT maturity scores.

Strengths and Limitations

The strengths of the study involve exploring ICT maturity in Indonesia, engaging several health care service sites (ie, *Puskesmas*, clinics, and health departments). Furthermore, the results of the FGDs provided a more detailed overview of the challenges related to the four assessed components, thus highlighting areas for future improvements to support the ongoing digital transformation of health care.

The limitations of the study are as follows: (1) the involvement of public health centers and primary clinics in the study was limited in each province and perhaps not entirely representative, although health departments involved in the study represented 80% - 90% of the 9 targeted provinces and districts; (2) other health care facilities, such as hospitals, laboratories, and pharmacies, were not involved in this study, and as such, additional aspects of needs or challenges may exist that have not been highlighted as yet; this may have implications for the representativeness of health care data users by volume; (3) the representativeness based on infrastructure distribution and regional characteristics (eg, urban, rural, remote, and very remote regions) should be viewed only as indicative, as health care services from more remote rural regions are less likely to have access to stable internet connection, and thus, unable to complete such questionnaires distributed by digital channels; (4) the respondents’ capacity can influence data completeness, and thus, it will be useful for future iterations if staff representing more functions of the health care centers are able to complete the questionnaire.

Conclusions

This study investigated for the first time the variations of ICT maturity across the health care systems (eg, public health centers, primary clinics, and district health offices) in 9 provinces in Indonesia, underscoring the diversity in ICT implementation and readiness. The maturity of ICT use was influenced by several critical components, specifically enhancing human resources, ensuring infrastructure, the availability of supportive hardware, and optimizing information systems. The findings of this study are in line with similar studies in other middle-income countries in the world. Our results demonstrate that to attain ICT maturity in health care services in Indonesia, it is imperative to address all of the above aspects, as each

represents ongoing needs and has been shown to be equally important and necessary in the field.

Conflicts of Interest

Although LM and DNA are affiliated with Aceso Global Health Consultants Pte Limited, which is a private company, the authors declare that this research project does not receive funding from Aceso Global Health Consultants. The company does not have a role in the study design, data collection, and analysis; decision to publish; or preparation of the manuscript. Where authors are identified as personnel of the International Agency for Research on Cancer or the World Health Organization (WHO), the authors alone are responsible for the views expressed in this paper, and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer or WHO. All authors declared no other conflicts of interest.

Multimedia Appendix 1
Modified information and communication technology (ICT) maturity assessment.
[DOCX File, 441 KB - medinform_v12i1e55959_app1.docx]

Multimedia Appendix 2
Questionnaire and in-depth interview guidelines.
[DOCX File, 220 KB - medinform_v12i1e55959_app2.docx]

References

1. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2023-11-12]
2. Woods L, Eden R, Pearce A, et al. Evaluating digital health capability at scale using the digital health indicator. Appl Clin Inform 2022 Oct;13(5):991-1001. [doi: 10.1055/s-0042-1757554] [Medline: 36261114]
3. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022;2022. [doi: 10.31478/202206e] [Medline: 36177208]
4. ICT trends. Asian and Pacific Training Centre for Information and Communication Technology for Development (APCICT). 2017. URL: https://www.unapcict.org/resources/publications/ict-trends [accessed 2023-11-12]
5. Pitaloka AA, Nugroho AP. Digital transformation in Indonesia health care services: social, ethical and legal issues. J STI Policy Manage 2021;6(1):51-66. [doi: 10.14203/STIPM.2021.301]
6. Saputra YE, Worsito SB, Firdaus DS, Listiyandini RA. Bridging a resilient recovery post-pandemic through digital health transformation. In: Indonesia Post-Pandemic Outlook: Rethinking Health and Economics Post-COVID-19: Overseas Indonesian Student's Association Alliance & BRIN Publishing; 2022. [doi: 10.55981/brin.537.c516]
7. Blueprint of Digital Health Transformation Strategy 2024. Indonesia Ministry of Health. 2021. URL: https://dto.kemkes.go.id/ENG-Blueprint-for-Digital-Health-Transformation-Strategy-Indonesia%202024.pdf [accessed 2023-11-03]
8. Zaeid ANH, Khairalla FA, Al-Rashed W, Zaied H. Assessing e-readiness in the Arab countries: perceptions towards ICT environment in public organisations in the state of Kuwait. Electro J e-Govern 2007;5(1):77-86 [FREE Full text]
9. Strategi e-kesehatan nasional. Indonesia Ministry of Health. 2017. URL: https://peraturan.bpk.go.id/Details/139565/permenkes-no-46-tahun-2017 [accessed 2023-11-03]
10. Chanyagorn P, Kungwannarongkun B. ICT readiness assessment model for public and private organizations in developing country. IJIET 2011;1(2):99-106. [doi: 10.7763/IJIET.2011.V1.17]
11. Pedoman dan standar etik penelitian dan pengembangan kesehatan nasional [Article in Indonesian]. Indonesia Ministry of Health. URL: https://kepk.poltekkestasikmalaya.ac.id/wp-content/uploads/2018/05/2017-KEPPKN-Standar-dan-Pedoman-.pdf [accessed 2024-07-15]
12. UNICEF's approach to digital health. UNICEF Health Section Implementation Research and Delivery Science Unit and the Office of Innovation Global Innovation Centre. 2018. URL: https://www.unicef.org/innovation/reports/unicefs-approach-digital-health%E2%80%8B%E2%80%8B [accessed 2023-11-12]
13. Shaygan A, Daim T. Technology management maturity assessment model in healthcare research centers. Technovation 2023 Feb [FREE Full text] [doi: 10.1016/j.technovation.2021.102444]
14. Sari EN. Navigating the landscape of digital health landscape, Indonesia. Health Intervention and Technology Assessment Program (HITAP) Ministry of Public Health (MoPH) Thailand. 2022. URL: https://www.hitap.net/en/documents/187528 [accessed 2023-11-03]
15. Amin M. ICT for rural area development in Indonesia: a literature review. JITU 2018;1(2):32. [doi: 10.30818/jitu.1.2.1881]
16. Handayani PW, Yazid S, Bressan S, Sampe AF. Information and communication technology recommendations for the further development of a robust national electronic health strategy for epidemics and pandemics. J Sistem Inf (J Inf Sys) 2020 Oct;16(2):31-42 [FREE Full text] [doi: 10.21609/jsi.v16i2.979]
17. Measuring the information society report: volume 1. International Telecommunication Union. 2017. URL: https://www.itu.int/en/ITU-D/Statistics/Pages/publications/misr2018.aspx [accessed 2023-11-03]

18. Empowering the health workforce: strategies to make the most of the digital revolution. OECD. 2019. URL: https://www.oecd.org/publications/empowering-the-health-workforce-to-make-the-most-of-the-digital-revolution-37ff0eaa-en.htm [accessed 2023-11-20]
19. Karamagi HC, Muneene D, Droti B, et al. eHealth or e-chaos: the use of digital health interventions for health systems strengthening in sub-Saharan Africa over the last 10 years: a scoping review. J Glob Health 2022 Dec 3;12:04090. [doi: 10.7189/jogh.12.04090] [Medline: 36462201]
20. Ajadi S. Digital health: a health system strengthening tool for developing countries. GSM Association. 2020. URL: https://www.gsma.com/mobilefordevelopment/resources/digital-health-a-health-system-strengthening-tool-for-developing-countries/ [accessed 2023-11-20]
21. Strategy on human resources for universal access to health and universal health coverage. PAHO. 2017. URL: https://www.paho.org/en/documents/strategy-human-resources-universal-access-health-and-universal-health-coverage-csp2910 [accessed 2023-11-20]
22. ChePa N, Md Jasin N, Abu Bakar NA. Information system implementation failure in Malaysian government hospitals: how change management helps? J Telecommun Electron Comput Eng 2018;10(1-11):69-75 [FREE Full text]
23. Myyrä N. Digital health interventions for employees: are digital health interventions able to improve a company's performance? Helsinki Metropolia University of Applied Sciences. 2016. URL: https://www.theseus.fi/handle/10024/115843 [accessed 2023-11-20]
24. Ndou V. E-government for developing countries: opportunities and challenges. E J Info Sys Dev Countries 2004 Jun;18(1):1-24 [FREE Full text] [doi: 10.1002/j.1681-4835.2004.tb00117.x]
25. Delponte L, Grigolini M, Moroni A, Vignetti S. ICT in the developing world: in-depth analysis. Sci Technol Options Assess 2015:6-9. [doi: 10.2861/61950]
26. Omotosho A, Ayegba P, Emuoyibofarhe J, Meinel C. Current state of ICT in healthcare delivery in developing countries. Int J Onl Eng 2019 May;15(8):91. [doi: 10.3991/ijoe.v15i08.10294]
27. Djawad YA, Suhaeb A, Mustakim R, Jaya H, et al. The development of an intelligent e-health mobile application in Indonesia: a preliminary study. INSIST 2019 Oct;4(2):240-245.
28. Ebnehoseini Z, Tabesh H, Deldar K, Mostafavi SM, Tara M. Determining the hospital information system (HIS) success rate: development of a new instrument and case study. Open Access Maced J Med Sci 2019 May 15;7(9):1407-1414. [doi: 10.3889/oamjms.2019.294] [Medline: 31198444]
29. Afrizal SH, Handayani PW, Hidayanto AN, Eryando T, Budiharsana M, Martha E. Barriers and challenges to primary health care information system (PHCIS) adoption from health management perspective: a qualitative study. Informatics in Medicine Unlocked 2019;17:100198. [doi: 10.1016/j.imu.2019.100198]
30. Health Metrics Network. Framework and Standards for Country Health Information Systems: World Health Organization; 2007.
31. Sayyadi Tooranloo H, Sepideh S, Arezoo Sadat A. Evaluation of failure causes in employing hospital information systems. J Syst Manag 2021;6(3):31-76. [doi: 10.30495/JSM.2021.678894]
32. Atashi A, Khajouei R, Azizi A, Dadashi A. User interface problems of a nationwide inpatient information system: a heuristic evaluation. Appl Clin Inform 2016;7(1):89-100. [doi: 10.4338/ACI-2015-07-RA-0086] [Medline: 27081409]
33. Yoshiura VT, de Azevedo-Marques JM, Rzewuska M, et al. A web-based information system for a regional public mental healthcare service network in Brazil. Int J Ment Health Syst 2017;11:1. [doi: 10.1186/s13033-016-0117-z] [Medline: 28053659]
34. Odekunle FF, Odekunle RO, Shankar S. Why sub-Saharan Africa lags in electronic health record adoption and possible strategies to increase its adoption in this region. Int J Health Sci (Qassim) 2017 Oct;11(4):59-64. [Medline: 29085270]
35. Sukums F, Mensah N, Mpembeni R, et al. Promising adoption of an electronic clinical decision support system for antenatal and intrapartum care in rural primary healthcare facilities in sub-Saharan Africa: the QUALMAT experience. Int J Med Inform 2015 Sep;84(9):647-657. [doi: 10.1016/j.ijmedinf.2015.05.002] [Medline: 26073076]
36. Chukwu E, Garg L, Foday E, Konomanyi A, Wright R, Smart F. Electricity, computing hardware, and internet infrastructures in health facilities in Sierra Leone: field mapping study. JMIR Med Inform 2022 Feb 3;10(2):e30040. [doi: 10.2196/30040] [Medline: 35113026]

Abbreviations

BPM: private midwife
DPM: private general practitioner
FGD: forum group discussion
ICT: information and communications technology
MoH: Ministry of Health
WHO: World Health Organization

Edited by C Lovis; submitted 30.12.23; peer-reviewed by F Lau; revised version received 31.05.24; accepted 02.06.24; published 18.07.24.

Please cite as:
Aisyah DN, Setiawan AH, Lokopessy AF, Faradiba N, Setiaji S, Manikam L, Kozlakidis Z
The Information and Communication Technology Maturity Assessment at Primary Health Care Services Across 9 Provinces in Indonesia: Evaluation Study
JMIR Med Inform 2024;12:e55959
URL: https://medinform.jmir.org/2024/1/e55959
doi:10.2196/55959

Original Paper

PCEtoFHIR: Decomposition of Postcoordinated SNOMED CT Expressions for Storage as HL7 FHIR Resources

Tessa Ohlsen[1,2], MSc; Josef Ingenerf[1,2], Prof Dr; Andrea Essenwanger[3], MSc; Cora Drenkhahn[2], MSc

[1]IT Center for Clinical Research, University of Luebeck, Luebeck, Germany

[2]Institute of Medical Informatics, University of Luebeck, Luebeck, Germany

[3]mio42 GmbH, Berlin, Germany

Corresponding Author:
Tessa Ohlsen, MSc
IT Center for Clinical Research
University of Luebeck
Ratzeburger Allee 160
Luebeck, 23562
Germany
Phone: 49 45131015623
Email: t.ohlsen@uni-luebeck.de

Abstract

Background: To ensure interoperability, both structural and semantic standards must be followed. For exchanging medical data between information systems, the structural standard FHIR (Fast Healthcare Interoperability Resources) has recently gained popularity. Regarding semantic interoperability, the reference terminology SNOMED Clinical Terms (SNOMED CT), as a semantic standard, allows for postcoordination, offering advantages over many other vocabularies. These postcoordinated expressions (PCEs) make SNOMED CT an expressive and flexible interlingua, allowing for precise coding of medical facts. However, this comes at the cost of increased complexity, as well as challenges in storage and processing. Additionally, the boundary between semantic (terminology) and structural (information model) standards becomes blurred, leading to what is known as the TermInfo problem. Although often viewed critically, the TermInfo overlap can also be explored for its potential benefits, such as enabling flexible transformation of parts of PCEs.

Objective: In this paper, an alternative solution for storing PCEs is presented, which involves combining them with the FHIR data model. Ultimately, all components of a PCE should be expressible solely through precoordinated concepts that are linked to the appropriate elements of the information model.

Methods: The approach involves storing PCEs decomposed into their components in alignment with FHIR resources. By utilizing the Web Ontology Language (OWL) to generate an OWL ClassExpression, and combining it with an external reasoner and semantic similarity measures, a precoordinated SNOMED CT concept that most accurately describes the PCE is identified as a Superconcept. In addition, the nonmatching attribute relationships between the Superconcept and the PCE are identified as the "Delta." Once SNOMED CT attributes are manually mapped to FHIR elements, FHIRPath expressions can be defined for both the Superconcept and the Delta, allowing the identified precoordinated codes to be stored within FHIR resources.

Results: A web application called PCEtoFHIR was developed to implement this approach. In a validation process with 600 randomly selected precoordinated concepts, the formal correctness of the generated OWL ClassExpressions was verified. Additionally, 33 PCEs were used for two separate validation tests. Based on these validations, it was demonstrated that a previously proposed semantic similarity calculation is suitable for determining the Superconcept. Additionally, the 33 PCEs were used to confirm the correct functioning of the entire approach. Furthermore, the FHIR StructureMaps were reviewed and deemed meaningful by FHIR experts.

Conclusions: PCEtoFHIR offers services to decompose PCEs for storage within FHIR resources. When creating structure mappings for specific subdomains of SNOMED CT concepts (eg, allergies) to desired FHIR profiles, the use of SNOMED CT Expression Templates has proven highly effective. Domain experts can create templates with appropriate mappings, which can then be easily reused in a constrained manner by end users.

(JMIR Med Inform 2024;12:e57853) doi:10.2196/57853

KEYWORDS

SNOMED CT; HL7 FHIR; TermInfo; postcoordination; semantic interoperability; terminology; OWL; semantic similarity

Introduction

Background

The growing digitization of medical records has led to an increase in patient data available for health care analysis. These data must be utilized to improve medical care and offer more personalized treatments. However, to accomplish this, the ability to automatically exchange and process data between different systems is essential. This requires not only technical compatibility but also semantic interoperability, ensuring that the data's meaning is preserved when transferred to another system. The ability to exchange and utilize data across different systems is crucial for fully leveraging digital medical records and improving patient care [1].

To ensure semantic interoperability, it is essential to follow both structural and semantic standards. Structural standards specify the syntax for accessing data fields within information models. In recent years, the newly developed HL7 (Health Level 7) standard, FHIR (Fast Healthcare Interoperability Resources), has gained international recognition due to its emphasis on simplified implementation and the use of modern technologies [2]. Semantic standards, by contrast, involve terminologies that use language-independent codes to represent the meaning of data in an interoperable manner. SNOMED CT is widely recognized as the most comprehensive medical terminology for enhancing semantic interoperability [3]. In 2021, Germany acquired a national license for SNOMED CT, leading to increased interest and usage of the terminology. SNOMED CT concepts are now being integrated into data modeling efforts, such as the Medical Information Objects by the German National Association of Statutory Health Insurance Physicians (NASHIP) [4] and the core data set of the Medical Informatics Initiative (MII) [5]. The use of SNOMED CT, which contains over 350,000 concepts, aims to provide a machine-readable interlingua that minimizes coding issues specific to different countries and medical fields. However, due to the complexity of natural language, not all medical situations can be accurately coded using SNOMED CT's extensive set of precoordinated concepts. To address this and avoid a rapid increase in the number of new concepts, SNOMED CT, unlike many other vocabularies, supports postcoordination. This feature allows the combination of precoordinated concepts into new expressions using a formal grammar.

Therefore, postcoordination is a unique feature that significantly enhances the precision of medical documentation and greatly increases SNOMED CT's expressive power. However, the adoption of postcoordination has been slow due to various challenges. While some of these obstacles have already been addressed [6-9], integrating postcoordinated expressions (PCEs) into the electronic health records of legacy hospital information systems remains difficult for several reasons (Textbox 1).

Textbox 1. Obstacles to integrating postcoordinated SNOMED Clinical Terms expressions into electronic health records.

- **Adherence to familiar data structures**

Medical circumstances are usually documented using individual codes, and there are established methods for storing and processing these codes. While restrictions such as length-limited data types can be managed with simple codes, they pose challenges when dealing with arbitrarily large formal expressions, such as postcoordinated expressions (PCEs). This has led to concerns about the practicality of using PCEs [10].

- **Lack of technical support**

The technical handling of PCEs is inherently complex, requiring at a minimum a description logic reasoner and the implementation of several formal specifications, such as the Concept Model, Compositional Grammar, and Expression Constraint Language, as defined by SNOMED International [3]. While specialized terminology servers can help alleviate the implementation burden, currently only the CSIRO (Commonwealth Scientific and Industrial Research Organization) Ontoserver [11] supports postcoordination [12].

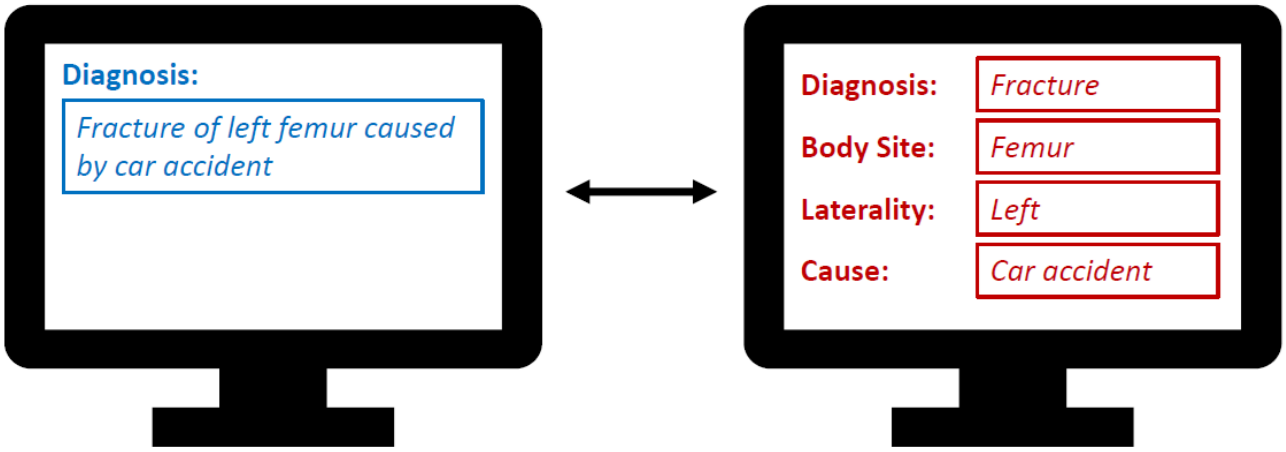
- **Difficulties with FHIR (Fast Healthcare Interoperability Resources) search**

When searching for information in FHIR resources, postcoordination is supported only if the exact same PCE has been explicitly and previously defined in a FHIR CodeSystem supplement. These supplements enable the extension of the standard FHIR CodeSystem for SNOMED CT with a collection of PCEs (eg, for value set definitions), but they do not support the recording of PCEs in a patient's electronic health record.

Given that these challenges are unlikely to be resolved in the short term, a different approach is needed to ensure interoperability between systems that support postcoordination and those that do not. Consequently, this paper will present an alternative representation of PCEs.

It has long been recognized that there is significant overlap between the scopes of structural and semantic standards, leading to unclear responsibilities and potentially ambiguous representations of medical facts, as well as inconsistent redundancies. This issue, known as the TermInfo problem [3,13-15], largely arose from the independent development of these standards, which resulted in mutual coverage of the same data elements (Figure 1). SNOMED CT's postcoordination capability further blurs the distinction between structure and semantics, potentially exacerbating the existing problem. However, PCEs are fully interpretable, allowing the information components they contain to be identified and flexibly disassembled. Building on extensive knowledge from previous projects [6,16-18] and the current integration of SNOMED CT with FHIR, the authors propose PCEtoFHIR—an application designed to decompose PCEs for storage as FHIR resources in a manner that preserves their meaning.

Figure 1. TermInfo: The same medical fact can be represented using either the terminology (left side) or the information model (right side) more heavily.



Related Work

Although there is a growing body of literature on the postcoordination of SNOMED CT concepts [6,16,19-21], a literature search revealed no existing publications specifically addressing the storage of PCEs in FHIR resources. However, a Health Level Seven International (HL7) working group is addressing this topic and provides several resources. For selected FHIR resources, such as Condition and Observation, mappings of some SNOMED CT attributes to FHIR are currently available (see [22] and [23]). Additionally, the Confluence pages "SNOMED on FHIR" ("Bindings to FHIR Clinical Resources") [24] outline various options for mapping SNOMED CT attributes to FHIR while avoiding semantic overlaps. The information from these documents is considered in our work.

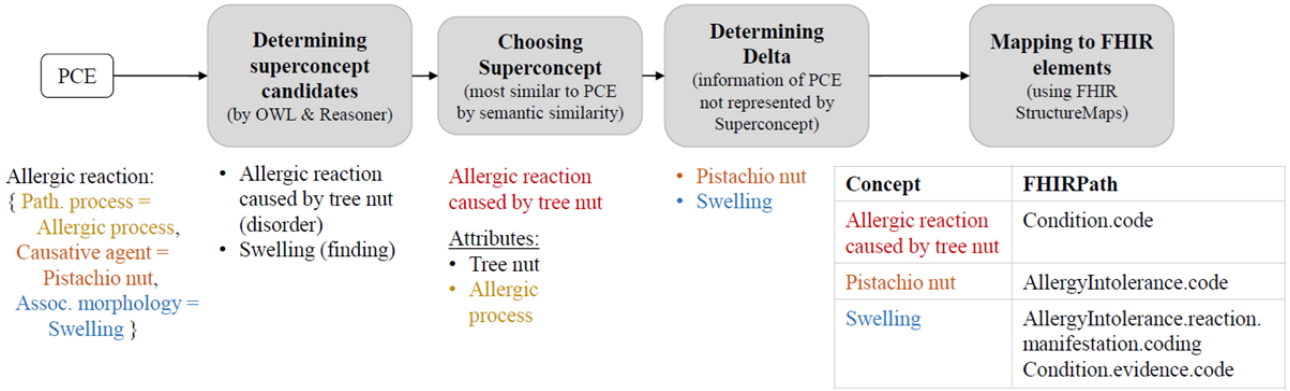
Additionally, some papers discuss the use of SNOMED CT in combination with standardized information models based on the HL7 Reference Information Model (RIM), HL7 Clinical Document Architecture, and HL7 FHIR resources. For instance, a project by Perez-Rey et al [25] focused on linking the normal form of precoordinated SNOMED CT concept definitions, normalizing SNOMED CT concepts, and binding them to HL7 RIM classes. This approach could potentially be extended to PCEs. However, the HL7 version 3 standard has not been widely adopted due to its complexity [1]. A project by Arguello-Casteleiro et al [26] addressed mapping precoordinated SNOMED CT concepts or PCEs from Consolidated Clinical Document Architecture to FHIR resources. While the objectives of this approach are similar to those of our work, Arguello-Casteleiro et al focused heavily on ontology. By contrast, our work primarily utilizes the widely adopted FHIR and SNOMED CT native specifications. Additionally, the specific ocular diseases examined by Arguello-Casteleiro et al pertain to a very narrow subset of SNOMED CT expressions. Further, the publication by Arguello-Casteleiro et al does not provide information on the extent to which all the data included in the PCEs can be transferred into the FHIR representation.

Methods

Overview

This work aims to develop and implement an approach that enables the storage of a SNOMED CT PCE within FHIR resources using only precoordinated codes. In this alternative representation, the PCE will be decomposed into precoordinated concepts, which can then be stored in appropriate elements of corresponding FHIR resources. An overview of the proposed approach is shown in Figure 2.

Figure 2. The decomposition of a PCE into elements of (profiled) FHIR resources consists of 4 steps. FHIR: Fast Healthcare Interoperability Resources; OWL: Web Ontology Language; PCE: postcoordinated expression.



A PCE, once verified for syntactic and semantic correctness, serves as the input. This PCE can be classified within SNOMED CT using the Web Ontology Language (OWL) and a reasoner, allowing for the identification of its direct supertype ancestors. Among these concepts, the most similar one to the PCE is selected as the Superconcept. The Delta is then calculated

between the Superconcept and the PCE, encompassing all the information in the PCE that is not represented by the Superconcept. In the final step, suitable elements of corresponding FHIR resources must be identified to store the information of the Superconcept and the Delta. To facilitate this, FHIR StructureMaps that define these associations on a general level need to be created in advance.

Validation of PCE

To ensure that flexible PCEs can be accurately interpreted and evaluated, they must adhere to the syntactic requirements of the Compositional Grammar and the semantic rules of the Concept Model defined by SNOMED International. In the initial step, the input PCE is checked for syntactic and semantic correctness using the HL7 FHIR service *$validate-code* and Ontoserver. Ontoserver, a FHIR-based terminology server provided by the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO), supports and facilitates working with key coding systems such as SNOMED CT and LOINC (Logical Observation Identifiers Names and Codes) [11]. Additionally, Ontoserver provides an integrated description logic reasoner and full support for SNOMED CT postcoordination, enabling PCE validation directly on the Ontoserver [11]. Only if a PCE is both syntactically and semantically correct will the process automatically proceed to the next steps.

Determining Superconcept Candidates

OWL serves as an exchange format for ontologies and is used here to determine the ancestors of a PCE. OWL allows for the representation of complex knowledge about concepts and their interrelationships [27,28]. SNOMED CT can also be represented as an OWL ontology, with the definitions of individual SNOMED CT concepts provided as OWL expressions in the monthly release packages of SNOMED CT, alongside other formats [29].

To obtain SNOMED CT as an OWL ontology, the SNOMED OWL Toolkit [30] was used in the Release Format 2 of the International Edition, dated 2023-04-30 [31]. The generated ontology includes, among other components, precoordinated SNOMED CT concepts and their definitions in functional syntax.

PCEs, by contrast, are based on the syntax of the Compositional Grammar. Therefore, a PCE must be transformed into an OWL ClassExpression for further processing. For each component of the PCE, the corresponding OWL counterpart is determined based on the existing concept definitions, as shown in Table 1. An algorithm utilizing the Java (Oracle Corporation) library OWL API [32] is developed to automate this transformation. The result of the algorithm is an OWL ClassExpression structured similarly to the OWL ClassExpressions created by SNOMED International for SNOMED CT concept definitions. An exemplary OWL expression for a PCE is shown in Figure 3, with Fully Specified Names added for improved readability. The symbol ":" serves as a placeholder for the defined namespace, which in this case is "http://snomed.info/id/" [29].

Table 1. Various OWL[a] constructs are used for the representation of the components of a PCE[b]. While the components in the first 2 rows are exclusively used in the native OWL ontology of SNOMED CT[c], the OWL constructs below are also used for the transformation of PCEs.

PCE component	OWL construct
SNOMED CT concept	OWL Class
SNOMED CT attribute, attribute value: SNOMED CT concept	OWL ObjectProperty
Linking of ungrouped attribute relationships	OWL ObjectIntersectionOf
Linking between individual attribute relationships in a Role Group	OWL ObjectIntersectionOf
Linking between focus concept and all attribute relationships of ungrouped attributes and Role Groups	OWL ObjectIntersectionOf
Attribute relationship of attribute and SNOMED CT Identifier as attribute value	OWL ObjectSomeValuesFrom
All grouped attribute relationships and Role Group Identifier	OWL ObjectSomeValuesFrom

[a]OWL: Web Ontology Language.

[b]PCE: postcoordinated expression.

[c]SNOMED CT: SNOMED Clinical Terms.

Figure 3. PCE on the left side and the associated OWL ClassExpression based on functional syntax on the right side. OWL: Web Ontology Language; PCE: postcoordinated expression.

PCE

```
419076005 |Allergic reaction| :
{ 370135005 |Pathological process| =
    472964009 |Allergic process| ,
  246075003 |Causative agent| =
    227512001 |Pistachio nut| ,
  116676008 |Associated morphology| =
    442672001 |Swelling|
}
```

Associated OWL Class Expression

```
ObjectIntersectionOf(
 : 419076005 |Allergic reaction|
 ObjectSomeValuesFrom(
  : 609096000 |Role group|
  ObjectIntersectionOf(
   ObjectSomeValuesFrom(
    : 370135005 |Pathological process|
    : 472964009 |Allergic process|
   )
   ObjectSomeValuesFrom(
    : 246075003 |Causative agent|
    : 44257100012418 |Tree nut|
   )
   ObjectSomeValuesFrom(
    : 116676008 |Associated morphology|
    : 442672001 |Swelling|
   )
  )
 )
)
```

Output (direct superconcepts):

- 65124004 |Swelling (finding)|
- 15920521000119105 |Allergic reaction caused by tree nut (disorder)|

In the next step, the previously formed OWL ontology of SNOMED CT, the created PCE-specific OWL ClassExpression, and a description logic reasoner are used to classify the PCE within the existing SNOMED CT hierarchy. A reasoner generates new knowledge through logical inferences from the existing content of an ontology, such as determining superrelationships between concepts [27]. In this work, the ELK reasoner is used because it is one of the few reasoners capable of handling the extensive SNOMED CT ontology [33].

The direct ancestors of PCEs are determined using the ELK reasoner with the Java library elk-reasoner [33]. This process identifies at least one OWL Class corresponding to a precoordinated SNOMED CT concept. All identified precoordinated SNOMED CT concepts are considered potential Superconcept candidates.

Choosing Superconcept

Overview

One of the previously identified Superconcept candidates must be selected as the Superconcept. The Superconcept is the precoordinated SNOMED CT concept that most closely resembles the PCE, covering the largest portion of the information contained within the PCE compared with any other precoordinated SNOMED CT concept.

Semantic Similarity Measure

To identify the most similar concept, the semantic similarity between concepts is assessed. This measure calculates the taxonomic proximity between two elements within a knowledge base, such as SNOMED CT. Higher semantic similarity indicates that the two elements are more closely related [34].

While several semantic similarity measures have been proposed in the literature, this work uses a path-based approach developed by Sánchez and Batet [34]. This method is specifically designed for large knowledge bases with a subtype-relationship–based polyhierarchy, such as SNOMED CT [34]. The equation for the calculation is as follows:

where $T(c_i)$ is the set of all ancestors of a concept c_i and the concept c_i itself. To calculate the semantic similarity between two concepts c_1 and c_2, the ratio between the set of nonshared ancestors (nominator) and the union of all ancestors of both concepts (denominator) is considered [34].

In our work, the semantic similarity between the classified PCE and each of its Superconcept candidates needs to be calculated. Therefore, the equation is modified as follows:

with c_i being one of the Superconcepts. The calculated similarities for the exemplary PCE are shown in Table 2.

Table 2. To choose the most fitting Superconcept, the semantic similarity between the PCE[a] and each of its Superconcept candidates is calculated using the measure proposed by Sánchez and Batet [34]. The previously introduced exemplary PCE shares a larger ratio of ancestors with *Allergic reaction caused by tree nut* than with *Swelling*. Thus, the former leads to a higher semantic similarity and is determined as Superconcept.

Superconcept candidates	Nonshared ancestors	Union of all ancestors	Semantic similarity
Allergic reaction caused by tree nut	8	30	2.10
Swelling	29	30	0.10

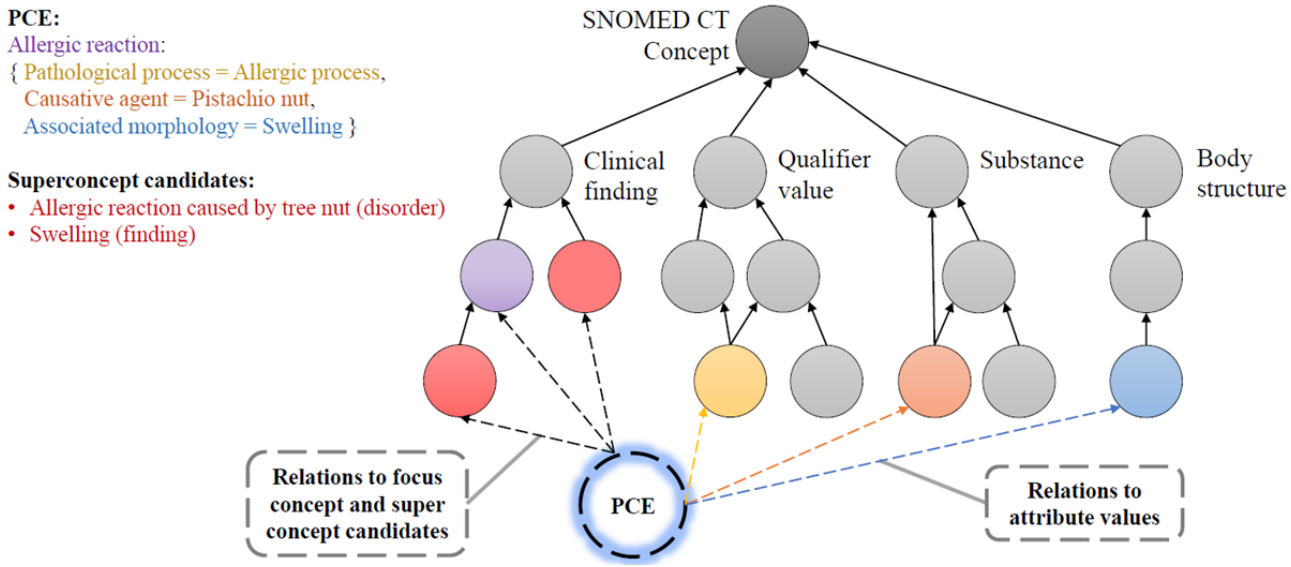
[a]PCE: postcoordinated expression.

Implementation

An algorithm was developed to select the most suitable (ie, the most semantically similar) SNOMED CT concept from the Superconcept candidates. The calculation of semantic similarity is based on the graph-based approach by Sánchez and Batet [34], as described above. To achieve this, SNOMED CT must first be transformed into a directed acyclic graph (DAG). The DAG was constructed by algorithmically processing the Release Format 2, version 2023-04-30 [31], of SNOMED CT using the Python (Python Foundation) library NetworkX [35]. In the resulting graph, SNOMED CT concepts are represented as nodes, while their relations, including relation types, form the connecting edges.

To enable semantic similarity calculation, the PCE must be temporarily inserted into the DAG. For this purpose, a node called “pce” is introduced into the graph as a subnode of its focus concept and the previously determined Superconcept candidates (Figure 4). The algorithm iterates over the individual attribute values of the PCE and inserts edges between these attributes and the “pce” node.

Figure 4. The PCE is represented as a subconcept of its focus concept and the Superconcept candidates. It also has edges to its attribute values. PCE: postcoordinated expression; SNOMED CT: SNOMED Clinical Terms.

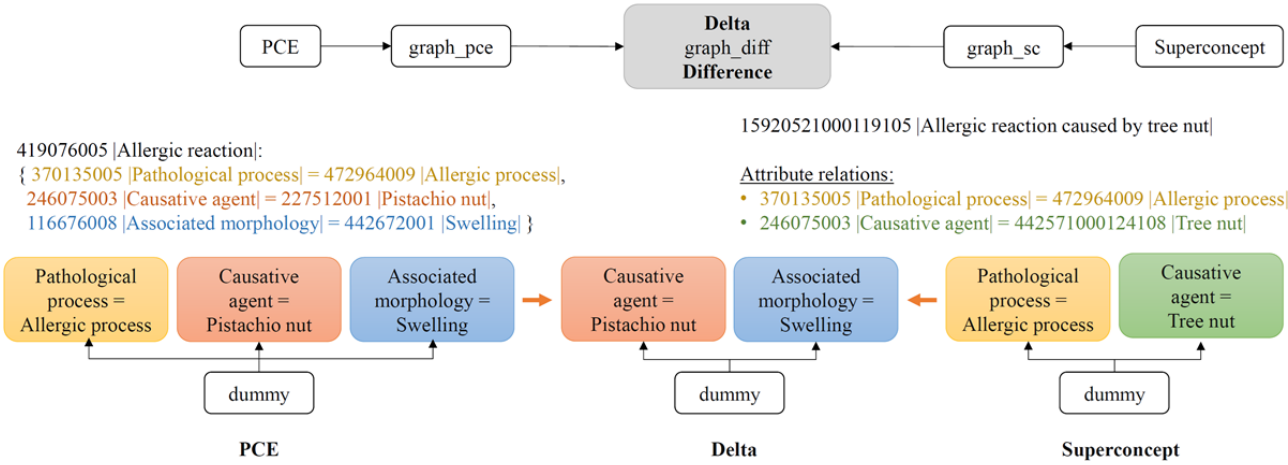


Next, the semantic similarity between the PCE and all potential Superconcept candidates is calculated based on the DAG. The Superconcept candidate with the highest semantic similarity describes the PCE most accurately among all precoordinated SNOMED CT concepts. This concept is then defined as the Superconcept.

Determining Delta

Although the Superconcept is the most similar precoordinated concept available in SNOMED CT, it does not cover all the information contained within the PCE. Therefore, the missing parts of information, referred to as the “Delta” between the PCE and the Superconcept, need to be determined. This is achieved using a graph-based approach, where both the attribute relations of the PCE and those of the Superconcept are represented in separate graphs. To facilitate merging later, a dummy root node with the same name is introduced into both graphs. Next, the attribute relations of the PCE and the Superconcept are added as nodes and connected via edges, as shown on either side of Figure 5. The Delta is then calculated by subtracting the edges of the Superconcept graph from the PCE graph. This process eliminates all equivalent components present in both graphs, leaving only the attribute relationships that are not or not precisely represented in the Superconcept graph (see the Delta graph in the middle of Figure 5).

Figure 5. When calculating the delta, the graph of the Superconcept is subtracted from that of the PCE. The determined edges contain attribute relations that were not covered by the Superconcept but are necessary for the precise representation of the PCE. PCE: postcoordinated expression.



Mapping to FHIR Elements

In the final step, the Superconcept and the Delta must be stored as precoordinated concepts in suitable FHIR elements. This can be accomplished using either the base FHIR resources from HL7 or more specific FHIR profiles. FHIR profiles are customized versions of FHIR resources tailored to national characteristics, legislation, or specific use cases [35].

In our work, 2 sets of FHIR profiles were considered as target representations for the mapping:

- profiles of the NASHIP (version 1.4.0, based on FHIR R4) [4] and
- profiles of the core data set of the German MII (version 1.0, based on FHIR R4) [5].

To accurately map a decomposed PCE to these profiled FHIR resources, mapping rules are required to align its Superconcept and Delta with the appropriate FHIR elements. As PCEs are often highly individualized, they can be categorized according to their focus concept, attribute relationships, and SNOMED CT Expression Templates. Consequently, it is not necessary to define a mapping rule for each PCE; instead, PCEs with similar content can be mapped using a uniform set of rules.

To demonstrate the applicability of our approach, we consider 5 content categories of PCEs from the top-level hierarchies *Procedure* and *Clinical finding*. These categories include *General procedures* for *Procedure* and *Allergies, Diseases due to allergy, Allergic reactions*, and *General clinical findings* for *Clinical finding*. To ensure the unambiguous assignment of PCEs, SNOMED CT's Expression Constraint Language (ECL) is used to formally define the scope of each category, creating pairwise disjoint partitions of the two hierarchies.

For each combination of 1 of the 5 content categories and either of the 2 sets of target representations, mapping rules are manually defined based on the respective Expression Templates [36], concept definitions of related precoordinated concepts, existing documents (as described in the "Related Work" section), and the author's expertise. These rules are formally transcribed into FHIR StructureMaps (R4) [37] using the Java library HAPI FHIR [38]. The resulting 10 FHIR StructureMaps are stored on a local HAPI FHIR server [39].

An excerpt of the FHIR StructureMap for the category "Allergic reaction" and the NASHIP profiles is shown in Figure 6. Each StructureMap contains exactly a single "rule" entry for the Superconcept and each relevant attribute relationship for the category. The mapping rule's "source" is either the string "Superconcept" or the SNOMED CT identifier of the respective attribute. For the "target" element, FHIRPath is used to traverse the FHIR resources and explicitly identify the correct subelement [2]. Illustrated in orange in Figure 6, the mapping of the attribute "Causative agent" ("246075003") to the FHIRPath *AllergyIntolerance.code* is exemplified. This specifies that the attribute value associated with the PCE should be stored in the code element of the *AllergyIntolerance* resource.

Figure 6. Extract of the FHIR StructureMap for allergic reactions in JSON format. It contains an entry for each SNOMED CT attribute and the Superconcept as well as references and names of the profiles. FHIR: Fast Healthcare Interoperability Resources; SNOMED CT: SNOMED Clinical Terms.


```
{
 "resourceType": "StructureMap",
 "id": "KBV-Mio-AllergicDisease",
  …
 "group": [ {
   …
    "rule": [… ,{
          "name": "CausativeAgent",
          "source": [{
              "context": "source",
              "element": "246075003"
           }],
           "target": [ {
              "context": "target",
              "contextType": "variable",
              "element": "AllergyIntolerance.code",
              …
            }]
        }, …{
          "name": "References",
          "source": [ {
              "context": "source",
              "element": "references"
            }],
          "target": [{
              "context": "target",
              "contextType": "variable",
              "element": "Condition.evidence.detail->AllergyIntolerance",
            }]
        },{
          "name": "NamesProfiles",
          "source": [{
              "context": "source",
              "element": "namesprofiles"
            }],
          "target": [{
              "context": "target",
              "contextType": "variable",
              "element":  "Condition->KBV_PR_Base_Condition_Diagnosis",
            }]
}]}]}
```


Mapping rule for each attribute or superconcept

References between FHIR resources

Names of FHIR profiles for respective resources

Apart from the central mapping rules, some further information is included in the FHIR StructureMaps.

If necessary, references between FHIR elements and resources are indicated (Figure 6, in pink). Additionally, the FHIR resources associated with the FHIRPath entries are mapped to the names of the respective profiles (Figure 6, in green).

With these general preparations completed, individual decomposed PCEs can now be mapped to appropriate FHIR resources. First, the PCE is assigned to 1 of the 5 content categories by identifying the subsuming subset using the ECL definitions and Ontoserver. Based on the category and desired target representation, the correct FHIR StructureMap is selected. The mapping rules in the StructureMap are then automatically

applied to the PCE's Superconcept and Delta, determining the combinations of FHIR elements and precoordinated SNOMED CT concepts needed for the alternative representation.

Ethics Approval

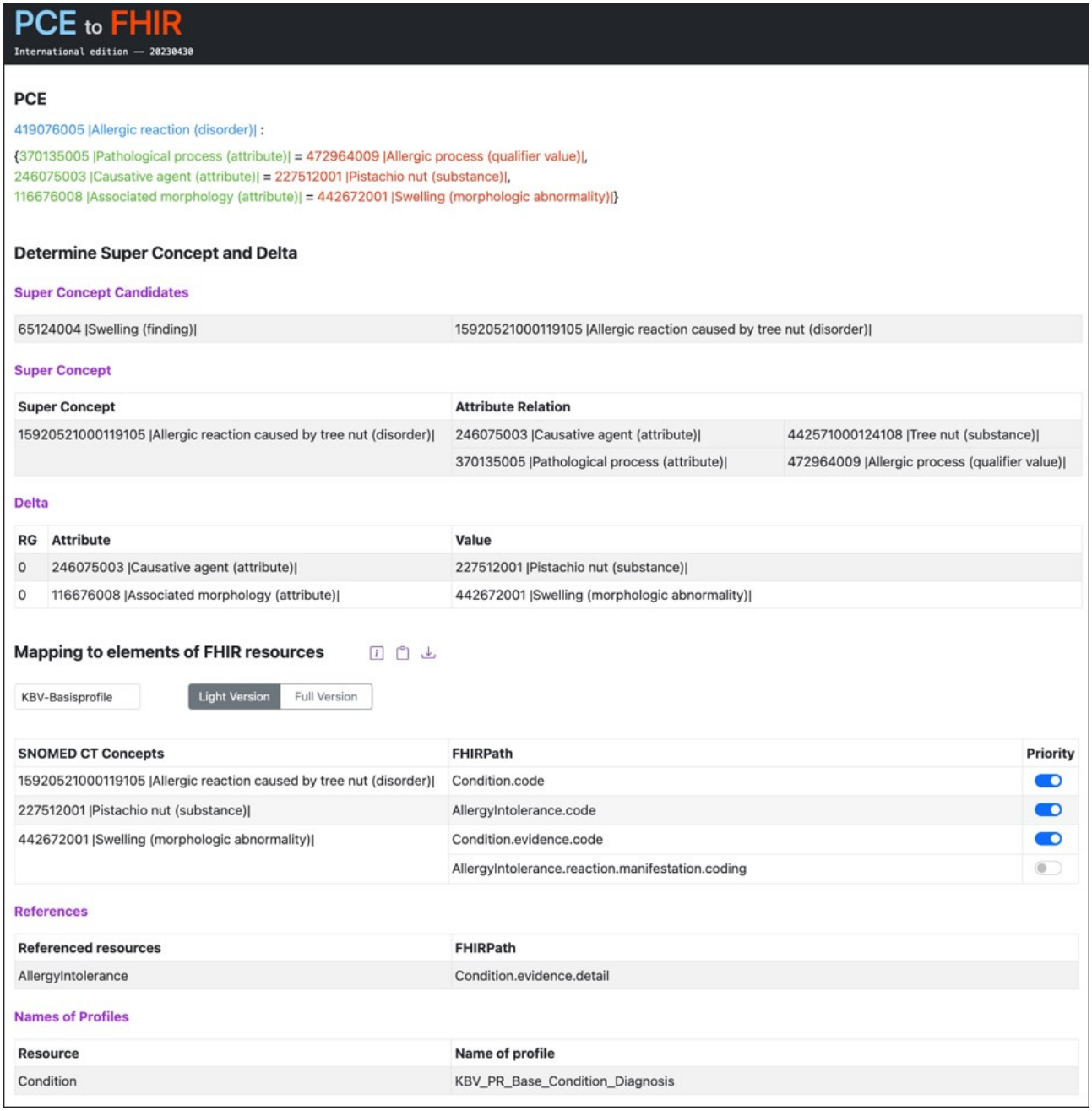
This research neither involves human nor animal subjects so ethics approval was not required.

Results

Web Application

Bringing all the previously explained preliminary considerations and processing steps together, a web application called "PCEtoFHIR" was developed (Figure 7). This single-page web application was built using Angular (version 15.2.2; Google LLC/Alphabet Inc.) and the Java framework Spring Boot (version 2.7.2; Java version 17).

Figure 7. Excerpt of the web application PCEtoFHIR. FHIR: Fast Healthcare Interoperability Resources; PCE: postcoordinated expression.



An excerpt of the web application is shown in Figure 7. After the entered PCE has been checked for syntactic and semantic correctness, the Superconcept candidates, the Superconcept, and the Delta are determined automatically. This information is displayed to the user in the "Determine Super Concept and Delta" section. The Superconcept and Delta are then mapped to FHIR elements according to the category-specific FHIR StructureMaps. The appropriate StructureMap is automatically determined for the PCE based on its content category and the desired mapping target is selected in a combo box. Figure 6 illustrates the target mapping category "KBV-Basisprofile" (KBV Base Profiles; Kassenärztliche Bundesvereinigung [KBV]

is the German abbreviation for the NASHIP). Using the StructureMap's mapping rules, the FHIRPaths, corresponding SNOMED CT concepts, required references, and profile names are displayed and can be copied to the clipboard or downloaded as a text file. This information is shown in Figure 7 in the section "Mapping to elements of FHIR resources."

The source code of PCEtoFHIR, the created FHIR StructureMaps, and all validation results as described in the following paragraphs are available on GitHub [40].

Determining Superconcept Candidates

To ensure the correct functionality of the algorithm for generating OWL ClassExpressions, validation was performed with 600 randomly selected precoordinated SNOMED CT concept definitions. These definitions, from the monthly release 2023-04-30 [31], were initially examined for diverse structures and then transformed into OWL ClassExpressions using PCEtoFHIR's regular algorithm. The generated OWL ClassExpressions was compared with the official OWL ClassExpressions of the concepts, available as part of the same SNOMED CT release, as reference data.

The comparison involved checking whether the OWL ClassExpressions matched syntactically and semantically. Focus concepts were excluded from the semantic validation because the reference data's OWL ClassExpressions are based on Stated Concept Definitions, while our approach uses Inferred Concept Definitions. Thus, the validation focused solely on whether the focus concepts were syntactically in the correct position within the OWL ClassExpression.

In summary, the comparison revealed no discrepancies between the OWL ClassExpressions generated by PCEtoFHIR and those from the reference data.

Choosing Superconcept

As described previously, a path-based measure by Sánchez and Batet [34] is applied to calculate semantic similarity in our work. A preliminary analysis was conducted to ensure both the theoretical and practical applicability of this measure.

SNOMED CT's most striking characteristics include its reliance on subtype relationships and the resulting, heavily interwoven polyhierarchy. Using a path-based approach that incorporates each concept's ancestors, these central features are prioritized. While several path-based semantic similarity measures are available [34], they mostly rely on the shortest path between concepts [41-44]. Sánchez and Batet [34] argue that in large knowledge bases such as SNOMED CT, a concept inherits information from several hierarchies and is connected to many concepts simultaneously. Therefore, considering only the shortest path is insufficient.

In their proposed measure, Sánchez and Batet [34] consider the ratio between the set of nonshared ancestors and the set of all ancestors of both concepts. Adapted to our approach, this means that for the PCE and each of its Superconcept candidates, all ancestors in the SNOMED CT hierarchy are determined, and the ratio is calculated. Thus, this measure is applicable in principle.

To evaluate if the calculated values are reasonable beyond that, a practical validation by means of an exemplary sample was done in succession. For 33 PCEs taken from a publication by Kate [45] (see detailed explanation in the "Overall Evaluation With Existing PCEs" section), the most similar Superconcept candidate was determined manually and compared with the Superconcept calculated by PCEtoFHIR via Sánchez and Batet's semantic similarity measure. In 76% (25/33) of the cases, the same Superconcept was chosen [34]. An analysis of the remaining 24% (8/33) revealed that different information components within the PCE were prioritized during Superconcept selection (eg, favoring localization over procedure type), but the divergent choices made by the algorithm were considered plausible. Consequently, the measure by Sánchez and Batet [34] was found to be feasible for Superconcept determination and is therefore used to calculate semantic similarity in this work.

Mapping to FHIR Elements

As previously described, the Basisprofile of the NASHIP (version 1.4.0) [4] and the profiles of the Core Data Set of the MII (version 1.0) [5] were used for mapping to the FHIR profiles. To illustrate the mapping, 5 categories were considered. For each category and profile type, a FHIR StructureMap was created, resulting in a total of 10 FHIR StructureMaps. Table 3 shows the FHIR resources used for the mapping.

Table 3. Utilization of different FHIR[a] resources per profile type, organized by content category.

Category and profile type	FHIR resources
Allergies	
NASHIP[b]	AllergyIntolerance
MII[c]	Condition, Observation
Disease due to allergies	
NASHIP	AllergyIntolerance, Condition
MII	Condition, Observation
Allergic reaction	
NASHIP	AllergyIntolerance, Condition
MII	Condition, Observation
Clinical finding (general)	
NASHIP	AllergyIntolerance, Condition
MII	Condition, Observation
Procedure	
NASHIP	Procedure
MII	Procedure

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]NASHIP: National Association of Statutory Health Insurance Physicians.

[c]MII: Medical Informatics Initiative.

For the category of general *Clinical finding*, the FHIR StructureMaps include mappings for the Superconcept and the following 5 SNOMED CT attributes: *Causative agent*, *Finding site*, *Associated morphology*, *Pathological process*, and *Clinical course*. In addition to the 5 SNOMED CT attributes, the following attributes were considered for the remaining *Clinical finding* categories focusing on allergies: *Has realization*, *Occurrence*, and *Due to*. Furthermore, for the hierarchy *Procedure*, mapping rules for the Superconcept and the following 12 additional SNOMED CT attributes are established: *Method*, *Procedure site—Direct*, *Procedure site—Indirect*, *Direct substance*, *Direct morphology*, *Using substance*, *Using device*, *Using access device*, *Has intent*, *Access*, *Surgical approach*, and *Has focus*.

Table 4 shows an example of the mapping of SNOMED CT elements to FHIRPath for Allergic reactions. The complete set of mapping rules and the associated StructureMaps are available online [46].

Table 4. The SNOMED CT elements and the associated FHIRPath for the category "Allergic reaction" based on the profiles of the NASHIP[a].

SNOMED CT element	FHIRPath in NASHIP profiles
Super concept	• Condition.code
Causative agent	• AllergyIntolerance.code
Finding site	• Condition.bodySite
Associated morphology	• AllergyIntolerance.reaction.manifestation.coding:snomed • Condition.evidence.code
Pathological process	• AllergyIntolerance.reaction.manifestation.coding:snomed • Condition.evidence.code
Has realization	• AllergyIntolerance.reaction.manifestation.coding:snomed • Condition.evidence.code
Occurrence	• AllergyIntolerance.extension: abatement-lebensphase-von [47]
Clinical course	• Extension of HL7: Condition.condition-diseaseCourse [48]
Due to	• Extension of HL7: Condition.condition-dueTo [49]

[a]NASHIP: National Association of Statutory Health Insurance Physicians.

To ensure their correctness, the created FHIR StructureMaps were validated by author AE, who had not been involved in the PCEtoFHIR project up to that point. AE possesses in-depth knowledge of FHIR and SNOMED CT and is profoundly familiar with the MII and NASHIP profiles due to her former and current work. She validated the FHIR StructureMaps based on the definitions of the profiles used, ensuring the correct choice of profiles, the accurate mapping of SNOMED CT elements to FHIRPaths, and appropriate references. The validation yielded the following results: the choice of profiles and references was found to be entirely correct. The mapping rules between SNOMED CT elements and FHIR paths were largely correct as well; however, 8 suggestions for improvement were provided. These suggestions were reviewed by the other authors, and the FHIR StructureMaps were updated accordingly based on their agreement.

Lastly, the finalized FHIR StructureMaps were analyzed for their coverage of the SNOMED CT attributes listed above. Table 5 illustrates the number of attributes per category that could be successfully mapped in the respective profile (mappable). Depending on the category, up to 4 attributes could not be mapped to the native profiles (unmappable). For some of these attributes, existing FHIR extensions, such as those from HL7, can be introduced into the profiles to achieve a more complete mapping (with extension). Overall, between 56% (5/9) and 92% (12/13) of attributes (76.1% on average) can be mapped without modifications, depending on the category and profile. By introducing these extensions, coverage could be increased to an average of 93.5% (ranging from 67% [6/9] to 100% [eg, 9/9, 6/6, 13/13, for individual combinations]).

Table 5. The number of SNOMED CT elements that can be mapped directly to the profiles are represented by “mappable,” whereas elements that cannot be mapped are shown as "ummappable". FHIR[a] offers extensions to map items that are not mappable by default, which could reduce the number of unmappable elements. The number of unmappable elements to represent by extension is shown in the last column.

Category: number of elements (total)	Profile type	Number of mappable elements (total)	Number of unmappable elements	
			Total	Could be mapped using extensions
Allergies: 9 elements				
	NASHIP[b]	5	4	1
	MII[c]	5	3	2
Disease due to allergies: 9 elements				
	NASHIP	7	2	2
	MII	7	2	2
Allergic reaction: 9 elements				
	NASHIP	7	2	2
	MII	7	2	1
Clinical finding (general): 6 elements				
	NASHIP	4	2	2
	MII	4	2	2
Procedure: 13 elements				
	NASHIP	12	1	1
	MII	12	1	1

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]NASHIP: National Association of Statutory Health Insurance Physicians.

[c]MII: Medical Informatics Initiative.

Overall Evaluation With Existing PCEs

After validating several steps individually, the entire process of PCEtoFHIR was evaluated. To achieve a realistic scenario, 35 existing PCEs from the publication “Automatic Full Conversion of Clinical Terms into SNOMED CT Concepts” by Kate et al [45] were used, which are available online [50]. This publication presents a method for converting clinical texts into SNOMED CT PCEs, with the 35 PCEs manually created from clinical terms for a small-scale evaluation (see the “4.3. Evaluation Methodology” section in [45]).

To use this data set as input for PCEtoFHIR, the 35 provided PCEs were manually reviewed and automatically checked for syntactic and semantic correctness, as described previously. As a result, 2 PCEs were excluded from further processing: 1 for violating cardinality restrictions of the Concept Model and 1 for being equivalent to another. The remaining 33 PCEs include 23 from the top-level hierarchy *Clinical finding* and 10 from the top-level hierarchy *Procedure.* The following SNOMED CT attributes are used:

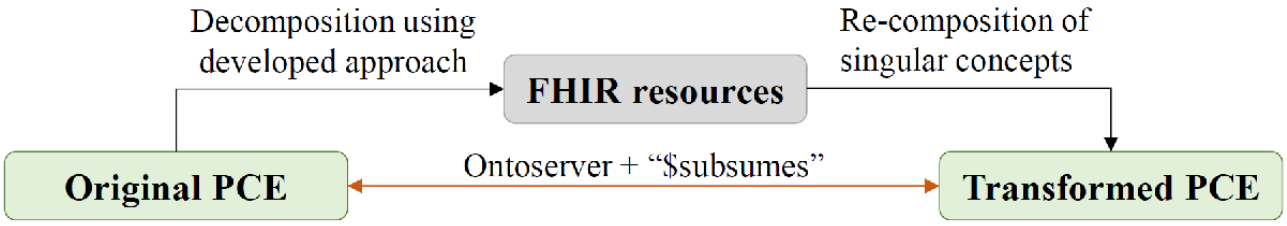
- *Clinical finding: Associated morphology, Finding site, Causative agent, Clinical course, Finding method,* and *Pathological process*
- *Procedure: Method, Procedure site—Direct, Using device,* and *Using substance*

These 33 PCEs were imported into PCEtoFHIR in bulk, bypassing the web front end but using the developed algorithm as usual. Each PCE was decomposed into a Superconcept and Delta, with FHIRPaths determined using the appropriate FHIR StructureMap. Based on these FHIRPaths, the corresponding attribute values were populated into FHIR resources, which had been prefilled with additional required data elements (such as references to a FHIR resource “Patient”). The resulting FHIR resources for each PCE were then stored on the local HAPI FHIR server [39].

To determine whether all information contained within the original PCE has been successfully translated into the FHIR resources, the process shall now be reversed as shown in Figure 8. Based on the separate precoordinated SNOMED CT concepts spread across multiple FHIR elements, a PCE will be recomposed and compared with the original. For this, the values of the stored FHIR resources were extracted, and the respective StructureMap was applied in the reverse direction. The recomposed PCE uses the Superconcept as the focus concept, while its refinement consists of attribute relationships with the extracted SNOMED CT concepts as attribute values. For each of the extracted concepts, the corresponding SNOMED CT attribute must be determined according to the StructureMap’s mapping rules. As most mappings are inherently unidirectional [3], using them in the reverse direction can lead to several possible SNOMED CT attributes for some FHIRPaths. In these

cases, the 3 rules listed in Textbox 2 are applied successively until a clear distinction is achieved.

Figure 8. Validation process: The original PCE is decomposed by PCEtoFHIR as usual, and accordingly stored in FHIR resources. Based on this representation, the singular concepts are recomposed into a second PCE. The original and the recomposed PCE are compared. FHIR: Fast Healthcare Interoperability Resources; PCE: postcoordinated expression.



Textbox 2. Rules applied for case distinction.

- **By attribute value range**

For each of the attributes in question, a value range is defined by the SNOMED Clinical Terms Concept Model. This allows checking if the extracted concept falls within the respective value range and is thus a valid value for the corresponding attribute. If only a single value range matches, the correct attribute is identified.

- **By attribute hierarchy**

Some SNOMED CT attributes are organized hierarchically (eg, *Procedure device—Using device—Using access devices*). If multiple attributes from the same hierarchy are available, the most general attribute within this hierarchy is selected (in this case: *Procedure device*).

- **By occurrence heuristic**

The concept definitions of precoordinated concepts are analyzed to identify which SNOMED CT attribute is most frequently used for the given concept. The attribute with the highest statistical occurrence is selected.

Finally, each recomposed PCE was compared with the corresponding original PCE as a reference by testing their subsumption relationship via Ontoserver and the FHIR service $subsumes. For 32 of the 33 comparisons, the PCEs were evaluated as equivalent, indicating that no information was lost during the process. In the remaining comparison, the recomposed PCE was found to be a subclass of the original. Further analysis revealed that while some semantic precision was lost due to the required case distinctions, the FHIR representation achieved through PCEtoFHIR remained semantically equivalent to the original PCE. The approach thus proved successful, demonstrating the preservation of a PCE's content.

Discussion

Principal Findings

This work aimed to develop an algorithm for generating alternative representations of SNOMED CT PCEs. The "Introduction" section discussed various reasons for the reluctance to use PCEs, including the preference for editing and saving individual codes. Precoordinated concepts offer the advantage of a single code representing a medical circumstance, along with a human-readable description. These aspects are seen as advantages when using precoordinated concepts but are not present in postcoordination. However, by identifying a Superconcept using OWL and semantic similarity, a precoordinated SNOMED CT concept that most accurately describes the PCE can be determined. The PCE, as a subconcept of this Superconcept, can then be effectively displayed and analyzed. In addition, the decomposition of the PCE results in individual codes that users are familiar with from practical experience. FHIR, recognized as a central interoperable data standard in health care, is gaining increasing importance both in Germany [51] and globally [51-53].

The calculation of semantic similarity is crucial for identifying the concept most similar to a PCE. In this work, the measure by Sánchez and Batet [34] was selected due to its suitability for SNOMED CT's polyhierarchies and multiple inheritances. While this measure has been validated as effective, other approaches to semantic similarity calculation could also be considered. Although choosing an appropriate Superconcept is important, any remaining information is represented through the Delta, so alternative semantic similarity measures would have minimal impact on the overall functionality of PCEtoFHIR. However, PCEtoFHIR could be enhanced by allowing users to manually select or adjust the Superconcept.

As described above, the Superconcept and the Delta are stored in the relevant FHIR profiles, within their corresponding elements. For this purpose, the profiles of the NASHIP [4] and the profiles of the core data set of the MII [5] were utilized. However, the approach can also be adapted to other profiles. As shown in Table 5, depending on the content category and profile type, between 56% (5/9) and 92% (12/13) of attributes could be mapped directly to the respective profile without modifications, demonstrating a high coverage of content but also highlighting some gaps.

An analysis of the unmappable attributes revealed that these relations mostly concern very specific details that occur infrequently in precoordinated concept definitions. As PCEs are generally constructed based on existing concept definitions,

they are unlikely to regularly utilize these highly specific attributes. SNOMED CT, however, provides the capability to include medical facts at such a granular level. By contrast, HL7 FHIR was designed with a pragmatic approach, focusing on the most prevalent information. This fundamental design discrepancy accounts for the majority of mapping challenges.

Nevertheless, as intended by the FHIR standard, suitable extensions from HL7 are available to appropriately represent some of these attribute relations (see Tables 4 and 5) and could be integrated into the profiles to extend their coverage of PCE content. Currently, the core data set lacks a profile for the FHIR resource AllergyIntolerance, which limits the representability of this content category. However, because the core data set is an ongoing modeling initiative, such a profile may be added in the future.

Apart from the unmappable attributes, the proposed mapping rules cannot always ensure an exact translation. In some cases, there is no precisely matching FHIR element for a specific attribute, or multiple attributes must be mapped to the same element (eg, *Using access device* and *Using device* can both only be recorded via the FHIR element *Device.type*, which is then referenced via *Procedure.usedReference*). As a result, some semantic precision may be lost. Therefore, it is advisable to store the original PCE in the metadata of the FHIR representation to ensure the preservation of the original meaning.

Nevertheless, the outlined difficulties in achieving a semantically equivalent representation through FHIR elements highlight the precision attainable through SNOMED CT's postcoordination, underscoring its importance for detailed medical data description.

Several validations were conducted to ensure the correctness of both the individual processing steps and the overall functionality of PCEtoFHIR. Although a large data set from SNOMED International's releases could be used to firmly validate the OWL expression generation, other evaluations required preexisting real-world PCEs of specific content categories, which are not readily available. The employed set of 35 PCEs meets these criteria and effectively facilitates our validation approaches. However, the limited number and semantic variance of these exemplary PCEs suggest that incorporating additional reference data could enhance the significance of the results.

Another validation included the manual review of the FHIR StructureMaps by a FHIR expert. This review revealed only minor inaccuracies, which were corrected in the current version of StructureMaps. Using these StructureMaps, the mapping for the considered categories was successfully completed and can be extended to other subhierarchies of SNOMED CT without difficulty. Hence, further possibilities of application may be considered, such as addressing the TermInfo problem. As mentioned, the choice between representing medical facts in a terminology or an information model often varies and depends on the intended use. For example, "Fracture of the left femur" can either be represented using a single PCE as the FHIR element *Condition.code* like

- *Condition.code: 71620000:{363698007=722738000}*
  (Fracture of femur : {Finding site = Structure of bone of left femur})

or by splitting the semantic meaning up into two precoordinated SNOMED CT concepts using further element-code-combinations of the FHIR resource *Condition*:

- *Condition.code: 71620000 |Fracture of femur (disorder)|*
- *Condition.bodySite: 722738000 |Bone structure of left femur (body structure)|*.

This variability in expressing medical facts was leveraged in the presented approach, enabling flexible transformation between terminology-focused and information model–focused representations. This allows for an alternative when replacing PCEs and may help address some challenges related to the TermInfo problem. By enabling flexible switching between different expression methods, semantic interoperability is maintained regardless of the representation paradigm used in an electronic health record. Additionally, this approach facilitates the plausibility check of recorded medical information by allowing the integration of disjointed elements (eg, scattered across various FHIR elements) into a single *interpretable* expression.

Despite the general ambiguity regarding the scope of semantic versus structural standards, some specific recommendations do exist, such as the suggestion that "contextual meaning should rather be represented via the information model" [54]. Contrarily, concepts within the SNOMED CT hierarchy *Situation with explicit context* encompass contextual information that extends beyond the typical scope of a terminology, such as suspected diagnoses, procedures not done, or family history facts. This additional contextual information can affect logical conclusions. An explanation of the reasons behind this, involving epistemological versus ontological components of meaning, is beyond the scope of this paper (see [3,13]). Based on the logical definitions of concepts directly related to PCEs, the approach presented in this paper enables the extraction of problematic pieces of information and their storage within separate elements of the information model. As a result, a concept like *165008002 |Allergy testing not done (situation)|* could be represented by separating the epistemological aspect "not done" into the suitable FHIR element *Procedure.status*:

- *Procedure.code: 252512005 |Allergy test (procedure)|*
- *Procedure.status: not-done* (according to the required HL7 FHIR ValueSet).

In this way, the integrity of SNOMED CT's hierarchies may be preserved.

Conclusions

The use of PCEs greatly enhances SNOMED CT's capacity to capture medical details comprehensively. However, despite its advantages, postcoordination has not yet been widely adopted in routine data collection. To address this, PCEtoFHIR offers a solution to ensure semantic interoperability between systems that are adept at postcoordination and those that are not, by leveraging the globally accepted HL7 FHIR standard to provide an alternative representation. State-of-the-art techniques in description logic and terminology services are integrated into

a largely automated web application that decomposes PCEs into their core components. Validation of both individual steps and the overall process confirms the approach's functionality. PCEtoFHIR is designed and implemented modularly, positioning it for future adaptations in the evolving landscape of health informatics. In addition to straightforward extensions to other SNOMED CT hierarchies or FHIR profiles by adding more StructureMaps, the algorithm can be adapted to work with other information models, such as openEHR or relational databases. The approach also holds the potential for addressing further challenges in semantic and structural standards, such as the TermInfo problem. By reversing the processing direction—from FHIR elements back to PCE—meaningful SNOMED CT-based analyses could be facilitated.

Acknowledgments

This work is funded by the German Federal Ministry of Education and Research (BMBF) as part of the Medical Informatics Initiative Germany (grant number 01ZZ2312A).

Conflicts of Interest

None declared.

References

1. Benson T, Grieve G. Principles of Health Interoperability: FHIR, HL7 and SNOMED CT. Cham, Switzerland: Springer; 2021.
2. Braunstein M. Health Informatics on FHIR: How HL7's API is Transforming Healthcare. Cham, Switzerland: Springer International Publishing; 2022.
3. Ingenerf J, Drenkhahn C. Referenzterminologie SNOMED CT: Interlingua zur Gewährleistung semantischer Interoperabilität in der Medizin. Cham, Switzerland: Springer; 2024.
4. Kassenärztliche Bundesvereinigung (KBV). KBV-Basis-Profile. KBV. URL: https://simplifier.net/organization/kassenrztlichebundesvereinigungkbv [accessed 2023-11-23]
5. Medizininformatik Initiative. Der Kerndatensatz der Medizininformatik-Initiative. Medizininformatik Initiative. 2023 Nov 29. URL: https://www.medizininformatik-initiative.de/de/der-kerndatensatz-der-medizininformatik-initiative [accessed 2023-11-29]
6. Drenkhahn C, Ohlsen T, Wiedekopf J, Ingenerf J. WASP—a web application to support syntactically and semantically correct SNOMED CT postcoordination. Applied Sciences 2023 May 16;13(10):6114. [doi: 10.3390/app13106114]
7. Karlsson D, Nyström M, Cornet R. Does SNOMED CT post-coordination scale? Stud Health Technol Inform 2014;205:1048-1052. [Medline: 25160348]
8. Cornet R, Nyström M, Karlsson D. User-directed coordination in SNOMED CT. Stud Health Technol Inform 2013;192:72-76. [Medline: 23920518]
9. Lopetegui M, Mauro A. A novel approach to create a machine readable concept model for validating SNOMED CT concept post-coordination. Stud Health Technol Inform 2015;216:1087. [Medline: 26262386]
10. Wardle M. Health informatics and information technology | Semantic interoperability: SNOMED CT, post-coordination and the model. Wardle (blog). URL: https://wardle.org/terminology/2018/10/27/snomed-postcoordination-1.html [accessed 2023-11-29]
11. Metke-Jimenez A, Steel J, Hansen D, Lawley M. Ontoserver: a syndicated terminology server. J Biomed Semantics 2018 Sep 17;9(1):24 [FREE Full text] [doi: 10.1186/s13326-018-0191-z] [Medline: 30223897]
12. CSIRO. Ontoserver: SNOMED CT post-coordination support. CSIRO. 2023 Nov 29. URL: https://ontoserver.csiro.au/docs/6/postcoordination.html [accessed 2023-11-29]
13. Schulz S, Schober D, Daniel C. Bridging the semantics gap between terminologies, ontologies, and information models. Studies in health technology and informatics 2010;160(2):1000-1004. [doi: 10.3233/978-1-60750-588-4-1000]
14. Cheetham E, Dolin R, Markwell D. Using SNOMED CT in HL7 version 3; implementation guide, release 1.5. HL7 International. 2024 Apr 26. URL: https://www.hl7.org/v3ballotarchive/v3ballot/html/infrastructure/terminfo/terminfo.html [accessed 2024-04-26]
15. TermInfo Project. HL7 International. 2024 Apr 22. URL: https://www.hl7.org/Special/committees/terminfo/overview.cfm [accessed 2024-04-22]
16. Ohlsen T, Kruse V, Krupar R, Banach A, Ingenerf J, Drenkhahn C. Mapping of ICD-O tuples to OncoTree codes using SNOMED CT post-coordination. Studies in Health Technology and Informatics 2022:294-311. [doi: 10.3233/shti220464]
17. Vogl K, Ingenerf J, Kramer J, Chantraine C, Drenkhahn C. LUMA: a mapping assistant for standardizing the units of LOINC-coded laboratory tests. Applied Sciences 2022 Jun 08;12(12):5848. [doi: 10.3390/app12125848]
18. Rinaldi E, Drenkhahn C, Gebel B, Saleh K, Tönnies H, von Loewenich FD, et al. Towards interoperability in infection control: a standard data model for microbiology. Sci Data 2023 Sep 23;10(1):654. [doi: 10.1038/s41597-023-02560-x] [Medline: 37741862]

19. Green JM, Wilcke JR, Abbott J, Rees LP. Development and evaluation of methods for structured recording of heart murmur findings using SNOMED-CT(R) post-coordination. Journal of the American Medical Informatics Association 2006 May 01;13(3):321-333. [doi: 10.1197/jamia.m1973]
20. Miñarro-Giménez JA, Martínez-Costa C, López-García P, Schulz S. Building SNOMED CT post-coordinated expressions from annotation groups. Stud Health Technol Inform 2017;235:446-450. [Medline: 28423832]
21. Castell-Díaz J, Miñarro-Giménez JA, Martínez-Costa C. Supporting SNOMED CT postcoordination with knowledge graph embeddings. J Biomed Inform 2023 Mar;139:104297 [FREE Full text] [doi: 10.1016/j.jbi.2023.104297] [Medline: 36736448]
22. HL7 International. Resource condition - mappings. HL7 International. 2023 Nov 29. URL: https://hl7.org/fhir/R4/condition-mappings.html [accessed 2023-11-29]
23. HL7 International. Resource observation - mappings. HL7 International. 2023 Nov 23. URL: https://hl7.org/fhir/R4/observation-mappings.html [accessed 2023-11-23]
24. SNOMED International. SNOMED on FHIR; bindings to FHIR clinical resources. SNOMED International. URL: https://confluence.ihtsdotools.org/display/FHIR/Bindings+to+FHIR+Clinical+Resources [accessed 202-04-26]
25. Perez-Rey D, Alonso-Calvo R, Paraiso-Medina S, Munteanu CR, Garcia-Remesal M. SNOMED2HL7: a tool to normalize and bind SNOMED CT concepts to the HL7 reference information model. Comput Methods Programs Biomed 2017 Oct;149:1-9 [FREE Full text] [doi: 10.1016/j.cmpb.2017.06.020] [Medline: 28802325]
26. Arguello CM, Martinez-Costa C, Des-Diu J, Maroto N, Fernandez-Prieto M. From SNOMED CT expressions to an FHIR RDF representation: exploring the benefits of an ontology-based approach. 2019 Presented at: The Joint Onthology Workshop; September 23-25, 2019; Graz, Austria URL: https://ceur-ws.org/Vol-2518/paper-ODLS1.pdf
27. Hitzler P, Krötzsch M, Rudolph S, Sure Y. Semantic Web. Berlin, Heidelberg: Springer; 2008.
28. Staab S, Studer R. Handbook on Ontologies. Berlin, Heidelberg: Springer; 2009.
29. International Health Terminology Standards Development Organisation. SNOMED CT OWL Guide. International Health Terminology Standards Development Organisation. URL: https://confluence.ihtsdotools.org/display/DOCOWL/SNOMED+CT+OWL+Guide?preview=/64265998/142119608/SNOMED%20CT%20OWL%20Guide-v19-20220128.pdf [accessed 2023-11-29]
30. International Health Terminology Standards Development Organisation. Snomed OWL Toolkit. GitHub. 2016. URL: https://github.com/IHTSDO/snomed-owl-toolkit [accessed 2024-01-04]
31. International Health Terminology Standards Development Organisation. SNOMED CT International Edition. National Library of Medicine. URL: https://www.nlm.nih.gov/healthit/snomedct/international.html [accessed 2023-11-29]
32. Horridge M, Bechhofer S. The OWL API: a Java API for OWL ontologies. Semantic Web 2011;2(1):11-21. [doi: 10.3233/sw-2011-0025]
33. Kazakov Y, Krötzsch M, Simancik F. ELK reasoner: architecture and evaluation. 2012 Presented at: The 1st International OWL Reasoner Evaluation Workshop; July 1, 2012; Manchester, UK p. 1-12.
34. Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J Biomed Inform 2011 Oct;44(5):749-759 [FREE Full text] [doi: 10.1016/j.jbi.2011.03.013] [Medline: 21463704]
35. Hagberg A, Swart P, Chult D. Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy2008). 2008 Presented at: 7th Annual Python in Science Conference; August 19-24, 2008; Pasadena, CA p. 11-15.
36. International Health Terminology Standards Development Organisation. Template syntax DRAFT specification. SNOMED Confluence. URL: https://confluence.ihtsdotools.org/display/DOCSTS?preview=/45529301/115875508/doc_TemplateSyntax_v1.1.1-en-US_INT_20201020.pdf [accessed 2023-11-29]
37. HL 7 International. Resource StructureMap - content. HL 7 International. URL: https://hl7.org/fhir/R4/structuremap.html [accessed 2023-11-29]
38. University Health Network. HAPI FHIR - the open source FHIR API for Java. University Health Network. URL: http://hapifhir.io/ [accessed 2024-04-03]
39. Smile Digital Health. HAPI FHIR. URL: https://hapifhir.io/hapi-fhir [accessed 2024-01-04]
40. Ohlsen T. PCEtoFHIR. Github. URL: https://github.com/itcr-uni-luebeck/pce-to-fhir [accessed 2024-08-30]
41. Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 1989;19(1):17-30. [doi: 10.1109/21.24528]
42. Leacock C, Chodorow M. Combining Local Context and WordNet Similarity for Word Sense Identification. WordNet: an electronic lexical database MIT Press 1998:265-283. [doi: 10.7551/mitpress/7287.003.0018]
43. Yuhua Li, Bandar Z, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 2003 Jul;15(4):871-882. [doi: 10.1109/tkde.2003.1209005]
44. Choi I, Kim M. Topic distillation using hierarchy concept tree. New York, NY: ACM; 2003 Presented at: SIGIR03: The 26th ACM/SIGIR International Symposium on Information Retrieval; July 28, 2003; Toronto, ON p. 371-372. [doi: 10.1145/860435.860506]
45. Kate RJ. Automatic full conversion of clinical terms into SNOMED CT concepts. J Biomed Inform 2020 Nov;111:103585 [FREE Full text] [doi: 10.1016/j.jbi.2020.103585] [Medline: 33011295]

46. Ohlsen T. PCEtoFHIR - FHIR StructureMaps. Github. URL: https://itcr-uni-luebeck.github.io/pce-to-fhir/ [accessed 2024-08-01]
47. Kassenärztliche Bundesvereinigung. KBV-Basis-Allergie/Unverträglichkeit, 1.5.1.2 BIS. KBV. URL: https://mio.kbv.de/display/BASE1X0/1.5.1.2+bis [accessed 2023-11-29]
48. HL7 International. StructureDefinition-condition-diseaseCourse. HL7 International. URL: https://fhir.org/guides/stats2/structuredefinition-hl7.fhir.uv.extensions.r4-condition-diseasecourse.html [accessed 2024-01-04]
49. HL7 International. StructureDefinition-condition-dueTo. HL7 International. URL: https://fhir.org/guides/stats2/structuredefinition-hl7.fhir.uv.extensions.r4-condition-dueto.html [accessed 2023-11-29]
50. Rohit K. Clinical terms to SNOMED CT concepts. University of Wisconsin Milwaukee. URL: https://sites.uwm.edu/katerj/conversion/ [accessed 2024-05-13]
51. Lau T. Gesundheitsdaten: FHIR wird europaweiter Standard. Ärzteblatt. URL: https://www.aerzteblatt.de/nachrichten/142159/Gesundheitsdaten-FHIR-wird-europaweiter-Standard [accessed 2024-01-04]
52. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
53. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
54. Schulz S. Kontroversen in der Medizinischen Informatik. Wozu benötigen wir standardisierte Terminologien wie SNOMED CT? Swiss Med Informatics 2011 Nov 25:27-32. [doi: 10.4414/smi.27.00272]

Abbreviations

CSIRO: Commonwealth Scientific and Industrial Research Organization
DAG: directed acyclic graph
ECL: Expression Constraint Language
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level 7
KBV: Kassenärztliche Bundesvereinigung
LOINC: Logical Observation Identifiers Names and Codes
MII: Medical Informatics Initiative
NASHIP: National Association of Statutory Health Insurance Physicians
OWL: Web Ontology Language
PCE: postcoordinated expression
RIM: Reference Information Model
SNOMED CT: SNOMED Clinical Terms

Edited by C Lovis; submitted 28.02.24; peer-reviewed by C Lien, S Lee; comments to author 21.04.24; revised version received 14.06.24; accepted 22.07.24; published 17.09.24.

Please cite as:
Ohlsen T, Ingenerf J, Essenwanger A, Drenkhahn C
PCEtoFHIR: Decomposition of Postcoordinated SNOMED CT Expressions for Storage as HL7 FHIR Resources
JMIR Med Inform 2024;12:e57853
URL: https://medinform.jmir.org/2024/1/e57853
doi:10.2196/57853
PMID:

Event Analysis for Automated Estimation of Absent and Persistent Medication Alerts: Novel Methodology

Janina A Bittmann[1,*], Dr sc hum; Camilo Scherkl[2,*]; Andreas D Meid[2], PD Dr sc hum; Walter E Haefeli[1], Prof Dr med; Hanna M Seidling[1], Prof Dr sc hum

[1]

[2]

[*]these authors contributed equally

Corresponding Author:
Hanna M Seidling, Prof Dr sc hum

Abstract

Background: Event analysis is a promising approach to estimate the acceptance of medication alerts issued by computerized physician order entry (CPOE) systems with an integrated clinical decision support system (CDSS), particularly when alerts cannot be interactively confirmed in the CPOE-CDSS due to its system architecture. Medication documentation is then reviewed for documented evidence of alert acceptance, which can be a time-consuming process, especially when performed manually.

Objective: We present a new automated event analysis approach, which was applied to a large data set generated in a CPOE-CDSS with passive, noninterruptive alerts.

Methods: Medication and alert data generated over 3.5 months within the CPOE-CDSS at Heidelberg University Hospital were divided into 24-hour time intervals in which the alert display was correlated with associated prescription changes. Alerts were considered “persistent” if they were displayed in every consecutive 24-hour time interval due to a respective active prescription until patient discharge and were considered “absent” if they were no longer displayed during continuous prescriptions in the subsequent interval.

Results: Overall, 1670 patient cases with 11,428 alerts were analyzed. Alerts were displayed for a median of 3 (IQR 1-7) consecutive 24-hour time intervals, with the shortest alerts displayed for drug-allergy interactions and the longest alerts displayed for potentially inappropriate medication for the elderly (PIM). Among the total 11,428 alerts, 56.1% (n=6413) became absent, most commonly among alerts for drug-drug interactions (1915/2366, 80.9%) and least commonly among PIM alerts (199/499, 39.9%).

Conclusions: This new approach to estimate alert acceptance based on event analysis can be flexibly adapted to the automated evaluation of passive, noninterruptive alerts. This enables large data sets of longitudinal patient cases to be processed, allows for the derivation of the ratios of persistent and absent alerts, and facilitates the comparison and prospective monitoring of these alerts.

(JMIR Med Inform 2024;12:e54428) doi:10.2196/54428

KEYWORDS

clinical decision support system; CDSS; medication alert system; alerting; alert acceptance; event analysis

Introduction

Computerized physician order entry (CPOE) systems with integrated clinical decision support systems (CDSS) can reduce medication errors by highlighting critical medication constellations [1]. To realize their full potential, medication alerts must be recognized and followed by users. Hence, measuring “alert acceptance” is a key prerequisite for evaluating the effectiveness of a CDSS.

In principle, two methods can estimate alert acceptance: (1) in-dialog analysis where users interactively click to accept or override displayed alerts; and (2) event analysis where the medication chart and associated documentation are reviewed for evidence of alert acceptance through further actions (“events”) responsive to the alert (eg, discontinued medication orders), which often requires extensive manual screening [2]. Most studies addressing alert acceptance used in-dialog analyses because the display of alerts, especially in English-speaking countries, is part of the technical architecture of the CPOE-CDSS [3]. There is limited evidence on how to perform event analyses because it is uncertain whether the prescribing behavior is influenced by alerts or other clinical therapeutic circumstances (eg, scheduled end of treatment) [2]. Moreover, the manual screening of the medication documentation is a time-consuming process [4], especially when administrative

processes such as changing wards and the simultaneous transfer of physicians' responsibility for the medication are considered in the alert presentation.

As CDSS installations presenting passive, noninterruptive alerts become increasingly popular in European countries [5,6], the need for developing and validating techniques for automatic event analyses is increasing. This is particularly important when considering all alerts throughout the inpatient stay.

We present a new approach to perform an automated event analysis, which was applied to a large data set of medication alerts.

Methods

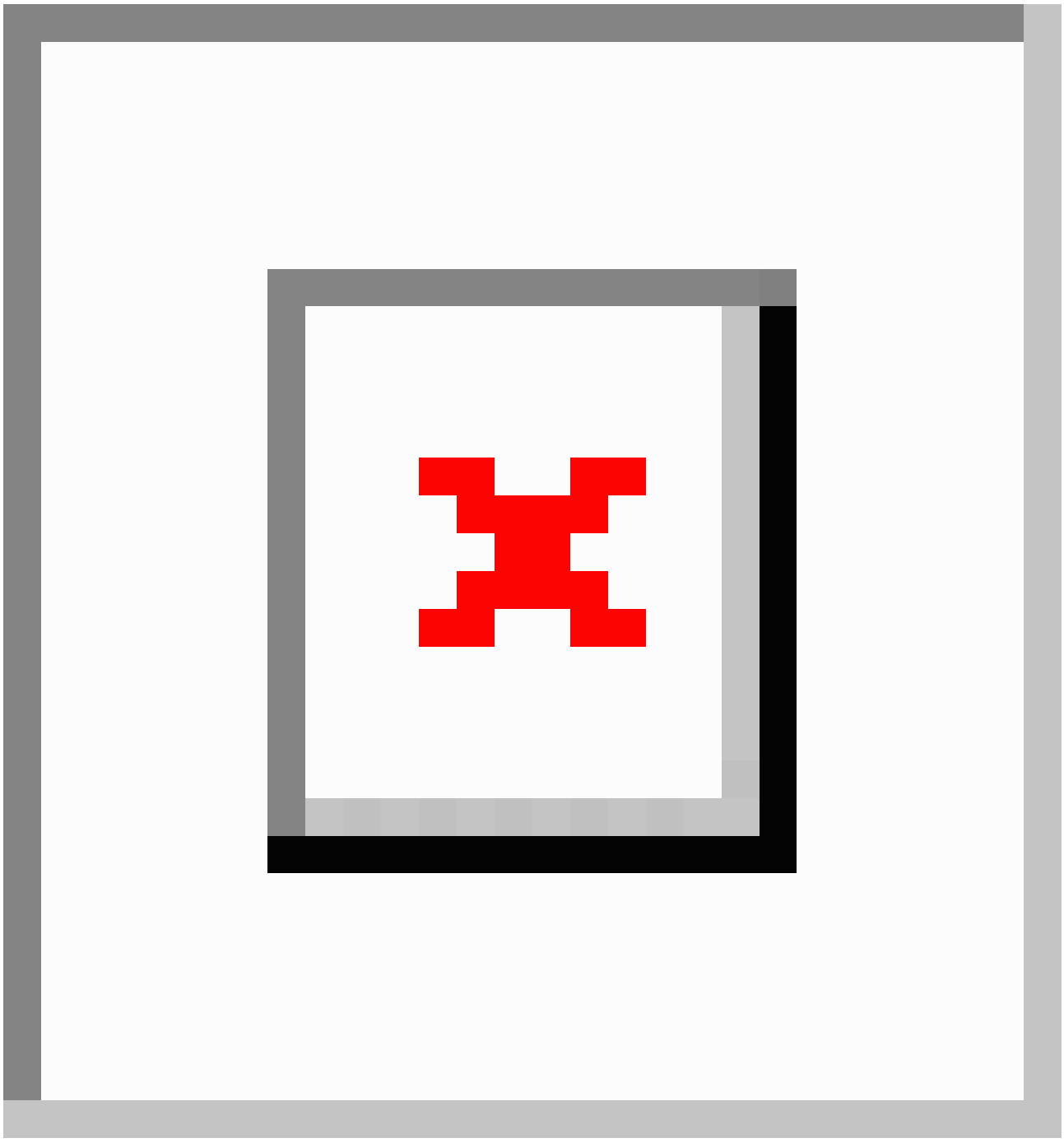
Ethical Considerations

Study approval was granted by the responsible Ethics Committee of the Medical Faculty of Heidelberg University (S-467/2020) and by the local data protection officer for the data protection concept. Human subjects were not directly involved; all data were pseudonymized and could neither be linked to individual patients nor to prescribers.

Setting

We analyzed the prescription and alert data issued over 3.5 months during routine care at Heidelberg University Hospital (a 2500-bed tertiary care hospital) within the local CPOE (*i.s.h.med Smart Medication*, Oracle Cerner, North Kansas City, USA) with an integrated CDSS (AiD*Klinik*, Dosing GmbH, Heidelberg, Germany). To view the presented passive and noninterruptive alerts, users actively navigate from their prescription screen to a separate window that opens upon request. In this window, all alerts are displayed in a single table sorted by severity and presented with a brief summary (Figure 1). Users are required to click on each alert to access more detailed information. The system does not recognize whether an alert has been viewed. Additionally, users are not obliged to directly flag alerts as accepted or overwritten. Therefore, these data are not available in our CPOE-CDSS. Implemented alert types comprised checking for drug-drug interactions (DDIs), drug-allergy interactions (DAIs), duplicate prescriptions (DPs), advanced dosing recommendations for potentially inappropriate medication for the elderly (aged ≥65 years, PIM), or prescriptions exceeding the maximum recommended daily dose (PE-MDDs).

Figure 1. Schematic display of an exemplary alert window listing all alerts for a patient in a table. Each alert is presented in a separate line, sorted by severity, with the most severe alerts listed first. The first column displays the alert type in a color-coded scheme (black=contraindicated, red=severe, orange=moderate), followed by an explanation for the severity, a brief description of the problem, and the name of the causative drug.



Data Collection

The relevant parameters extracted from the CPOE-CDSS were information on prescriptions, issued alerts, administrative patient data, and setting data. Prescription schedules with regimen changes were documented as separate entries so that prescriptions potentially resulting from previous prescriptions (eg, because of dose reduction or conversion of fixed to as-needed prescriptions) could be linked retroactively. Follow-up prescriptions were defined as prescriptions of the same drug and administration route when the previous prescription ended and the subsequent one started within 10 minutes.

Alert Management

In this CDSS, prescriber review of alerts may result in alerts disappearing due to prescription changes and adaptations or in alerts being continuously displayed for unchanged prescriptions.

In this methodology, alerts are defined as “absent” when they disappear during continuous prescriptions for which underlying risk constellations no longer exist (eg, dose reduction of an overdosed prescription but the prescription itself remains continuous). In contrast, alerts consistently displayed until patient transfer, discharge, or end of the prescription are categorized as “persistent” (eg, the prescription remains valid

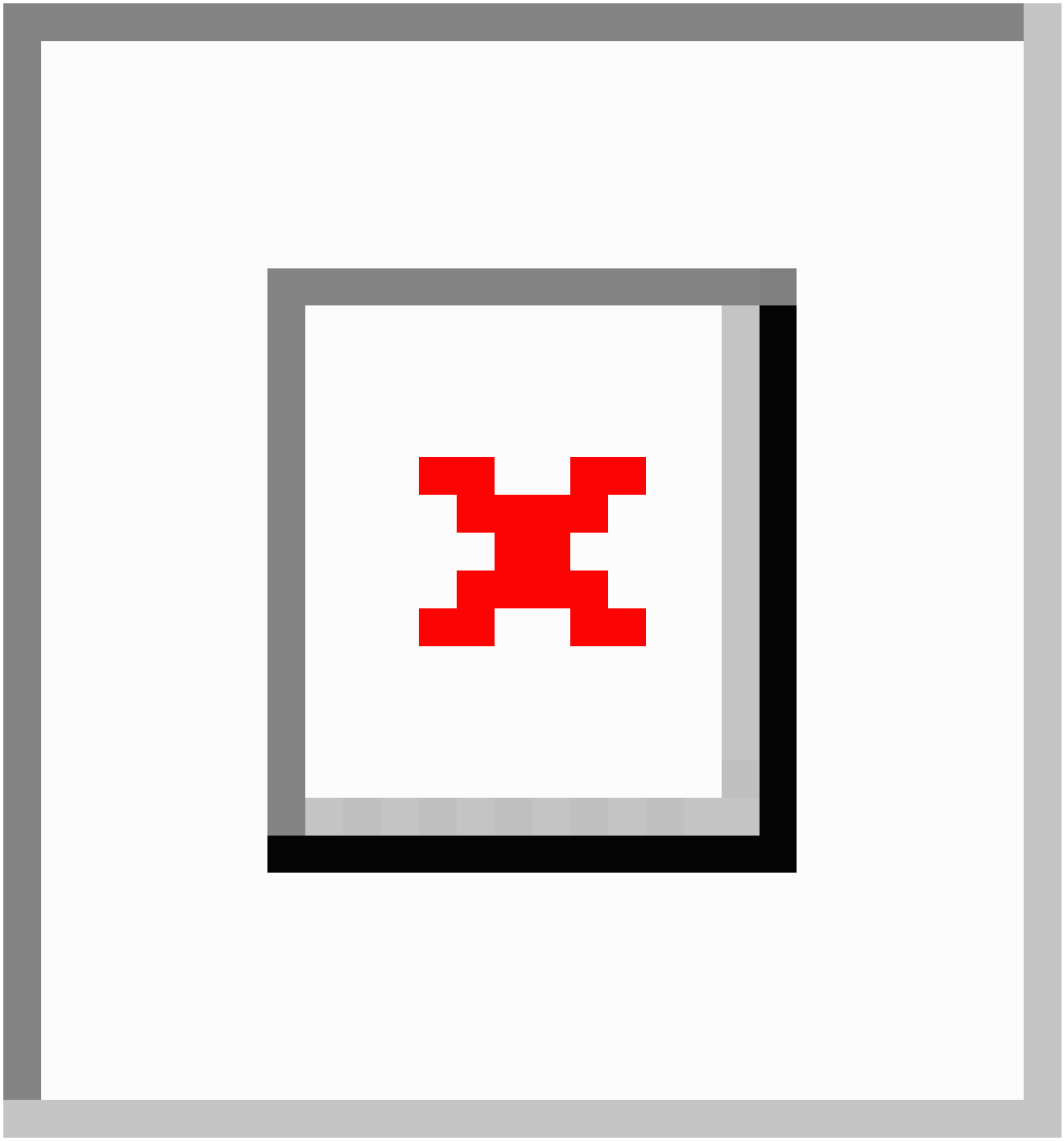
even though the prescribed active ingredient is alerted by a DAI).

Data Analysis

To automatically identify absent alerts, the medication and corresponding alert data were divided into 24-hour time intervals. Since reproducible time stamps were lacking in the database, selection of this conservative interval allowed the retroactive linkage of alert display and associated changes in the corresponding prescription. Therefore, two consecutive time intervals were compared according to whether or not the alerts were continuously displayed. Alerts were excluded if (1) they were first displayed on the discharge day (Figure 2, Alert 4); (2) required user interaction (eg, answering questions to decide whether conditions for alerts are present); or (3) were triggered by one-time prescriptions, in which case the alert response cannot be assessed the next day (Supplementary Methods section A and Figure S1 in Multimedia Appendix 1). The display duration of alerts (DDoA) was calculated from the time interval between the first and last alert display (Figure 2). Further details on the development of the 24-hour time intervals can be found in Supplementary Methods section B in Multimedia Appendix 1; examples of data analysis of different alert types are provided in Tables S1-S3 and Figure S2 of Multimedia Appendix 1.

Based on the exploratory rates of absent alerts, basic descriptive statistics were applied. The χ^2 test was performed to evaluate whether absent alerts differed stratified by alert types considering a two-tailed P value <.05 as significant (IBM SPSS Statistics version 25, Ehningen, Germany). The R packages *gpmodels* [7], *survival*, and *ggplot2* (version 4.1.2, R Foundation for Statistical Computing, Vienna, Austria) were used for data analysis and visualization.

Figure 2. Proposed methodology for identification of absent alerts, exemplified for a 5-day inpatient stay. Each midnight (dotted lines), all alerts displayed within the last 24-hour time interval were identified. Alert 1 is displayed between day 1 (admission) and day 4; the display duration of alerts (DDoA) is 4 days. Alert 2 is displayed from day 2 until discharge; the DDoA is 4 days. Alert 3 is displayed only on day 3; the DDoA is 1 day. Alert 4 is displayed for the first time on the day of discharge and remained until discharge; the DDoA is 1 day. Alert 4 was excluded from the analysis because, due to the discharge, there is no subsequent (sixth) 24-hour time interval with which the fifth interval could have been compared to evaluate the alert display. Each alert could be identified because of a unique alert ID code. Using this identification concept, alert IDs detected within a 24-hour time interval could be compared to alerts detected in the previous 24-hour time interval. Therefore, it was possible to automatically classify which alerts were (1) newly displayed (no matching ID in the previous interval: Alert 1, Day 1), (2) displayed for more than 24 hours (matching ID in consecutive intervals; Alert 1, Days 2-4), or (3) absent (no matching ID in the current 24-hour time interval: Alert 1, Day 4).



Results

Alert Display and Composition

We considered the data of 1670 patient cases (Figure S3 in Multimedia Appendix 1) with a median hospital stay of 7 days (IQR 4-13). During this time, 13,979 alerts were displayed. Because 2284 alerts (16.3%) were triggered by one-time prescriptions and 267 alerts (1.9%) were first displayed on the discharge day, the remaining 11,428 alerts (81.8%) formed the basis for analysis. The alert types triggering the alerts are shown in Table 1.

The median DDoA was 3 days (IQR 1-7) and varied by alert type, with alerts for DAIs showing the shortest DDoA (Table 1).

Table . Alert types triggering the alerts, corresponding rates of absence, and display duration of the alerts.

Alert type	Triggered alerts, n (%)[a]	Absent alerts, n (%)[b]	Display duration of alerts (days), median (IQR; range)
Alerts for duplicate prescriptions	7643 (66.9)	3674 (48.1)	3 (1-8; 1-31)
Alerts for drug-drug interactions	2366 (20.7)	1915 (80.9)	2 (1-5; 1-31)
Alerts for drug-allergy interactions	517 (4.5)	416 (80.5)	1 (1-2; 1-24)
Alerts for potentially inappropriate medication for the elderly	499 (4.4)	199 (39.9)	4 (2-8; 1-31)
Alerts for prescriptions exceeding the maximum recommended daily dose	403 (3.5)	209 (51.9)	3 (1-6; 1-30)
Total number of alerts	11,428	6413	3 (1-7; 1-31)

[a]Percentages are based on the total number of analyzed alerts (N=11,428).

[b]Percentages are based on the number of analyzed alerts for each alert type.

Absent and Persistent Alerts

From all 11,428 analyzed alerts, 43.9% (n=5015) persisted and 56.1% (n=6413) were absent, with alerts for DDIs showing the highest rate of absence (80.9%) and PIMs the lowest (39.9%) (Table 1).

The proportions of absent alerts differed significantly between the individual alert types (P_{χ^2}<.001), except for DDI alerts compared to DAI alerts (P_{χ^2}=.80) and for DP alerts compared to alerts for PE-MDDs (P_{χ^2}=.14). The proportion of absent alerts in relation to the DDoA was the highest for DAI alerts and the lowest for alerts for PIMs in the first 24 hours after admission (Figures S4-S5 in Multimedia Appendix 1).

Discussion

Principal Findings

A new methodological approach for routine care data was applied performing an automated event analysis that is transferable to other CPOE-CDSS with passive, noninterruptive alerts. In previous studies using event analyses, alert acceptance rates were identified at the drug administration level [8], prescription level [9], or at both levels [10]. There is general consensus that alert acceptance rates vary widely depending on the measuring method and study setting, resulting in different and incomparable rates [11]. Since in-dialog analysis is often not possible in a European CPOE-CDSS, this new methodology adapted to the technical structures of such a CPOE-CDSS is needed.

A key strength of the proposed method is that it variably adjusts the time intervals and consequently the lookback windows underlying the method's programming. Thus, temporary changes in prescriptions within the determined time interval (here 24 hours) are considered persistent alerts; however, this CPOE-CDSS interrupts the alert display in certain cases, such as when patients change wards and responsibility for medication is handed over to another physician. This transfer results in automatic prescription pauses that are actively suspended by physicians, technically leading to the redisplay of alerts. Without the definition of this time interval, these pauses would incorrectly increase the overall number of alerts when reappearing and the rate of absent alerts as they disappear for a few hours during valid prescriptions. Hence, this method considers administrative processes of the daily clinical routine and guarantees that only alerts of real prescription changes are evaluated. For retrospectively matching the time-dependent correlations of the alerts over time and in the clinical routine, it is essential to consider alerts throughout the inpatient stay and our proposed method meets this need.

However, according to the technical architecture of this CPOE-CDSS, there is no obvious link between reviewing alerts and possible resulting changes in prescription data. Therefore, various assumptions had to be made for this data evaluation. Alerts were categorized as either persistent or absent based on the assumptions that alerts were regularly checked and that alerts disappeared because underlying risk constellations no longer existed due to previously displayed alerts. This general assumption may overestimate the rate of actual alert acceptance, as a prescribed medication could be switched based on patient conditions (eg, adverse events, intolerance) or treatment schedules. As it remains unclear whether the change in drug prescriptions was caused by the alert display or due to other variables and because no control group was available due to the retrospective design, caution is required when interpreting absolute numbers and comparing the proportion of absent alerts to previously published acceptance rates. In the future, this retrospective method will need to be prospectively evaluated including validity measurements by comparing the results of this automated approach with those derived from manual screening. Another limitation is that this study was conducted in a single center with a CPOE-CDSS that is highly specific and strongly adapted to workflows and care processes at our institution. This analysis only considered alerts at the prescribing level and did not measure whether the respective drugs were indeed administered. For instance, many of the alerts for DPs were triggered by drugs that were prescribed as as-needed prescriptions. Hence, these DPs tended to indicate a variety of treatment options rather than actually being administered together. This might have contributed to the reduced occurrence of the absence of alerts for DPs on a prescribing level compared to other alerts. However, in our CPOE-CDSS, it is unalterably

stipulated that medication alerts are implemented in a passive and noninterruptive way. While it may be challenging to transfer this complex method to systems with differing data infrastructures, to our knowledge, this is the first automated method for processing persistent and absent medication alerts in a system with passive, noninterruptive alerts. Additionally, since this method was programmed in a modular way, it seems feasible to transfer and adapt it to other settings.

Conclusions

A methodology was applied to an automatic event analysis in a CPOE-CDSS with passive, noninterruptive alerting. This enables the processing of large data sets of longitudinal periods of inpatient stays and can be used to automatically derive the percentage of absent alerts. Once implemented, this analysis can be repeated at any time and one could even imagine that real-time monitoring of persistent alerts in daily clinical routines could be set up using these data for future optimization of the CPOE-CDSS.

Acknowledgments

We thank Sonja Baumann, Silvia Kugler, Michael Metzner, Larissa Schiller, and Andreas Wirtherle for initial data extraction, preparation, and maintenance.

Authors' Contributions

JAB planned the study, was involved in the development of the method and data evaluation, and wrote the manuscript. CS was involved in the development of the method and data evaluation and wrote the manuscript. ADM was involved in the development of the method and data evaluation, wrote parts of the manuscript and critically revised it. WEH wrote parts of the manuscript and critically revised it. HMS planned the study, was involved in the development of the method and data evaluation, and wrote and critically revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Description of one-time prescription and illustration of its impact on the alert display (Figure S1). Description of the development of 24-hour time intervals, with exemplary data sets for fixed variables (Table S1), time-dependent variables (Table S2), and an exemplary time frame of processed longitudinal alert data (Table S3). Data analysis examples for different alert types (Figure S2). CONSORT (Consolidated Standards of Reporting Trials) diagram for included patient cases (Figure S3). Proportions of absent (Figure S4) and persistent (Figure S5) alerts stratified by the alert type.
[DOCX File, 353 KB - medinform_v12i1e54428_app1.docx]

References

1. Wolfstadt JI, Gurwitz JH, Field TS, et al. The effect of computerized physician order entry with clinical decision support on the rates of adverse drug events: a systematic review. J Gen Intern Med 2008 Apr;23(4):451-458. [doi: 10.1007/s11606-008-0504-5] [Medline: 18373144]
2. Hussain MI, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc 2019 Oct 1;26(10):1141-1149. [doi: 10.1093/jamia/ocz095] [Medline: 31206159]
3. Duke JD, Li X, Dexter P. Adherence to drug-drug interaction alerts in high-risk patients: a trial of context-enhanced alerting. J Am Med Inform Assoc 2013 May 1;20(3):494-498. [doi: 10.1136/amiajnl-2012-001073] [Medline: 23161895]
4. Sethuraman U, Kannikeswaran N, Murray KP, Zidan MA, Chamberlain JM. Prescription errors before and after introduction of electronic medication alert system in a pediatric emergency department. Acad Emerg Med 2015 Jun;22(6):714-719. [doi: 10.1111/acem.12678] [Medline: 25998704]
5. Carli-Ghabarou D, Seidling HM, Bonnabry P, Lovis C. A survey-based inventory of clinical decision support systems in computerised provider order entry in Swiss hospitals. Swiss Med Wkly 2013;143:w13894. [doi: 10.4414/smw.2013.13894] [Medline: 24338034]
6. Ploegmakers KJ, Medlock S, Linn AJ, et al. Barriers and facilitators in using a clinical decision support system for fall risk management for older people: a European survey. Eur Geriatr Med 2022 Apr;13(2):395-405. [doi: 10.1007/s41999-021-00599-w] [Medline: 35032323]
7. R package gpmodels. GitHub. URL: https://github.com/ML4LHS/gpmodels [accessed 2023-11-16]
8. Muylle KM, Gentens K, Dupont AG, Cornu P. Evaluation of an optimized context-aware clinical decision support system for drug-drug interaction screening. Int J Med Inform 2021 Apr;148:104393. [doi: 10.1016/j.ijmedinf.2021.104393] [Medline: 33486355]

9. Slight SP, Beeler PE, Seger DL, et al. A cross-sectional observational study of high override rates of drug allergy alerts in inpatient and outpatient settings, and opportunities for improvement. BMJ Qual Saf 2017 Mar;26(3):217-225. [doi: 10.1136/bmjqs-2015-004851] [Medline: 26993641]
10. Muylle KM, Gentens K, Dupont AG, Cornu P. Evaluation of context-specific alerts for potassium-increasing drug-drug interactions: a pre-post study. Int J Med Inform 2020 Jan;133:104013. [doi: 10.1016/j.ijmedinf.2019.104013] [Medline: 31698230]
11. Kannry J. Alert acceptance: are all acceptance rates the same? J Am Med Inform Assoc 2023 Sep 25;30(10):1754. [doi: 10.1093/jamia/ocad151] [Medline: 37535817]

Abbreviations

CDSS: clinical decision support system
CPOE: computerized physician order entry
DAI: drug-allergy interaction
DDI: drug-drug interaction
DDoA: display duration of alert
DP: duplicate prescription
PE-MDD: prescription exceeding the maximum recommended daily dose
PIM: potentially inappropriate medication for the elderly

Edited by C Lovis; submitted 17.11.23; peer-reviewed by A Simona, D Malone; revised version received 19.03.24; accepted 07.04.24; published 04.06.24.

Please cite as:
Bittmann JA, Scherkl C, Meid AD, Haefeli WE, Seidling HM
Event Analysis for Automated Estimation of Absent and Persistent Medication Alerts: Novel Methodology
JMIR Med Inform 2024;12:e54428
URL: https://medinform.jmir.org/2024/1/e54428
doi:10.2196/54428

Accelerating Evidence Synthesis in Observational Studies: Development of a Living Natural Language Processing–Assisted Intelligent Systematic Literature Review System

Frank J Manion[1], PhD; Jingcheng Du[1], PhD; Dong Wang[2], PhD; Long He[1], MS; Bin Lin[1], MS; Jingqi Wang[1], PhD; Siwei Wang[1], MS; David Eckels[2], BA; Jan Cervenka[2]; Peter C Fiduccia[2], PhD; Nicole Cossrow[2]; Lixia Yao[2], PhD

1

2

Corresponding Author:
Dong Wang, PhD

Abstract

Background: Systematic literature review (SLR), a robust method to identify and summarize evidence from published sources, is considered to be a complex, time-consuming, labor-intensive, and expensive task.

Objective: This study aimed to present a solution based on natural language processing (NLP) that accelerates and streamlines the SLR process for observational studies using real-world data.

Methods: We followed an agile software development and iterative software engineering methodology to build a customized intelligent end-to-end living NLP-assisted solution for observational SLR tasks. Multiple machine learning–based NLP algorithms were adopted to automate article screening and data element extraction processes. The NLP prediction results can be further reviewed and verified by domain experts, following the human-in-the-loop design. The system integrates explainable articificial intelligence to provide evidence for NLP algorithms and add transparency to extracted literature data elements. The system was developed based on 3 existing SLR projects of observational studies, including the epidemiology studies of human papillomavirus–associated diseases, the disease burden of pneumococcal diseases, and cost-effectiveness studies on pneumococcal vaccines.

Results: Our Intelligent SLR Platform covers major SLR steps, including study protocol setting, literature retrieval, abstract screening, full-text screening, data element extraction from full-text articles, results summary, and data visualization. The NLP algorithms achieved accuracy scores of 0.86-0.90 on article screening tasks (framed as text classification tasks) and macroaverage F1 scores of 0.57-0.89 on data element extraction tasks (framed as named entity recognition tasks).

Conclusions: Cutting-edge NLP algorithms expedite SLR for observational studies, thus allowing scientists to have more time to focus on the quality of data and the synthesis of evidence in observational studies. Aligning the living SLR concept, the system has the potential to update literature data and enable scientists to easily stay current with the literature related to observational studies prospectively and continuously.

(JMIR Med Inform 2024;12:e54653) doi:10.2196/54653

KEYWORDS

machine learning; deep learning; natural language processing; systematic literature review; artificial intelligence; software development; data extraction; epidemiology

Introduction

Systematic literature reviews (SLRs) are widely recognized as a robust method to identify and summarize evidence from published sources [1]. However, conducting an SLR can be a complex, time-consuming, labor-intensive, and expensive task, depending on the breadth of the topic, level of granularity, or resolution of the review needed [2,3]. One recent study estimated the time and cost required to conduct an SLR can be as high as 1.72 person-years of scientist effort and approximately $140,000 per review [4]. Because SLRs are so resource intensive, it is difficult to stay up to date, and once an SLR is complete and new literature is published, the SLR may become incomplete and obsolete as time goes by.

Natural language processing (NLP) refers to artificial intelligence (AI) technologies that can extract structured information from textual documents such as medical charts, lab results, and many other types of unstructured text. NLP has significantly advanced a variety of biomedical applications in recent years. There is considerable community interest in using AI such as machine learning (ML) and NLP to improve automation in aspects of literature reviews [2,5-7]. For example,

Thomas et al used NLP to identify randomized controlled trials for Cochrane reviews, and Wallace et al developed methods to extract sentences from literature related to clinical trial reports. There are also some SLR management software, such as Raynan.ai [8], which leverages NLP to expedite certain SLR steps, including article screening.

Despite these existing efforts, there is a lack of systematic and integrated NLP solutions for SLR to cover its full aspects, preventing the wide adoption of such tools in SLR projects.

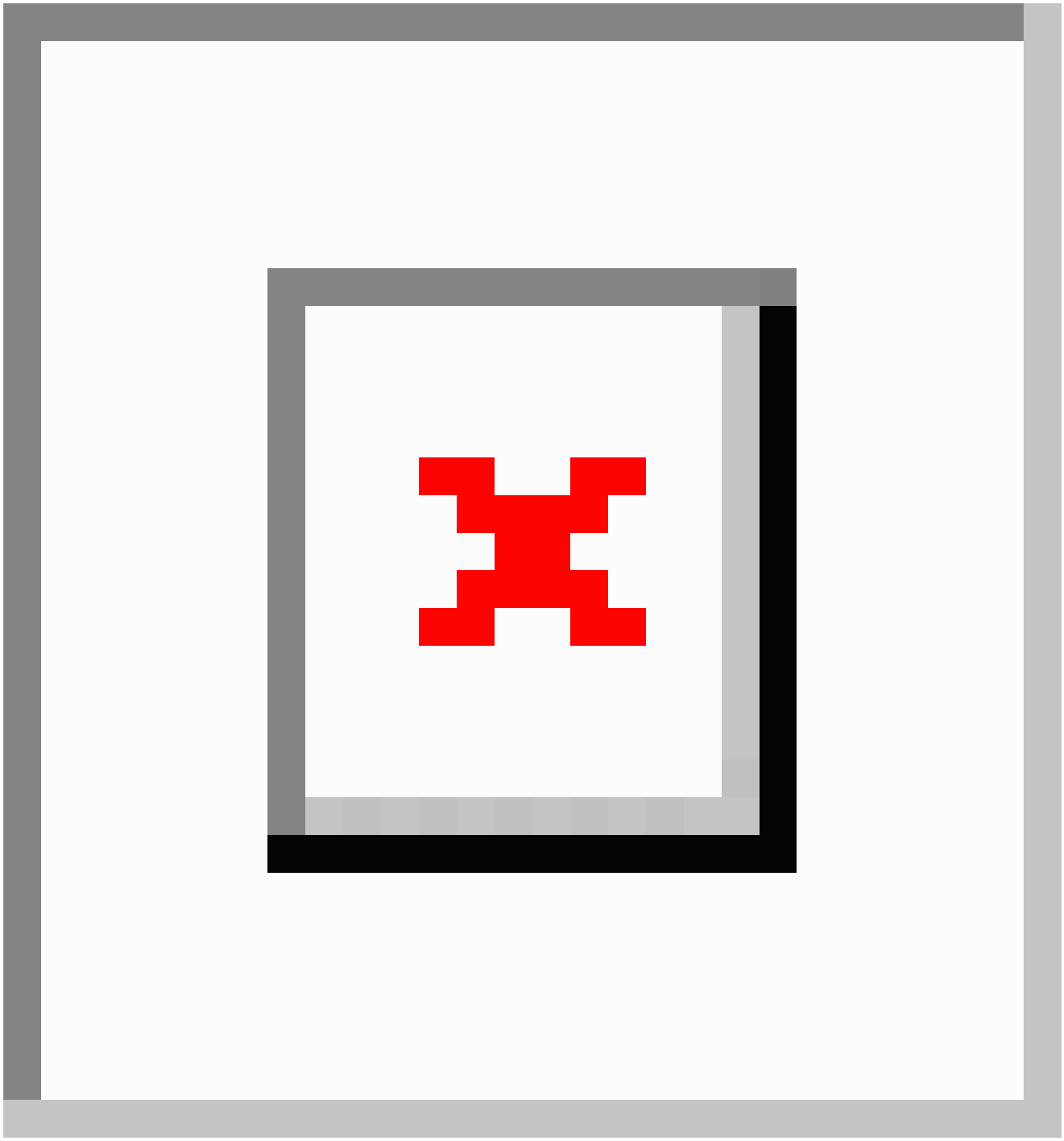
Thus, in this study, we evaluated an intelligent SLR system (hereinafter referred to as ISLR) for observational SLR tasks. The use of NLP improves efficiency, while the human-in-the-loop approach improves accuracy and reduces errors. The system uses cutting-edge NLP tools that employ ML and deep learning (DL) approaches to expedite the time-consuming processes involved in an SLR by making a series of learned recommendations to the end user. The purpose of this study is to evaluate an AI tool that accelerates and streamlines the SLR process and to demonstrate the validity of this tool in 3 use cases.

Methods

Workflow and System Architecture

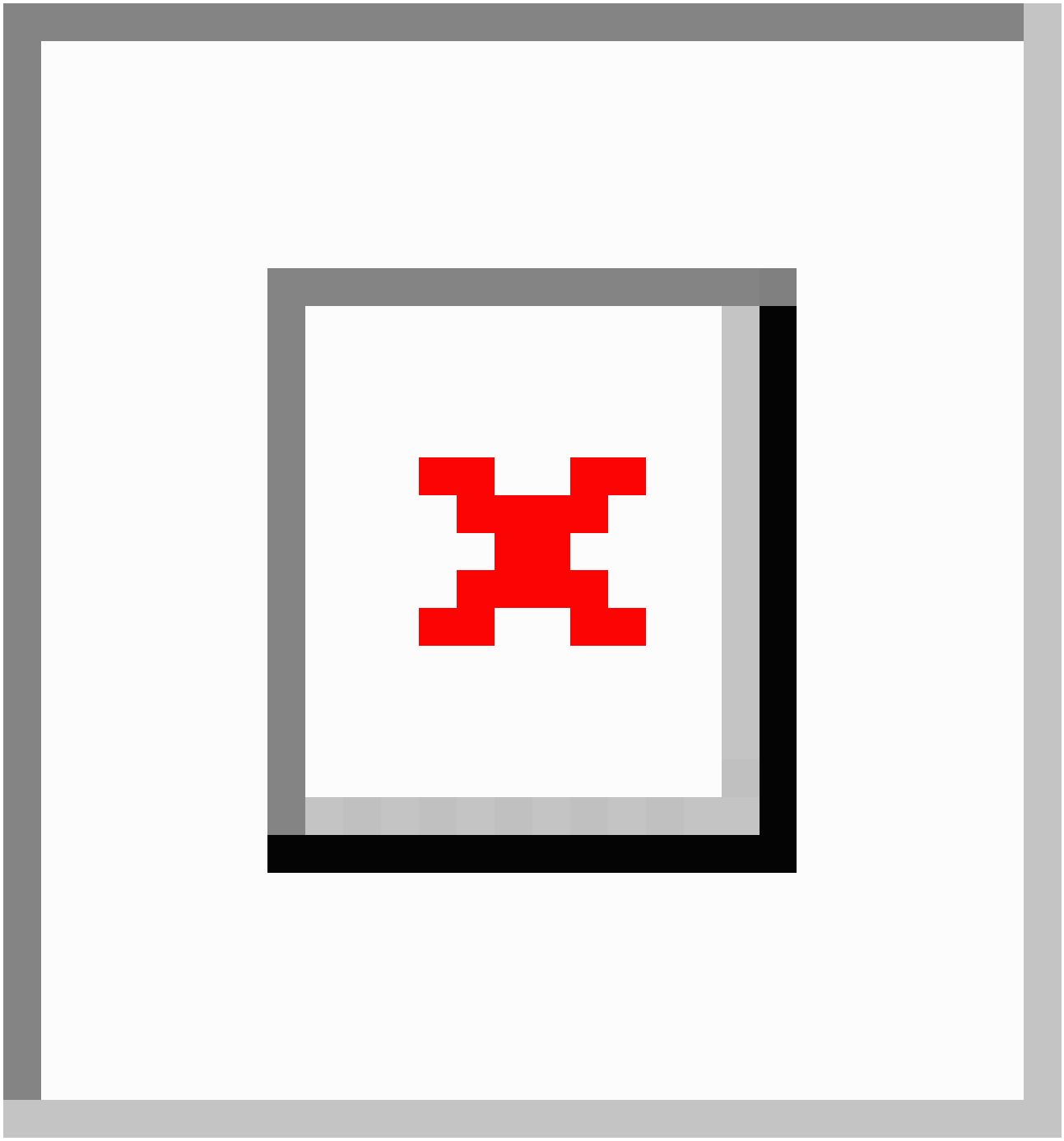
ISLR has 2 major views that target 2 types of users in the observational studies in an SLR lifecycle: (1) an intelligent SLR workbench for literature reviewers who conduct routine literature reviews, and (2) a living literature data dashboard for researchers and analysts who focus on analyzing SLR data and keep up to date on new evidence. Figure 1 shows the overview architecture, including the 2 major views and data flow of the SLR system. ISLR integrates AI technologies and an SLR workflow management system to support literature collection, screening, and data extraction. The living literature dashboard continuously searches and updates the SLR, allowing users to interactively navigate the updated literature and develop new insights.

Figure 1. Overall data flow architecture of ISLR demonstrating the 2 major views. AI: artificial intelligence; ISLR: intelligent systematic literature review; SLR: systematic literature review.



Reliable NLP systems depend heavily on the development of a reasonable workflow, user interfaces, and high-performance NLP algorithms. To develop the system and define the system workflow and user interfaces, we collaborated with end users who are experts in SLR using an iterative approach that employed industry-standard agile methodology. The team identified 6 major functional areas that were essential for the application: (1) protocol specification assistance, (2) literature search and indexing, (3) abstract screening with NLP assistance, (4) support for full-text searching, uploading, and screening, (5) full-text data element extraction using NLP assistance to identify and extract relevant data elements from full-text and embedded tables, and (6) literature data visualization to enable users to assess the SLR results and perform data discovery. Figure 2 shows the system workflow and the embedded NLP services to expedite two of the most time-consuming steps, which are article screening and data element extraction.

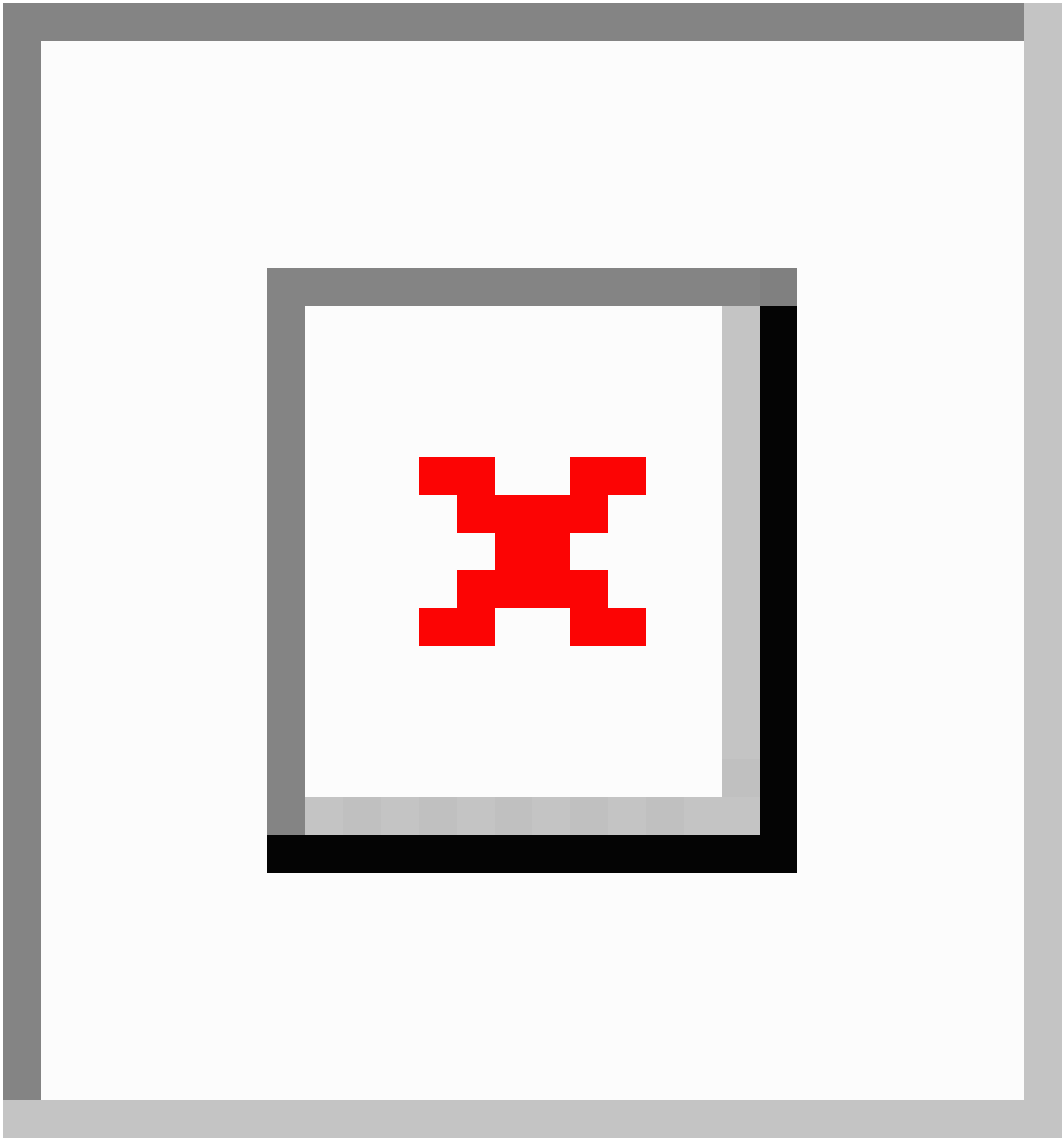
Figure 2. ISLR workflow and embedded NLP engines. ISLR: intelligent systematic literature review; NLP: natural language processing.



Development and Validation of NLP Algorithms

As mentioned earlier, 2 sets of NLP algorithms are required for a specific SLR project, including abstract screening and full-text data element extraction. Figure 3 outlines the NLP algorithm development process for these 2 steps separately. For abstract screening, the first step is to annotate and build a corpus that includes the abstract text, citation metadata, and inclusion/exclusion status. Once the corpus is prepared, NLP algorithm training, evaluation, and selection can be performed, and the best-performing algorithms will be chosen for deployment.

Figure 3. SLR NLP algorithm development steps. NLP: natural language processing; SLR: systematic literature review.



Similar to abstract screening, the NLP algorithm for the full-text data element extraction also requires a complete NLP development lifecycle. Unlike abstract screening, where labeled corpora may be available from previous SLR projects, data annotation is required to curate a labeled data set for training and evaluating NLP algorithms. The best-performing algorithms will be selected for deployment after evaluation. The following figure describes details on NLP algorithm development and validation process for SLR projects.

Three previously completed SLRs were used to guide and validate NLP development. These 3 projects included: (1) the prevalence of human papillomavirus (HPV) detected in head and neck squamous cell carcinomas (referred to as *HPV Prevalence*); (2) the epidemiology of the pneumococcal disease (referred to as *Pneumococcal Epidemiology*), and (3) the economic burden of pneumococcal disease (referred to as *Pneumococcal Economic Burden*). The inclusion and exclusion criteria for these 3 SLRs can be found in Table S1 in Multimedia Appendix 1.

Developing the Abstract Screening Corpora

Abstract screening was treated as a binary document classification task, ie, inclusion or exclusion of the article based on the abstract. Consequently, it was necessary to select and train NLP models for the task that demonstrated adequate performance and that had a reasonable computational time. The annotated screening literature sets from the 3 previous SLRs were used as the gold standard to train and validate models,

including 1697, 207, and 421 articles for *HPV Epidemiology*, *Pneumococcal Epidemiology*, and *Pneumococcal Economic Burden*, respectively. The corpora contained citation metadata, including title, authors, Medical Subject Heading terms [9], and the text of the corresponding abstracts.

Developing the Full-Text Data Element Extraction Corpora

We selected 190, 25, and 24 full-text articles for *HPV Prevalence*, *Pneumococcal Epidemiology*, and *Pneumococcal Economic Burden* for annotation, respectively. Based on the key outcome variables defined in the 3 SLRs, we annotated 12 types of data elements, covering information related to general observational studies, such as *Study Population*, to disease-specific information such as *HPV Lab Technique* and *Pneumococcal Disease Type*.

Abstract Screening NLP Algorithms

For abstract screening, the NLP model classifies each article for its relevance based on its title, abstract, and other citation meta data. To build the abstract screening module, we evaluated 4 traditional ML-based document classification algorithms, XGBoost [10], support vector machines [11], logistic regression [12], and random forest [13] on the binary inclusion/exclusion classification task for abstract screening. The abstract screening corpora were used to evaluate NLP models by calculating standard metric of *precision (fraction of relevant instances among the retrieved instances, also called positive predictive value*), *recall (fraction of relevant instances that were retrieved, also called sensitivity*), *accuracy*, and *F1 scores* (the harmonic mean of precision and recall). The full features include title, abstract, authors, keywords, journal, Medical Subject Heading term, and publication types. We concatenated all features and extracted the term frequency-inverse document frequency vector as feature representation.

Data Element Extraction NLP Algorithms

To construct the module for data element extraction, we treated the problem of data element recognition and extraction as a named entity recognition (NER) problem, which aims to recognize the mentions of entities from the text [14]. We evaluated a series of NLP algorithms consisting of ML and DL algorithms to recognize and extract data elements from full text, including (1) conditional random fields (CRFs), a classic statistical sequence modeling algorithm that has been widely applied to NER tasks [15,16]; (2) long short-term memory (LSTM), a variation of recurrent neural networks that has achieved remarkable success in NER tasks [17,18]; and (3) "Clinical BERT (Bidirectional Encoder Representations from Transformers)" [19], a novel transformer-based DL model. Standard metrics, including *precision*, *recall*, *accuracy*, and *F1 scores*, were calculated.

Ethical Considerations

This is not applicable as this study is not human subjects research.

Results

Here, we report the results of the construction of the annotation corpora and the results of the NLP algorithm for abstract screening and data element extraction, respectively.

Abstract Screening Corpora Description

The *HPV Prevalence* corpus we constructed from the existing SLR project contained 1697 total citations, of which 538 were included, and 1159 were excluded due to study criteria. The constructed *Pneumococcal Epidemiology* contained 207 citations, of which 85 were included and 122 were excluded. The constructed *Pneumococcal Economic Burden* corpus contained 421 citations, of which 79 were included, and 342 were excluded.

Abstract Screening NLP Evaluation Results

Extensive studies have shown the superiority of transformer-based DL models for many NLP tasks [20-23]. Based on our experiments, however, adding features to the pretrained language models did not significantly boost their performance. The performance comparison results for each task are shown in Table 1. XGBoost achieved the highest accuracy on *HPV Prevalence* and *Pneumococcal Economic Burden* tasks, while a support vector machine achieved the highest accuracy on *Pneumococcal Epidemiology* task. XGBoost was ultimately chosen for deployment due to its better generalizability.

Table . Comparison of article screening natural language processing model performance.

Task and algorithm	F1 score	Precision	Recall	Accuracy
HPV Prevalence **(n=1697)**				
XGBoost	0.808	0.769	0.851	0.888
Support vector machine	0.727	0.781	0.681	0.859
Logistics regression	0.684	0.897	0.553	0.859
Random forest	0.523	0.944	0.362	0.818
Pneumococcal Economic Burden **(n=421)**				
XGBoost	0.750	0.857	0.667	0.907
Support vector machine	0.533	0.667	0.444	0.667
Logistics regression	0.333	0.667	0.222	0.831
Random forest	0.429	0.600	0.333	0.814
Pneumococcal Epidemiology **(n=207)**				
XGBoost	0.667	0.533	0.889	0.619
Support vector machine	0.667	0.667	0.667	0.861
Logistics regression	0.429	0.600	0.333	0.619
Random forest	0.615	1.000	0.444	0.762

Full-Text Data Element Extraction Corpora Description

The human annotators annotated 190, 25, and 24 full-text articles for the *HPV Prevalence*, *Pneumococcal Epidemiology*, and *Pneumococcal Economic Burden* tasks, respectively. Among these full-text articles, 4498, 579, and 252 entity mentions were annotated for 3 projects, respectively. However, the distribution of annotated entities is highly imbalanced. For example, data elements like *HPV Lab Technique* and *HPV Sample Type* were very prevalent, but data elements like *Maximum/Minimum Age in Study Cohort* were rarely annotated in the corpora.

Results of the Full-Text Screening and Data Element Extraction NLP Methods

Tables 2 and 3 show the comparison of NLP performance among CRFs, LSTM, and BERT on the 3 corpora. For each of the 3 corpora used to train the NLP models, LSTM demonstrated superiority over the conventional ML algorithm (ie, CRF) on entity recognition. Among DL models, we did not observe significant improvement in F1 scores by use of the BERT model. The BERT model achieved similar or worse performance on most data elements. The performance across different tasks varies, primarily due to the availability of annotated data. For example, on average, models' performance on *HPV Prevalence* is higher than *Pneumococcal Epidemiology* and *Pneumococcal Economic Burden*, as *HPV Prevalence* has the largest annotated data. Due to the highly imbalanced distribution of annotated entities, we observe a variation in performance across different data elements for the same model. For example, in the *Pneumococcal Epidemiology* task, the LSTM model has achieved 0.412 in the identification of the *Study Cohort* and 0.768 in the identification of the *Pneumococcal Disease Type*.

Table . Overall performance comparison for the named entity recognition task in the 3 natural language processing training corpora. Scores averaged across all 12 extracted data elements. Measured in lenient F1 score.

Measure	*HPV Prevalence*			*Pneumococcal Epidemiology*			*Pneumococcal Economic Burden*		
	CRF[a]	LSTM[b]	Clinical BERT[c]	CRF	LSTM	Clinical BERT	CRF	LSTM	Clinical BERT
Microaverage (global average that uses the total number of true positives, false positives, and false negatives)	0.856	0.890	0.782	0.571	0.646	0.444	0.609	0.615	0.478
Macroaverage score (arithmetic mean of all the per-entity type scores)	0.522	0.674	0.685	0.270	0.295	0.227	0.216	0.238	0.231

[a]CRF: conditional random field.

[b]LSTM: long short-term memory.

[c]BERT: Bidirectional Encoder Representations from Transformers.

Table . Performance comparison for the named entity recognition task on selected data elements. Measured in lenient F1 score.

Measure	*HPV Prevalence*			*Pneumococcal Epidemiology*			*Pneumococcal Economic Burden*		
	CRF[a]	LSTM[b]	Clinical BERT[c]	CRF	LSTM	Clinical BERT	CRF	LSTM	Clinical BERT
Study Cohort	0.482	0.695	0.727	—[d]	0.412	0.278	—	—	—
Study Location	0.434	0.520	0.574	0.514	0.508	0.546	0.586	0.484	0.497
Study Type	0.733	0.760	0.753	0.364	0.525	—	—	0.328	0.299
Pneumococcal Disease Type	—	—	—	0.725	0.768	0.526	0.644	0.715	0.523
Incidence or Prevalence	0.986	0.983	0.924	—	—	—	—	—	—
Study Time	0.714	0.888	0.930	0.222	0.636	0.328	—	—	—

[a]CRF: conditional random field.

[b]LSTM: long short-term memory.

[c]BERT: Bidirectional Encoder Representations from Transformers.

[d]Not applicable.

Final NLP Algorithm Selection

NLP algorithms were needed for the 2 tasks, abstract screening, and data element extraction, in the ISLR system. The abstract screening was treated as a classification task. Based on our experimental results, XGBoost was selected for this task due to good performance on our document classification experiments and less computational complexity than DL-based models. For the data element extraction task, LSTM was selected over CRF and BERT for the same reasons.

ISLR System Components

Study Protocol Specification

Study protocol specification is one of the first steps in an SLR project. Users can upload a PDF document to the system that describes the SLR study protocol for reference. The SLR system has a default list of data elements with their descriptions and answer types (eg, free text, multiple choice, and checkbox), which will be extracted from full-text PDFs of articles. The system also allows users to create and modify the list. At the

end of the project, all the extracted data elements can be exported in a structured format.

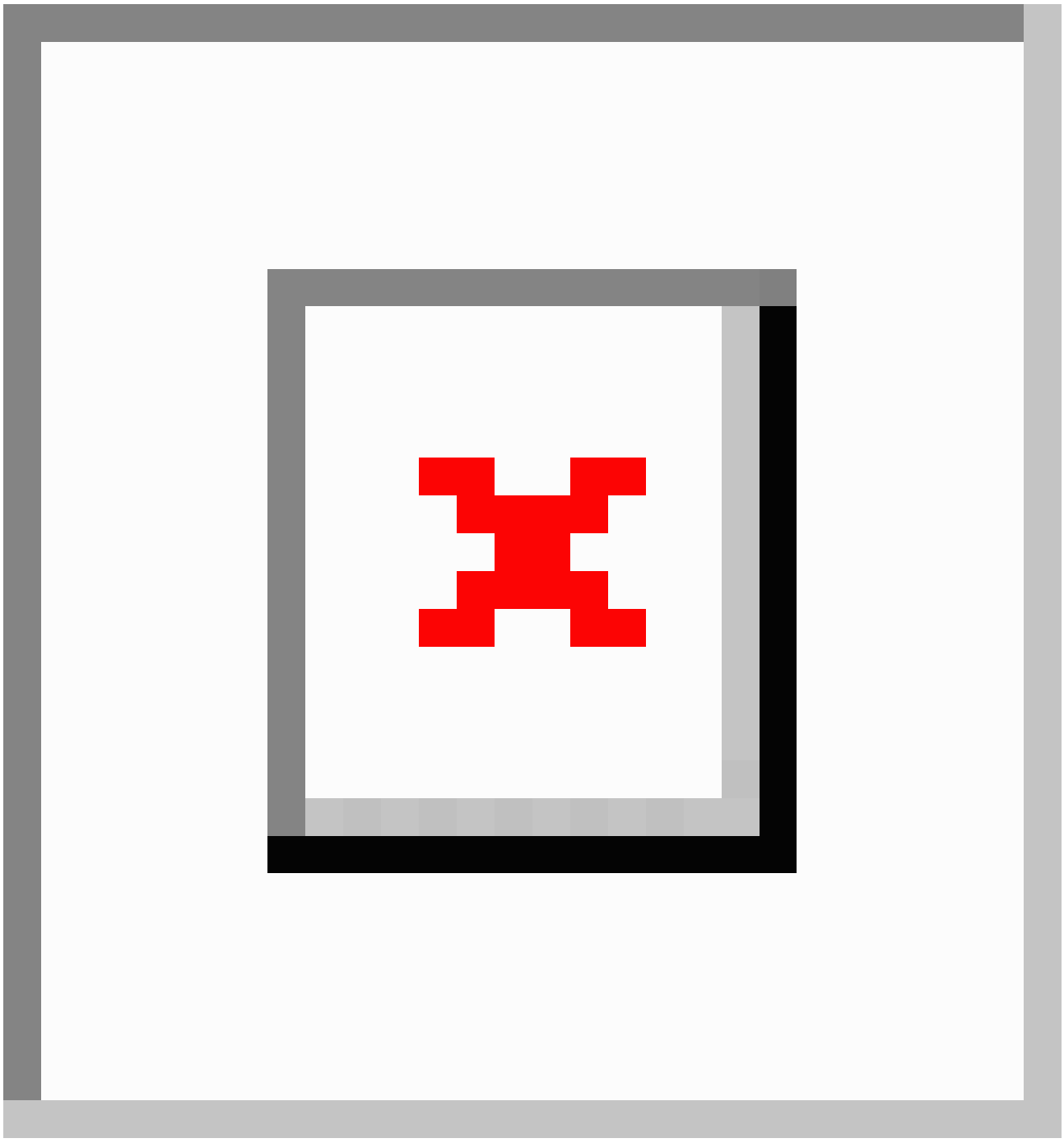
Literature Search

The ISLR system is integrated with the PubMed E-utilities application programming interface, which enables users to perform direct searches on PubMed. Citation metadata such as abstracts, titles, journals, and authors can be retrieved from PubMed and indexed in the system for further screening and data element extraction. Additionally, the system provides an option for users to retrieve this citation metadata by uploading a list of individual PubMed IDs.

Abstract Screening

The purpose of abstract screening is to review collected articles' relevance based on their title, abstract, and other relevant metadata, such as journal names, article types, and keywords. The relevant articles will be included for the following full-text screening and data element extraction steps. NLP services are provided at this step to make recommendations on whether a particular article should be included for full-text review. The supporting information (eg, salient words that are impactful to inclusion and exclusion) for the NLP recommendation will also be shown to provide explainable evidence. Human experts can further review the predictions for each article and decide on abstract screening status (keep or exclude). Figure 4 shows the abstract screening interface demonstrating prediction results and relevant terms discovered by the NLP algorithms.

Figure 4. Abstract screening interface. Terms that support inclusion in the finalized cohort of relevant articles are shown in green, while terms that detract from inclusion are shown in red. The scale of the colors shows how significantly one term can impact prediction decisions (eg, darker color indicates higher impact).



Full-Text Searching, Uploading, and Screening

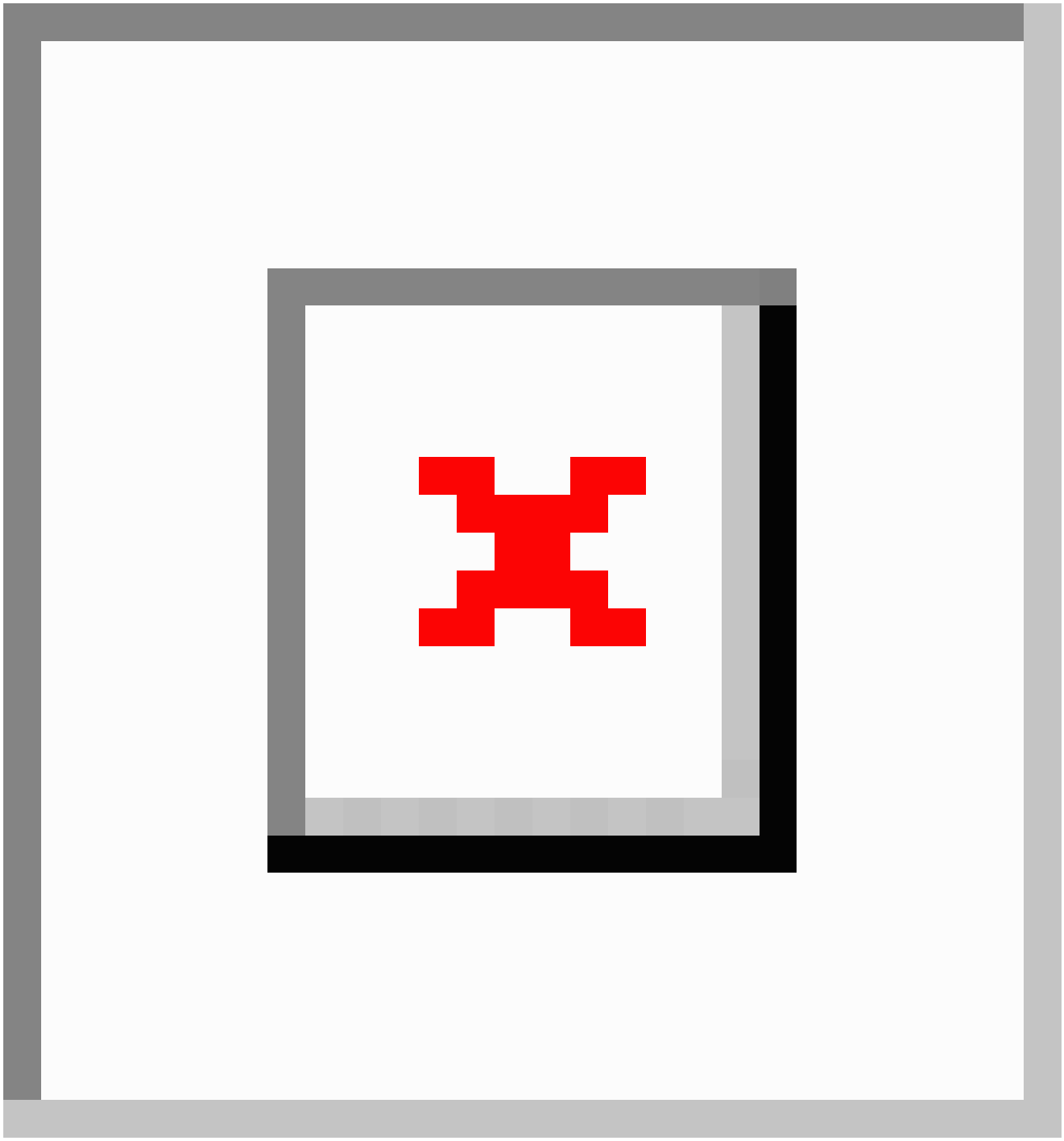
This step aims to identify full-text PDF documents for each included article and further screen their relevance based on the SLR study protocol. Only the articles that are deemed relevant after this stage will be included in the final full-text data element extraction step. The process of locating full-text PDF documents for each article can be time-consuming. The ISLR system integrates with PubMed Central to automatically find and collect full-text PDFs if they are publicly available. However, for articles whose full-text PDFs are not publicly available, users need to manually locate the articles through publishers and upload the corresponding PDFs to the system through the provided user interface.

Full-Text Data Element Extraction

Extracting full-text data elements is a time-consuming process in SLR projects. It requires reviewing the full-text article and extracting multiple relevant pieces of information defined in the study protocol. These data elements are often found in various sections of an article, including tables. The ISLR system uses Amazon Textract [24] for optical character recognition to extract text and tables from PDF files, followed by NLP services to further extract information from both text and tables. The NLP services can recommend potential answers for each data

element, and human experts can review, select, and modify the extracted information. Figure 5 shows a screenshot of the user interface for this step.

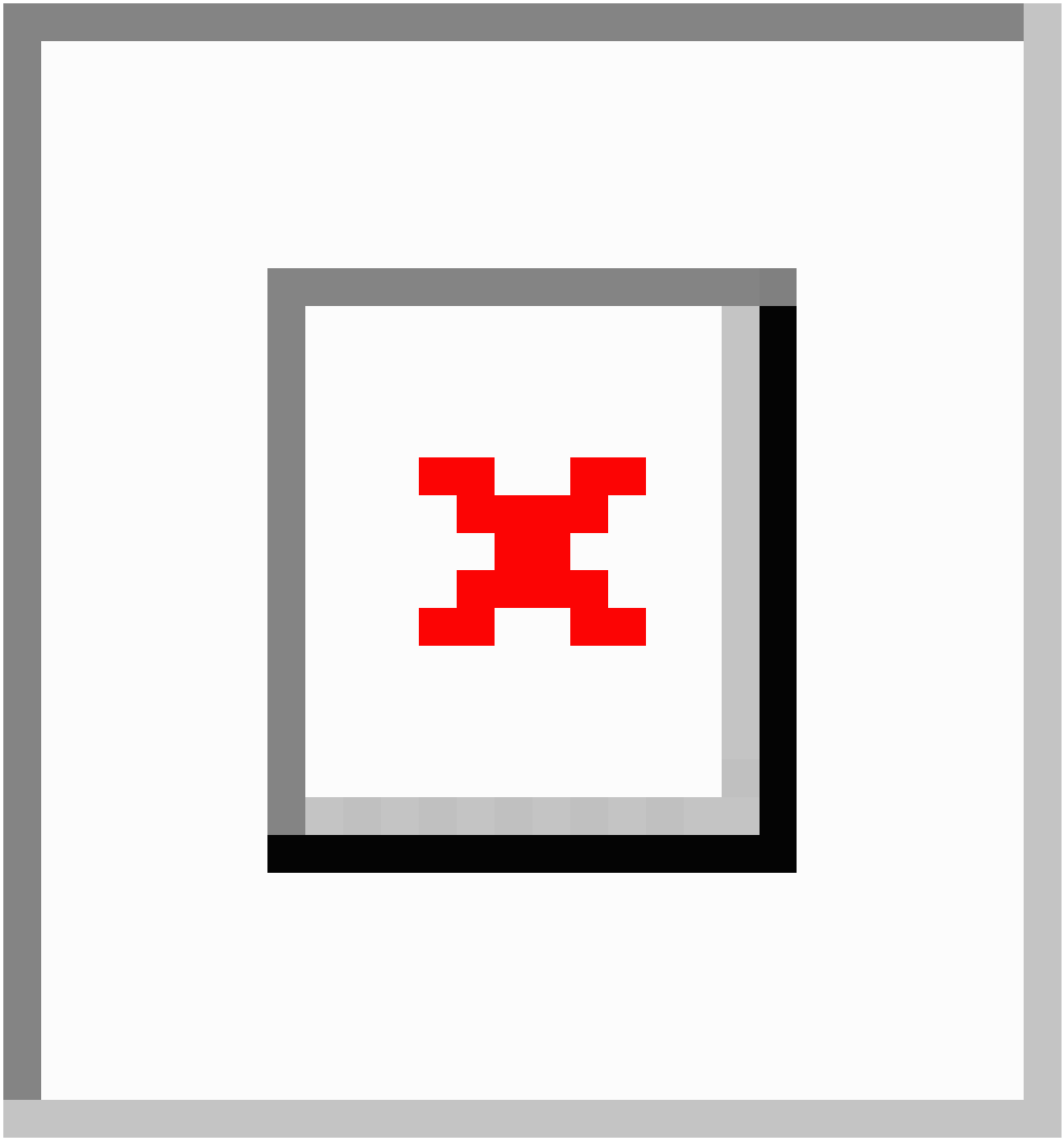
Figure 5. Full-text data element extraction user interface. Data elements from the article extracted by the NLP algorithms are color-coded and highlighted in the PDF. Highlight colors in the PDF text are linked to the data elements as shown in the right-hand frame. For the data element list on the right side, all the extracted data elements can pop up as candidates for the users to choose from. NLP: natural language processing.



Data Summary and Visualization

The ISLR system offers interactive dashboards to end users, such as researchers, for exploring the SLR results and data. These dashboards allow users to apply data filters, such as study location and cohort size, to refine their search results. For each data element extracted from full-text articles, users can click on the element to navigate to the corresponding article, ensuring traceability and appropriate references to source documents in the SLR project. Additionally, the dashboards recommend recent relevant articles and suggest articles that may require full-text screening. Figure 6 displays the major functions and screenshots of the dashboard.

Figure 6. Interactive visualization of existing SLR data, lists of relevant publications, and data exportation control. SLR: systematic literature review.



Discussion

Principal Findings

As described in the introduction, conducting an SLR is complex and expensive. There is also a rapid growth of the available number of publications and other data, such as clinical trial reports used in the article search and screening processes, with an average annual growth rate for the life sciences of around 5% [25]. Consequently, there is considerable community interest in applying various types of automation, including AI, DL, and NLP, to the multiple tasks required for producing an accurate SLR [2,5-7].

An important consideration for using the results of an SLR is how often the SLR is updated and hence how timely and complete these data are with respect to the real-world evidence. “Living” ISLR system addresses the difficulty of updating an SLR by providing an automated workflow including review tools to detect when new data are available and to trigger at least a semi-automated update process for the expedited review. The system is also expandable to cover additional data elements of interest by updating existing NLP pipelines.

The major accomplishments of this ISLR system include improving the time, efficiency, cost, completeness of evidence, and error avoidance through techniques to assist researchers with decision-making (so-called human-in-the-loop). The ISLR system is aligned with the living SLR concept, as it supports a

rapid update of existing literature data. Additionally, since the classification and data element extraction tasks are maintained by the system, results can be used for retraining the classification and NLP algorithms on a routine basis. Consequently, the performance of the system should improve over time.

The focus of this work was to evaluate an intelligent system that includes all major steps of an SLR with humans in the loop. The corpora evaluated in this study mostly focus on health economics and outcomes research in specific therapeutical areas. The generalizability of the learning algorithms to another domain will benefit from further formal examination. Since we have not yet conducted a time analysis of an SLR study conducted both manually and with this tool, we are unable to precisely quantify the time savings from the ISLR system. In addition, our NLP technologies limit to the extraction of relevant information directly from the text but are not able to conduct reasoning with long context to support complex data element extraction, such as GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) or RoB2 (Risk of Bias 2). The recent advances in large language models, such as generative pretrained transformer 4, bring NLP technologies expert-level performance on various professional and academic benchmarks. Given its high performance, generalizability, and reasoning capacity, it would be interesting to further assess the efficacy and accuracy of large language models in various SLR tasks and complex data element extraction.

As an early and innovative attempt to automate SLR lifestyle through NLP technologies, ISLR does not fully support PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting yet. We plan to continuously iterate ISLR to cover the PRISMA checklist and report generation in the future. In addition, we have not yet conducted formal usability studies of the user interface, although agile methods involving iterative refinement of the interface through input from domain experts in SLR were employed throughout the software development process.

Conclusions

Our ISLR system is a user-centered, end-to-end intelligent solution to automate and accelerate the SLR process and supports “living” SLRs with humans in the loop. The system integrates cutting-edge ML- and DL-based NLP algorithms to make recommendations on article screening and data element extraction, which allow the system to prospectively and continuously update relevant literature in a timely fashion. This allows scientists to have more time to focus on the quality of data and the synthesis of evidence and to stay current with literature related to observational studies.

Acknowledgments

This research was supported by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, NJ.

Disclaimer

The content is the sole responsibility of the authors and does not necessarily represent the official views of Merck & Co, Inc, Rahway, NJ, or Melax Tech.

Data Availability

The annotated corpora underlying this article are available on GitHub [25].

Authors' Contributions

Study concept and design: JD and LY. Corpus preparation: DW, JD, and LY. Experiments: JD and BL. Draft of the manuscript: FJM, JD, DW, NC, and LY. Acquisition, analysis, or interpretation of data: JD, DW, NC, and LY. Critical revision of the manuscript for important intellectual content: all authors. Study supervision: JD, LY, and NC.

Conflicts of Interest

DW, JC, DE, NC, PCF, and LY are employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA. JD, BL, SW, XW, LH, JW, and FJM are employees of IMO.

Multimedia Appendix 1
Inclusion and exclusion criteria for 3 systematic literature review projects.
[DOCX File, 24 KB - medinform_v12i1e54653_app1.docx]

References

1. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Methodol 2018 Jan 10;18(1):5. [doi: 10.1186/s12874-017-0468-4] [Medline: 29316881]
2. Tsafnat G, Glasziou P, Choong MK. Systematic review automation technologies. Syst Rev 2014;3(74) [FREE Full text] [doi: 10.1186/2046-4053-3-74]

3. Higgins J, Thomas J, editors. Cochrane Handbook for Systematic Reviews of Interventions, Version 65 2024. URL: https://training.cochrane.org/handbook/current [accessed 2024-10-17]
4. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun 2019 Dec;16:100443. [doi: 10.1016/j.conctc.2019.100443] [Medline: 31497675]
5. Michelson M, Ross M, Minton S. AI2 leveraging machine-assistance to replicate a systematic review. V H 2019 May;22:S34. [doi: 10.1016/j.jval.2019.04.006]
6. Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes RB. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study. J Med Internet Res 2018 Jun 25;20(6):e10281. [doi: 10.2196/10281] [Medline: 29941415]
7. Elliott JH, Turner T, Clavisi O, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med 2014 Feb;11(2):e1001603. [doi: 10.1371/journal.pmed.1001603] [Medline: 24558353]
8. Rayyan - Intelligent systematic review. Rayyan. 2021. URL: https://www.rayyan.ai/ [accessed 2024-04-23]
9. Medical Subject Headings. National Library of Medicine. 2024. URL: https://www.nlm.nih.gov/mesh/meshhome.html [accessed 2022-05-30]
10. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Presented at: KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
11. Noble WS. What is a support vector machine? Nat Biotechnol 2006 Dec;24(12):1565-1567. [doi: 10.1038/nbt1206-1565] [Medline: 17160063]
12. Kleinbaum DG, Klein M. Logistic Regression: A Self-Learning Text: Springer; 2010. URL: https://link.springer.com/book/10.1007/978-1-4419-1742-3 [accessed 2022-05-30]
13. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens 2005;26(1):217-222. [doi: 10.1080/01431160412331269698]
14. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvist Investig 2007 Aug 15;30(1):3-26. [doi: 10.1075/li.30.1.03nad]
15. Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. 2001 Presented at: CML '01: Proceedings of the Eighteenth International Conference on Machine Learning; Jun 28 to Jul 1, 2001; San Francisco, CA p. 282-289 URL: http://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf
16. Lin S, Ng JP, Pradhan S, et al. Extracting formulaic and free text clinical research articles metadata using conditional random fields. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents: Association for Computational Linguistics; 2010:90-95 URL: https://aclanthology.org/W10-1114 [accessed 2022-08-07]
17. Chiu JPC, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. arXiv. Preprint posted online on Nov 26, 2015 URL: https://arxiv.org/abs/1511.08308 [accessed 2024-10-17]
18. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. arXiv. Preprint posted online on Mar 4, 2016 URL: https://arxiv.org/abs/1603.01360 [accessed 2024-10-17]
19. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv. Preprint posted online on Apr 6, 2019 URL: https://arxiv.org/abs/1904.03323 [accessed 2024-10-17] [doi: 10.18653/v1/W19-1909]
20. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2019 URL: https://arxiv.org/abs/1810.04805 [accessed 2024-10-17]
21. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
22. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2022 Jan 31;3(1):1-23. [doi: 10.1145/3458754]
23. Chen Q, Du J, Allot A, et al. LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. arXiv. Preprint posted online on Apr 19, 2022 URL: https://arxiv.org/abs/2204.08649 [accessed 2024-10-17]
24. Amazon Textract. Amazon Web Services. URL: https://aws.amazon.com/textract/ [accessed 2022-08-08]
25. Merck/NLP-SLR-corpora. GitHub. URL: https://github.com/Merck/NLP-SLR-corpora [accessed 2024-10-17]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
CRF: conditional random field
DL: deep learning
GRADE: Grading of Recommendations, Assessment, Development, and Evaluation
HPV: human papillomavirus

ISLR: intelligent systematic literature review
LSTM: long short-term memory
ML: machine learning
NER: named entity recognition
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RoB2: Risk of Bias 2
SLR: systematic literature review

Edited by C Perrin; submitted 17.11.23; peer-reviewed by S Matsuda, S Zhou; revised version received 24.04.24; accepted 23.07.24; published 23.10.24.

Please cite as:
Manion FJ, Du J, Wang D, He L, Lin B, Wang J, Wang S, Eckels D, Cervenka J, Fiduccia PC, Cossrow N, Yao L
Accelerating Evidence Synthesis in Observational Studies: Development of a Living Natural Language Processing–Assisted Intelligent Systematic Literature Review System
JMIR Med Inform 2024;12:e54653
URL: https://medinform.jmir.org/2024/1/e54653
doi:10.2196/54653

Leveraging Artificial Intelligence and Data Science for Integration of Social Determinants of Health in Emergency Medicine: Scoping Review

Ethan E Abbott[1,2,3,4,*], MSCR, DO; Donald Apakama[1,2,3,4,*], MS, MD; Lynne D Richardson[1,2,3], MD; Lili Chan[4,5,6], MD; Girish N Nadkarni[3,4,5,6,7], MPH, MD

1

2

3

4

5

6

7

[*]these authors contributed equally

Corresponding Author:
Ethan E Abbott, MSCR, DO

Abstract

Background: Social determinants of health (SDOH) are critical drivers of health disparities and patient outcomes. However, accessing and collecting patient-level SDOH data can be operationally challenging in the emergency department (ED) clinical setting, requiring innovative approaches.

Objective: This scoping review examines the potential of AI and data science for modeling, extraction, and incorporation of SDOH data specifically within EDs, further identifying areas for advancement and investigation.

Methods: We conducted a standardized search for studies published between 2015 and 2022, across Medline (Ovid), Embase (Ovid), CINAHL, Web of Science, and ERIC databases. We focused on identifying studies using AI or data science related to SDOH within emergency care contexts or conditions. Two specialized reviewers in emergency medicine (EM) and clinical informatics independently assessed each article, resolving discrepancies through iterative reviews and discussion. We then extracted data covering study details, methodologies, patient demographics, care settings, and principal outcomes.

Results: Of the 1047 studies screened, 26 met the inclusion criteria. Notably, 9 out of 26 (35%) studies were solely concentrated on ED patients. Conditions studied spanned broad EM complaints and included sepsis, acute myocardial infarction, and asthma. The majority of studies (n=16) explored multiple SDOH domains, with homelessness/housing insecurity and neighborhood/built environment predominating. Machine learning (ML) techniques were used in 23 of 26 studies, with natural language processing (NLP) being the most commonly used approach (n=11). Rule-based NLP (n=5), deep learning (n=2), and pattern matching (n=4) were the most commonly used NLP techniques. NLP models in the reviewed studies displayed significant predictive performance with outcomes, with F1-scores ranging between 0.40 and 0.75 and specificities nearing 95.9%.

Conclusions: Although in its infancy, the convergence of AI and data science techniques, especially ML and NLP, with SDOH in EM offers transformative possibilities for better usage and integration of social data into clinical care and research. With a significant focus on the ED and notable NLP model performance, there is an imperative to standardize SDOH data collection, refine algorithms for diverse patient groups, and champion interdisciplinary synergies. These efforts aim to harness SDOH data optimally, enhancing patient care and mitigating health disparities. Our research underscores the vital need for continued investigation in this domain.

(JMIR Med Inform 2024;12:e57124) doi:10.2196/57124

KEYWORDS

data science; social determinants of health; natural language processing; artificial intelligence; NLP; machine learning; review methods; review methodology; scoping review; emergency medicine; PRISMA

Introduction

Medical care, while crucial, contributes to only about 20% of the modifiable factors influencing a population's health outcomes, while 80% are influenced by genetics, individual behaviors, and socioeconomic factors. The latter two form the social determinants of health (SDOH) [1] that operate at various levels. From macroeconomic policies of nations to public education and housing policies, these structural factors shape resource distribution and societal positions. Consequently, they influence living conditions, access to essential resources, and daily life circumstances, ultimately molding health and health disparities [2]. Every patient's health trajectory is influenced by SDOH, which can manifest positively (eg, high income, food security) or adversely [3]. The negative aspects can be categorized into social risks, conditions linked to poor health, and social needs, which are individual preferences for assistance [4]. These determinants, especially when adverse, can hinder optimal care and impact clinical outcomes [5]. While not the focus of this review, health-related social needs reflect individual level social needs, and are inextricably related to the impacts and conditions of SDOH.

Emergency medicine (EM) is a unique medical specialty: it can both identify and address adverse SDOH, making it a pivotal setting for intervention. The high prevalence of social needs among emergency department (ED) patients, especially those with low socioeconomic status, housing insecurity, or limited access to care, underscores the potential of ED-based SDOH interventions [6,7]. However, there are significant challenges; comprehensive social risk screenings in the ED are often impractical due to patient volume, acuity, and health system financial constraints. Relying solely on electronic health records (EHRs) is time-consuming and inconsistent. Furthermore, the scattered and unstructured nature of SDOH data in EHRs makes it difficult for ED physicians to identify patients with adverse SDOH [8-10]. However, advancements in techniques in the data sciences provide novel methods to address these issues.

Social informatics, a subfield of medical informatics, refers to the usage of information technology to better harmonize social and health data through improved capture and usage. It bridges the gap between the technical and social worlds, offering insights into how technology and societal factors interplay. The vast possibilities of social informatics lie in its potential to reshape how we understand, interpret, and act on social data in health care settings. By integrating social data with health data, it aims to enhance clinical care and overall health outcomes [11]. Techniques like natural language processing (NLP), artificial intelligence (AI), and machine learning (ML) are being harnessed to extract, use, and model SDOH data effectively [12-14]. AI represents an umbrella term that broadly encompasses computer systems that can achieve human-level performance on cognitive tasks [15], while NLP, a subfield of AI, is a field of computational linguistics that involves analyzing and understanding human language. This entails using a combination of statistical approaches, including ML, to extract structure and meaning from language [16]. Pertinent to this review, ML can also be used as an analytic technique for predictive modeling and better understanding patterns in data sets.

While existing literature has touched upon SDOH in the ED, a comprehensive review focusing on the application of data sciences in this context is lacking. This scoping review aims to map the current literature, pinpoint areas for future research, and highlight the transformative potential of integrating data sciences into EM SDOH research.

Methods

Data Sources and Literature Search Strategy

To capture the evolving role of AI and advancing data science techniques for patients seen in the ED, particularly pertaining to SDOH, we searched the literature from 2015 to 2022, a period marked by rapid advancements in AI applications in health care. We included articles from databases such as Medline (Ovid), Embase (Ovid), CINAHL, Web of Science, and ERIC, prioritizing research that melded data science with emergency care settings.

Our search encompassed terms related to SDOH, data science techniques such as ML algorithms, NLP, AI, and EM (see Multimedia Appendix 1 for search terms used). We chose to include EM patient populations to best understand this clinical context. While there is no uniform definition for data science, for this review we have used the definition proposed by Mike and Hazzan [17] that considers data science as a research method by "integration of research tools and methods taken from statistics and computer science that can be used to conduct research in various application domains, such as social science and digital humanities".

Article Selection

We focused our review on studies leveraging data science techniques to extract or model SDOH data in EM. Recognizing the paucity of EM-specific research using AI/ML algorithms, we also considered studies on emergency-related conditions that might be seen in other clinical non-ED settings. This included conditions such as opioid use disorder (OUD), HIV, and epilepsy. We intentionally excluded COVID-19 studies, given their unique characteristics and sheer volume of literature. This exclusion ensured a more focused review with current relevance. Two independent reviewers (DA and EA) assessed titles and abstracts for final inclusion. Any disagreements were resolved through joint discussions. We used Covidence (Covidence systematic review software; Veritas Health Innovation), a standardized systematic review software.

Data Extraction

We extracted data from the selected studies using a standardized form, ensuring uniformity. We captured study objectives, methods, clinical care setting, ML algorithms, modeling approaches, and specified outcomes (see Multimedia Appendix 2). We focused on broad data science techniques, including subfields of AI such as ML algorithms, NLP key SDOH domains, and overall clinical outcomes. While our focus remains descriptive, we abstained from a quality assessment, aligning with standard scoping review guidelines.

Results

Overall Study Characteristics

After screening 1047 studies, 26 met our final inclusion criteria (Figure 1 and Multimedia Appendix 1). We excluded a significant number of studies because they did not focus on EM conditions or complaints and did not use AI/ML techniques in the overall approach to the study question. Most studies were published after 2020 (Figure 2) and included patient populations focusing exclusively on the ED (n=9), pediatric patients (n=2), patients treated by emergency medical services (n=2), and US veterans (n=2), among other examples.

Figure 1. Preferred reporting for systematic reviews and meta-analysis flow chart (PRISMA) diagram: 2015 - 2022 search of Medline (Ovid), Embase (Ovid), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Web of Science, and Education Resource Information Center (ERIC).

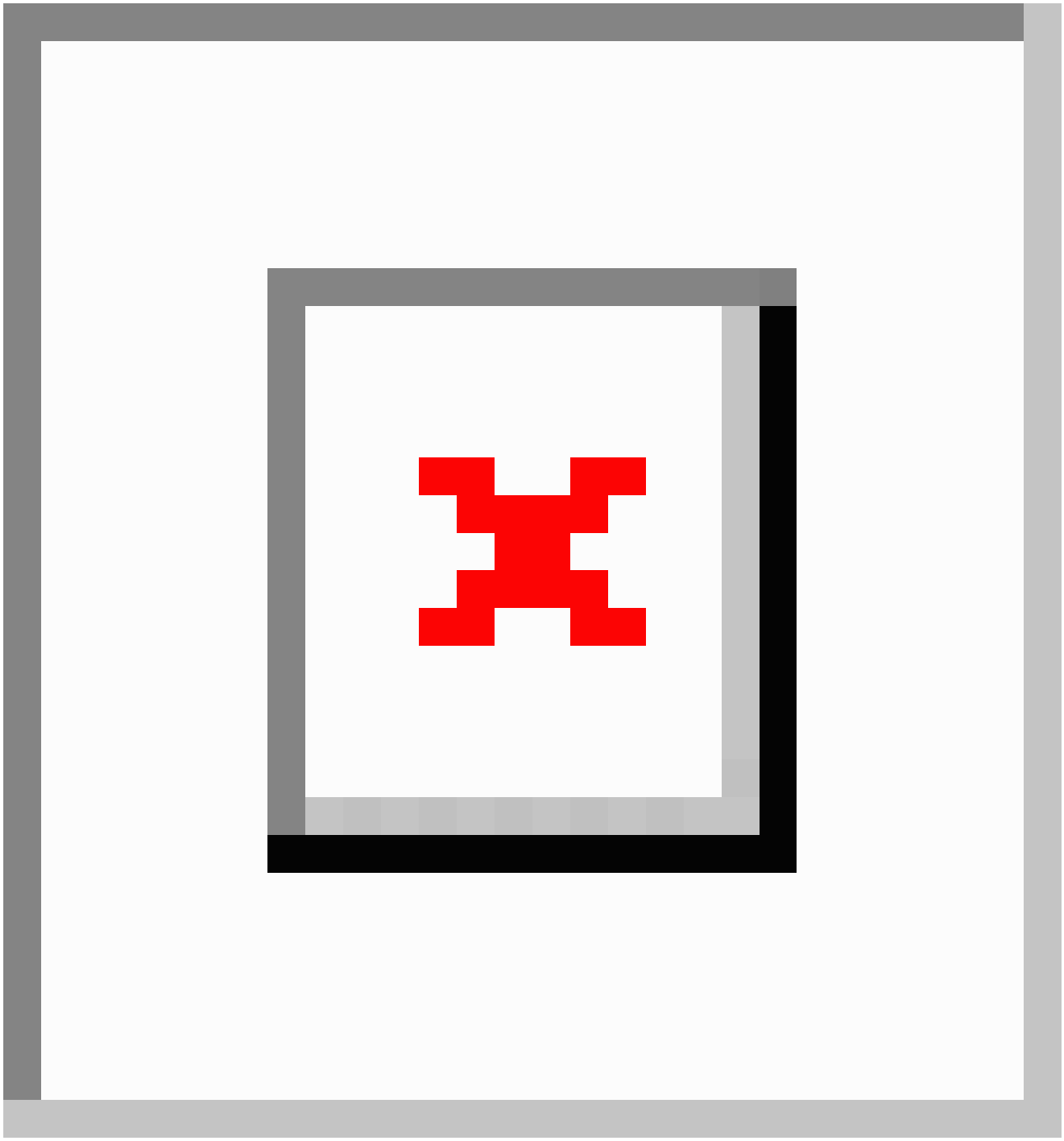
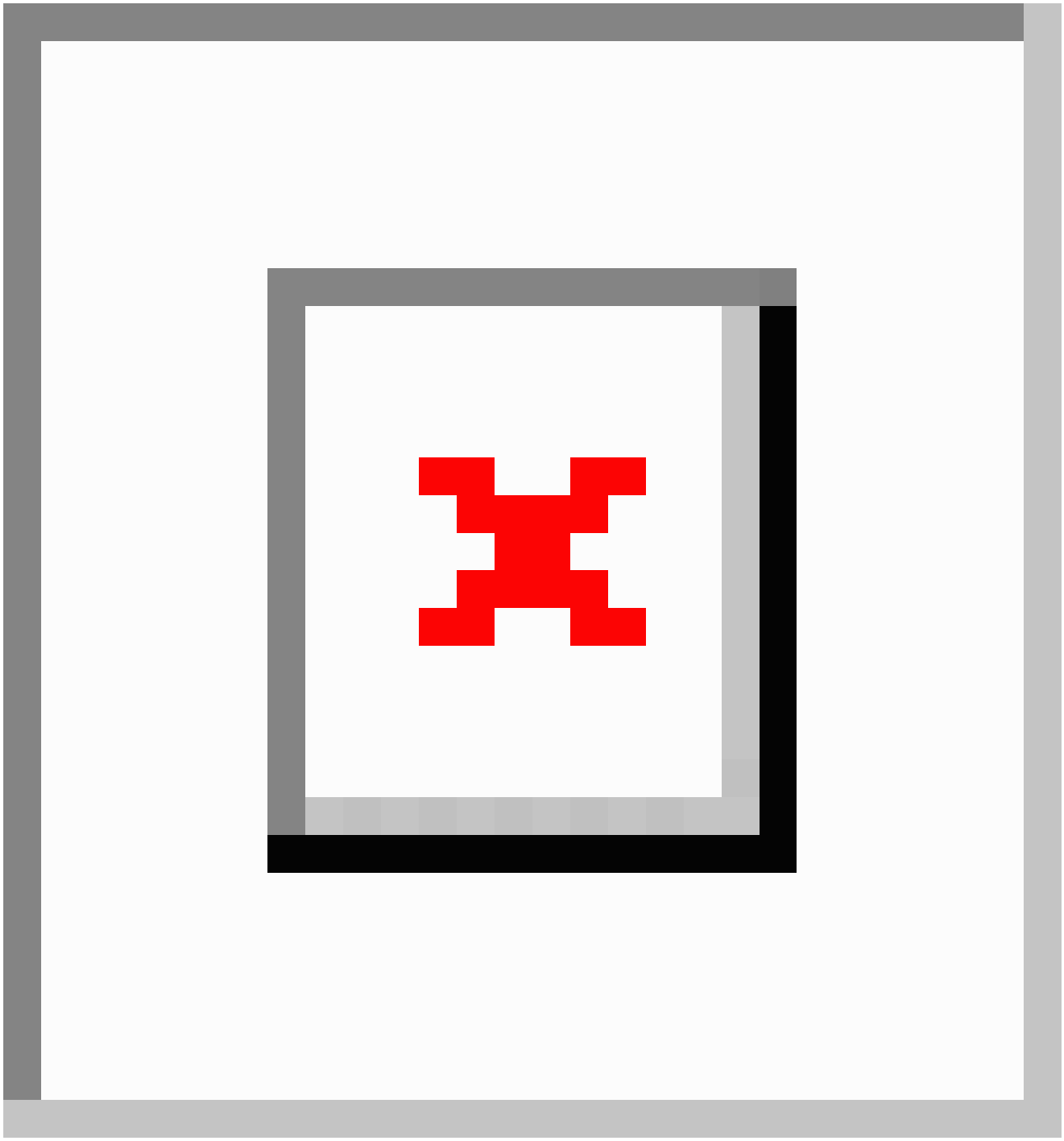


Figure 2. Number of publications by year (2015 - 2022) identified in this review. Word clouds show main themes of papers for the corresponding year.



SDOH Domains

Over 60% of studies we identified (n=16) used a broad array of data science techniques such as ML or NLP to use, model, or extract features across more than one of the major SDOH domains, resulting in significant overlap across publications included in this review (Figure 2). These domains included housing insecurity and homelessness, neighborhood and built environment, income and socioeconomic status, employment, family and social support, food insecurity, insurance status and stability, and history of incarceration. While individual level SDOH data was the prevalent unit of analysis, 5 studies used area level data or aggregated measures such as the Social Deprivation Index, Area Deprivation Index, or the Gini coefficient. Housing insecurity and homelessness emerged as the most predominant SDOH domains assessed among 23% of the studies identified (n=6). The domain of neighborhood and built environment was also present across multiple studies and the focus of several publications (n=4). Exposure to or history of incarceration (n=2) as well as OUD (n=2) were also notable.

Exploration of Emergency Medicine Conditions

The scope of EM clinical conditions and complaints that were studied were broad, including sepsis, acute myocardial infarction, heart failure, asthma, diabetes, chest pain, and epilepsy (Multimedia Appendix 3). Sepsis was the only specific EM condition we identified in more than one study (n=2). Several studies focused on all-cause ED revisits (n=2),

"preventable visits" and admissions (n=2), and ED utilization (n=2).

AI and ML Algorithms

AI techniques were used in 23 studies, encompassing methods like random forest, classification and regression trees, support vector machines, neural networks, and NLP (Figure 3 and Table 1). Of these, random forest emerged as the most common (n=13), closely followed by NLP (n=11). Key algorithms are discussed in further detail.

Figure 3. Overall counts of AI/ML algorithms used. AI: artificial intelligence; ML: machine learning; NLP: natural language processing; CART: classification and regression trees; LSTM: long short-term memory.

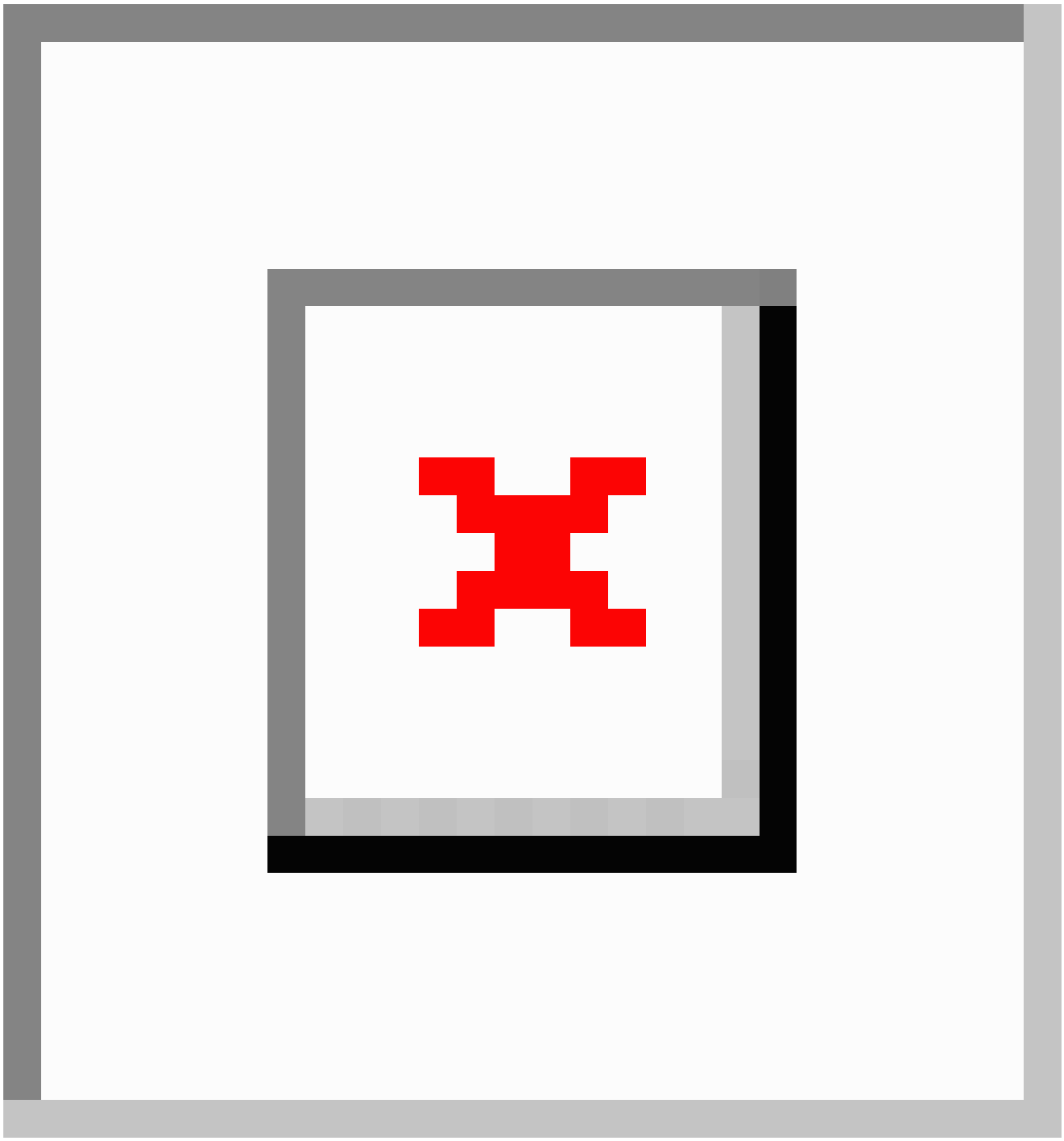


Table . Emergency medicine applications of random forest techniques used for extraction of social determinants of health data.

Category	Instances (n, %)
Identifying and highlighting pivotal SDOH[a] variables	2 (15.4)
Predictive modeling for potential health trajectories	1 (7.7)
Imputation using random forest techniques	3 (23.1)
Data integrity and robustness evaluation	5 (38.5)
Tree representations	2 (15.4)

[a]SDOH: Social Determinants Of Health

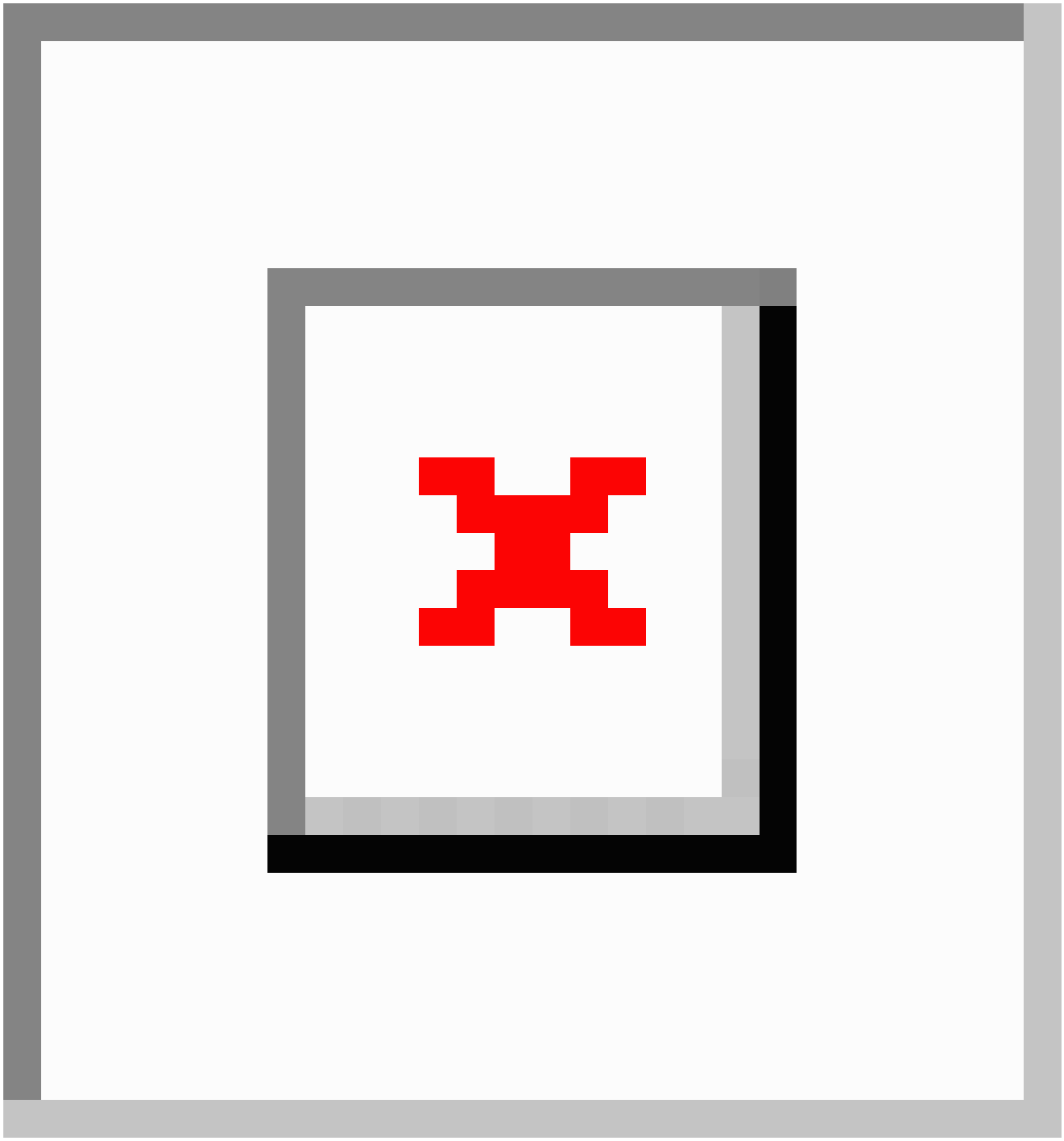
Random Forest in SDOH Variable Classification and Data Management

Random forest, an advanced ensemble ML method, was notably present across multiple studies identified in this review. This technique was used to discern and highlight pivotal SDOH variables. Its ability to create predictive models, offering foresight into potential health trajectories based on the subtleties of SDOH indicators was also evident. For example, in "Improving Fairness in the Prediction of Heart Failure Length of Stay and Mortality by Integrating Social Determinants of Health," the authors used random forest classifiers on sex, ethnoracial, and insurance for study subpopulations, finding differences in underdiagnosis and overdiagnosis of heart failure [18]. The most common use of random forest, however, was to address missing EHR data through imputation, ensuring the integrity and robustness of analyses. Beyond its analytical capabilities, it was also used to create insightful visual representations, offering a comprehensive view into the intricate web of variables, their interactions, and their overarching impact on disease states like acute coronary syndrome, epilepsy, asthma, and OUD.

NLP for SDOH Data Extraction

The integration of NLP in the realm of EM offers a promising avenue for the precise extraction of SDOH from EHRs. Traditional methodologies, such as manual reviews and rudimentary keyword searches, are increasingly recognized for their inherent limitations, particularly in the context of vast and intricate EHR data sets. Our scoping review elucidated the prevalence of several NLP techniques in the field. Text representation methods like term frequency-inverse document frequency (n=4), Bag of Words (n=3), and Word2Vec (n=1) were prominently used, underscoring their fundamental role in converting textual data into computationally amenable formats (Figure 4). In the realm of topic modeling and semantic analysis, latent Dirichlet allocation was noted in 2 studies, highlighting its potential in discerning latent topics within medical records. Approaches favored rule-based methodologies, found in 5 studies, followed by deep learning (n=2) particularly structures like bidirectional long short-term memory and pattern matching (n=4). From a software perspective, both proprietary (n=5) and open source (n=6) tools were harnessed, reflecting the diverse ecosystem of NLP tools available for research. These NLP methodologies, especially the dominant ones like Bag of Words, term frequency-inverse document frequency, and deep learning structures, have demonstrated notable efficacy. However, to achieve the pinnacle of precision in SDOH data extraction, it is imperative to continually refine these NLP techniques. Collaborative endeavors involving domain experts, iterative model training, and the assimilation of multifaceted data sources are paramount to enhancing the accuracy and relevance of extracted insights.

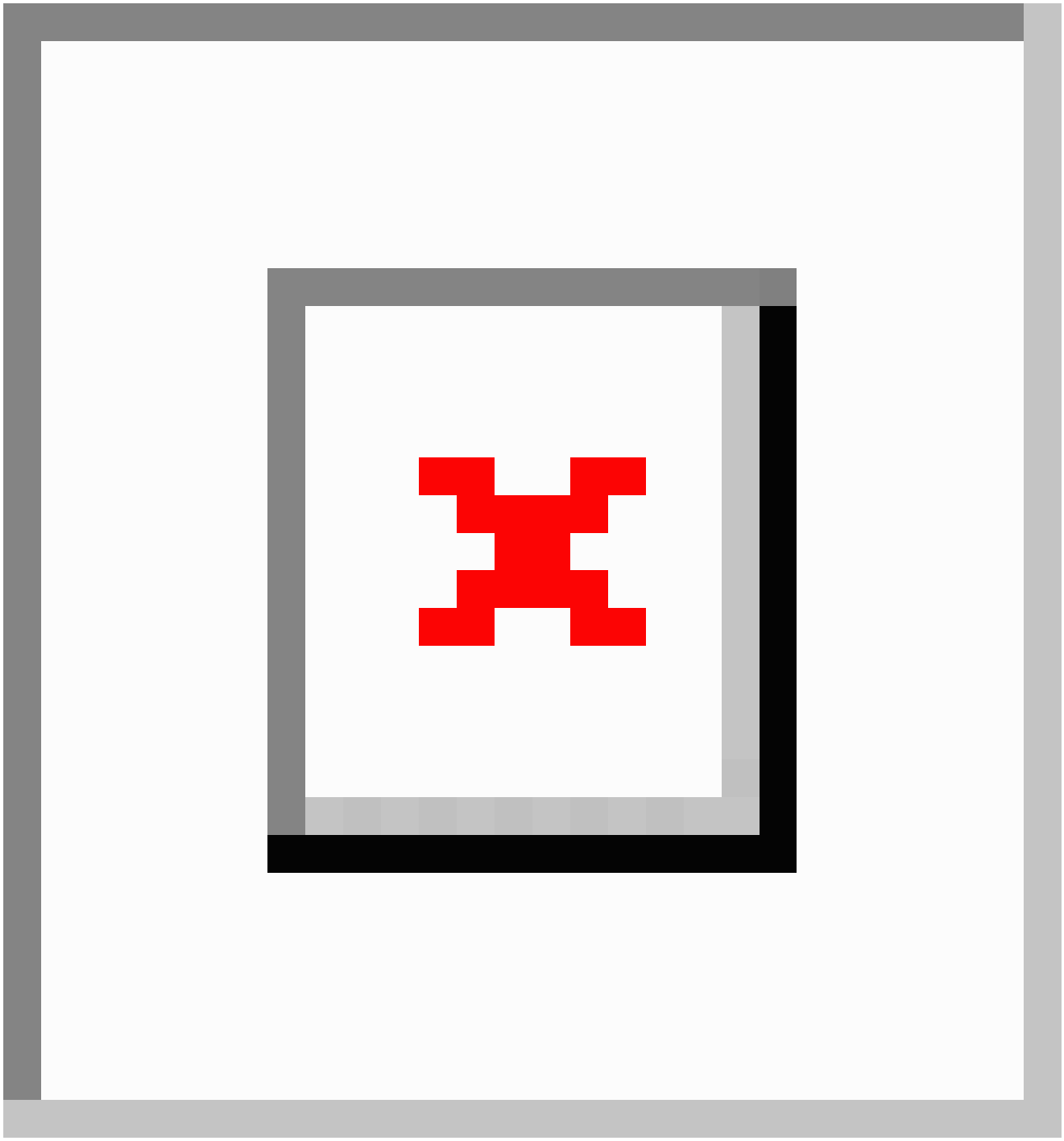
Figure 4. Overall counts of NLP techniques used for extraction of SDOH data including natural language processing (NLP), social determinants of health (SDOH), term frequency-inverse document frequency (TF-IDF) .



Health Information Exchange (HIE) for SDOH Data Aggregation

Health information exchange (HIE), featured in 4 out of 26 studies (Figure 5), aggregates patient data across health care entities, providing a comprehensive view of a patient's clinical journey. With 40% of patients in one study having encounters at multiple organizations, the importance of HIE in reflecting the distributed nature of ED care becomes evident. HIE can aid in analyzing care transitions and augmenting the sample size and diversity for SDOH research. However, challenges like data sharing, data quality, and privacy regulations need to be addressed. In essence, HIE holds immense potential for EM research, offering both multi-organizational and community-level insights.

Figure 5. Overall counts for identified articles by domain and/or task.



Predictive Applications

ML techniques can be important tools to leverage for clinical outcome predictions, and we identified a significant number of studies in our review. A total of 19 studies, representing a diverse array of ML strategies were identified, each tailored for distinct predictive outcomes. The efficacy of predictive models was a pivotal outcome in 10 studies. Performance metrics for these models, when used as a primary outcome, showcased a range of F1-scores from 0.4 to 0.75, indicating varied precision and recall across studies. These ranged from supervised models like random forest classifiers for acute coronary syndrome predictions to neural networks targeting sepsis-related readmission risks. Notably, ensemble methods were adeptly used to discern primary risk factors for OUDs. Within this realm, NLP proved instrumental, particularly in classification tasks, risk stratification, and fortifying clinical decision-making processes. Beyond this, key outcomes highlighted patterns in ED revisits (n=6), in-hospital mortality (n=2), algorithmic bias (n=1), and mismatches between physician annotations and claims data (n=1).

Data Quality, Privacy, and Algorithmic Bias

In the comprehensive spectrum of studies analyzed, an intriguing observation surfaced: a noticeable gap in the examination of data quality and privacy concerns within the realm of SDOH and EM. Despite sifting through a multitude of research articles, we did not encounter any study that directly tackled these pivotal

issues. This omission underlines the potential vulnerabilities in the application of AI and ML techniques, especially when handling sensitive patient data. Furthermore, a singular study broached the topic of algorithmic bias, a topic of paramount importance given the potential repercussions on health care outcomes and equity. The underrepresentation of these themes hints at uncharted territories that warrant meticulous exploration in future research endeavors, ensuring that the integration of AI and ML in EM is both robust and ethically sound.

Discussion

Principal Findings

The leveraging of SDOH data within EM research and clinical care is pivotal for gaining new insights, improving patient outcomes, and optimizing health care delivery. Our scoping review highlights the transformative role of data science, chiefly AI/ML, and the subdiscipline of NLP, in improving SDOH data integration and modeling within EM. Emerging from our review is an increase toward using data science to harness, operationalize, and model SDOH in emergency care settings. This progression signifies a shift: a pivot toward a comprehensive, data-infused approach that addresses not just emergent conditions but also the intricate web of social and economic determinants impacting health. NLP excels at extracting SDOH information from the unstructured text of EHRs, while ML's predictive strength can transform these insights into actionable predictions. Such models, equipped with SDOH data, can catalyze precision interventions, potentially identifying mechanisms for ED revisits, in-hospital mortality, and readmissions.

Delving deeper into the SDOH domains, housing insecurity and the neighborhood environment emerged as primary determinants, witnessing significant attention across the studies. Their frequent appearance in the research landscape underscores their profound impact on health outcomes within emergency settings. While these domains were at the forefront, other determinants like education, employment, and social networks were also featured, albeit to a lesser extent. The emphasis on these SDOH domains, especially housing insecurity, suggests a pressing need for targeted interventions and policies within emergency care settings. As the health care sector continues to evolve, understanding these predominant SDOH domains and harnessing the power of data science will be pivotal in offering a more holistic, patient-centric approach to emergency care.

Our comprehensive review, while offering insights, bears certain limitations warranting acknowledgment. The time frame for our study, confined to 2015-2022, captures most contemporary advancements but might inadvertently omit foundational studies predating this period, potentially offering evolutionary insights. While we highlighted NLP and other ML techniques, the vast expanse of data science boasts other emerging tools such as the recent and rapid development of large language modeling (LLM) such as ChatGPT (OpenAI), which were absent in our review. The potential applications of LLM in harnessing and modeling SDOH within the EM setting are rapidly emerging and will likely expand the possibilities for improving health outcomes and disparities.

Second, the encompassing scope of our review, spanning diverse SDOH domains and emergency conditions, enriches the study's comprehensiveness. Yet, the scope also poses challenges in distilling specific conclusions regarding the utility of data science techniques across distinct SDOH or EM conditions. Comparative analysis across studies was hampered by the varied outcome measures adopted. Although a significant number focused on the performance of predictive models, this only scratches the surface of data science's potential impact on SDOH within EM. Lastly, our review refrained from assessing the methodological quality of the incorporated studies. This approach aligns with scoping review guidelines but omits considerations of each study's methodological soundness during our synthesis.

Amidst these intricacies and challenges, there are still other pressing concerns. Data quality, privacy concerns, and algorithmic biases are potential hurdles that merit attention. In particular, the limited exploration and assessment of algorithmic bias in our reviewed studies, given its potential to perpetuate health care disparities, suggests an urgent avenue for further investigation. Without careful assessment of algorithmic biases, there exists the potential for reinforcing racial and ethnic discrimination and institutional racism. Only one study we identified in our review specifically assessed ML model bias and fairness in the context of heart failure outcomes [18]. ML algorithmic biases are critical to address in the context of SDOH research, as prior studies have demonstrated the potential for reinforcement of pre-existing racial, ethnic, and socioeconomic disparities [19].

Future Works and Recommendations

Potential Areas of Exploration

Our review has illuminated the significance of certain domains like housing insecurity, within the context of SDOH in EM. However, the vast landscape of SDOH offers numerous other domains that remain relatively unexplored:

1. **Education:** Investigating the role of educational attainment and access to quality education can provide insights into its impact on health outcomes. For instance, understanding how literacy levels influence patient adherence to medical advice in emergency settings could be pivotal. AI techniques could be used to better understand the complexity of both current educational attainment and future or desired educational needs through careful extraction or modeling of data.
2. **Employment:** Employment status, job security, economic mobility, and workplace conditions can have profound effects on mental and physical health. Exploring these factors can shed light on stress-related conditions or injuries that present in EDs.
3. **Social networks:** The influence of social support systems, community engagement, and familial ties can play a crucial role in patient recovery and mental well-being. Delving into these aspects can offer a holistic view of a patient's environment and its implications for health.

With these potential areas in mind, it becomes evident that a multi-faceted approach to SDOH within EM is the way forward.

Building on these areas of exploration, we propose several recommendations to harness the full potential of SDOH in EM.

Recommendations

1. **Establishing gold standard metrics:** For the evolution and standardization of emergency SDOH research, it is essential to define and adopt gold standard metrics. These metrics should be robust, universally accepted, and tailored to capture the nuances of SDOH in emergency settings. Collaborative efforts among researchers, clinicians, and policy makers should be made to create these benchmarks.
2. **Innovative data capture and implementation:** The high-paced nature of emergency settings necessitates innovative solutions for capturing SDOH data. Leveraging AI-assisted tools or predictive algorithms based on existing patient data could offer one approach. Creating automated workflows to allow for capture and implementation of SDOH data at the patient encounter level will be important for addressing social risks and needs.
3. **Algorithmic innovation:** The prominence of ML and NLP in our findings suggests a horizon brimming with algorithmic advancements and adaptation for EM. As these tools evolve and new tools emerge, crafting and evaluating interventions tailored to specific SDOH is crucial.
4. **Connecting SDOH with clinical outcomes:** Beyond identifying SDOH, understanding their tangible impact on patient outcomes is vital. A concerted effort in this direction can revolutionize our care approach. The intersection of SDOH and clinical outcomes is recognized by the Centers for Medicare and Medicaid Services, which requires health care systems to screen for key SDOH domains.
5. **Interdisciplinary collaboration:** The confluence of expertise, from clinicians to data scientists, will be instrumental in harnessing the full potential of SDOH data.
6. **Addressing algorithmic bias:** As we increasingly rely on algorithms, it is imperative to ensure they are free from biases that could perpetuate or exacerbate health disparities. Rigorous testing, validation, and refinement of algorithms, with a focus on fairness and equity, should be prioritized.

Conclusions

This scoping review underscores the transformative potential of data science in elevating the understanding and application of SDOH within EM. Through the adept integration of data science methodologies, particularly ML and NLP, we are poised to redefine the way SDOH data is adopted within EM. This offers a broader and more data-informed approach to influencing critical patient outcomes. The literature landscape indicates a promising embrace of this cross-disciplinary synergy, manifesting in an increasing number of studies that deploy data science methodologies to unearth, interpret, or model SDOH within emergency care contexts. Such a trajectory not only affirms the growing acknowledgment of these methodologies' efficacy but also underlines the health care sector's commitment to delivering more holistic care.

Nevertheless, our review also pinpoints avenues that warrant deeper exploration. Despite the expansive focus on various SDOH domains, certain determinants like housing insecurity and the neighborhood environment have garnered disproportionate attention. A more balanced exploration across SDOH domains would provide a richer, more comprehensive insight into their collective and individual impacts on patient trajectories. Moreover, while the current trend leans heavily on ML and NLP, there exists a vast expanse of data science techniques yet to be fully leveraged like LLM. Diving into these untapped methodologies might further refine our capabilities in SDOH identification and intervention.

In conclusion, the fusion of data science with EM marks the dawn of a new health care epoch. It envisions a future where EDs transcend their traditional roles, evolving into hubs that address the foundational SDOH challenges within communities. As we navigate this promising trajectory, the potential to revolutionize EM and fortify patient-centric care is immense.

Acknowledgments

Dr. Abbott was supported by National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (1K08HL169980-01A1).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Finalized search strategies.
[DOCX File, 17 KB - medinform_v12i1e57124_app1.docx]

Multimedia Appendix 2
Finalized complete list of literature included in the review.
[DOCX File, 18 KB - medinform_v12i1e57124_app2.docx]

Multimedia Appendix 3
Emergency medicine conditions word cloud. Size of word cloud represents aggregated counts of terms from titles of articles identified for this review (Python (version 3.7), matplot library).

[PNG File, 422 KB - medinform_v12i1e57124_app3.png]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis flow chart) checklist.
[PDF File, 319 KB - medinform_v12i1e57124_app4.pdf]

References

1. Hood CM, Gennuso KP, Swain GR, Catlin BB. County health rankings: relationships between determinant factors and health outcomes. Am J Prev Med 2016 Feb;50(2):129-135. [doi: 10.1016/j.amepre.2015.08.024] [Medline: 26526164]
2. Shah R, Della Porta A, Leung S, et al. A scoping review of current social emergency medicine research. West J Emerg Med 2021 Oct 27;22(6):1360-1368. [doi: 10.5811/westjem.2021.4.51518] [Medline: 34787563]
3. Malecha PW, Williams JH, Kunzler NM, Goldfrank LR, Alter HJ, Doran KM. Material needs of emergency department patients: a systematic review. Acad Emerg Med 2018 Mar;25(3):330-359. [doi: 10.1111/acem.13370] [Medline: 29266523]
4. Alderwick H, Gottlieb LM. Meanings and misunderstandings: a social determinants of health lexicon for health care systems. Milbank Q 2019 Jun;97(2):407-419. [doi: 10.1111/1468-0009.12390] [Medline: 31069864]
5. Hosseinpoor AR, Bergen N, Schlotheuber A. Promoting health equity: WHO health inequality monitoring at global and national levels. Glob Health Action 2015;8:29034. [doi: 10.3402/gha.v8.29034] [Medline: 26387506]
6. Kangovi S, Barg FK, Carter T, Long JA, Shannon R, Grande D. Understanding why patients of low socioeconomic status prefer hospitals over ambulatory care. Health Aff (Millwood) 2013 Jul;32(7):1196-1203. [doi: 10.1377/hlthaff.2012.0825] [Medline: 23836734]
7. Samuels-Kalow ME, Ciccolo GE, Lin MP, Schoenfeld EM, Camargo CA Jr. The terminology of social emergency medicine: measuring social determinants of health, social risk, and social need. J Am Coll Emerg Physicians Open 2020 Oct;1(5):852-856. [doi: 10.1002/emp2.12191] [Medline: 33145531]
8. Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. N Engl J Med 2015 Feb 19;372(8):698-701. [doi: 10.1056/NEJMp1413945] [Medline: 25693009]
9. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. J Am Med Inform Assoc 2020 Nov 1;27(11):1764-1773. [doi: 10.1093/jamia/ocaa143] [Medline: 33202021]
10. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc 2021 Dec 28;29(1):187-196. [doi: 10.1093/jamia/ocab199] [Medline: 34664641]
11. Pantell MS, Adler-Milstein J, Wang MD, Prather AA, Adler NE, Gottlieb LM. A call for social informatics. J Am Med Inform Assoc 2020 Nov 1;27(11):1798-1801. [doi: 10.1093/jamia/ocaa175] [Medline: 33202020]
12. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying patients with significant problems related to social determinants of health with natural language processing. Stud Health Technol Inform 2019 Aug 21;264:1456-1457. [doi: 10.3233/SHTI190482] [Medline: 31438179]
13. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. J Am Med Inform Assoc 2021 Nov 25;28(12):2716-2727. [doi: 10.1093/jamia/ocab170] [Medline: 34613399]
14. Conway M, Keyhani S, Christensen L, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. J Biomed Semantics 2019 Apr 11;10(1):6. [doi: 10.1186/s13326-019-0198-0] [Medline: 30975223]
15. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol 2019 Apr;28(2):73-81. [doi: 10.1080/13645706.2019.1575882] [Medline: 30810430]
16. Harrison CJ, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction to natural language processing. BMC Med Res Methodol 2021 Jul 31;21(1):158. [doi: 10.1186/s12874-021-01347-1] [Medline: 34332525]
17. Mike K, Hazzan O. What is data science? Commun ACM 2023;66(2):12-13. [doi: 10.1145/3575663]
18. Li Y, Wang H, Luo Y. Improving fairness in the prediction of heart failure length of stay and mortality by integrating social determinants of health. Circ Heart Fail 2022 Nov;15(11):e009473. [doi: 10.1161/CIRCHEARTFAILURE.122.009473] [Medline: 36378761]
19. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]

Abbreviations

AI: artificial intelligence
ED: emergency department
EHR: electronic health record
EM: emergency medicine
HIE: health information exchange

LLM: large language modeling
ML: machine learning
NLP: natural language processing
OUD: opioid use disorder
SDOH: social determinants of health

Edited by C Lovis; submitted 05.02.24; peer-reviewed by B Senst, BG Patra, J Bell, PBDL Vega; revised version received 10.07.24; accepted 21.07.24; published 30.10.24.

Please cite as:
Abbott EE, Apakama D, Richardson LD, Chan L, Nadkarni GN
Leveraging Artificial Intelligence and Data Science for Integration of Social Determinants of Health in Emergency Medicine: Scoping Review
JMIR Med Inform 2024;12:e57124
URL: https://medinform.jmir.org/2024/1/e57124
doi:10.2196/57124

Original Paper

Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study

Boshu Ru[1], PhD; Arthur Sillah[1], MPH, PhD; Kaushal Desai[1], MS, PhD; Sheenu Chandwani[1], MPH, PhD; Lixia Yao[1], PhD; Smita Kothari[1], MBA, PhD

Center for Observational and Real-world Evidence (CORE), Merck & Co, Inc, West Point, PA, United States

Corresponding Author:
Boshu Ru, PhD
Center for Observational and Real-world Evidence (CORE)
Merck & Co, Inc
770 Sumneytown Pike
WP37A
West Point, PA, 19486
United States
Phone: 1 215 652 4301
Email: boshu.ru@merck.com

Abstract

Background: The importance of real-world evidence is widely recognized in observational oncology studies. However, the lack of interoperable data quality standards in the fragmented health information technology landscape represents an important challenge. Therefore, adopting validated systematic methods for evaluating data quality is important for oncology outcomes research leveraging real-world data (RWD).

Objective: This study aims to implement real-world time to treatment discontinuation (rwTTD) for a systemic anticancer therapy (SACT) as a new use case for the Use Case Specific Relevance and Quality Assessment, a framework linking data quality and relevance in fit-for-purpose RWD assessment.

Methods: To define the rwTTD use case, we mapped the operational definition of rwTTD to RWD elements commonly available from oncology electronic health record–derived data sets. We identified 20 tasks to check the completeness and plausibility of data elements concerning SACT use, line of therapy (LOT), death date, and length of follow-up. Using descriptive statistics, we illustrated how to implement the Use Case Specific Relevance and Quality Assessment on 2 oncology databases (*Data sets A and B*) to estimate the rwTTD of an SACT drug (*target SACT*) for patients with advanced head and neck cancer diagnosed on or after January 1, 2015.

Results: A total of 1200 (24.96%) of 4808 patients in Data set A and 237 (5.92%) of 4003 patients in Data set B received the target SACT, suggesting better relevance of the former in estimating the rwTTD of the target SACT. The 2 data sets differed with regard to the terminology used for SACT drugs, LOT format, and target SACT LOT distribution over time. Data set B appeared to have less complete SACT records, longer lags in incorporating the latest data, and incomplete mortality data, suggesting a lack of fitness for estimating rwTTD.

Conclusions: The fit-for-purpose data quality assessment demonstrated substantial variability in the quality of the 2 real-world data sets. The data quality specifications applied for rwTTD estimation can be expanded to support a broad spectrum of oncology use cases.

(JMIR Med Inform 2024;12:e47744) doi:10.2196/47744

KEYWORDS

data quality assessment; real-world data; real-world time to treatment discontinuation; systemic anticancer therapy; Use Case Specific Relevance and Quality Assessment; UReQA framework

Introduction

Background

The importance of real-world evidence drawn from real-world data (RWD) is widely recognized in oncology research [1-5]. Over the past decade, federal legislation and incentives promoting the secondary use of RWD in the United States [6-8], coupled with advances in health information technology, have resulted in an explosion of RWD sources and a complex RWD ecosystem [1]. However, this rich data landscape can also pose challenges in identifying fit-for-purpose RWD to meet biopharma research needs.

Two key obstacles to identifying high-quality data are the fragmentation of RWD sources and the lack of interoperable data quality standards. These obstacles are particularly pertinent in the United States, where progress is slow in reaching full interoperability of data sourced from thousands of providers who customized their electronic health record (EHR) systems from solutions provided by >40 different EHR software vendors [9]. Therefore, adopting validated systematic methods for evaluating data quality is important for research leveraging RWD [10-12].

In 2016, an expert panel proposed the concepts of *conformance*, *completeness*, and *plausibility* as 3 categories (with subcategories) to describe the intrinsic data quality of EHR databases and to serve as a framework for assessing data quality that could then be verified (with organizational data) or validated using an accepted gold standard [13]. Several working groups and authors have applied these terms or proposed others for defining research data quality [14-16], and multiple initiatives in the United States, both public and private, have developed frameworks and tools to evaluate and improve the quality of EHR data sets [17-21] and to implement model-driven, quantitative approaches to address RWD completeness and plausibility issues [22-25]. However, there is no single RWD source that can fit the needs of all studies, and the selection of RWD to support an individual use case must also consider data relevance and measurement thresholds in addition to data quality.

Objective

In a previous study, we introduced the Use Case Specific Relevance and Quality Assessment (UReQA) framework, an RWD quality framework that combines both the data quality and the relevance aspects of assessing RWD, with the goal of developing data quality assessment specifications tailored to use cases [3]. In this study, we aimed to implement this framework in the use case for estimating real-world time to treatment discontinuation (rwTTD) in oncology. Our work had two main components: (1) to design comprehensive data quality assessment checks for estimating rwTTD for a systemic anticancer therapy (SACT) and (2) to illustrate how these quality checks can be used to evaluate EHR-derived RWD products.

We selected rwTTD as the first use case to implement the UReQA framework because of its high utility as a pragmatic real-world effectiveness end point for continuously administered SACTs (such as immunotherapies) and its known correlation with overall survival [26-28]. Moreover, the estimation of rwTTD requires information on medication use patterns, mortality, and follow-up. These data elements are foundational to outcomes research. Therefore, implementation of the rwTTD use case can be expanded to other use cases in or beyond oncology, as well as different data sources, such as claims databases.

Methods

Ethical Considerations

This study used 2 commercially licensed deidentified structured secondary data sources accessible to the study team. It was exempted from institutional review board review because of the following: (1) each data source contains a significant level of protection against the release of personal information to outside entities and (2) the use of such databases presents the lowest risk to potential subjects because the analysis involves only anonymous data; hence, conducting the study will not place the subjects at risk.

Study Overview

This study comprised four main steps: (1) conceptual definition of the rwTTD use case; (2) mapping of the rwTTD use case definition to RWD elements (operational definition); (3) identifying data quality checks for the required data elements to determine rwTTD for an SACT, designated the “target SACT”; and (4) implementing the UReQA framework [3] in assessing the RWD fitness for estimating rwTTD. The data quality assessment was undertaken on 2 US EHR-based oncology databases for estimating rwTTD for a target SACT, an immunotherapy drug that is administered intravenously in advanced-stage head and neck cancer (HNC). The targeted SACT received approval in 2016 for the treatment of previously treated advanced HNC and in 2019 for its use as a first-line therapy in advanced HNC. The focus of this study is on designing data quality assessment methods that are tailored for specific use cases, rather than calculating rwTTD for a particular medication. Therefore, we mask the name of the actual drug product.

Step 1: Conceptual Definition of the rwTTD Use Case

The end point, rwTTD, is defined as the length of time from initiation to discontinuation of a medication (*[date of last recorded dose – date of first recorded dose] + 1 d*), with discontinuation defined as the date of the last dose if a patient died during therapy or initiated a new treatment or if there is a gap of ≥120 days between the last recorded dose and last recorded activity in a data set. Patients who do not meet the discontinuation criteria are censored at the last medication use [26-28].

Step 2: Mapping of the rwTTD Use Case Definition to RWD Elements

Owing to the variations in data element definition and data structures between real-world EHR databases, we need to operationalize the concept of rwTTD by deconstructing its definition and mapping it to four sets of required data elements that are commonly available from oncology EHR–derived data

sets: (1) SACT, (2) line of therapy (LOT) specifying the regimen names and sequence of current treatment in the treatment plan [29,30], (3) mortality status, and (4) follow-up time, as summarized in Table 1. Although SACT, mortality status, and follow-up time are often recorded directly as procedure, prescription, and administrative events in raw EHR databases, the LOT was often derived from raw EHR by the algorithm.

Table 1. Required data elements for determining real-world time to treatment discontinuation (rwTTD) for a systemic anticancer therapy (SACT) drug in a specific line of therapy (LOT).

Operational steps to ascertain rwTTD and type of data category	Commonly used data elements
Identify records of the drug of interest	
SACT drug	Drug_name, NDC[a], HCPCS[b] code, RxNorm code
SACT administration	Drug administration date[c]
SACT order	Drug order date[d]
Identify discontinuation date from subsequent LOT start date	
LOT	LOT name[e]
LOT	LOT number
LOT	LOT start date
LOT	LOT end date
If no subsequent LOT, identify discontinuation date from patient death record during treatment	
Mortality status	Vital status or date of death
If no date of death, identify discontinuation date by last follow-up date subheading	
Last follow-up	Date of last follow-up[f]

[a]NDC: National Drug Code.

[b]HCPCS: Healthcare Common Procedure Coding System.

[c]The drug administration date is defined as the date of receiving medication at a health care facility as a medical service, often applicable to an intravenous drug.

[d]The drug order date is defined as the order date for drugs used at home, often applicable to an oral drug.

[e]The LOT name is determined by the combination of SACT drugs administered or ordered from the LOT start to end dates.

[f]The date of last follow-up is defined as the last documented clinic visit or procedure in the electronic health record.

We defined SACT as any systemic anticancer medication received by the patient, documented as given either by a health care provider at the site of care (eg, by infusion), with the date defined as the "administration" date, or as a prescription to take or apply at home, with the date defined as the "drug order" date. The number of refills (or alternative data elements such as days of supply or expected medication end date) was used to determine the last use of oral drugs (Table 1).

LOT was defined as the sequence of the SACT regimens prescribed for an individual patient, as previously described in detail [29,30]. In brief, the first LOT (line 1 [1L]) begins with the first SACT initiated after a study index date (often the advanced or metastatic cancer diagnosis date), and any other drug introduced within the next 28 days is considered part of that LOT [29]. We defined the start of a new LOT when a new SACT not belonging to the prior LOT was introduced or if a new SACT was initiated after a ≥120-day gap in therapy.

Because the target SACT was administered intravenously, we omitted 2 tasks applicable only to oral target SACTs: the check of patient numbers with target drug order date after the index date (Multimedia Appendix 1, task 6 [13]) and the check for distribution of gaps between drug order dates (Multimedia Appendix 1, task 9).

The patient mortality status was determined based on the recorded dates of death. For patients who were still alive at data cutoff, the date of the last follow-up was defined as the last documented clinical activity date in the EHR (Table 1).

Step 3: Identifying Data Quality Checks for Required Data Elements

For each of the required data elements, we identified corresponding verification checks to assess data quality at both the variable level and the cohort level. A total of 20 data quality checks (tasks) were identified and categorized into the quality dimensions of conformance, completeness, and plausibility, as per the harmonized data quality assessment terms and framework developed by Kahn et al [13] (Multimedia Appendix 1). Our goal in creating these tasks was to develop a comprehensive toolbox for assessing data quality for the rwTTD use case. However, when adapting them to a specific RWD database and a SACT drug of interest, not every task and check would be necessary. For example, the checks for LOT, mortality, and follow-up are not needed if a data set already provides the reason for discontinuation and censored status for each drug exposure. In addition, tasks 3-5 were applicable to cancer therapies received in hospitals or clinics as intravenous or infusion procedures, whereas tasks 4-9 were dedicated to oral

cancer therapies that were mostly self-administrated at home. As tracking the actual time patients took oral therapies was infeasible, researchers examined days supply and refill records to estimate the drug exposure period. Therefore, when investigating the rwTTD of an oral SACT drug, it is necessary to check the completeness of these oral therapy–specific data elements (task 7).

Step 4: Implementing the rwTTD Use Case for Assessing 2 RWD Sets

Data Set Preassessment

We followed the preassessment step in UReQA [3] to identify 2 anonymized, commercially available US real-world oncology databases, designated as *Data set A* and *Data setB* in this report, which included patients with advanced (metastatic or unresectable, recurrent) HNC. Both databases contained data elements sourced from structured and unstructured information captured within health care providers' EHR systems as part of routine cancer care.

Cohort Selection and Patient Characteristics

Data set A was commercialized and included patients with advanced HNC, whereas Data set B included patients with all stages of HNC. To align the 2 patient populations as having advanced HNC, we restricted Data set B to the subset of patients with HNC and a record of the American Joint Committee on Cancer stage IV and *International Classification of Diseases* (ICD), *revision 9 or 10* (ICD-9 or ICD-10) code for metastatic tumor (ICD-9 codes 196.x, 197.x, and 198.x and ICD-10 codes C76.x, C77.x, and C78.x). The distributions of the patient characteristics were then tabulated for the 2 data sets.

Data Elements Harmonization

In Data set A, the names of SACT medications were harmonized from clinic formulary information and medical service records to standard generic drug names in a commercial drug database along with drug category information. In Data set B, all medication records in the raw EHR data were harmonized into the RxNorm code [31]; however, drug category information was not available. To harmonize all SACT medication in Data set B, we retrieved the RxNorm codes for generic names of all SACT medications using the RxNav software developed by and available from the US National Library of Medicine [32].

The LOT information was previously derived by both data providers but was presented differently in the 2 data sets. In Data set A, the LOT table provided the LOT number, LOT regimen name, LOT start date, and LOT end date, with a flag indicative of maintenance therapy, as appropriate. Instead, Data set B included only the LOT number and LOT start date. Therefore, to evaluate the LOT information in Data set B, we indirectly deduced the end date of each LOT as the date before the start of the next LOT or as the data cutoff date for the last LOT in the data set. Then, all individual SACT medications administered or ordered between the LOT start and end dates were combined to serve as the LOT regimen name. This approach was a necessary but imperfect solution because the LOT end date and the LOT regimen name should ideally be generated using a more rigorous algorithm [29,30].

The date of death was provided at the month and day levels in Data set A, whereas in Data set B, the death date was aggregated by year. Given the relatively short length of survival of many patients with advanced HNC [33-36], the allocation of death dates by year was not sufficiently granular for accurate rwTTD calculation; better precision (ie, month of death) would be needed for accurate rwTTD calculation. Consequently, quality assessment tasks related to mortality variables were omitted (task 17) for Data set B.

Reporting the Verification Results

Descriptive statistics were used to summarize the results of implementing rwTTD data quality checks on Data sets A and B. We used frequencies to summarize categorical variables and mean (SD) and median (IQR or range) to summarize continuous variables. The study index date was the date of first advanced HNC diagnosis, and the cutoff date was November 25, 2019.

All analyses were conducted using SAS Studio release 3.8 (Basic Edition; SAS Institute, Inc).

Results

Patient Characteristics

Data set A included 7366 patients with advanced HNC, and we identified 11,386 patients in Data set B with advanced HNC. The median patient age at the first advanced HNC diagnosis was 65 (IQR 58-72) years in Data set A and 61 (IQR 54-68) years in Data set B, and the percentages of male individuals were 74.16% (5643/7366) and 69.97% (7967/11386), respectively (Table 2), similar to the HNC population data from the United States [33,37].

Table 2. Baseline characteristics of patients with advanced head and neck cancer (HNC) included in 2 data sets under evaluation[a].

Characteristic	Data set A (n=7366)	Data set B (n=11,386)
Sex, n (%)		
Female	1723 (23.4)	3408 (29.9)
Male	5643 (76.6)	7967 (70)
Missing or unknown	0 (0)	11 (0.1)
Age at first advanced HNC diagnosis (y), median (IQR)	65 (58-72)	61 (54-68)
Age at first advanced HNC diagnosis (y), n (%)		
<18	0 (0)	31 (0.27)
18-44	187 (2.53)	688 (6.04)
45-64	3402 (46.19)	5955 (52.3)
65-88	3777 (51.28)	4111 (36.11)
≥89	0 (0)	6 (0.05)
Missing or unknown	0 (0)	595 (5.23)
Race or ethnicity, n (%)		
American Indian or Alaska Native	N/A[b]	40 (0.35)
Asian	103 (1.4)	165 (1.45)
Black or African American	487 (6.61)	1250 (10.98)
Hispanic or Latino	13 (0.18)	0 (0)
Native Hawaiian or other Pacific Islander	N/A	6 (0.05)
White	4939 (67.05)	9239 (81.14)
Missing	650 (8.82)	686 (6.02)
Other race	1174 (15.94)	0 (0)
AJCC[c] stage at first HNC diagnosis, n (%)		
0	2 (0.03)	28 (0.25)
I	419 (5.69)	603 (5.3)
II	505 (6.86)	542 (4.76)
III	929 (12.61)	798 (7.01)
IV	4330 (58.78)	4978 (43.72)
Missing or unknown	1181 (16.03)	4437 (38.97)
Year of first advanced HNC diagnosis, n (%)		
Before 2006	0 (0)	1245 (10.9)
2006-2009	0 (0)	1537 (13.5)
2010-2012	1068 (14.5)	2721 (23.9)
2013-2018	5435 (73.8)	5577 (49.0)
2019 or later	863 (11.7)	306 (2.7)

[a]Percentages may not add up to 100% because of rounding.

[b]N/A: not applicable.

[c]AJCC: American Joint Committee on Cancer.

SACT Data Checks

Overall, 75.91% (5592/7366) and 38.74% (4411/11386) of the patients in Data sets A and B, respectively, had a recorded drug administration or drug order for any SACT (Table 3, task 1). A complete start date (y, mo, and d) was recorded for all SACT administrations and orders in both data sets (Table 3, tasks 4 and 8).

Table 3. Data quality assessment of SACT[a] administration and order records after the advanced HNC[b] diagnosis[c].

SACT data quality checks	Data set A	Data set B
Task 1: patients with any SACT drug administration or order record after the advanced HNC diagnosis date, n (%)[d]	5592 (75.9)	4411 (38.7)
Task 2: SACT drug records with missing drug identity (name and code) information		
Value, n (%)	0 (0)	0 (0)
Normalization of medication name	Normalized generic name	RxNorm ingredient level
Task 3: patients with target SACT administration date after the advanced HNC diagnosis date, 2015 onward, % (n/N)[e]	24.96 (1200/4808)	5.92 (237/4003)
Task 4: SACT drug administration records with complete administration date, n (%)	425,505 (100)	37,662 (100)
Task 5: gap (in d) between the target SACT drug administration dates, median (IQR; range)	21 (21-21; 1-113)	21 (11-21; 1-824)
Task 6: patients with target SACT order date after the advanced HNC diagnosis date	N/A[f,g]	N/A[g]
Task 7 SACT drug order records with complete days supply and refill information, n (%)	1732 (53.4)	N/A[h]
Task 8: SACT drug order records with complete order date, n (%)[i]	3241 (100)	8380 (100)
Task 9: distribution of gaps (in d) between target SACT drug order dates, normalized by days supply, refill, and cancelation record	N/A[g]	N/A[g]

[a]SACT: systemic anticancer therapy.

[b]HNC: head and neck cancer.

[c]Drug *administration* refers to drugs administered by health care providers at the site of care, whereas drug *order* refers to prescriptions for drugs used at home.

[d]Task 1 was applied to the full data sets, including 7366 and 11,386 patients in Data sets A and B, respectively.

[e]Task 3 was applied for patients with the first advanced HNC diagnosis on or after January 1, 2015, including 4808 and 4003 patients in Data sets A and B, respectively.

[f]N/A: not applicable.

[g]Tasks 6 and 9 were not conducted because they apply to an oral target SACT.

[h]Information about the number of refills, days supply, or alternative data elements was not available in Data set B.

[i]The total number of drug order records in Data set A (3241) and Data set B (8380) was used as the denominator in task 8.

We determined that 4808 (65.27%) of the 7366 patients in Data set A and 4003 (35.16%) of the 11,386 patients in Data set B had a first advanced HNC diagnosis on or after January 1, 2015, the timeline we applied for the study index date as it covered the key diagnostic and therapeutic timeline of the target SACT (first approved in 2016). A total of 1200 (24.96%) of the 4808 patients meeting this timeline in Data set A and 237 (5.92%) of the 4003 patients meeting this timeline in Data set B had a record of receiving the target SACT (Table 3, task 3).

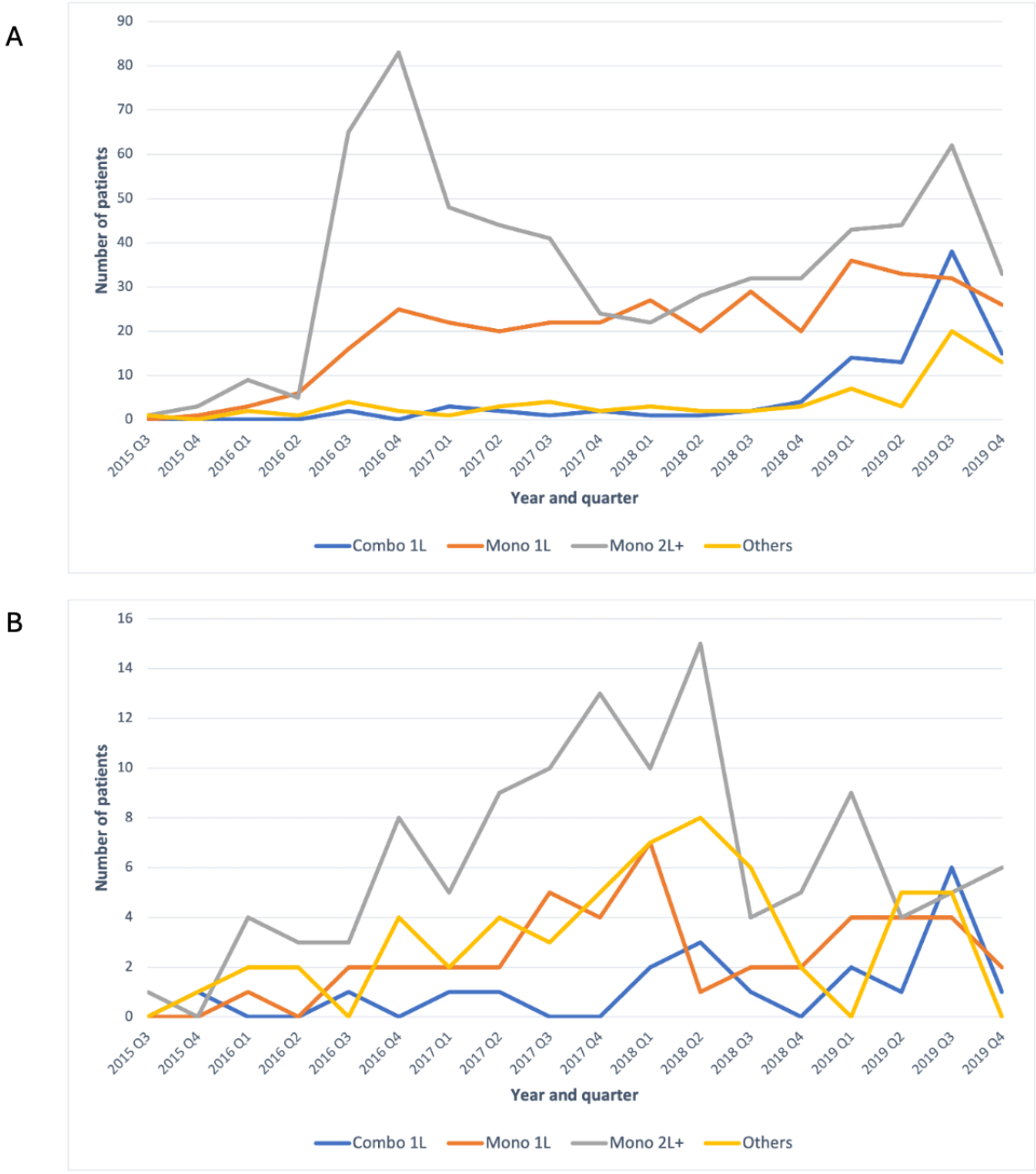
The median length of the gap between target SACT administrations was 21 days in both the data sets, which aligned with the expected dose schedule for the target SACT (Table 3, task 5). However, the range of the gap was considerably shorter in Data set A (1-113 d) than in Data set B (1-824 d), suggesting incomplete target SACT administration records in Data set B.

For the oral SACT records, Data set A included the number of refills and a flag for canceled medication orders, whereas Data set B did not provide refill information (Table 3, task 7). This could impact the accuracy of calculating rwTTD for an orally dispensed SACT because the drug orders for patients remaining on treatment through refills would not be recorded in the database.

LOT Data Checks

The 2 data sets differed in terms of the target SACT LOT distribution over time. The cumulative frequency of target SACT initiation, including as monotherapy or combination therapy and in any LOT, tended to be greater in later years in Data set A, peaking in the third quarter (Q3) of 2019, than in Data set B, peaking in the first and second quarters of 2018 (Figure 1, task 10). In Data set A, a greater frequency of target SACT initiation as second-line or later monotherapy was consistent with approval timing in this setting (2016), which preceded the first-line approvals (2019). The later time points for second-line or later monotherapy initiation in Data set B suggest the possibility of longer data lags than for Data set A.

Figure 1. Task 10: number of patients initiating the target systemic anticancer therapy (SACT) by year and quarter in (A) Data set A and (B) Data set B. Note: Y-axis heights in panels A and B differ but were selected to best depict the patient numbers in Data sets A and B. 1L: first-line therapy; 2L+: second-line or later therapy; combo: target SACT in any combination therapy (approved or not approved); mono: target SACT monotherapy; Q1: first quarter; Q2: second quarter; Q3: third quarter; Q4: fourth quarter.



In both data sets, we observed the inclusion of patients who initiated the target SACT therapy before the applicable first-line or second-line or later US Food and Drug Administration approval dates. We believe that these are true real-world findings, which do not always correspond to recommended or approved indications, rather than data quality issues.

In Data set B, only 40.3% (4589/11386) of patients had SACT LOT records (Table 4, task 11), which coincides with the finding of lower-than-expected SACT drug administration and order rates (Table 3, task 1).

Table 4. LOT[a] rules for SACT[b] and mortality information.

Task	Data set A	Data set B
Task 10: number of patients initiating the target SACT by year and quarter	Figure 1A	Figure 1B
Task 11: completeness of LOT information, n (%)[c]		
Patients with complete line number	5594 (75.94)	4589 (40.3)
Patients with complete line name	5594 (75.94)	N/A[d,e]
Patients with complete line start date	5594 (75.94)	4589 (40.3)
Patients with complete line end date	5594 (75.94)	N/A[e]
Task 12: patients for whom the first LOT number after the advanced HNC[f] diagnosis date was not 1, n (%)[c]	0 (0)	434 (3.81)
Task 13: distribution of LOT number at target SACT initiation		
Patients who received the target SACT, n	1200	237
Line 1, n (%)	481 (40.08)	65 (27.43)
Line 2, n (%)	486 (40.5)	92 (38.82)
Line 3, n (%)	161 (13.42)	54 (22.78)
Line 4, n (%)	46 (3.83)	16 (6.75)
Line 5, n (%)	13 (1.83)	10 (4.22)
Lines 6-10, n (%)	13 (1.83)	0 (0)
Task 14: use of target SACT in 1L[g] before approval date		
1L monotherapy		
Patients who received target SACT, % (n/N)	78.37 (377/481)	67.69 (44/65)
First administration date[h] in database	July 15, 2015	November 10, 2014
Cutoff date for the earliest 5% receipt	August 29, 2016	February 4, 2016
Cutoff date for the earliest 10% receipt	November 3, 2016	September 23, 2016
Cutoff date for the earliest 25% receipt	June 28, 2017	April 27, 2017
Approved 1L combination		
Patients who received the target SACT in approved 1L combination, % (n/N)	7.69 (37/481)	6.15 (4/65)
First administration date in database	December 18, 2018	July 1, 2019
Cutoff date for the earliest 5% receipt	February 18, 2019	N/A[i]
Cutoff date for the earliest 10% receipt	April 2, 2019	N/A[i]
Cutoff date for the earliest 25% receipt	July 9, 2019	N/A[i]
Task 15: patients with death record, n (%)[c]	4695 (63.74)	3531 (31)
Task 16: patients with multiple death records on different dates, n (%)	0 (0)	N/A[j]
Task 17: patients with clinical records showing health care activity after death date (n=4695), n (%)	1497 (31.88)	N/A[j]
≥1 d after date of death	1436 (30.59)	N/A[j]
≥3 d after date of death	1290 (27.48)	N/A[j]
≥7 d after date of death	1002 (21.34)	N/A[j]
≥30 d after date of death	79 (1.68)	N/A[j]

[a] LOT: line of therapy.

[b]SACT: systemic anticancer therapy.

[c]Tasks 11, 12, and 15 were applied to the full data sets, including 7366 and 11,386 patients in Data sets A and B, respectively.

[d]N/A: not applicable.

[e]Line name and line end date were not available in Data set B.

[f]HNC: head and neck cancer.

[g]1L: first line of therapy after the advanced HNC diagnosis date.

[h]Dates are written as month/day/year.

[i]Not calculated as only 4 patients received the target SACT in a 1L combination LOT.

[j]Only the year of death was available in Data set B.

The LOT start date in both data sets included year, month, and day, and the minimum LOT number started from 1 (first line) after the earliest advanced HNC diagnosis date for all but 3.81% (434/11386) of the patients in Data set B (Table 4, task 12). A line number other than 1 after the advanced HNC diagnosis date suggests that either a definition different from the commonly used definition [29,30] was used or that there was an earlier advanced HNC diagnosis date that was not documented.

In Data set A, 40.08% (481/1200) of patients received the target SACT in first-line therapy and 59.91% (719/1200) in second-line or later therapy, including 13.42% (161/1200) in third-line therapy (Table 4, task 13). In Data set B, 27.4% (65/237) of patients received the target SACT in first-line therapy, and 72.6% (172/237) received it in the second-line or later therapy, with frequent third-line receipt (54/237, 22.8%). Therefore, LOT rules may have been applied differently in Data set A and Data set B.

Complete information about the start date of first-line target SACT drug administration (as both monotherapy and combination therapy) was available for 377+37=414 (86.1%) of 481 patients in Data set A and 44+4=48 (74%) of 65 patients in Data set B (Table 4, task 14). In Data set A, first-line target SACT monotherapy was initiated for the first time in 2015, and approximately 5% of first-line monotherapy initiation dates fell on or before 2016, when the target SACT was approved for second-line or later therapy. Target SACT in combination therapy was first initiated in late 2018, with approximately 25% of the initiation dates falling before the start of Q3 in 2019, shortly after the approval of first-line combination therapy. In Data set B, first-line target SACT monotherapy initiation was first recorded in the fourth quarter in 2014, earlier than in Data set A, and close to 10% of initiation dates occurred before the end of Q3 in 2016. Instead, the approved target SACT combination therapy was first initiated at the start of Q3 in 2019, in line with the approval date for this indication.

Mortality Data

Among 7366 and 11,386 patients in Data sets A and B, 4695 (63.74%) and 3531 (31%), respectively, had a recorded date of death (Table 4, task 15); and 4427 (60%) and 3093 (27%) patients, respectively, had death records within 3 years after the date of advanced HNC diagnosis. These percentage differences indicate that Data set B may have incomplete mortality records (or a high loss to follow-up).

In Data set A, one-third of patients (1497/4695, 31.88%) with a recorded date of death had clinical records recorded after the death date (Table 4, task 17), with a median of 11 days from the death date to the last activity date. Thus, clinical records could be entered into the health information system after the reported death date, but extreme values (eg, >30 d after the death date) might indicate integrity issues in collecting mortality data. This information was not available for Data set B, in which the dates of death were recorded only by year.

Follow-Up Data

In Data set A, most patients (5840/7366, 79.28% to 7269/7366, 99.86%) had recorded data for diagnosis, drug records, laboratory results, facility visits, and vital sign measurements (Table 5, task 18). Similarly, in the subset of 7754 patients in Data set B whose advanced HNC diagnosis date was on or after January 1, 2011, the earliest date in Data set A, these data categories were also recorded for most patients (6123/7754, 78.97% to 6893/7754, 88.7%). Records of medical procedures not related to drug administration and genomic testing were not available in Data set A, which could result in inaccurate estimates of follow-up times.

Table 5. Unique number of patients and patient-date pairs after the advanced HNC[a] diagnosis date (task 18): follow-up data for patients with advanced HNC diagnosis on or after January 1, 2011.

Variable	Data set A (n=7366)			Data set B (n=7754)		
	Value, n (%)	Unique patient-date pairs, n	Pairs per patient, n	Value, n (%)	Unique patient-date pairs, n	Pairs per patient, n
Diagnosis	6567 (89.15)	60,178	9.2	6893 (88.9)	370,671	53.8
Drug records[b]	5840 (79.28)	113,948	19.5	6802 (87.72)	269,225	39.6
Laboratory records	6860 (93.13)	179,177	26.1	6403 (82.58)	147,314	23
Facility visit	7269 (98.68)	274,714	37.8	6838 (88.19)	392,175	57.4
Vital sign measurements	7254 (98.48)	233,623	32.2	6123 (78.97)	217,797	35.6
Nondrug medical procedure	N/A[c]	N/A	N/A	6740 (86.92)	390,556	57.9
Genomic test	N/A	N/A	N/A	118 (1.52)	208	1
Biomarker test	440 (5.97)	469	1.1	N/A	N/A	N/A
ECOG PS[d]	5416 (73.53)	100,607	17.7	N/A	N/A	N/A

[a]HNC: head and neck cancer.

[b]Any drug, not just systemic anticancer therapies.

[c]N/A: not applicable.

[d]ECOG PS: Eastern Cooperative Oncology group performance status.

The median frequency of visits (normalized by length between first and last target SACT administration) for patients who received the target SACT was somewhat less in Data set A, varying from 0.05 to 0.12, depending on treatment line, than in Data set B, in which it varied from 0.14 to 0.18 (Table 6, task 19). This might indicate that more clinical activities were recorded in Data set B during treatment.

Table 6. Follow-up data for patients with advanced HNC[a] diagnosis on or after January 1, 2011.

Task	Data set A (n=7366)			Data set B (n=7754)		
	Value, n	Value, median (IQR; range)	Value, mean (SD)	Value, n	Value, median (IQR; range)	Value, mean (SD)
Task 19: frequency of visits during target SACT[b,c]						
1L[d] combination therapy	101	0.11 (0.07-0.16; 0.02-0.33)	0.13 (0.07)	19	0.17 (0.11-0.24; 0-0.36)	0.17 (0.09)
1L monotherapy	358	0.05 (0.05-0.08; 0.01-0.50)	0.07 (0.05)	44	0.18 (0.09-0.22; 0-0.48)	0.16 (0.11)
2L+[e] monotherapy	634	0.06 (0.05-0.10; 0.01-0.95)	0.08 (0.06)	106	0.13 (0.07-0.25; 0-0.75)	0.17 (0.14)
All other	104	0.12 (0.09-0.17; 0.02-0.48)	0.14 (0.08)	76	0.14 (0.09-0.21; 0-1.3)	0.17 (0.17)
Task 20: for patients still alive, gap (in d) from the last target SACT administration and last visit[f]	708	28 (6-187; 0-1118)	128 (199)	167	70 (29-223; 0-1755)	159 (215)

[a]HNC: head and neck cancer.

[b]SACT: systemic anticancer therapy.

[c]Frequency defined as number of visits between the first and last target SACT administration dates within the same LOT number and name, divided by number of days between the last and first target SACT administration.

[d]1L: first line of therapy after the advanced HNC diagnosis date.

[e]2L+: second-line or later therapy.

[f]Limited to patients who (1) were still alive ≥180 days after last receipt of target SACT and (2) received last dose of target SACT ≥180 days before data cutoff on November 25, 2019 (thus on or before May 29, 2019).

Discussion

Principal Findings

This study identified 20 data quality assessment tasks for the use case of estimating the rwTTD of an SACT. By executing the 18 tasks pertinent to the intravenously administered target SACT, we demonstrated that the UReQA framework for the rwTTD use case can be implemented to generate descriptive summary statistics and charts. These visualizations provide additional insights into the relevance and quality of 2 US EHR-based oncology RWD. The approach is generalizable to implement for other SACT and databases.

Both data sets in the evaluation provided all the required data elements; however, verification checks revealed that Data set B might not be suitable for analyzing rwTTD for the target SACT because (1) the large decrease in patient receiving the target SACT in recent years suggests longer lags in incorporating the most recent data and (2) the completeness and plausibility issues in the SACT, LOT, and mortality data could cause faulty determination of treatment discontinuation date and status of censoring.

The fact that Data set B included a lower percentage of patients receiving the target SACT (237/4003, 5.9% vs 1200/4808, 24.96% in Data set A) limited the utility of the data for determining the rwTTD. This finding highlights the need and importance of conducting a rigorous and use case–specific data quality assessment in the planning stage of RWD studies. In addition, for Data set B, findings of extremely low and high gaps between target SACT administration dates would warrant further investigation of each patient's trajectory to verify the specific data quality issue before taking proper data quality improvement actions such as removing the patient or the SACT record as outliers.

Limitations

This study has several limitations that require further discussion. First, adequately assessing the reasons for missingness across different RWD sources is challenging. In particular, the data feeds and capture of elements across different data sources are variable. A lack of transparency and consistency means that different RWD sources are often not fully interoperable [38]. In this study, we applied cohort attrition steps to align populations represented in the 2 data sets and imputed the LOT end date and LOT name that were missing in Data set B. However, a major remaining roadblock was the vendor's privacy-preserving aggregation, which does not allow data sources to be adequately reviewed on more granular level to understand the reason behind missing data, data quality issues, or data discrepancies.

Second, the implementation of data quality checks for new RWD sources, especially for those with data table structures that differ from those of prior data sets, requires customization and reconfiguration that are often time consuming. We are developing a data dashboard tool that can accelerate this process for both raw data and a common data model such as that of the Observational Health Data Sciences and Informatics [17,18].

Third, use case–specific data quality assessment checks often provide only a limited view of the comparative validity of the RWD under consideration, particularly when a well-recognized gold standard is absent. The paucity of data often limits an effective comparison with the distribution of key data elements in the general population (external validity). In this study, we set a priori metrics for these checks by using domain knowledge such as HNC prevalence [33] and regulatory approval timelines. It would be interesting for future studies to validate and update these metrics.

Comparison With Prior Work

Prior studies have evaluated rwTTD, also known as the duration of therapy and real-world time on treatment, for immuno-oncology agents used in treating recurrent or metastatic HNC [39], advanced non-small cell lung cancer [28,40-42], and other solid cancers [42]. In contrast to this study, these studies drew on research-ready databases (as would be identified in the preassessment step of UReQA [3]), and the actions taken to ensure RWD fitness and quality were limited to aligning patient eligibility criteria (the cohort definition step of UReQA [3]).

New use cases can be created for other medication-related outcomes or therapeutic areas by following the first 3 steps of implementing the rwTTD use case in this study. In addition, the data quality checks that we identified and created for the rwTTD use case can be used for other types of use cases. For example, checks on medication identification and dates can also be used to evaluate the fitness of RWD sources for studying medication adherence. The checks on mortality and follow-up visits could validate the applicability of an RWD source for survival analyses.

Future Work

We selected 2 US EHR-based oncology databases to implement the UReQA use case of rwTTD. These were the only 2 databases the research team had access to that provided both oncology treatment and LOT information during the time of study execution. Each database may have its own bias in representing the overall advanced HNC population in the United States. Future work could implement (1) evaluation of more US EHR-based oncology databases to bring more impactful findings and (2) investigating the associations between rwTTD calculation and quantitative data quality assessment for various medications of interest and cancer types.

Conclusions

The fit-for-purpose quality assessment demonstrated the high level of variability in quality of the 2 real-world data sets for estimating the rwTTD of an SACT for advanced HNC. This study illustrates the application and value of use case–specific data assessment tasks in identifying high-quality RWD for research studies. The data quality specifications supporting this comprehensive use case can be expanded to other use cases in oncology outcomes research. Incorporating such comprehensive data quality assessment could help the study team select the most suitable database in the planning stage of a real-world evidence study. In addition, understanding data quality concerns particularly relevant to research questions can provide additional insights for properly preparing data in full study execution.

Acknowledgments

This work was supported by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States. Medical writing and editorial assistance were provided by Elizabeth V Hillyer, DVM (freelance). This assistance was funded by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States.

Data Availability

The data sets generated during and/or analyzed during this study are not publicly available as they were data vendors' proprietary assets provided to the study team under commercial licenses but are available from the corresponding author on reasonable request and permission from the data vendor.

In addition, we cannot disclose the identities of the data vendors, as doing so would inevitably promote the business of 1 data vendor and may violate data use agreements.

Authors' Contributions

BR, AS, KD, and SC conceptualized and designed the study. BR and AS contributed to data acquisition and data analysis. BR, AS, KD, SC, LY, and SK contributed to the interpretation of results. BR and AS drafted the manuscript. BR, AS, KD, SC, LY, and SK contributed to manuscript revision. All authors approved the publication of the manuscript.

Conflicts of Interest

BR, KD, and SK report employment with Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, NJ, United States, and stock ownership of Merck & Co, Inc, Rahway, NJ, United States. AS reports employment with Real World Evidence, Epidemiology, Medical Affairs and Value Statistics (REM) Data Science department, Jazz Pharmaceutical. SC reports employment with ConcertAI. LY reports employment with and ownership of Polygon Health Analytics LLC. AS, SC, and LY were employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, NJ, United States, when they worked on this study.

Multimedia Appendix 1
Data checks comprising 20 tasks assessing conformance, completeness, or plausibility.
[DOCX File , 28 KB - medinform_v12i1e47744_app1.docx]

References

1. Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. J Natl Cancer Inst 2017 Nov 01;109(11):1-5. [doi: 10.1093/jnci/djx187] [Medline: 29059439]
2. Miksad RA, Samant MK, Sarkar S, Abernethy AP. Small but mighty: the use of real-world evidence to inform precision medicine. Clin Pharmacol Ther 2019 Jul;106(1):87-90 [FREE Full text] [doi: 10.1002/cpt.1466] [Medline: 31112289]
3. Desai KD, Chandwani S, Ru B, Reynolds MW, Christian JB, Estiri H. Fit-for-purpose real-world data assessments in oncology: a call for cross-stakeholder collaboration. Value Outcomes Spotlight 2021 Jun;24:S25 [FREE Full text] [doi: 10.1016/j.jval.2021.04.129]
4. Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of real-world evidence to drive drug development strategy and inform clinical trial design. Clin Pharmacol Ther 2022 Jan;111(1):77-89 [FREE Full text] [doi: 10.1002/cpt.2480] [Medline: 34839524]
5. Lakdawalla DN, Shafrin J, Hou N, Peneva D, Vine S, Park J, et al. Predicting real-world effectiveness of cancer therapies using overall survival and progression-free survival from clinical trials: empirical evidence for the ASCO value framework. Value Health 2017;20(7):866-875 [FREE Full text] [doi: 10.1016/j.jval.2017.04.003] [Medline: 28712615]
6. Framework for FDA's real-world evidence program. U.S. Food & Drug Administration (FDA). 2018. URL: https://www.fda.gov/media/120060/download [accessed 2023-03-10]
7. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. U.S. Department of Health and Human Services. 2021. URL: https://www.fda.gov/media/152503/download [accessed 2023-03-10]
8. Submitting documents using real-world data and real-world evidence to FDA for drug and biological products: guidance for industry. U.S. Food & Drug Administration (FDA). 2022 Sep. URL: https://www.regulations.gov/document/FDA-2019-D-1263-0014 [accessed 2023-03-10]
9. Snapshot: healthcare data ecosystem (2023). DATAVANT. URL: https://datavant.com/health-data-ecosystem/ [accessed 2023-10-13]

10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
11. Assessing data quality for healthcare systems data used in clinical research. NIH Pragmatic Trials Collaboratory. 2014. URL: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf [accessed 2023-03-10]
12. Franklin JM, Liaw KL, Iyasu S, Critchlow CW, Dreyer NA. Real-world evidence to support regulatory decision making: new or expanded medical product indications. Pharmacoepidemiol Drug Saf 2021 Jun;30(6):685-693. [doi: 10.1002/pds.5222] [Medline: 33675248]
13. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
14. Mahendraratnam N, Silcox C, Mercon K, Kroetsch A, Romine M, Harrison N, et al. Determining real-world data?s fitness for use and the role of reliability: Duke-Margolis Center for Health Policy. Duke Margolis Center for Health Policy. 2019. URL: https://healthpolicy.duke.edu/publications/determining-real-world-datas-fitness-use-and-role-reliability [accessed 2023-03-10]
15. Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, et al. A comparison of data quality assessment checks in six data sharing networks. EGEMS (Wash DC) 2017 Jun 12;5(1):8 [FREE Full text] [doi: 10.5334/egems.223] [Medline: 29881733]
16. Reynolds MW, Bourke A, Dreyer NA. Considerations when evaluating real-world data quality in the context of fitness for purpose. Pharmacoepidemiol Drug Saf 2020 Oct;29(10):1316-1318 [FREE Full text] [doi: 10.1002/pds.5010] [Medline: 32374042]
17. OHDSI: data quality dashboard. GitHub. URL: https://github.com/OHDSI/DataQualityDashboard [accessed 2023-03-10]
18. Home page. Observational Health Data Sciences and Informatics (OHDSI). URL: https://www.ohdsi.org/ [accessed 2023-03-10]
19. PEDSnet/data quality analysis. GitHub. URL: https://github.com/PEDSnet/Data-Quality-Analysis/blob/master/Data/DQACatalog/DQA_Check_Type_Inventory.csv [accessed 2023-03-10]
20. Home page. The National Patient-Centered Clinical Research Network (PCORnet). URL: https://pcornet.org/ [accessed 2023-03-10]
21. DQUEEN v 0.5 (data QUality assEssmENt and managing tool). GitHub. URL: https://github.com/ABMI/DQUEEN_OMOP_CDM_Version [accessed 2023-03-10]
22. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Med Inform Decis Mak 2019 Jul 23;19(1):142 [FREE Full text] [doi: 10.1186/s12911-019-0852-6] [Medline: 31337390]
23. Estiri H, Murphy SN. Semi-supervised encoding for outlier detection in clinical observation data. Comput Methods Programs Biomed 2019 Nov;181:104830 [FREE Full text] [doi: 10.1016/j.cmpb.2019.01.002] [Medline: 30658851]
24. Estiri H, Klann JG, Weiler SR, Alema-Mensah E, Joseph Applegate R, Lozinski G, et al. A federated EHR network data completeness tracking system. J Am Med Inform Assoc 2019 Jul 01;26(7):637-645 [FREE Full text] [doi: 10.1093/jamia/ocz014] [Medline: 30925587]
25. Huser V. Facilitating analysis of measurements data though stricter model conventions: exploring units variability across sites. Observational Health Data Sciences and Informatics. 2017. URL: https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:huser-2017-ohdsi-symp-units.pdf [accessed 2023-03-10]
26. Blumenthal GM, Gong Y, Kehl K, Mishra-Kalyani P, Goldberg KB, Khozin S, et al. Analysis of time-to-treatment discontinuation of targeted therapy, immunotherapy, and chemotherapy in clinical trials of patients with non-small-cell lung cancer. Ann Oncol 2019 May 01;30(5):830-838 [FREE Full text] [doi: 10.1093/annonc/mdz060] [Medline: 30796424]
27. Establishing a framework to evaluate real-world endpoints. Friends of Cancer Research. 2018. URL: https://friendsofcancerresearch.org/wp-content/uploads/RWE_FINAL-7.6.18_1.pdf [accessed 2023-03-10]
28. Stewart M, Norden AD, Dreyer N, Henk HJ, Abernethy AP, Chrischilles E, et al. An exploratory analysis of real-world end points for assessing outcomes among immunotherapy-treated patients with advanced non-small-cell lung cancer. JCO Clin Cancer Inform 2019 Jul;3:1-15 [FREE Full text] [doi: 10.1200/CCI.18.00155] [Medline: 31335166]
29. Meng W, Ou W, Chandwani S, Chen X, Black W, Cai Z. Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer. J Biomed Inform 2019 Dec;100:103335 [FREE Full text] [doi: 10.1016/j.jbi.2019.103335] [Medline: 31689549]
30. Meng W, Mosesso KM, Lane KA, Roberts AR, Griffith A, Ou W, et al. An automated line-of-therapy algorithm for adults with metastatic non-small cell lung cancer: validation study using blinded manual chart review. JMIR Med Inform 2021 Oct 12;9(10):e29017 [FREE Full text] [doi: 10.2196/29017] [Medline: 34636730]
31. RxNorm. NIH National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/rxnorm/index.html [accessed 2023-03-10]
32. RxNav-in-a-box. NIH National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/RxNav/applications/RxNav-in-a-Box.html [accessed 2023-03-10]

33. Cancer stat facts: oral cavity and pharynx cancer. National Cancer Institute: Surveillance, Epidemiology, and End Results Program (SEER). 2022. URL: https://seer.cancer.gov/statfacts/html/oralcav.html [accessed 2023-03-10]
34. Nadler E, Joo S, Boyd M, Black-Shinn J, Chirovsky D. Treatment patterns and outcomes among patients with recurrent/metastatic squamous cell carcinoma of the head and neck. Future Oncol 2019 Mar;15(7):739-751. [doi: 10.2217/fon-2018-0572] [Medline: 30511880]
35. Grünwald V, Chirovsky D, Cheung WY, Bertolini F, Ahn MJ, Yang MH, et al. Global treatment patterns and outcomes among patients with recurrent and/or metastatic head and neck squamous cell carcinoma: results of the GLANCE H and N study. Oral Oncol 2020 Mar;102:104526 [FREE Full text] [doi: 10.1016/j.oraloncology.2019.104526] [Medline: 31978755]
36. Mody MD, Rocco JW, Yom SS, Haddad RI, Saba NF. Head and neck cancer. Lancet 2021 Dec 18;398(10318):2289-2299. [doi: 10.1016/S0140-6736(21)01550-6] [Medline: 34562395]
37. Mourad M, Jetmore T, Jategaonkar AA, Moubayed S, Moshier E, Urken ML. Epidemiological trends of head and neck cancer in the united states: a SEER population study. J Oral Maxillofac Surg 2017 Dec;75(12):2562-2572 [FREE Full text] [doi: 10.1016/j.joms.2017.05.008] [Medline: 28618252]
38. Penberthy LT, Rivera DR, Lund JL, Bruno MA, Meyer AM. An overview of real-world data sources for oncology and considerations for research. CA Cancer J Clin 2022 May;72(3):287-300 [FREE Full text] [doi: 10.3322/caac.21714] [Medline: 34964981]
39. Ramakrishnan K, Liu Z, Baxi S, Chandwani S, Joo S, Chirovsky D. Real-world time on treatment with immuno-oncology therapy in recurrent/metastatic head and neck squamous cell carcinoma. Future Oncol 2021 Aug;17(23):3037-3050 [FREE Full text] [doi: 10.2217/fon-2021-0360] [Medline: 34044594]
40. Waterhouse D, Lam J, Betts KA, Yin L, Gao S, Yuan Y, et al. Real-world outcomes of immunotherapy-based regimens in first-line advanced non-small cell lung cancer. Lung Cancer 2021 Jun;156:41-49 [FREE Full text] [doi: 10.1016/j.lungcan.2021.04.007] [Medline: 33894493]
41. Horvat P, Gray CM, Lambova A, Christian JB, Lasiter L, Stewart M, et al. Comparing findings from a friends of cancer research exploratory analysis of real-world end points with the cancer analysis system in England. JCO Clin Cancer Inform 2021 Dec;5:1155-1168 [FREE Full text] [doi: 10.1200/CCI.21.00013] [Medline: 34860576]
42. Torres AZ, Nussbaum NC, Parrinello CM, Bourla AB, Bowser BE, Wagner S, et al. Analysis of a real-world progression variable and related endpoints for patients with five different cancer types. Adv Ther 2022 Jun;39(6):2831-2849 [FREE Full text] [doi: 10.1007/s12325-022-02091-8] [Medline: 35430670]

Abbreviations

EHR: electronic health record
HNC: head and neck cancer
ICD: International Classification of Diseases
LOT: line of therapy
Q3: third quarter
RWD: real-world data
rwTTD: real-world time to treatment discontinuation
SACT: systemic anticancer therapy
UReQA: Use Case Specific Relevance and Quality Assessment

Edited by Q Chen; submitted 30.03.23; peer-reviewed by HJ Kim, S Setia, T Royce; comments to author 30.06.23; revised version received 30.11.23; accepted 14.01.24; published 06.03.24.

Please cite as:
Ru B, Sillah A, Desai K, Chandwani S, Yao L, Kothari S
Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study
JMIR Med Inform 2024;12:e47744
URL: https://medinform.jmir.org/2024/1/e47744
doi:10.2196/47744
PMID:38446504

Distributed Statistical Analyses: A Scoping Review and Examples of Operational Frameworks Adapted to Health Analytics

Félix Camirand Lemyre[1,2,*], PhD; Simon Lévesque[1,2,3,*], MSc; Marie-Pier Domingue[1,2,4], BSc; Klaus Herrmann[2], PhD; Jean-François Ethier[1,3,5], MD, PhD

1

2

3

4

5

*these authors contributed equally

Corresponding Author:
Jean-François Ethier, MD, PhD

Abstract

Background: Data from multiple organizations are crucial for advancing learning health systems. However, ethical, legal, and social concerns may restrict the use of standard statistical methods that rely on pooling data. Although distributed algorithms offer alternatives, they may not always be suitable for health frameworks.

Objective: This study aims to support researchers and data custodians in three ways: (1) providing a concise overview of the literature on statistical inference methods for horizontally partitioned data, (2) describing the methods applicable to generalized linear models (GLMs) and assessing their underlying distributional assumptions, and (3) adapting existing methods to make them fully usable in health settings.

Methods: A scoping review methodology was used for the literature mapping, from which methods presenting a methodological framework for GLM analyses with horizontally partitioned data were identified and assessed from the perspective of applicability in health settings. Statistical theory was used to adapt methods and derive the properties of the resulting estimators.

Results: From the review, 41 articles were selected and 6 approaches were extracted to conduct standard GLM-based statistical analysis. However, these approaches assumed evenly and identically distributed data across nodes. Consequently, statistical procedures were derived to accommodate uneven node sample sizes and heterogeneous data distributions across nodes. Workflows and detailed algorithms were developed to highlight information sharing requirements and operational complexity.

Conclusions: This study contributes to the field of health analytics by providing an overview of the methods that can be used with horizontally partitioned data by adapting these methods to the context of heterogeneous health data and clarifying the workflows and quantities exchanged by the methods discussed. Further analysis of the confidentiality preserved by these methods is needed to fully understand the risk associated with the sharing of summary statistics.

(JMIR Med Inform 2024;12:e53622) doi:10.2196/53622

KEYWORDS

distributed algorithms; generalized linear models; horizontally partitioned data; GLMs; learning health systems; distributed analysis; federated analysis; data science; data custodians; algorithms; statistics; synthesis; review methods; searches; scoping

Introduction

Health Analytics at Scale

Learning health systems (LHSs) are coming of age and are being deployed to address important health challenges at different scales. The framework starts by leveraging health data created across various activities. It obviously includes data points from clinics and hospitals, but the perimeter of data required to meaningfully and optimally address important problems is much wider and includes research cohorts, biobanks, quantified self-data, environmental exposures, and social service delivery.

While some questions might be addressed at the scale of an individual organization, LHSs focus on system interactions and often require the analysis of processes and outcomes from various organizations. For example, to fully understand a cancer care trajectory, multiple data sources from multiple organizations will need to be examined to cover all relevant aspects (both within the traditional health system and in the community). This often implies organizations that are at least regional or of a wider scope (provinces, states, and countries),

such as in the context of the Health Data Research Network Canada or Health Data Research United Kingdom. Similarly, comparing various approaches is often a fruitful way to identify the best approaches and understand what works, why, and how to scale the promising projects. It can also be a way to amass a critical number of observations in the context of rarer diseases. Nevertheless, working with data from multiple sources, from multiple organizations, and located in multiple jurisdictions poses significant challenges.

Traditionally, the analytical methods used by researchers in health-related and other domains have relied on data pooling (sometimes referred to as data centralization)—all required data are physically copied to a single location where analysis can take place. However, when working with data from multiple jurisdictions (even when part of the same country, such as the Canadian provinces and territories), data pooling is often very difficult, if not impossible, for ethical, legal, and social acceptability reasons.

Therefore, there is a pressing need to offer analytical methods allowing for the analysis of such data without the need to physically copy the data in a central location.

The primary intent of this study was to lay the foundations for future practical assessments of the feasibility of conducting distributed statistical analyses in health-related contexts. We achieved this by reviewing the literature on existing methods, evaluating which ones could be applied to a class of widely used regression models, and precisely identifying their operational and information sharing requirements in a notational framework that allows for straight comparisons. This unified framework enables the description of methods' operational workflows, quantities exchanged, and algorithmic implementation and operates under assumptions commonly satisfied in health analytics.

This paper is structured as follows. We begin with a formal description of the distributed analytical framework considered in this study, followed by a discussion on the challenges associated with its implementation in health analytics. Next, we state our specific objectives and outline the methodology used to achieve them, including a scoping review and our approach to establishing the common notational framework for method description and comparison. After presenting the results pertaining to each objective, we discuss our findings and remaining challenges regarding distributed statistical analyses for health analytics.

Distributed Analysis

Overview

This study was concerned with frameworks in which the data needed for a statistical analysis consisted of the data about n individuals (referred to as the *analytical data set*), which are not all stored in a single source but are partitioned among K locations that will be called *nodes* hereafter. Therefore, the mereological sum of all the data held at each node forms the analytical data set. Data can be partitioned horizontally or vertically (or in a mixed way).

A horizontal partition implies that all data pertaining to a given individual can be found in a single node. If we assume that patients receive care only in 1 province, Canadian provincial health administrative data sets hosted by organizations such as Population Data BC, the Institute for Clinical Evaluative Sciences in Ontario, or the Manitoba Centre for Health Policy in Manitoba can be part of a horizontal partition. A clinical trial in which each recruiting site captures all data for a given participant is another example.

A vertical partition occurs when all data of a certain type are available in a single node for a group of individuals. A classic example is a hospital with its various information systems. All pathology results can be found in the pathology system, all billing information can be extracted from the finance system, and all x-rays are accessible in the picture archiving and communication system. However, to obtain the full picture of the care received by a patient, multiple systems need to be interrogated. Similarly, in the research setting, health administrative data may be in a provincial data center, and genomics data could be held in a research institute.

A mixed partition occurs when both principles partly apply—some individuals may have their data spread out across nodes, and different individuals may be present in different nodes.

Assumptions

The difficulties in conducting analyses on a large scale mentioned previously are often associated with horizontally partitioned data, and this work focused on this type of partition. Therefore, the methods presented in this paper might not be directly applicable to vertically partitioned data.

One group of approaches often labeled as *distributed analysis* involves calculations at each participating node and exchanges of the resulting aggregated statistics with a *coordinating center* (CC), which can itself also perform additional calculations based on the received aggregated statistics. The CC can be an organization not responsible for a data node or a data node taking on the additional role of CC for a given analysis.

It is important to note that, whether in the more traditional way of data pooling or using distributed approaches (in which the data are not copied centrally), data sources will be different on multiple levels. They will represent information using data models with significant variability in terms of structure and technology but also in terms of semantics. This situation also leads to heterogeneous data in which the presence of predictors and outcomes is likely to be different in different nodes. Different approaches (eg, data mediation or extract, transform, and load) have been developed to address these issues, and this work assumed that one of them was applied so that the data nodes mentioned hereafter are assumed to share the same structure, the same technological syntax, and the same semantics, as well as no missing data.

Horizontally Partitioned Statistical Analytics

In what follows, the field that pertains to the statistical analysis of horizontally partitioned and semantically homogeneous data

that cannot be consolidated into a central location will be called *horizontally partitioned statistical analytics* (HPSA).

Methodological contributions to this field have arisen from several streams of literature. Meta-analysis and meta-regression methods [1] can be viewed as part of HPSA (eg, by considering that each node-specific data set belongs to a different pseudostudy). However, their scope is narrower compared to that of HPSA because they typically assume that only established study-level estimates are available as data. Conversely, HPSA allows for the sharing of additional summary statistics between the nodes and the CC, such as gradients and Hessians, to ensure the best possible performance at the global level. As meta-analysis does not leverage any supplementary information that could be obtained from studies with access to patient-level data, it can be susceptible to biased estimation, especially in settings with rare outcomes or in the presence of data nodes with limited sample sizes [2]. As meta-analysis and meta-regression methods have been extensively covered in the literature, approaches specifically designed for the analysis of already established study-level estimates will not be discussed hereafter.

An important research community that has generated a significant amount of analytical contributions is concerned with the massive data setting. There, a data set often cannot be processed by a single server and, therefore, is split across multiple machines, which are then considered as nodes able to perform computations and send aggregated results to a CC tasked with fitting a global model from them. The methodological avenues proposed in this setting share similarities with the ones designed for the multi–research facility setting involved in LHSs but also have important differences. For example, in a massive data setting, the experimenter has control over the distribution of individuals across nodes, which is typically not the case in multi–research facility studies. Thus, while these approaches share mechanistic similarities and have been suggested as options to consider in the health domain, some hypotheses may not hold. In regression settings, it is often reasonable to assume that the regression link between the response and covariate predictors is the same across nodes. However, assuming that the sampling distribution of covariates involved is equal across nodes is unrealistic in the health domain, particularly due to the presence of data centers that may systematically involve different types of patients. For example, certain clinics may predominantly serve older individuals. While this may not affect the estimation of parameter values, it can have implications for computing CIs to ensure the validity of inferences.

So far, 2 reviews discussing methods applicable to horizontally partitioned data have been published [3,4]. However, their focus is on the massive data setting, which works almost invariably under the assumption of even sampling distribution of covariates and equal sample sizes across nodes, and statistical inference tasks beyond parameter estimation are barely covered. This makes them less helpful for health analytics purposes as most studies involving data analyses rely on CIs or hypothesis testing in settings in which predictors' distribution and sample sizes vary across nodes.

Contemporary Challenges in HPSA

Overview

The problem is 3-fold. First, there is a need to raise awareness regarding the existence of HPSA approaches among researchers aiming at undertaking statistical analyses from horizontally partitioned data, especially in health analytics. The reflex is often to request data pooling because it is perceived as the sole option. This has been the tendency of requests made by researchers to the Health Data Research Network Canada. Practitioners are usually concerned with finding the most appropriate statistical model that will take into account as many of the features of their specific context of application as possible. Consequently, a clear and unifying mapping of the state of the HPSA field is needed for them to be informed of the scope of existing methods available for their analyses to see whether alternatives to pooling exist.

Second, as underlined previously, methodological contributions came from research fields whose working assumptions can be fundamentally different from the ones researchers would be willing to make in health analytics. To ensure proper use of statistical inference techniques, it is necessary that the underlying assumptions of existing methods be adequately identified and understood. If necessary, these methods should be adapted to suit the specific requirements of health applications, thereby ensuring accurate and reliable results.

Third, data custodians have to be properly informed on data sharing requirements entailed by the use of a specific HPSA method applicable to a given research setting. While HPSA avoids the complexities of pooling data, there are still flows of information that have to be acceptable to data stewards. However, even in basic statistical scenarios, available methods are often presented in a way that makes them challenging to compare in terms of information sharing requirements and operational complexity. Therefore, there is a need for clearer and more accessible presentations of these methods to facilitate decision-making regarding data sharing and operational implementation.

Although it would be ideal to offer managers a comprehensive operational workflow for each identified method to evaluate the information shared and execution complexity, with their accompanying underlying modeling assumptions, the abundance and diversity of available approaches make it unfeasible to accomplish this in a single paper. In fact, methods often differ in terms of their targeted application beyond their distributed aspect. For example, differences may exist in the studied model (eg, linear, logistic or Cox regression, and additive models), the dimensionality and sparsity of the predictor variable space, the use of regularization or shrinkage, the presence of missingness, confounders, imbalances, and heterogeneity.

Objectives

The objectives of this study were as follows:

1. To identify and map, from the literature, methodological approaches that make it possible to perform CI estimation and hypothesis testing from a horizontally partitioned data set

2. Among the approaches identified, to describe the ones that allow for the conduct of generalized linear model (GLM) analyses and identify their distributional assumptions
3. On the basis of the approaches identified for GLM-based inferences, to present methods adapted to the setting of uneven sampling distributions across nodes and compare them in terms of information sharing requirements and operational complexity

A scoping review methodology was chosen to achieve objective 1 of mapping the state of the field of HPSA that pertains to inference procedures. For our second objective (objective 2), we identified from the articles selected from the literature search the ones that presented a methodological framework for conducting statistical inference procedures from a GLM with horizontally partitioned data. We then used these frameworks to derive and describe GLM estimators that are applicable to horizontally partitioned data sets. For each identified method, we analyzed and reported its communication workflow and the distributional assumptions. For our third objective (objective 3), we first used statistical theory to adapt the identified procedures to the unequal sample size and uneven covariate distribution setting. Algorithms and mathematical expressions for the quantities involved are reported. For conciseness, we present mathematical formulas for estimation procedures of CIs only. Expressions involved for hypothesis testing are similar and can be deduced following the close connection between CIs and hypothesis tests in GLMs (eg, see Agresti [5]).

The mathematical description of the GLM setting considered for this analysis is described in the following section along with the mathematical notations to be used.

Mathematical Foundations and Notation

Notation

In the following, lowercase letters in bold will represent vector-valued quantities, whereas uppercase letters in bold will denote matrices. The jth element of any vector $\boldsymbol{a} \in {}^{p}$ will be denoted as $[\boldsymbol{a}]_j$. Similarly, the entry at position (j, l) of any matrix $\boldsymbol{A} \in {}^{p \times p}$ will be denoted as $[\boldsymbol{A}]_{jl}$. If g is a real-valued and invertible function, we will use $g^{(-1)}$ to represent its inverse. In addition, if f_θ is a real-valued function that depends on a parameter vector θ and is twice continuously differentiable, $\nabla_\theta f_\theta$ and $\nabla^2_\theta f_\theta$ will, respectively, indicate the gradient and Hessian matrix of f_θ with respect to θ.

Model Mathematical Assumptions

A mathematical depiction of the horizontally partitioned data framework studied in this paper is as follows. There are n individuals horizontally partitioned across K data storage nodes. Each node's data set is denoted by $D_k=\{z_{ik}=x_{1ik},\ldots,x_{pik},y_{ik}\}_{i=1}^{n_k}$, where $1 \le k \le K$. Here, z_{ik} represents the measurements on the ith individual at node k, where $y_{i(k)} \in \mathbb{R}$ denotes their response variable and $[x_{1i(k)},\ldots,x_{pi(k)}] \in \mathbb{R}^p$ denotes their covariate vector. The total sample size at node k is denoted by $n^{(k)}$. The combined data set $D_{(1)},\ldots,D_{(K)}$ make up the whole data set without any duplicated individuals, indicating that $\sum_{k=1}^{K} n_k = n$.

Throughout the analysis, it is assumed that each z_{ik} is independent across $1 \le i \le n^{(k)}$ and $1 \le k \le K$ and there are no missing data. In addition, the size of the covariate space (ie, the dimension of $x_{1ik},\ldots,x_{pik}$, which is equal to p representing the number of features to include as predictors in the GLM) is assumed to be low, eliminating the need for regularization or variable selection. Finally, it is assumed that each node possesses a nonnegligible proportion of the whole data set. Specifically, for each $k \in \{1, \ldots, K\}$, the quantity $n^{(k)}/n$ is bounded away from 0 and 1 as the sample size n tends to infinity, denoted as $n^{(k)}/n \to p^{(k)} \in (0,1)$.

Mathematical Description of the GLM Framework

The formulation of the GLM considered in this paper encompasses various commonly used regression models, such as linear regression, logistic regression, Poisson regression, and probit models. It assumes that the density or probability mass function of each response variable (known as the random components) belongs to the exponential family of distributions. Within this formulation, the mean of the response variable is expressed as a function of a linear combination of the corresponding covariate vector. Formally, it assumes that there exist unknown parameters $\beta \in \mathbb{R}^{p+1}$ and $\varphi > 0$ and known model-specific functions b, c, g, and h such that, with $x_{ik}=x_{0ik},x_{1ik},\ldots,x_{pik}$ and $x_{0ik}=1$, $y_{i(k)} \mid x_{i(k)} \sim f(\cdot\,; x_{i(k)}, \beta, \varphi)$, where, for any $\beta=[\beta_0,\beta_1,\ldots,\beta_p] \in \mathbb{R}^{p+1}$ and φ,

$$f(y; x_{i(k)}, \beta, \varphi) = \exp[y h(\beta^{\top} x_{i(k)}) - b\{h(\beta^{\top} x_{i(k)})\}\varphi + c(y,\varphi)] \quad (1)$$

In formula 1, b is such that $b'\{h(\beta^{\top} x_{i(k)})\} = E(y_{i(k)} \mid x_{i(k)}) = g^{-1}(\beta^{\top} x_{i(k)})$, with $b'(x)=\partial b(x)/\partial x$. In this framework, g is called *link function*, the term $h(\beta^{\top} x_{i(k)})$ is usually referred to as the *natural parameter*, and b is referred to as the *cumulant* function. φ is often called the *dispersion parameter* and is either known (eg, with $\varphi=1$) or unknown. When $h(x)=x$ (ie, h is the identity function), the link g is called *canonical*.

The logistic regression model is obtained upon taking $\varphi=1$, $h(x)=x$, $b(x)=\log(1+e^x)$, $c_{y,\varphi}=0$, and $g(x)=\log\{x/(1-x)\}$. The linear regression model with homoscedastic residual error variance φ is derived upon setting $h(x)=x$, $b(x)=x^2/2$, $c_{y,\varphi}=-y^2/2\varphi-\log 2\pi\varphi/2$, and $g(x)=x$. Hence, both the logistic and the linear regression models rely on a canonical link function in the exponential family distribution.

Methods

Methodology Related to Objective 1

Overview

Scoping reviews are well suited to efficiently map key concepts within a research area [6]. They are widely acknowledged for their ability to clarify working definitions and conceptual boundaries in a specific topic or field [7], facilitating a shared understanding among researchers regarding the status of the research area. These considerations make the scoping review methodology well designed to achieve objective 1.

Scoping studies use systematic searches of relevant databases, using specific keywords to define the boundaries of the research

field. However, identifying these keywords can be challenging, particularly when relevant papers are scattered across different research streams or in independent clusters that do not reference each other. To address the risk of overlooking significant methodological contributions due to a limited number of keywords, a snowballing literature search was initially conducted to generate a comprehensive list of keywords related to HPSA. The scoping review then proceeded with a systematic literature search using the identified keywords. It is worth noting that, as the planning of the scoping review is independent of the search approach, the guidelines presented in the work by Arksey and O'Malley [6] are still appropriate.

Methodology Pertaining to the Snowballing Keyword Search

Snowballing is generally used as a literature search method aimed at identifying papers belonging to a given field [8]. It typically consists of three steps: (1) initiate searches in prominent journals or conference proceedings to gather an initial set of papers, (2) conduct a backward review by examining the reference lists of the relevant articles discovered in step 1 (continue iterating until no new papers are found), and (3) perform a forward search by identifying articles that cite the papers identified in the previous steps.

To avoid selection bias, the initial set of papers for the snowballing approach in step 1 is sometimes generated through a search in Google Scholar [9]. The latter strategy was used in this study, too.

As mentioned previously, in this review, the snowballing search strategy was used in preparation for the application of the scoping review protocol with the goal of identifying relevant keywords. Specifically, the starting set of papers was assembled by screening titles and abstracts from the first 50 papers generated through a Google Scholar search using the strings *distributed inference* and *federated inference*. The main inclusion criterion was "presents, applies or discusses a statistical inference method to analyse horizontally partitioned data." The backward and forward snowballing step approaches were then applied.

From the set of keywords found in the selected papers, a list of those relevant to HPSA but not directly associated with any specific method was retained for the scoping review step. It is worth noting that, as the objective of this scoping review was to identify statistical inference methods for horizontally partitioned data, keywords linked to method identifiers had to be excluded from the retained list to avoid preselection bias in the scoping review phase of this project.

The selected keywords that were identified from the snowballing literature search were *distributed algorithms*, *distributed estimation*, *distributed inference*, *distributed learning*, *distributed regression*, *federated inference*, *federated estimation*, *federated learning*, *privacy-protecting algorithm*, *privacy-preserving algorithm*, and *aggregated inference*.

Methodology Pertaining to the Scoping Review

The scoping review methodological framework by Levac et al [10] (see also the work by Arksey and O'Malley [6]) was followed. The steps are briefly described in this section. A detailed protocol is available in Multimedia Appendix 1 [2,4,6,10-19].

We conducted a comprehensive search across 4 bibliographic databases—MEDLINE, Scopus, MathSciNet, and zbMATH—to encompass the interdisciplinary nature of the topic and identify relevant research articles. Our search strategies were based on 2 key concepts: distributed data and statistical inference. In addition to the keywords obtained from the snowballing step, we incorporated terms such as *confidence interval* to target articles focusing specifically on statistical inference. To ensure the inclusion of recent advancements, our search was limited to papers published from 2000 onward. This cutoff date was chosen to account for the emergence of distributed data, the prevalence of massive data sets, and advancements in technology. It was set conservatively to capture any early developed methods and ensure comprehensive coverage of the topic.

After completing the primary search, a 2-stage selection process was used. Initially, 2 authors (MPD and FCL) collaborated to screen all articles identified through the search strategy based on their titles and abstracts. Subsequently, the full texts of the selected articles were independently reviewed by both authors to finalize the selection. This rigorous approach ensured a thorough evaluation of each article's relevance and eligibility for inclusion.

The primary inclusion criterion for the selection process was as follows: *presents a solution for conducting inferential statistics on horizontally partitioned data*. This criterion was used to ensure that the chosen articles specifically addressed methods associated with performing statistical inference on horizontally partitioned data.

The following exclusion criteria were derived directly from objective 1: (1) does not address inferential statistics, including CIs, hypothesis testing, or asymptotic normality; (2) does not provide a methodological contribution; and (3) presents a solution for encryption or secret sharing.

To ensure the inclusion of validated approaches, the selection process only considered published papers that had full-text availability in English or French. Discussion papers were excluded as they do not present novel methods or approaches.

Exclusion was considered if any of the exclusion criteria were met or if any of the inclusion criteria were not met.

Finally, the references of each included article from the databases were assessed to identify any relevant articles that may not have been captured during the initial screening due to specific keywords. This additional step in the selection process was necessary given the broad range of vocabulary used to describe applicable approaches in our context.

Data extraction for the included articles was conducted by one author (MPD) and followed a collectively developed data-charting form. Model type (*parametric regression*, *semiparametric regression*, *nonparametric regression*, or *not specific to regression*) and number of communications from the CC to the nodes (0 or ≥1) were among the data extracted. All

methods from the included articles were subsequently classified according to their specified characteristics, as outlined in the protocol. In addition, as part of the analysis, we conducted a screening of the general distributed approaches commonly used across all specific methods.

Methodology Related to Objective 2

Overview

To achieve objective 2, a total of 3 steps were taken. First, we identified methodological approaches from articles included in this scoping review that enable parameter and CI estimations from horizontally partitioned data within a standard GLM framework. Methods designed specifically for the particular cases of linear or logistic regression were also reported but were not analyzed in detail. Second, we extracted workflows for each approach to determine the information exchanged between data storage nodes and the CC. Third, we analyzed the mathematical assumptions necessary for parameter estimation and the consistency of CI procedures. We specifically reported the assumptions related to the distribution of node-specific covariates.

Identification of the Approaches

To identify approaches that enabled the fitting of any GLM using horizontally partitioned data, 2 authors (FCL and MPD) independently assessed all articles included in this scoping review. The reviewers specifically looked for articles that discussed approaches applicable to the GLM class described in the *Mathematical Foundations and Notation* section, including likelihood-based methods, M-estimation, and estimating equations. In addition, we identified and reported articles that specifically focused on regression settings for linear or logistic regression. However, unless the method described was considered easily adaptable to the GLM framework, these articles were not retained for detailed analysis.

A method was selected if it provided an algorithm for fitting GLMs using horizontally partitioned data, aligning with the characteristics outlined in the *Mathematical Foundations and Notation* section. In cases in which an article presented asymptotic normality results for the estimators but did not provide an estimator for the asymptotic variance-covariance matrix, the article was still retained, and an estimator for the asymptotic variance was derived using the available calculated quantities.

As our GLM framework assumes no missing values, low dimensionality, and a small number of nodes relative to the total sample size, any terms related to these specific conditions mentioned in an article's methodology were disregarded. Consequently, the calculations for CIs were adjusted accordingly. If an article solely focused on one of these aspects without contributing to the overall methodology, it was not included in the final selection.

Methodological components regarding parameter estimation and CI procedures were extracted from the screened articles. Specifically, the focus was on understanding how parameters should be estimated within a horizontally partitioned framework and how CIs should be computed for these parameters. For each article, the formulas related to quantities shared among the nodes and quantities calculated by the CC were derived and analyzed. These formulas were examined within a workflow that indicated the necessary circulation of information for the procedure to be executed. The derived workflows constituted the first part of our defined unified framework for HPSA approach comparisons in GLM settings.

For the reported results, the rationale behind each method that was deemed suitable for fitting GLMs was documented, along with the corresponding reference to the paper included in this scoping review in which the method was introduced or discussed.

Articles that discussed approaches specifically applicable to the cases of linear or logistic regression were also mentioned but not elaborated on in detail.

Methodology Related to Objective 3

In most statistical settings with horizontally partitioned data, it is commonly assumed that the sample sizes of the data nodes are equal and that the distribution of covariates is the same across all nodes. However, when the number of nodes is fixed and relatively small compared to the sample sizes, it is possible to adapt a particular approach to handle situations in which the sample sizes and covariate distributions vary across nodes. This can be achieved by combining the theoretical arguments presented in the original article on the method in question with the principles of asymptotic statistics theory concerning maximum likelihood estimation.

To adapt a given approach for situations in which sample sizes and covariate distributions differ across nodes, the following steps were taken:

1. The formulas for the relevant quantities were modified to emphasize the changes caused by this scenario. It was ensured that the adapted quantities were equivalent to their counterparts presented in the original article for an equal sample size setting.
2. Using asymptotic theory, an asymptotic normality result was derived for the estimators of interest considering a set of assumptions that accommodated potential variations in sample sizes and covariate sampling distribution across nodes while still enabling meaningful theoretical arguments.
3. Formulas for the asymptotic variances were derived. Statistical theory on maximum likelihood estimation was used to obtain consistent estimators for asymptotic variances. The latter estimators were derived under the constraint that they had to be calculated without requiring any additional communication round between the CC and the nodes. Thus, throughout the adaptation process, the communication workflow remained unchanged compared to the original method.

These steps ensured the mathematical correctness of adapting the approaches to handle different sample sizes and covariate distributions across nodes. Importantly, these adaptations maintained consistency with the original method's communication workflows. To adapt and compare the methods based on their information sharing requirements, common assumptions and a unified mathematical and algorithmic notation

were necessary. Ultimately, these assumptions and the notation, along with the workflow types derived for objective 2, enabled us to establish a unified notational framework for approach comparisons when performing HPSA based on GLMs in the context of health analytics.

We describe the statistical estimates of interest. A standard GLM typically includes 1 or 2 unknown parametric components. The first are the β parameters, which are commonly assumed to be unknown. The second component is the nuisance parameter φ, which can be either known (eg, in logistic models) or unknown (eg, in linear models). In practical applications, when φ is unknown, its estimated value is often not the main focus, although the latter is necessary to estimate the asymptotic variance of the β parameter estimates.

In the upcoming analysis, we will assume that the parameter φ is unknown and estimated using the recommended approach in the selected methods. However, in cases in which φ is known, the process becomes simpler. This involves substituting the known value of φ and disregarding the estimation step. It is important to highlight that estimating φ requires additional information to be shared between the nodes and the CC but it does not necessitate any extra communication round between them.

The estimation processes for both the β and φ parameters are discussed. In addition, we explain how to compute an estimator for the asymptotic variance specifically for the estimator of β. It is important to note that the results presented in the following section can be modified and extended to develop a similar procedure for estimating φ.

Using these results, based on an estimator of β (eg, $\hat{\beta}$) and a formula for the estimator of the asymptotic variance-covariance matrix involved in its associated asymptotic normality result (eg, $\hat{\Sigma}$), Wald-type $(1-\alpha)$ CIs can be computed for each component of β using the following formula:

$$[\hat{\beta}]_j \pm z_{1-\alpha/2}\sqrt{[\hat{\Sigma}]_{jj}/n} \text{ for } j \in \{1,\ldots,p+1\}$$

Regarding the reported results, for each approach considered, we present the formulas necessary to compute the final estimates of the β parameters and their corresponding CIs. The presentation of these formulas was designed to emphasize the communication workflow. Furthermore, a comprehensive algorithm is provided outlining the step-by-step process.

In addition, the asymptotic normality of the β parameter estimators is stated accompanied by the formula for the asymptotic variance and its consistent estimator. Detailed proofs for these results can be found in Multimedia Appendix 2 [20-22].

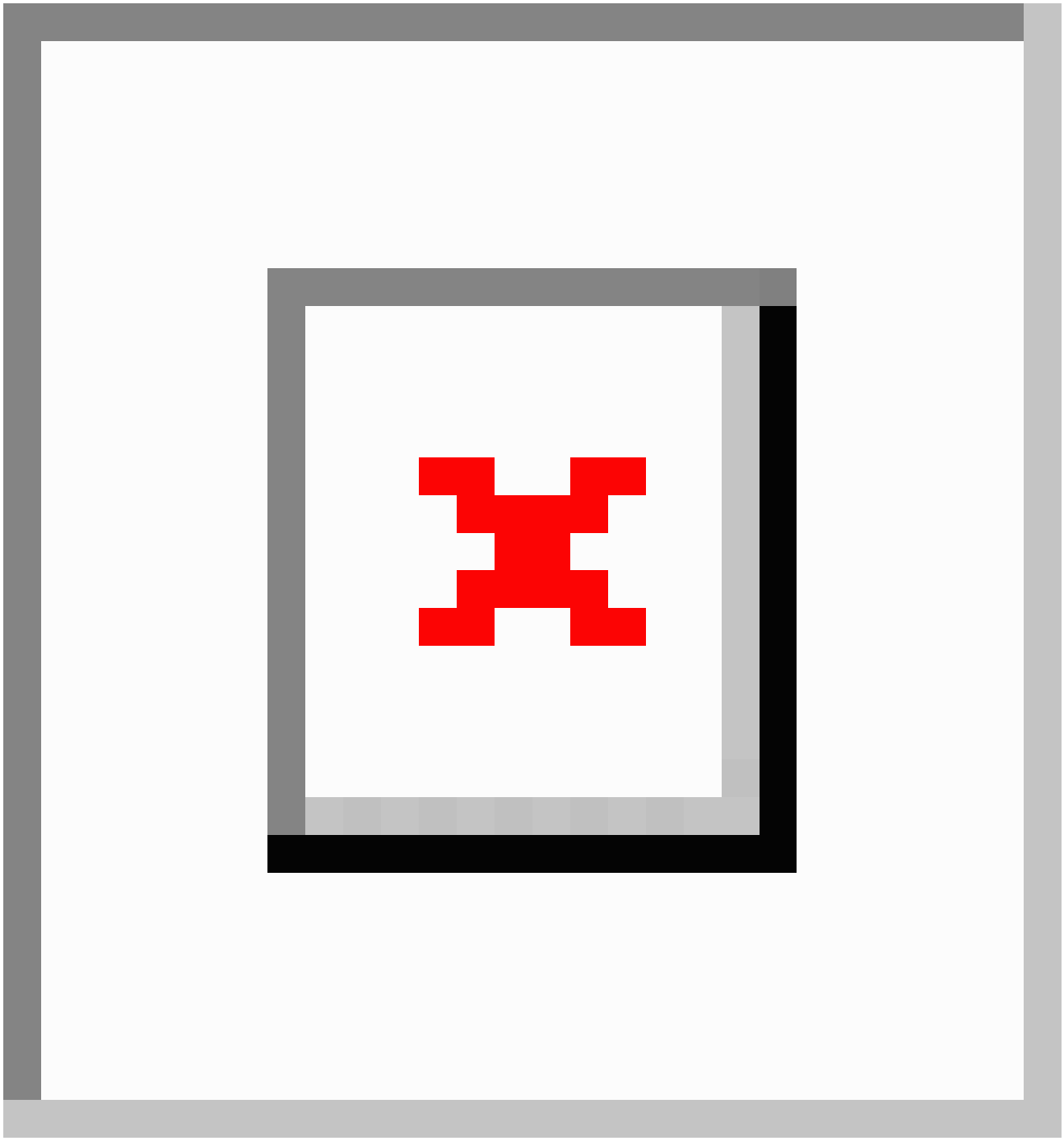
Results

Results Related to Objective 1

Search Outcomes From the Scoping Review

As presented in Figure 1, a total of 1407 articles were initially identified across all 4 databases after removing duplicates. Subsequently, most of these articles (1274/1407, 90.55%) were excluded based on the evaluation of titles and abstracts, leaving 9.45% (133/1407) of the articles for eligibility assessment through full-text review. Following this assessment, 29 articles were included from the databases. In addition, by reviewing the references of the included articles, 12 more articles were identified and added to the study.

Figure 1. Article selection process for the scoping review. Detailed inclusion and exclusion criteria are described in the text and in the protocol.



Regarding the additional 12 articles obtained through the assessment of references of the included articles, it was observed that most of them did not mention statistical inference or related terms in their abstracts [2,11,23]. Consequently, these articles were not captured in the initial database search results. Furthermore, some articles directly referred to the specific method used without including any keywords related to horizontally partitioned data in their abstracts or titles [24,25], which greatly reduced the chance of initially identifying them. However, during the process of reviewing the references of the included articles, all the relevant papers that were initially identified through the snowballing strategy were eventually retrieved either through the search strategy or the selection process based on the references of the included articles.

Results of the Scoping Review

Each article included in this scoping review put forth one or multiple methodological approaches pertaining to objective 1. The similarities and differences regarding the communication schemes involved and their background of origin are summarized in this section.

First, all the selected articles discussed one or more statistical procedures that operate on horizontally partitioned data using one of the communication schemes depicted in Figures 2-5.

Figure 2. Workflow I: each node calculates summary statistics from its own samples. Results are sent to the coordinating center, which combines the information provided by each node to produce the final estimates.

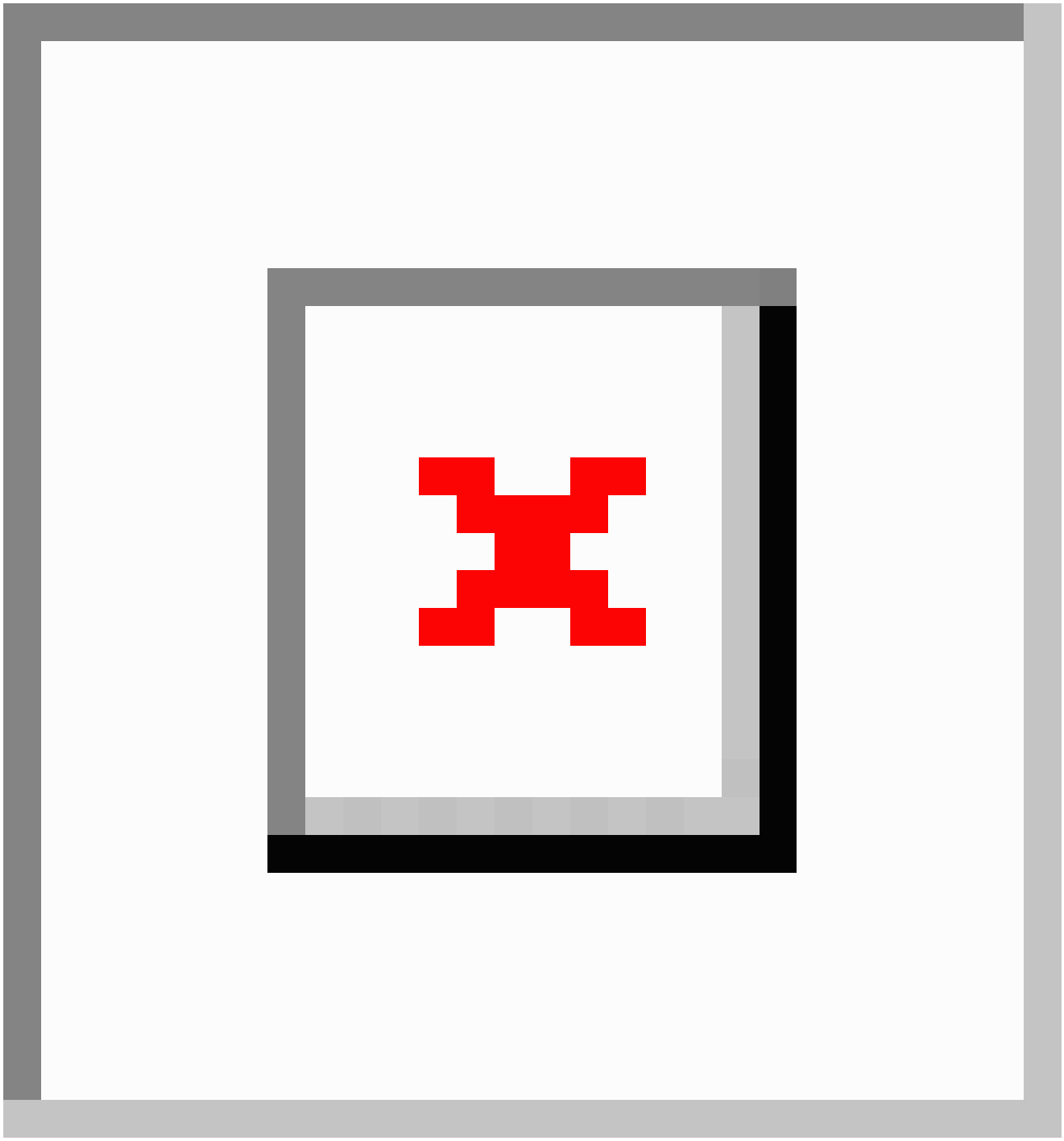


Figure 3. Workflow II: multiple communication rounds are allowed between the coordinating center and the data storage nodes.

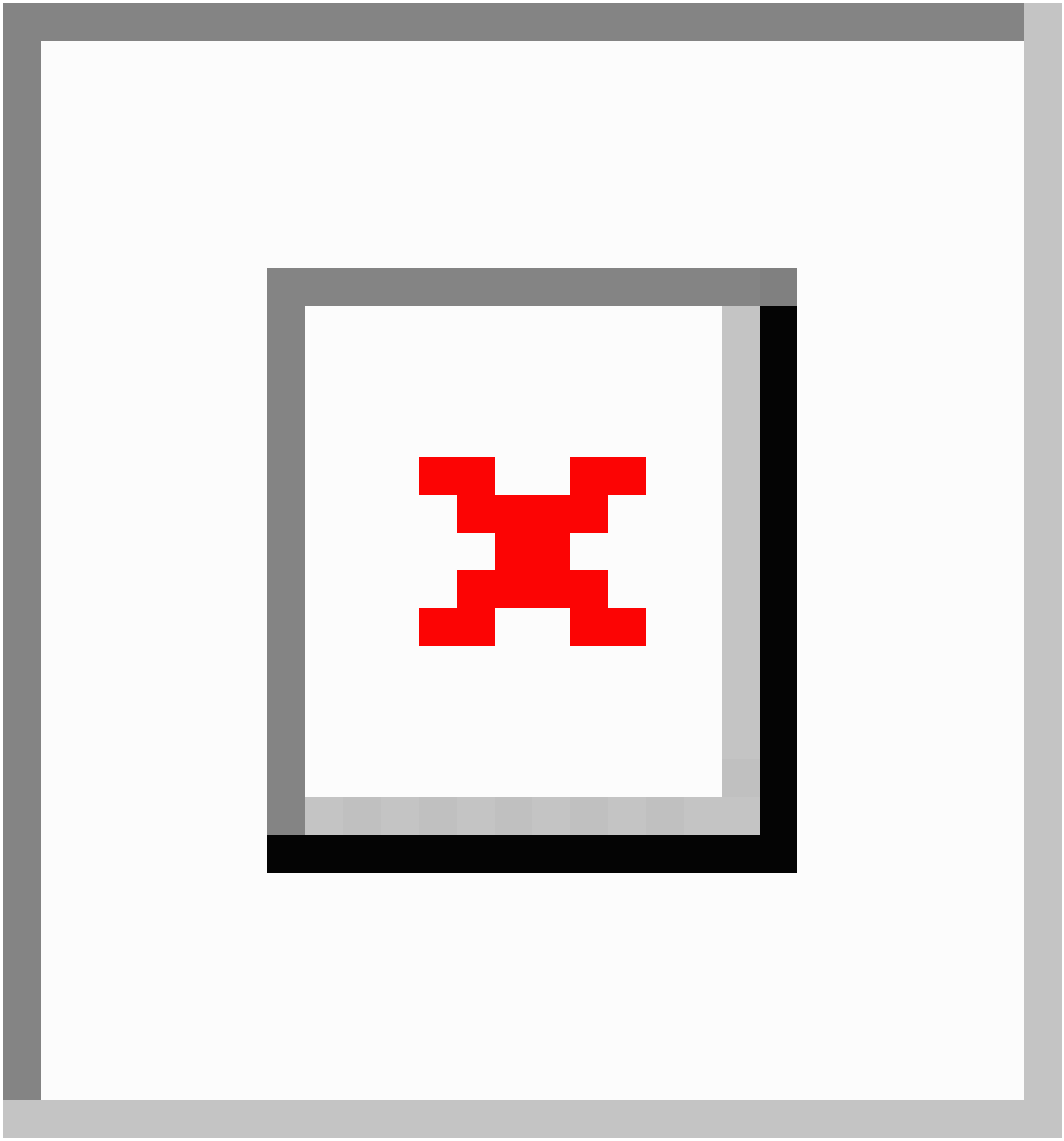


Figure 4. Workflow III: multiple communication rounds are allowed between the coordinating center and the data storage nodes, with node 1 following a distinct communication pattern compared to the other nodes.

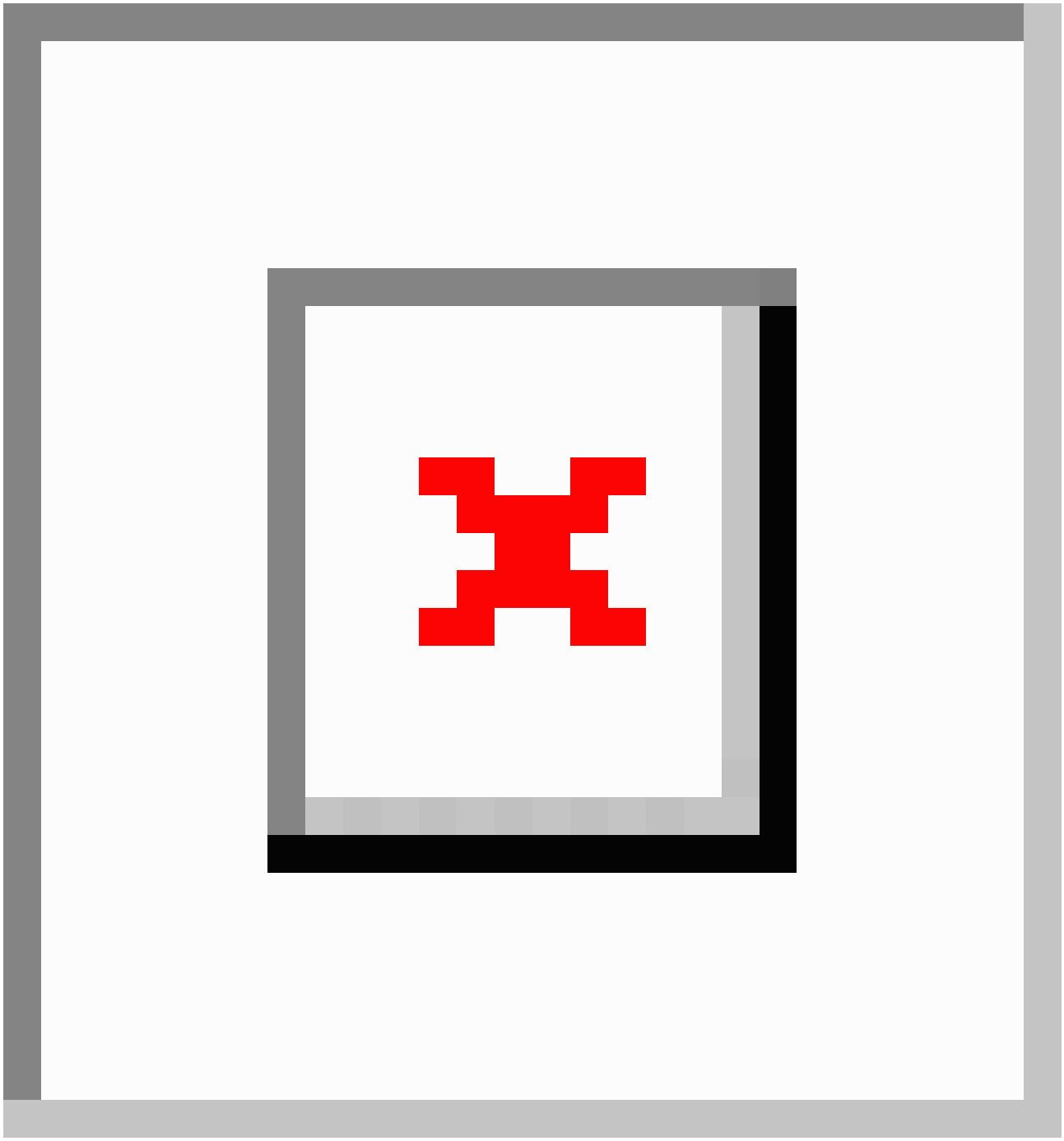
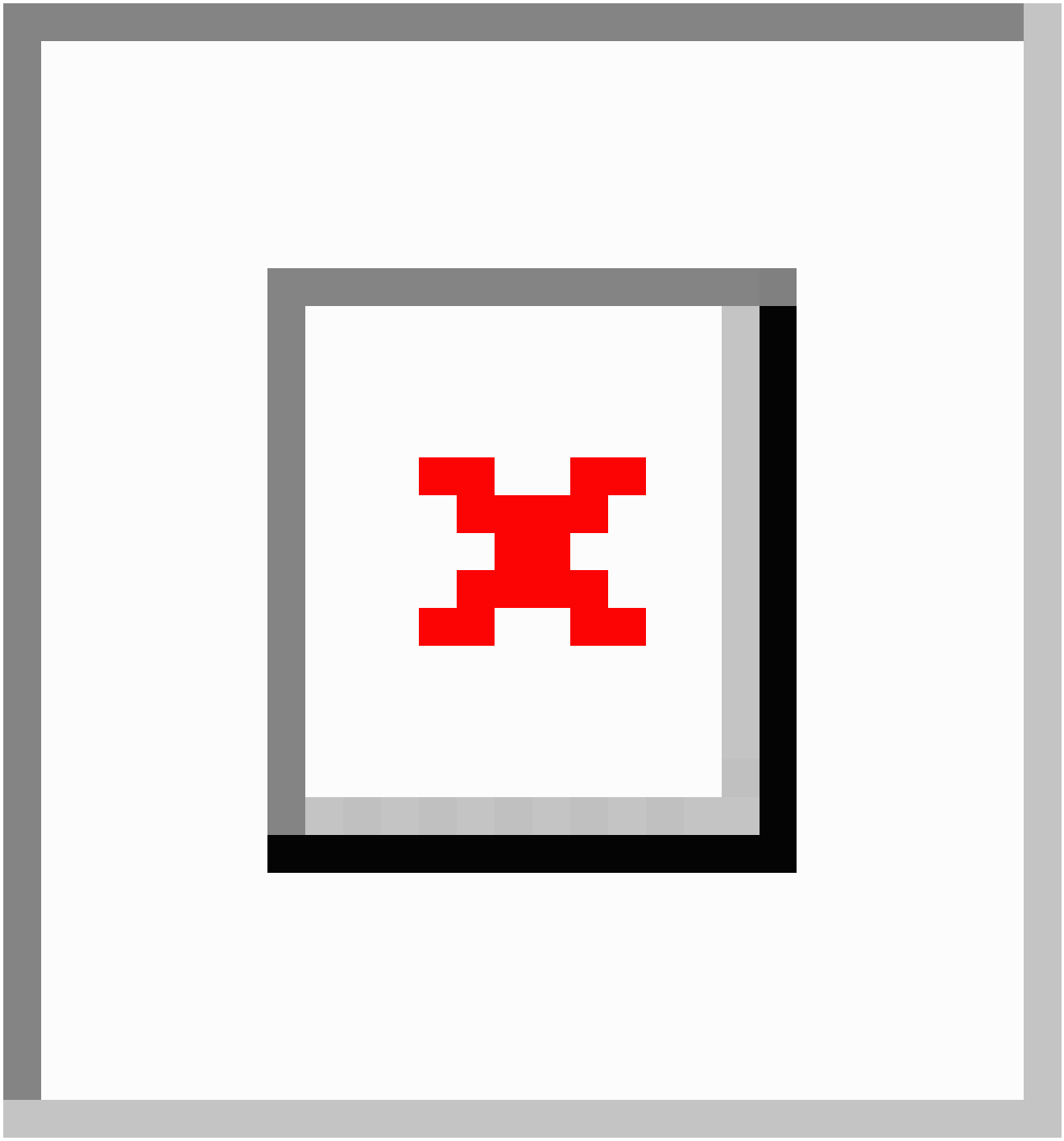


Figure 5. Workflow IV: multiple communication rounds are allowed between the coordinating center (CC) and the data storage nodes, with 2 back-and-forth distinct communication exchanges between each node and the CC at each iteration.



In workflow I, as shown in Figure 2, each node calculates summary statistics from its own samples, and the results are sent to the CC. The CC combines the information provided by each node to produce the final estimates. This communication approach is commonly referred to as "one-shot" or "noniterative" in the literature, although not always consistently.

In workflow II, as shown in Figure 3, multiple communication rounds are allowed between the CC and the data storage nodes. This allows for iterative interactions between the nodes and the CC to refine the estimates.

Some approaches fundamentally differ from the 2 previous workflows by assigning a different role to one of the nodes (eg, node 1) compared to the others. These approaches operate using workflow III, as illustrated in Figure 4, where node 1 follows a distinct communication pattern compared to the other nodes. In the papers included in this scoping review that discussed these approaches, node 1 was invariably designated as the CC. However, in the context of this paper, their roles were differentiated. The additional step performed by the CC, which involves data aggregation, can be particularly well suited for privacy protection purposes in practice.

The particular setting shown in workflow IV in Figure 5 requires 2 back-and-forth communication exchanges between each node and the CC at each iteration. This communication pattern distinguishes this workflow from the others.

In light of the preceding discussion, from an operational standpoint, 2 categories of workflows emerge. On the one hand, there are workflows that do not necessitate any communication from the CC to the nodes, which are captured in workflow I. On the other hand, there are workflows that involve one or more communication exchanges from the CC to the nodes, which are captured in workflows II, III, and IV.

To emphasize similarities among the methods presented in the articles in this review and facilitate the identification of methods suitable for specific purposes, a systematic classification is presented in Table 1. The articles are categorized based on the type of models used and the number of communications from the CC to the individual nodes.

Table . Classification of the articles included in this scoping review.

Type of model	0 communication from CC[a] to nodes	≥1 communication exchange from CC to nodes
Parametric regression	• Basiri et al [26] • Battey et al [27] • Fan et al [28] • Guo et al [29] • Chen and Xie [25] • Lin and Xi [30] • Rosenblatt and Nadler [23] • Zhang et al [31] • Chang et al [32] • Wu et al [33] • Hector and Song [34]	• Huang and Huo [12] • Jordan et al [13] • Mozafari-Majd and Koivunen [35,36] • Yue et al [37] • Duan et al [38] • Duan et al [2] • Tong et al [39] • Di et al [40] • Edmondson et al [41] • Luo and Li [42] • Shu et al [43]
Semiparametric regression	• Zhao et al [44] • Park et al [14]	• Luo et al [45] • Duan et al [11]
Nonparametric regression	• Liu et al [46] • Zhang et al [47] • Volgushev et al [48]	• Wang et al [49]
Not specific to regression	• Atta-Asiamah and Yuan [50] • Minsker [51] • Lin and Xi [52] • Bruce et al [53] • Chen and Peng [54] • Nezakati and Pircalabelu [55] • Banerjee et al [24] • Shi et al [56] • Wu et al [57]	• Lai et al [58]

[a]CC: coordinating center.

Most of the methods were published the big or massive data and multi-machine methodological setting, whereas some were reported within the context of health analytics. Within the big or massive data and multi-machine methodological setting, many methods involved an initial step of random data partitioning among multiple machines. However, certain methods assumed a scenario in which data were already stored on separate machines, as observed in the studies by Fan et al [28] and Jordan et al [13]. Furthermore, it is worth noting that no articles published before 2010 were included, aligning with our initial hypothesis regarding the identification of contemporary methodological settings. Most of the included articles (30/41, 73%) were published after the year 2018.

Most articles (33/41, 81%) addressed a setting in which a CC existed external to the nodes, as exemplified by articles such as those by Lin and Xi [52], Volgushev et al [48], and Yue et al [37]. In contrast, as mentioned previously, some studies (8/41, 19%) designate one of the nodes to assume this central role, as demonstrated in the study by Chang et al [32].

The methods identified through our search strategy shared a common characteristic of using a global model that incorporated population-level parameters. In some cases, these parameters may also include node-specific components to accommodate node-specific statistical heterogeneity in the outcome-predictors relationship, which captures deviations from the population-level conditional probability distribution of the outcome given the predictors.

A few of the reported methods had the capability to yield results identical to those obtained if the individual line data were pooled from all nodes [33,43].

While most articles (31/41, 76%) featured methods related to regression models, including semiparametric and nonparametric designs, a few (10/41, 24%) reported results for other modeling frameworks. These included methods for M-estimation, U-statistics, symmetrical statistics, and natural parameter estimation, some of which encompassed regression models as a specific instance.

Results Related to Objective 2

In total, 6 approaches were selected as applicable to the standard GLM framework. They all assumed that nodes had equal sample sizes and identical distributions for the covariates.

Simple Averaging

One of the simplest methods for horizontally partitioned data analysis, often referred to as the "simple averaging method" or the "divide-and-conquer" approach, has been extensively studied in the literature, for example, the studies by Zhang et al [31] and Shamir et al [59], which were included in our scoping review. This method operates through workflow I in Figure 2. In this approach, node-level model estimates are gathered and averaged at the CC to generate the final estimates.

In the context of GLM, each node is initially tasked with calculating the maximum likelihood estimator (MLE) of the β and φ parameters using their respective data. In addition, the Hessian matrix of the log-likelihood function with respect to the β parameters must be computed for constructing Wald-type CIs. The estimated parameters and the computed Hessian matrix are then transmitted to the CC.

The final parameter estimates of are obtained by averaging the node-specific estimates while the local Hessians and estimates of φ are used to compute an estimator for the asymptotic variance.

Single Distributed Newton-Raphson Updating

The single distributed Newton-Raphson updating method is an iterative procedure that includes an additional communication round between the CC and the nodes compared to the simple averaging method. It was originally proposed as the "distributed one-step" method in the study by Beyan et al [60], but in this study, it is referred to by a different term to avoid any confusion regarding communication complexity. This method operates using workflow II, as depicted in Figure 3, with t=1 (where T represents the number of cycles in the iteration scheme). It enhances the simple averaging estimators by incorporating a single distributed Newton-Raphson updating step.

In the context of GLM, each node first calculates the MLE of β and φ and transmits them to the CC. The CC aggregates these estimates using averaging and sends the result back to the nodes. The nodes then compute the gradient and the Hessian matrix of the log-likelihood function, evaluated at the received β and φ estimates. Subsequently, the gradient and the Hessian matrix are sent back to the CC, which averages them and computes a Newton-Raphson updating step based on the simple averaging estimates. An estimator for the asymptotic variance can be calculated by using the received Hessian matrices and the updated estimate of φ .

Multiple Distributed Newton-Raphson Updatings

The multiple distributed Newton-Raphson updating method leverages the fact that, for standard GLMs, the algorithm typically used to calculate the MLE of β and φ in a centralized pooled setting can be executed in a distributed manner without any loss of information. This is possible because the algorithm relies on Newton-Raphson updatings (or, sometimes, Fisher scoring updatings) that are expressed using 2 sums of node-specific summary statistics, namely, local gradients and local Hessian matrices of the log-likelihood function, evaluated at the β and φ estimates from the previous iteration. A version of this method is proposed in the study by Wu et al [33] under the logistic regression framework. It operates through workflow II in Figure 3 for a general $T \geq 1$.

Distributed Estimating Equation

The class of estimating equation estimators is vast and encompasses a broad range of statistical estimation techniques, including likelihood-based approaches that rely on searching for critical points. The fundamental idea behind estimating equation methods is to establish a system of equations that involve both the sample data and the unknown model parameters. These equations are then solved to determine the parameter estimates. MLEs, which are obtained by setting the gradient of the log-likelihood function with respect to the unknown parameters equal to 0, belong to the class of estimating equation estimators.

The distributed estimating equation approach involves gathering summary statistics from nodes at the CC level, enabling the reconstruction of the estimating equations, or, more commonly, an approximation of them that would have been obtained in a pooled centralized setting. This method is discussed in the study by Lin and Xi [30] and operates using workflow I, as depicted in Figure 2.

In the context of GLMs, the distributed estimating equation approach involves initially assigning each node the task of computing and sending their local MLEs and the Hessian matrix of their local log-likelihood, evaluated at those MLEs, to the CC. The CC uses these received quantities to reconstruct the global estimating equations or an approximation thereof. This reconstruction ultimately leads to an analytical solution for obtaining the resulting estimates. CIs are computed using a combination of the Hessian matrices and the final estimator of φ .

It is important to note that, when this approach is applied in the context of linear regression, it enables the acquisition of β parameter estimates that are identical to those obtained in a pooled centralized setting.

Distributed Estimation Using a Single Gradient-Enhanced Log-Likelihood

This method differs fundamentally from the ones discussed thus far as it involves a distinct role for one particular node in obtaining model parameter estimates. It operates using workflow III, as depicted in Figure 4, and was proposed in the study by Jordan et al [13] under the name "Surrogate likelihood." This approach relies on an approximation of the global likelihood by viewing it as an analytic function. It expands the global likelihood into an infinite series around an initial guess β^SGE,0 and replaces the higher-order derivatives (order of ≥ 2) of the global likelihood with those of a Taylor expansion of a node's (eg, node k=1) local likelihood around the same value. By following this procedure, the so-called surrogate likelihood can be solved using data from node k=1and aggregated gradients from nodes $k \in \{2, ..., K\}$.

In the context of GLM, the CC first collects the necessary information to compute initial estimates for the parameters β and φ . These initial estimates can be obtained through various approaches, such as a simple averaging estimator or the MLEs

computed using data from node 1. These initial estimates are then transmitted to nodes $k \in \{2, ..., K\}$. Each of these nodes calculates the gradient of the log-likelihood function, evaluated at the received estimates, and sends it back to the CC. The CC averages these gradients and sends the result to node 1. Node 1 solves a gradient-enhanced log-likelihood using its own data and the received average gradient. The resulting estimate is sent back to the CC as the final estimate. To compute CIs, each node must send the Hessian matrix of its local log-likelihood function, evaluated at the initial received estimate.

The steps related to estimation can be repeated multiple times.

Distributed Estimation Using Multiple Gradient-Enhanced Log-Likelihoods

This method is in the spirit of the *distributed estimation using a single gradient-enhanced log-likelihood* approach described previously except that all nodes have to solve a gradient-enhanced log-likelihood instead of only one of them. Results pertaining to statistical inference are discussed in the study by Fan et al [28] under a penalized setting. A nonpenalized version of this method was introduced in the study by Shamir et al [59], although the latter did not discuss CIs or hypothesis testing and, hence, was not included in our scoping review. It operates through workflow IV depicted in Figure 5.

The following subsection presents the 6 approaches described in objective 2 within a unified notational framework that accounts for the peculiarities commonly encountered in health analytics. Algorithms were derived using this common notation to rigorously describe the methods and enable their comparison. Theoretical results regarding the estimators involved are detailed and proven in Multimedia Appendix 2. While this section is necessary to increase trust in HPSA methods by transparently showing precisely what information is shared with the CC through comprehensive mathematical formulas, the summary provided in Table 2 suffices for a high-level understanding of the overall picture.

Table . Quantities shared in each adapted method's communication workflow.

Method number	Method	Exchanged quantities from nodes to CC[a]		Exchanged quantities from CC to nodes
		S0(k)	St(k), $t \geq 1$	C_t
1	Simple averaging	β^MLE(k); φ^MLE(k); VMLE(k)	—[b]	—
2	Single distributed Newton-Raphson updating	β^MLE(k); φ^MLE(k)	DNR,1(k); VNR,1(k); ENR,1(k); FNR,1(k)	β^NR,0; φ^NR,0
3	Multiple distributed Newton-Raphson updatings (with *T* Newton-Raphson updatings)	β^MLE(k); φ^MLE(k)	DNR,t(k); VNR,t(k); ENR,t(k); FNR,t(k)	β^NR,t-1; φ^NR,t-1
4	Distributed estimating equations	β^MLE(k); φ^MLE(k); VMLE(k); FMLE(k)	—	—
5	Distributed single gradient-enhanced log-likelihood	β^MLE(k); φ^MLE(k)	Nodes 2 to *K*: DSGE,1(k), VSGE,1(k), and ESGE,1(k); node 1: β^SGE,1, φ^SGE,1, and VSGE,1(1)	β^SGE,0; φ^SGE,0; to node 1 only: DSGE,1 and ESGE,1
6	Distributed multiple gradient-enhanced log-likelihood	β^MLE(k); φ^MLE(k)	DMGE,1(k); VMGE,1(k); EMGE,1(k); β^MGE,1(k); φ^MGE,1(k)	β^MGE,0; φ^MGE,0; D¯MGE,1; E¯MGE,1

[a]CC: coordinating center.

[b]No exchanged quantities.

Results Related to Objective 3

Notation for Shared Quantities

In what follows, let the log-likelihood of the data stored in node k (using $D^{(k)}$) be denoted by

$$l(k)(\beta,\varphi)=1n(k)\sum i=1n(k)\{yi(k)h(\beta xi(k))-b\{h(\beta xi(k))\}\varphi+c(yi(k),\varphi)\}.$$

In addition, let $\mathbf{D}^{(k)}(\beta) \in R^{p+1}$ be such that

$$(2)D(k)(\beta)=1n(k)\sum i=1n(k)xi(k)h'(\beta xi(k))[yi(k)-b'\{h(\beta xi(k))\}]$$

and define the $(p+1)\times(p+1)$ matrix $\mathbf{V}^{(k)}(\beta)$ as

$$(3)V(k)(\beta)=1n(k)\sum i=1n(k)xi(k)xi(k)h(\beta xi(k))2b'\{h(\beta xi(k))\}-h(\beta xi(k))yi(k)-b'\{h(\beta xi(k))\}$$

As $D(k)(\beta)=\varphi\nabla\beta l(k)(\beta,\varphi)$, solving the equation $D(k)(\beta)=0$ yields the node-specific MLE of β, denoted hereafter using β^MLE(k). The matrix $V(k)(\beta)$ is equal to $-\nabla\beta D(k)(\beta)$ and relates to the Fisher information matrix through the equation $V(k)(\beta)=-\varphi\nabla\beta 2l(k)(\beta,\varphi)$.

Finally, set

$$(4)E(k)(\varphi,\beta)=1n(k)\sum i=1n(k)yi(k)h(\beta xi(k))-b\{h(\beta xi(k))\}-\varphi 2n(k)\sum i=1n(k)\partial\partial\varphi c(yi(k),\varphi)$$

and

$$(5)F(k)(\varphi)=2\varphi n(k)\sum i=1n(k)\partial\partial\varphi c(yi(k),\varphi)+\varphi 2n(k)\sum i=1n(k)\partial 2\partial\varphi 2c(yi(k),\varphi).$$

Because $E(k)(\varphi,\beta)=-\varphi 2(\partial/\partial\varphi)l(k)(\beta,\varphi)$, when φ is unknown, solving the equation $E(k)(\varphi,\beta\text{^}MLE(k))=0$ for φ yields its node-specific MLE of φ . We have $F(k)(\varphi)=-(\partial/\partial\varphi)E(k)(\varphi,\beta)$.

Simple Averaging

The simple averaging method follows upon execution of algorithm 1 (Textbox 1). First, each data node computes their local maximum by solving successively D(k)(β)=0 and E(k)(φ,β^MLE(k))=0. To compute the CIs at the CC level, the entries of the (p+1)×(p+1) matrix VMLE(k)=V(k)(β^MLE(k)) have to be computed from formula 3 with β=β^MLE(k). Then, the set

(6)S0(k)={β^MLE(k),φ^MLE(k),VMLE(k)}

is sent to the CC. The parameter estimates are then aggregated by the CC through averaging. Specifically, the CC computes

(7)β^SA=∑k=1Kw(k)β^MLE(k)andφ^SA=∑k=1Kw(k)φ^MLE(k),

where $w^{(1)}$, ..., $w^{(K)}$ are weights (ie, $w^{(k)}\geq 0$ and ∑k=1Kw(k)=1) used to combine each node's contribution. Often, weights can be taken proportional to local sample sizes, leading to the choice $w^{(k)}=n^{(k)}/n$.

Textbox 1. Algorithm 1—simple averaging inference procedure.

Input at the coordinating center (CC) level: Weight $w^{(1)}$, ..., $w^{(K)}$ attributed to each node's contribution

Step required from each node $k \in \{1, \ldots, K\}$:

Using data in D(k), compute the following quantities:

- MLE β^MLE(k) of β by solving D(k)(β)=0;
- MLE φMLE(k) of φ by solving E(k)(φ,β^MLE(k))=0;
- VMLE(k)=V(k)(β^MLE(k)) using formula (3) with β=β^MLE(k)

Send to the CC: S0(k)={β^MLE(k),φ^MLE(k),VMLE(k)}.

Step required from the CC:

Using the received sets of quantities S0(1),…,S0(K), calculate

- The simple averaging estimators β^SA=∑k=1Kw(k)β^MLE(k) and φ^SA=∑k=1Kw(k)φ^MLE(k);
- The estimator of the variance-covariance matrix Σ^SA=φ^SA∑k=1Knw(k)2n(k)VMLE(k)

Output from the CC:

Final estimates: R={β^SA,Σ^SA}

Wald-type CIs for β can be constructed based on the fact that the sequence n(β^SA−β) converges in distribution to a centered normal random variable with covariance matrix

ΣSA=φ∑k=1Kw(k)2p(k)Tβ(k),whereTβ(k)=E{Vk(β)}.

As Tβ(k) is consistently estimated using VMLE(k) and φ by φ^SA, and as $p^{(k)}$ can be estimated using $n^{(k)}/n$, it follows that a consistent estimator for Σ_{SA} is given by

Σ^SA=φ^SA∑k=1Knw(k)2n(k)VMLE(k).

The simple averaging final estimates are then given by

R={β^SA,Σ^SA}.

Single Distributed Newton-Raphson Updating

The single distributed Newton-Raphson updating method follows upon execution of algorithm 2 (Textbox 2) with T=1. First, the CC gathers summary statistics to compute the simple averaging estimators of β and φ without their accompanying CI. Hence, for $k\in\{1, ..., K\}$, and with β^MLE(k) as mentioned previously (equation 6), node k sends to the CC the quantities

(8)S0(k)={β^MLE(k),φ^MLE(k)},

and the CC uses them to compute β^SA and φ^SA using the formulas in equation 7.

Textbox 2. Algorithm 2—distributed Newton-Raphson updatings procedure.

Input at the coordinating center (CC) level: Weight $w^{(1)}$, …, $w^{(K)}$ attributed to each node's contribution

Step required from each node $k \in \{1, \ldots, K\}$:

Using data in D(k), compute

- β^MLE(k) by solving $\boldsymbol{D}^{(k)}(\beta)=0$;
- φ^MLE(k) by solving E(k)(φ,β^MLE(k))=0

Send to the CC: S0(k)={β^MLE(k),φ^MLE(k)}.

Step required from the CC:

Using the received quantities S0(1),…,S0(K):

- Calculate β^SA=∑k=1Kw(k)β^MLE(k) and φ^SA=∑k=1Kw(k)φ^MLE(k);
- Initialize β^NR,t=0=β^SA and φ^NR,t=0=φ^SA.

Execute for *t*=1, ..., *T*:

Step required from the CC:

Broadcast to nodes: Ct={β^NR,t-1,φ^NR,t-1}

Step required from each node $k \in \{1,\ldots,K\}$:

Using data in $D^{(k)}$ and quantities in C_t, compute:

- DNR,t(k) using formula (2) with β=β^NR,t-1;
- VNR,t(k) using formula (3) with β=β^NR,t-1;
- ENR,t(k) using formula (4) with φ=φ^NR,t-1 and β=β^NR,t-1;
- FNR,t(k) using formula (5) with φ=φ^NR,t-1 and β=β^NR,t-1.

Send to the CC: St(k)={DNR,t(k),VNR,t(k),ENR,t(k),FNR,t(k)}.

Step required from the CC:

Using the quantities in St(k), compute

- D¯NR,t=∑k=1Kw(k)DNR,t(k)
- V¯NR,t=∑k=1Kw(k)VNR,t(k)
- E¯NR,t=∑k=1Kw(k)ENR,t(k)
- F¯NR,t=∑k=1Kw(k)FNR,t(k)

Using, β^NR,t-1, φ^NR,t-1 and the aggregated quantities, update previous parameter estimates

- β^NR,t=β^NR,t-1+V¯NR,t−1D¯NR,t
- φ^NR,t=φ^NR,t+E¯NR,tF¯NR,t

Step required from the CC:

Compute Σ^NR=(V¯NR,T)−1{φ^NR,T∑k=1Knw(k)2n(k)VNR,T(k)}(V¯NR,T)−1

Output from the CC:

Estimates R={β^NR,T,Σ^NR}

For reasons of convenience that will become clear later, the notation β^NR,0 and φ^NR,0 will be used instead of β^SA and φ^SA, respectively. In this notation, the set of values

C1={β^NR,0,φ^NR,0}

is broadcasted to data nodes, which are then tasked with computing and sending back the quantities

S1(k)={DNR,1(k),VNR,1(k),ENR,1(k),FNR,1(k)} ,

where, for any integer $t \geq 1$, one defines

[illegible]

Upon receiving the S1(k)s from each node, the CC calculates the following weighted averages:

[illegible]

This enables the CC to execute Newton-Raphson updates from β^NR,0 and φ^NR,0, respectively:

(10)β^NR,1=β^NR,0+V¯NR,1−1D¯NR,1φ^NR,1=φ^NR,0+F¯NR,1−1E¯NR,1.

It is shown in the Multimedia Appendix 2 that

n(β^NR,1−β)→N(0,ΣNR),whereΣNR=(∑k=1Kw(k)Tβ(k))−1{φ∑k=1Kw(k)2n(k)Tβ(k)}(∑k=1Kw(k)Tβ(k))−1

As Tβ (k) is consistently estimated using VNR,1(k) and φ by φ^NR,1, and as p(k) can be estimated using n(k)/n, it follows that a consistent estimator for ΣNR is given by

Σ^NR=(∑k=1KnV^NR,1(k))−1{φ^NR,1∑k=1KnW^(k)2n(k)V^NR,1(k)}(∑k=1KnV^NR,1(k))−1

The method's final estimates are then given by

R={β^NR,1,Σ^NR}.

Multiple Distributed Newton-Raphson Updatings

The multiple distributed Newton-Raphson updatings method follows upon execution of algorithm 2 with *T*>1.

The first communication cycle follows the same procedure described previously for the single distributed Newton-Raphson updating method. It involves distributively computing a simple averaging estimator and then performing a Newton-Raphson iteration starting from this estimator. The Newton descent is calculated as described in equation 10.

Formally, the algorithm begins with each data node *k* sending the set of quantities S0(k), as described in equation 8, to the CC. Next, the CC calculates the simple averaging estimators using formula 7 and uses them to initialize β^NR,Step=0=β^SA and φ^NR,Step=0=φ^SA.

The following steps are then repeated for a certain number of iterations. At iteration t, starting from t=1, the CC broadcasts the values Ct=(β^NR,t-1,φ^NR,t-1) to the data nodes. The data nodes compute the quantities DNR,t(k), VNR,t(k), ENR,t(k), and FNR,t(k) as defined in equation 9 and send them back to the CC.

The CC then uses these quantities to perform a Newton update. Specifically, it calculates β^NR,t=β^NR,t-1+VNR,t−1DNR,t and φ^NR,t=φ^NR,t-1+ENR,t/FNR,t.

If the iterative cycle is repeated until convergence, the resulting estimates of β are equivalent to the MLEs derived from pooled data. This is because, in GLMs, for MLEs, if both the pooled and distributed algorithms are initialized using the same values for β^NR,Step=0 and φ^NR,Step=0, then, at each subsequent iteration, the distributed Newton update computed by the CC will be identical to the update obtained in a pooled setting.

For the method to yield consistent estimates, it is not necessary to initialize it using simple averaging estimators. However, using simple averaging estimators as initialization may speed up convergence, particularly in large sample sizes, as these estimators are n-consistent.

Let β^NR,T denote the estimator obtained at convergence. As it is (nearly) equal to the pooled MLE of β , we can deduce from the Multimedia Appendix 2 that

(11)n(β^NR,T−β)→N(0,ΣNR),whereΣNR=(∑k=1Kw(k)Tβ(k))−1{φ∑k=1Kw(k)2n(k)Tβ(k)}(∑k=1Kw(k)Tβ(k))−1

Following the same reasoning used previously for the single distributed Newton-Raphson updating method, we can consistently estimate the variance-covariance matrix as

Σ^NR=(V¯NR,T)−1{φ^NR,T∑k=1Knw(k)2n(k)V^NR,T(k)}(V¯NR,T)−1.

Distributed Estimating Equation

The distributed estimating equation method follows upon execution of algorithm 3 (Textbox 3). First, each node is responsible for computing the MLEs β^MLE(k) and φ^MLE(k) of β and φ , respectively, using its own data. These estimators, along with the Hessian matrix VMLE(k)=V(k)(β^MLE(k)) and FMLE(k)=F(k)(φ^MLE(k)), are then sent to the CC. The set

(12)S0(k)={β^MLE(k),φ^MLE(k),VMLE(k),FMLE(k)}

is transmitted to the CC. The CC calculates the weighted average of the Hessians and the FMLE(k) values as follows:

(13)V¯EE=∑k=1Kw(k)VMLE(k)andF¯EE=∑k=1Kw(k)FMLE(k).

The parameter estimates can then be calculated as

(14)β^EE=V¯EE−1∑k=1Kw(k)VMLE(k)β^MLE(k)φ^EE=F¯EE−1∑k=1Kw(k)FMLE(k)φ^MLE(k)

It is shown in the Multimedia Appendix 2 that n(β^EE−β) converges in distribution to a centered normal random variable with the variance-covariance matrix given by

ΣEE=(∑k=1Kw(k)Tβ(k))−1{φ∑k=1Kw(k)2n(k)Tβ(k)}(∑k=1Kw(k)Tβ(k))−1

It can be consistently estimated by

Σ^EE=(V¯EE)−1{φ^EE∑k=1knw(k)2n(k)VMLE(k)}(V¯EE)−1

Textbox 3. Algorithm 3—distributed estimating equation inference procedure.

Input at the coordinating center (CC) level: Weights $w^{(1)}$, …, $w^{(K)}$ attributed to each node's contribution

Step required from each node $k \in \{1, \ldots, K\}$:

Using data in Dk, compute the following quantities:

- MLE β^MLE(k) of β by solving D(k)(β)=0;
- MLE φ^MLE(k) of φ by solving E(k)(φ,β^MLE(k))=0;
- VMLE(k)=V(k)(β^MLE(k)) using formula (3) with β=β^MLE(k);
- FMLE(k)=F(k)(φ^MLE(k)) using formula (5) with φ=φ^MLE(k).

Send to the CC: S0(k)={β^MLE(k),φ^MLE(k),VMLE(k),FMLE(k)}.

Step required from the CC:

Using the received sets of quantities S01,…,S0K, calculate

- Aggregated quantities $\bar{V}EE=\sum_{k=1}^{K} w(k)VMLE(k)$ and $\bar{F}EE=\sum_{k=1}^{K} w(k)FMLE(k)$
- EE estimators $\hat{\beta}EE=\bar{V}EE^{-1}\sum_{k=1}^{K} w(k)VMLE(k)\hat{\beta}MLE(k)$ and $\hat{\varphi}EE=\bar{F}EE^{-1}\sum_{k=1}^{K} w(k)FMLE(k)\hat{\varphi}MLE(k)$; the variance-covariance matrix $\hat{\Sigma}EE=(\bar{V}EE)^{-1}\{\hat{\varphi}EE\sum_{k=1}^{K} nw(k)2n(k)VMLE(k)\}(\bar{V}EE)^{-1}$.

Output from the CC:

Parameter estimates and CIs R={β^SA,Σ^EE}

Distributed Estimation Using a Single Gradient-Enhanced Log-Likelihood

Overview

This method operates through algorithm 4 (Textbox 4). First, the necessary information is collected by the CC to compute the initial estimates of β and φ, denoted as β^SGE,0 and φ^SGE,0. In what follows, we assume that these estimates are obtained using the simple averaging estimators calculated through algorithm 1.

Textbox 4. Algorithm 4—inference procedure based on the distributed estimation using a single gradient-enhanced log-likelihood method.

Input at the coordinating center (CC) level: Weights $w^{(1)}$, …, $w^{(K)}$ attributed to each node's contribution

Step required from each node $k \in \{1, \ldots, K\}$:

Using data in Dk, compute

- β^MLE(k) by solving D(k)(β)=0;
- φ^MLE(k) by solving E(k)(φ,β^MLE(k))=0

Send to the CC: S0(k)={β^MLE(k),φ^MLE(k)}.

Step required from the CC:

Using the received sets of quantities S01,…,S0K, calculate

- simple averaging estimators β^SA=∑k=1Kw(k)β^MLE(k) and φ^SA=∑k=1Kw(k)φ^MLE(k)
- initialize β^SGE,0=β^SA and φ^SGE,0=φ^SA.

Broadcast to nodes: C1,1={β^SGE,0,φ^SGE,0}.

Step required from each node $k \in \{2, \ldots, K\}$:

Using data in Dk, compute the following quantities:

- DSGE,1(k)=D(k)(β^SGE,0) using formula (2) with β=β^SGE,0;
- VSGE,1(k)=V(k)(β^SGE,0) using formula (3) with β=β^SGE,0;
- ESGE,1(k)=E(k)(φ^SGE,0,β^SGE,0) using formula (4) with (φ,β)=(φ^SGE,0,β^SGE,0)

Send to the CC: S1,1(k)={DSGE,1(k),VSGE,1(k),ESGE,1(k)}

Step required from the CC:

Using the received sets of quantities S1,12,…,S1,1K, calculate

- DSGE,1=∑k=2Kw(k)DSGE,1(k)
- ESGE,1=∑k=2Kw(k)ESGE,1(k)

Broadcast to node k=1: $C_{2,1} = \{\mathbf{D}_{\mathrm{SGE},1}, \boldsymbol{E}_{\mathrm{SGE},1}\}$

Step required from node k=1:

Using data in D1, calculate

- DSGE,1(1)=D(1)(β^SGE,0) using formula (2) with β=β^SGE,0;
- VSGE,1(1)=V(1)(β^SGE,0) using formula (3) with β=β^SGE,0;
- ESGE,1(1)=E(1)(φ^SGE,0,β^SGE,0) using formula (4) with (φ,β)=(φ^SGE,0,β^SGE,0)
- D¯SGE,1=DSGE,1+w(1)DSGE,1(1)
- E¯SGE,1=ESGE,1+w(1)ESGE,1(1)

Send to the CC: S2,1(1)={β^SGE,1,φ^SGE,1,VSGE,1(1)}

Step required from the CC:

Compute

- A^SGE,1(k)=w(k)Ip+1+{VSGE,1(1)−(∑k′=1Kw(k′)VSGE,1(k′))}(VSGE,1(k))−1 for $k \in \{1, \ldots, K\}$
- Σ^SGE,1=φ^SGE,1(VSGE,1(1))−1{∑k=1Knn(k)(A^SGE,1(k)) VSGE,1(k)A^SGE,1(k)}(VSGE,1(1))−1

Output from the CC:

Parameter estimates R={β^SGE,1,Σ^SGE,1}

Subsequently, the CC broadcasts C1,1={β^SGE,0,φ^SGE,0} to node $k \in \{1, \ldots, K\}$. Each node is then requested to compute and transmit back the following quantities:

S1,1(k)={DSGE,1(k),VSGE,1(k),ESGE,1(k)},

with DSGE,1(k)=D(k)(β^SGE,0), VSGE,1(k)=V(k)(β^SGE,0), and ESGE,1(k)=E(k)(φ^SGE,0,β^SGE,0).

The CC aggregates the D(k)s and the E(k)s using averaging by calculating

$$D_{SGE,1}=\sum_{k=2}^{K} w(k) D_{SGE,1}(k) \text{ and } E_{SGE,1}=\sum_{k=2}^{K} w(k) E_{SGE,1}(k)$$

.

The Vks are momentarily stored and will be used later to compute the estimator for the asymptotic variance-covariance matrix of the final estimator of β . The quantities

$$C_{2,1}=\{D_{SGE,1}, E_{SGE,1}\}$$

are then sent to node k=1. Node k=1 computes the global average of the Dks by adding its own counterpart; that is, it first computes

$$\bar{D}_{SGE,1}=D_{SGE,1}+w(1)D_{SGE,1}(1), \text{ and } \bar{E}_{SGE,1}=E_{SGE,1}+w(1)E_{SGE,1}(1),$$

and then solves the surrogate likelihood function. Formally, it finds successively the values $\hat{\beta}_{SGE,1}(1)$ and $\hat{\varphi}_{SGE,1}(1)$ that solve

$$D_{(1)}(\beta)+\bar{D}_{SGE,1}-D_{SGE,1}(1)=0 \quad E_{(1)}(\varphi,\hat{\beta}_{SGE,1})+\bar{E}_{SGE,1}-E_{SGE,1}(1)=0.$$

The results are sent back to the CC, along with $V_{SGE,1}(1)$, yielding

$$S_{2,1}(1)=\{\hat{\beta}_{SGE,1}, \hat{\varphi}_{SGE,1}, V_{SGE,1}(1)\}.$$

If simple averaging estimators for $\hat{\beta}_{SGE,0}$ and $\hat{\varphi}_{SGE,0}$ are chosen, then $n(\hat{\beta}_{SGE,1}-\beta)$ converges in distribution to a mean-zero multivariate normal random variable with a variance-covariance matrix given by

$$\Sigma_{SGE,1}=\varphi(T_\beta)^{-1}\sum_{k=1}^{K} w(k)\left(A_{(k)}(\beta)A_{(k)}(\beta)^{-1}\right)\left(A_{(k)}(\beta)A_{(k)}(\beta)^{-1}\right)^{T}\varphi(T_\beta)^{-1}+w(k)\left(A_{(k)}(\beta)-I_{p+1}\right)\left(T_\beta\right)^{-1}\left(A_{(k)}(\beta)-I_{p+1}\right)^{T}\left(T_\beta\right)^{-1},$$

with Ip+1, the *p*+1 square identity matrix, and $T_\beta=\sum_{k=1}^{K} w(k)T_\beta(k)$. The Multimedia Appendix 2 contains the proofs. The latter can be consistently estimated by

$$\hat{\Sigma}_{SGE,1}=\hat{\varphi}_{SGE,1}(V_{SGE,1}(1))^{-1}\sum_{k=1}^{K} w(k)\left(A_{SGE,1}(k)V_{SGE,1}(k)A_{SGE,1}(k)^{T}\right)(V_{SGE,1}(1))^{-1}$$

where

$$A_{SGE,1}(k)=n(k)nI_{p+1}+\{V_{SGE,1}(1)-T_{\hat{\beta}}\}(V_{SGE,1}(k))^{-1}$$

With $T_{\hat{\beta}}=\sum_{k=1}^{K} w(k)V_{SGE,1}(k)$.

Remark

In the study by Jordan et al [13], where the method described previously was originally proposed, the authors discuss a version in which the latter process is repeated multiple times. Their version assumes that the data are uniformly and randomly split across nodes. Under this assumption, the resulting estimator of β is asymptotically equivalent to the pooled estimator regardless of the number of iterations executed. This equivalence occurs because, when the predictors' distribution is the same across nodes and the node sample sizes are equal, then $T_\beta(k)\equiv T_\beta$ and $p(k)\equiv 1/K$. By choosing $w(k)=1/K$, it follows that $A_\beta(k)=I_{p+1}$, resulting in the following expression for $\Sigma_{SGE,1}$: $\Sigma_{SGE,1}=\varphi(T_\beta)^{-1}$. The aforementioned variance-covariance matrix is also the same as that of the simple averaging estimator in the setting of equal sample sizes and even predictor distributions. Consequently, at each iteration, the probability distribution of the resulting estimator remains unchanged. However, in a more general setting in which predictor distributions and sample sizes vary across nodes, these cancellations no longer occur. Therefore, in this case, the probability distribution of the obtained estimator changes after each iteration, and tracking these changes falls beyond the scope of objective 3 (see the Multimedia Appendix 2). Hence, this presentation focused on the case in which only 1 iteration is executed.

Distributed Estimation Using Multiple Gradient-Enhanced Log-Likelihood

This method operates through algorithm 5 (Textbox 5). First, the CC collects the necessary information to compute the initial estimates, denoted as $\hat{\beta}_{MGE,0}$ and $\hat{\varphi}_{MGE,0}$. In this case, we assume that these estimates are obtained using the simple averaging estimators calculated through algorithm 1.

Textbox 5. Algorithm 5—inference procedure based on the distributed estimation using a multiple gradient-enhanced log-likelihood method.

Input at the coordinating center (CC) level: Weight $w^{(1)}$, …, $w^{(K)}$ attributed to each node's contribution

Step required from each node $k \in \{1, \ldots, K\}$:

Using data in D(k), compute

- β^MLE(k) by solving D(k)(β)=0;
- φ^MLE(k) by solving E(k)(φ,β^MLE(k))=0

Send to the CC: S0(k)={β^MLE(k),φ^MLE(k)}.

Step required from the CC:

Using the received sets of quantities S0(1),…,S0(K), calculate

- simple averaging estimators β^SA=∑k=1Kw(k)β^MLE(k) and φ^SA=∑k=1Kw(k)φ^MLE(k)
- initialize β^MGE,0=β^SA and φ^MGE,0=φ^SA.

Broadcast to nodes: C1,1={β^MGE,0,φ^MGE,0}.

Step required from each node $k \in \{1, \ldots, K\}$:

Using data in Dk, calculate

- DMGE,1(k)=D(k)(β^MGE,0) using formula (2) with β=β^MGE,0;
- VMGE,1(k)=V(k)(β^MGE,0) using formula (3) with β=β^MGE,0;
- EMGE,1(k)=E(k)(φ^MGE,0,β^MGE,0) using formula (4) with (φ,β)=(φ^MGE,0,β^MGE,0).

Send to the CC: S1,1(k)={DMGE,1(k),VMGE,1(k),EMGE,1(k)}

Step required from the CC:

Using the received sets of quantities S1,11,…,S1,1K, calculate

- D¯MGE,1=∑k=1Kw(k)DMGE,1(k)
- E¯MGE,1=∑k=1Kw(k)EMGE,1(k)

Broadcast to nodes: C2,1={D¯MGE,1,E¯MGE,1}.

Step required from each node $k \in \{1, \ldots, K\}$:

Using data in Dk, calculate

- β^MGE,1(k) that solves D(k)(β)+D¯MGE,1−DMGE,1(k)=0
- φ^MGE,1(k) that solves E(k)(φ,β^MGE,1)+E¯MGE,1−EMGE,1(k)=0

Send to the CC: S2,1(k)={β^MGE,1(k),φ^MGE,1(k)}

Step required from the CC:

Compute

- β^MGE,1=∑k=1Kw(k)β^MGE,1(k)
- φ^MGE,1=∑k=1Kw(k)φ^MGE,1(k)
- T^β =∑k=1Kw(k)VMGE,1(k)
- U^β =∑k=1Kw(k)(VMGE,1(k))−1
- Σ^MGE,1=φ^MGE,1∑k=1Knw(k)2n(k)[U^β VMGE,1(k)+(Ip+1−U^β T^β)]×[U^β +(Ip+1−U^β T^β)(VMGE,1(k))−1]

Output from the CC:

Parameter estimates R={β^MGE,1,Σ^MGE,1}

Subsequently, the CC broadcasts C1,1={β^MGE,0,φ^MGE,0} to each node, which is then requested to compute and transmit back the following quantities:

Here DMGE,1(k)=D(k)(β^MGE,0), VMGE,1(k)=V(k)(β^MGE,0), and EMGE,1(k)=E(k)(φ^MGE,0,β^MGE,0).

S1,1(k)={DMGE,1(k),VMGE,1(k),EMGE,1(k)}.

The CC aggregates the Dks and the Eks using averaging by calculating $\bar{D}_{MGE,1}=\sum_{k=1}^{K} w(k) D_{MGE,1}(k)$ and $\bar{E}_{MGE,1}=\sum_{k=1}^{K} w(k) E_{MGE,1}(k)$.

The CC then broadcasts $C_{2,1}=\{\bar{D}_{MGE,1}, \bar{E}_{MGE,1}\}$ to each node, which is then tasked with solving the surrogate likelihood function. Formally, they find successively the value $\hat{\beta}_{MGE,1}(k)$ and $\hat{\varphi}_{MGE,1}(k)$ that solves

$$D_1(k)(\beta)+\bar{D}_{MGE,1}-D_{MGE,1}(k)-E_1(k)(\varphi,\hat{\beta}_{MGE,1}(k))+\bar{E}_{MGE,1}-E_{MGE,1}(k)=0$$

Each node then transmits their set of local surrogate likelihood estimators to the CC:

$$S_{2,1}(k)=\{\hat{\beta}_{MGE,1}(k), \hat{\varphi}_{MGE,1}(k)\}.$$

Using the received sets of quantities $S_{2,11},\ldots,S_{2,1K}$, the CC aggregates them through averaging using the following formulas:

$$\hat{\beta}_{MGE,1}=\sum_{k=1}^{K} w(k)\hat{\beta}_{MGE,1}(k) \text{ and } \hat{\varphi}_{MGE,1}=\sum_{k=1}^{K} w(k)\hat{\varphi}_{MGE,1}(k).$$

It is shown in the Multimedia Appendix 2 that $n(\hat{\beta}_{MGE,1}-\beta)$ converges in distribution to a multivariate normal random variable with mean 0 and a variance-covariance matrix given by

$$\Sigma_{MGE,1}=\varphi\sum_{k=1}^{K} w(k)[\ddot{\varphi}_1(k)U_\beta^{-1}T_\beta(k)+(\dot{\varphi}+1-U_\beta^{-1}T_\beta)\dot{\varphi}_1(k)][\ddot{\varphi}_1(k)U_\beta^{-1}+(\dot{\varphi}+1-U_\beta^{-1}T_\beta)\dot{\varphi}_1(k)]^{-1}$$

where $U_\beta=\sum_{k=1}^{K} w(k)(T_\beta(k))^{-1}$. The latter can be consistently estimated using

$$\hat{\Sigma}_{MGE,1}=\hat{\varphi}_{MGE,1}\sum_{k=1}^{K} w(k)[\ddot{\varphi}_1(k)\hat{U}_\beta^{-1}V_{MGE,1}(k)+(\dot{\varphi}+1-\hat{U}_\beta^{-1}\hat{T}_\beta)\dot{\varphi}_1(k)][\ddot{\varphi}_1(k)\hat{U}_\beta^{-1}+(\dot{\varphi}+1-\hat{U}_\beta^{-1}\hat{T}_\beta)\dot{\varphi}_1(k)]^{-1}$$

where $\hat{T}_\beta=\sum_{k=1}^{K} w(k)V_{MGE,1}(k)$ and $\hat{U}_\beta=\sum_{k=1}^{K} w(k)(V_{MGE,1}(k))^{-1}$

Summary of Quantities Exchanged in the Adapted Methods

Table 2 presents a summary of the quantities exchanged between the nodes and the CC in both directions. Table 2 demonstrates that the quantities involved in exchanges from the nodes to the CC consist of parameter estimates, gradients (Dk vectors), Hessians (Vk matrices), and real numbers (Ek and Fk). On the other hand, the quantities shared from the CC to the nodes primarily consist of parameter estimates. Notably, methods 5 and 6 differentiate themselves by requiring the sharing of aggregated gradient vectors and Hessian matrices as well.

Comparison of Adapted Methods

Table 3 compares the main adapted HPSA methods regarding the quantities shared between the CC and the nodes and the operational complexity of the procedures. Methods 1 and 4 require only 1 communication from the data nodes to the CC and no communication back from the CC to the nodes. These so-called one-shot methods have the lowest operational complexity. Method 4 requires the additional quantity FMLEk to be transmitted from each node to the CC.

Table . Comparison of adapted methods.

Method number	Method	Information shared		Number of communications		Workflow
		From nodes to CC[a]	From CC to nodes	From nodes to CC	From CC to nodes	
1	Simple averaging	Local parameter estimates; Hessian matrix of log-likelihood (with respect to β only)	None	1	0	I in Figure 2
2	Single distributed Newton-Raphson updating	Local parameter estimates; gradient and Hessian of log-likelihood	Simple averaging aggregated estimates of parameters	2	1	II in Figure 3 with *T*=1
3	Multiple distributed Newton-Raphson updatings (with *T* Newton-Raphson updatings)	Local parameter estimates; *T* × gradient and Hessian of log-likelihood	Simple averaging aggregated estimates of parameters; *(T-1)* × Newton-updated parameter estimates	*T*+1	*T*	II in Figure 3 with T>1
4	Distributed estimating equations	Local parameter estimates; Hessian of log-likelihood	None	1	0	I in Figure 2
5	Distributed single gradient-enhanced log-likelihood	From all nodes: local parameter estimates and Hessian of log-likelihood (with respect to β only); from nodes 2 to *K*: gradient of log-likelihood; from node 1 only: gradient-enhanced parameter estimates	To all nodes: simple averaging aggregated estimates of parameters; to node 1 only: average of local gradients and Hessians	All nodes: 2	Nodes 2 to *K*: 1; node 1: 2	III in Figure 4 with *T*=1
6	Distributed multiple gradient-enhanced log-likelihood	Local parameter estimates; gradient of log-likelihood; Hessian of log-likelihood (with respect to β only); gradient-enhanced parameter estimates	Simple averaging aggregated estimates of parameters; average of local gradients and Hessians	3	2	IV in Figure 5 with *T*=1

[a]CC: coordinating center.

Methods 2 and 3 perform Newton-Raphson updating using some initial estimator as a basis, usually the simple averaging estimator. While method 2 requires this initial estimator to be n-consistent, if *T* is large enough, any initial value will work for method 3 (although convergence may be slower). Both methods require D(k),V(k),E(k), and Fk to be evaluated and sent to the CC *T* times, with *T*=1 for method 2. Compared to method 1, method 2 requires the additional quantities Dk, Ek, and Fk, and method 3 further requires these quantities to be evaluated and communicated multiple times.

Method 5 relies on an approximation of the log-likelihood function. It requires an initial estimator, usually the simple averaging estimator. This approach treats node 1 differently, making it solve the surrogate log-likelihood using aggregates from the other nodes and its own data. The CC sends the initial estimator to each node and then requires them to evaluate Dk,Vk, and Ek and send the result back to the CC once. It averages the results and then communicates them to node 1, which solves the surrogate log-likelihood and sends its results back to the CC.

Method 6 applies method 5 to every node, making each node solve the surrogate log-likelihood function using its own data before averaging the resulting local estimators.

Discussion

Summary of Findings

The first objective of this study (objective 1) aimed to identify and map the methodological approaches used and developed in the literature regarding HPSA. To achieve this, we conducted a scoping review, which included 41 articles following our protocol. These articles were categorized based on the types of models and communication schemes involved, as presented in Table 1. The analysis revealed that most methods included in this scoping review focused on methodological settings associated with massive data. The communication schemes of

these methods were demonstrated through workflows I, II, III, and IV.

The second objective of this study (objective 2) aimed to describe the approaches that can be used for basic GLM regression analyses and identify the distributional assumptions they require. To accomplish this, we identified 6 approaches and classified them within workflows I to IV, enabling their comparison in terms of operational communication protocols within a unified framework. However, a limitation of these methods is that they assume identical node sample sizes and node covariate distributions. This assumption reduces their suitability in settings commonly encountered in health analytics, where data-collecting nodes are prone to generating different covariate distributions.

The third objective of this study (objective 3) was to present methods that relaxed these assumptions by adapting the approaches identified in objective 2 to the unequal sample sizes and nonidentical covariate sample distribution setting. In addition, we compared these methods in terms of the information shared and operational complexity. This involved adapting the quantities and estimators described in the original articles and deriving new asymptotic results with relaxed assumptions. We defined a unified framework to describe inference procedures using these methods. This unified framework encompasses common hypotheses and notation and allows for both estimation and the construction of CIs, providing detailed steps for both the data nodes and the CC.

Challenges and Opportunities

Work pertaining to objective 1 illustrated why it is so challenging for researchers and data custodians alike to find information regarding HPSA. While the HPSA literature is very recent (all the included articles were published in 2010 or later), the literature is nonhomogeneous, and it has not come to a consensus on nomenclature. No universal terminology exists, and different terms are used in the different fields developing and applying HPSA methods. Many specific methods introduced in applied contexts are special cases of more general methods that may or may not be cited. These characteristics make finding useful and efficient keywords arduous. This required adapting our search strategy.

This difficulty is compounded by the fact that statistical inference is not the main focus of most of the HPSA literature. Most published work falls within the prediction, learning, and optimization contexts. As a result, method assumptions are rarely discussed. This can be a problem when adapting these methods for inference. Furthermore, the methodological setting is often assumed to be in the massive data context, in which data are randomly distributed between nodes. This allows the authors to make strong assumptions on node sample sizes and covariate distributions that may be unrealistic in the confidential data context using multiple data sources. These methods cannot be used directly for inference using confidential health data. While some work remains to be done when the structure of association between the covariate and the outcome is heterogeneous between nodes, we adapted widely used methods for when the distribution of covariates and sample sizes between nodes is not identical.

Table 1 illustrates how most HPSA methods are focused on parametric models. Some work has also been done for semiparametric and nonparametric regression, and some methods are introduced outside of the regression framework (although they can also be applied to regression). Many methods do not require communication of quantities from the CC to the data nodes—they only require 1 transmission from the nodes to the CC. Given the lack of awareness regarding HPSA, starting by implementing lower–operational complexity methods while providing useful results offers a promising path.

LHSs seek to improve health by generating knowledge during practice and making use of that knowledge to improve practice. This requires analyzing data from a variety of sources. HPSA methods offer a way to tackle the challenges that come with multi-jurisdictional data.

The methods can be implemented “manually” (eg, via email exchanges), but platforms enabling semiautomated distributed fittings of statistical models have been proposed in the literature [60]. As LHSs work by continuously monitoring and analyzing data rather than through cross-sectional studies only, automated platforms are necessary. Ideally, these platforms should be able to link and standardize data sources from multiple jurisdictions as well as perform HPSA. On the other hand, explicit descriptions of these methods’ algorithms and the quantities exchanged are not always easily accessible, and this complicates the evaluation of the tools by data custodians and researchers.

This is especially important as it is essential to clarify here that operating an HPSA algorithm does *not* ensure confidentiality in and of itself.

For example, it is known that sharing sample moments can compromise confidentiality. It can be shown that a set of n observations is uniquely determined by its first n sample moments [61]. This could prove problematic for methods that rely on sharing the first few moments of each node’s sample, especially if the number of observations is low, as the sample could be partially reconstructed by the CC.

The results presented in this paper contribute to this objective by clarifying the workflows and quantities exchanged in each method. Nevertheless, further analysis of the confidentiality preserved by HPSA methods is needed to fully understand the risk associated with the sharing of summary statistics, especially as more rounds of communication between the CC and data nodes are completed. The framework of differential privacy has been used to guarantee the preservation of confidentiality in a few HPSA methods, but a wider application of differential privacy to existing and popular methods has yet to be explored.

Acknowledgments

The authors acknowledge the Health Data Research Network Canada, Natural Sciences and Engineering Research Council of Canada, Fonds de recherche du Québec – Nature et technologies, the Chaire en informatique de la santé de l'Université de Sherbrooke, and the Chaire MEIE (Ministry of Economy, Innovation and Energy) Québec—le numérique au service des systèmes de santé apprenants. The authors would like to thank the GRIIS (Groupe de recherche interdisciplinaire en informatique de la santé [Interdisciplinary Research Group in Health Informatics]) members who enriched this work via multiple conversations over the last few months and kept the authors going. They would also like to thank Professor Kim McGrail for her very insightful comments on this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed protocol for the scoping review.
[DOCX File, 51 KB - medinform_v12i1e53622_app1.docx]

Multimedia Appendix 2
Mathematical derivations pertaining to objective 3.
[DOCX File, 102 KB - medinform_v12i1e53622_app2.docx]

Checklist 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File, 217 KB - medinform_v12i1e53622_app3.pdf]

References

1. Sinha BK, Hartung J, Knapp G. Statistical Meta-Analysis with Applications: John Wiley & Sons; 2011.
2. Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. J Am Med Inform Assoc 2020 Mar 1;27(3):376-385. [doi: 10.1093/jamia/ocz199] [Medline: 31816040]
3. Gao Y, Liu W, Wang H, Wang X, Yan Y, Zhang R. A review of distributed statistical inference. Stat Theory Relat Fields 2022 May 27;6(2):89-99. [doi: 10.1080/24754269.2021.1974158]
4. Huo X, Cao S. Aggregated inference. WIREs Comp Stats 2019 Jan;11(1):e1451. [doi: 10.1002/wics.1451]
5. Agresti A. Foundations of Linear and Generalized Linear Models: John Wiley & Sons; 2015.
6. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
7. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050]
8. Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. Presented at: EASE '14: 18th International Conference on Evaluation and Assessment in Software Engineering; May 13-14, 2014; London, United Kingdom p. 1-10 URL: https://dl.acm.org/doi/proceedings/10.1145/2601248 [doi: 10.1145/2601248.2601268]
9. Jalali S, Wohlin C. Systematic literature studies: database searches vs. backward snowballing. Presented at: SEM '12: 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement; Sep 19-20, 2012; Lund, Sweden p. 29-38. [doi: 10.1145/2372251.2372257]
10. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69. [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
11. Duan R, Luo C, Schuemie MJ, et al. Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. J Am Med Inform Assoc 2020 Jul 1;27(7):1028-1036. [doi: 10.1093/jamia/ocaa044] [Medline: 32626900]
12. Huang C, Huo X. A distributed one-step estimator. Math Program 2019 Mar;174(1-2):41-76. [doi: 10.1007/s10107-019-01369-0]
13. Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. J Am Stat Assoc 2019 Apr 3;114(526):668-681. [doi: 10.1080/01621459.2018.1429274]
14. Park JA, Kim TH, Kim J, Park YR. WICOX: Weight-Based Integrated Cox model for time-to-event data in distributed databases without data-sharing. IEEE J Biomed Health Inform 2023 Jan;27(1):526-537. [doi: 10.1109/JBHI.2022.3218585] [Medline: 36318551]
15. Lu CL, Wang S, Ji Z, et al. WebDISCO: A web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015 Nov 1;22(6):1212-1219. [doi: 10.1093/jamia/ocv083]

16. Toh S, Wellman R, Coley RY, et al. Combining distributed regression and propensity scores: A doubly privacy-protecting analytic method for multicenter research. Clin Epidemiol 2018;10:1773-1786. [doi: 10.2147/CLEP.S178163] [Medline: 30568510]
17. Xiong R, Koenecke A, Powell M, Shen Z, Vogelstein JT, Athey S. Federated causal inference in heterogeneous observational data. arXiv. Preprint posted online on Aug 10, 2021. [doi: 10.48550/arXiv.2107.11732]
18. Vo TV, Hoang TN, Lee Y, Leong TY. Federated estimation of causal effects from observational data. arXiv. Preprint posted online on 2021. [doi: 10.48550/arXiv.2106.00456]
19. Li W, Tong J, Anjum MM, Mohammed N, Chen Y, Jiang X. Federated learning algorithms for generalized mixed-effects model (GLMM) on horizontally partitioned data from distributed sources. BMC Med Inform Decis Mak 2022 Oct 16;22(1):269. [doi: 10.1186/s12911-022-02014-1] [Medline: 36244993]
20. Wedderburn RWM. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika 1976;63(1):27-32. [doi: 10.1093/biomet/63.1.27]
21. Van der Vaart AW. Asymptotic Statistics: Cambridge University Press; 2000.
22. Van Der Vaart AW, Wellner JA. Weak Convergence and Empirical Processes: With Applications to Statistics: Springer; 1996.
23. Rosenblatt JD, Nadler B. On the optimality of averaging in distributed statistical learning. Inf inference 2016 Dec;5(4):379-404. [doi: 10.1093/imaiai/iaw013]
24. Banerjee M, Durot C, Sen B. Divide and conquer in nonstandard problems and the super-efficiency phenomenon. Ann Statist 2019;47(2):720-757. [doi: 10.1214/17-AOS1633]
25. Chen X, Xie M. A split-and-conquer approach for analysis of extraordinarily large data. Stat Sin 2014;24:1655-1684. [doi: 10.5705/ss.2013.088] [Medline: 25076817]
26. Basiri S, Ollila E, Koivunen V. Robust, scalable, and fast bootstrap method for analyzing large scale data. IEEE Trans Signal Process 2016;64(4):1007-1017. [doi: 10.1109/TSP.2015.2498121]
27. Battey H, Fan J, Liu H, Lu J, Zhu Z. Distributed testing and estimation under sparse high dimensional models. Ann Stat 2018 Jun;46(3):1352-1382. [doi: 10.1214/17-AOS1587] [Medline: 30034040]
28. Fan J, Guo Y, Wang K. Communication-efficient accurate statistical estimation. J Am Stat Assoc 2023;118(542):1000-1010. [doi: 10.1080/01621459.2021.1969238] [Medline: 37347088]
29. Guo G, Sun Y, Jiang X. A partitioned quasi-likelihood for distributed statistical inference. Comput Stat 2020 Dec;35(4):1577-1596. [doi: 10.1007/s00180-020-00974-4]
30. Lin N, Xi R. Aggregated estimating equation estimation. Stat Interface 2011;4(1):73-83. [doi: 10.4310/SII.2011.v4.n1.a8]
31. Zhang Y, Duchi JC, Wainwright MJ. Communication-efficient algorithms for statistical optimization. Presented at: 2012 IEEE 51st IEEE Conference on Decision and Control (CDC); 2012; Maui, HI, USA p. 6792. [doi: 10.1109/CDC.2012.6426691]
32. Chang C, Bu Z, Long Q. CEDAR: communication efficient distributed analysis for regressions. Biometrics 2023 Sep;79(3):2357-2369. [doi: 10.1111/biom.13786] [Medline: 36305019]
33. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012;19(5):758-764. [doi: 10.1136/amiajnl-2012-000862] [Medline: 22511014]
34. Hector EC, Song PK. Joint integrative analysis of multiple data sources with correlated vector outcomes. Ann Appl Stat 2022;16(3):1700-1717. [doi: 10.1214/21-AOAS1563]
35. Mozafari-Majd E, Koivunen V. Two-stage robust and sparse distributed statistical inference for large-scale data. IEEE Trans Signal Process 2022;70:5351-5365. [doi: 10.1109/TSP.2022.3216704]
36. Mozafari-Majd E, Koivunen V. Robust variable selection and distributed inference using t-based estimators for large-scale data. Presented at: 2020 28th European Signal Processing Conference (EUSIPCO); Jan 18-21, 2021; Amsterdam, Netherlands p. 2453-2457 URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=9287308 [doi: 10.23919/Eusipco47968.2020.9287773]
37. Yue X, Kontar RA, Gómez AME. Federated data analytics: A study on linear models. IISE Trans 2024 Jan 2;56(1):16-28. [doi: 10.1080/24725854.2022.2157912]
38. Duan R, Ning Y, Chen Y. Heterogeneity-aware and communication-efficient distributed statistical inference. Biometrika 2022 Feb 1;109(1):67-83. [doi: 10.1093/biomet/asab007]
39. Tong J, Duan R, Li R, Scheuemie MJ, Moore JH, Chen Y. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. Pac Symp Biocomput 2020;25:695-706. [doi: 10.1142/9789811215636_0061] [Medline: 31797639]
40. Di F, Wang L, Lian H. Communication-efficient estimation and inference for high-dimensional quantile regression based on smoothed decorrelated score. Stat Med 2022 Nov 10;41(25):5084-5101. [doi: 10.1002/sim.9555] [Medline: 36263919]
41. Edmondson MJ, Luo C, Nazmul Islam M, et al. Distributed Quasi-Poisson regression algorithm for modeling multi-site count outcomes in distributed data networks. J Biomed Inform 2022 Jul;131:104097. [doi: 10.1016/j.jbi.2022.104097] [Medline: 35643272]
42. Luo L, Li L. Online two-way estimation and inference via linear mixed-effects models. Stat Med 2022 Nov 10;41(25):5113-5133. [doi: 10.1002/sim.9557] [Medline: 35983945]

43. Shu D, Young JG, Toh S. Privacy-protecting estimation of adjusted risk ratios using modified Poisson regression in multi-center studies. BMC Med Res Methodol 2019 Dec 5;19(1):228. [doi: 10.1186/s12874-019-0878-6] [Medline: 31805872]
44. Zhao T, Cheng G, Liu H. A partially linear framework for massive heterogeneous data. Ann Stat 2016 Aug;44(4):1400-1437. [doi: 10.1214/15-AOS1410] [Medline: 28428647]
45. Luo J, Sun Q, Zhou WX. Distributed adaptive Huber regression. Comput Stat Data Anal 2022 May;169:107419. [doi: 10.1016/j.csda.2021.107419]
46. Liu M, Shang Z, Cheng G. Nonparametric distributed learning under general designs. Electron J Statist 2020;14(2):3070-3102. [doi: 10.1214/20-EJS1733]
47. Zhang L, Castillo ED, Berglund AJ, Tingley MP, Govind N. Computing confidence intervals from massive data via penalized quantile smoothing splines. Comput Stat Data Anal 2020 Apr;144:106885. [doi: 10.1016/j.csda.2019.106885]
48. Volgushev S, Chao SK, Cheng G. Distributed inference for quantile regression processes. Ann Statist 2019;47(3):1634-1662. [doi: 10.1214/18-AOS1730]
49. Wang X, Yang Z, Chen X, Liu W. Distributed inference for linear support vector machine. J Mach Learn Res 2019;20:1-41. [doi: 10.48550/arXiv.1811.11922]
50. Atta-Asiamah E, Yuan M. Distributed inference for degenerate u-statistics. Stat (Int Stat Inst) 2019;8(1):e234. [doi: 10.1002/sta4.234]
51. Minsker S. Distributed statistical estimation and rates of convergence in normal approximation. Electron J Statist 2019;13(2):5213-5252. [doi: 10.1214/19-EJS1647]
52. Lin N, Xi R. Fast surrogates of U-statistics. Comput Stat Data Anal 2010 Jan;54(1):16-24. [doi: 10.1016/j.csda.2009.08.009]
53. Bruce S, Li Z, Yang HC, Mukhopadhyay S. Nonparametric distributed learning architecture for big data: algorithm and applications. IEEE Trans Big Data 2019;5(2):166-179. [doi: 10.1109/TBDATA.2018.2810187]
54. Chen SX, Peng L. Distributed statistical inference for massive data. Ann Statist 2021;49(5):2851-2869. [doi: 10.1214/21-AOS2062]
55. Nezakati E, Pircalabelu E. Unbalanced distributed estimation and inference for the precision matrix in Gaussian graphical models. Stat Comput 2023 Apr;33(2):47. [doi: 10.1007/s11222-023-10211-9]
56. Shi C, Lu W, Song R. A massive data framework for m-estimators with cubic-rate. J Am Stat Assoc 2018;113(524):1698-1709. [doi: 10.1080/01621459.2017.1360779] [Medline: 30739966]
57. Wu S, Xu Y, Feng Z, Yang X, Wang X, Gao X. Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. BMC Bioinformatics 2012 Dec 2;13(1):320. [doi: 10.1186/1471-2105-13-320] [Medline: 23198695]
58. Lai RCS, Hannig J, Lee TCM. Method G: uncertainty quantification for distributed data problems using generalized fiducial inference. J Comput Graph Stat 2021 Oct 2;30(4):934-945. [doi: 10.1080/10618600.2021.1923514]
59. Shamir O, Srebro N, Zhang T. Communication-efficient distributed optimization using an approximate Newton-type method. Presented at: Proceedings of the 31st International Conference on Machine Learning; Jun 21-26, 2014; Beijing, China p. 1000-1008. [doi: 10.48550/arXiv.1312.7853]
60. Beyan O, Choudhury A, van Soest J, et al. Distributed analytics on sensitive medical data: the personal health train. Data Intell 2020 Jan;2(1-2):96-107. [doi: 10.1162/dint_a_00032]
61. Provost SB, Zareamoghaddam H, Ahmed SE, Ha HT. The generalized Pearson family of distributions and explicit representation of the associated density functions. Commun Stat Theory Methods 2022 Aug 18;51(16):5590-5606. [doi: 10.1080/03610926.2020.1843680]

Abbreviations

CC: coordinating center
GLM: generalized linear model
HPSA: horizontally partitioned statistical analytics
LHS: learning health system
MLE: maximum likelihood estimator

Edited by A Benis; submitted 12.10.23; peer-reviewed by L Guo, Y Wang; revised version received 10.07.24; accepted 19.07.24; published 14.11.24.

Please cite as:
Camirand Lemyre F, Lévesque S, Domingue MP, Herrmann K, Ethier JF
Distributed Statistical Analyses: A Scoping Review and Examples of Operational Frameworks Adapted to Health Analytics
JMIR Med Inform 2024;12:e53622
URL: https://medinform.jmir.org/2024/1/e53622
doi:10.2196/53622

Knowledge Graph for Breast Cancer Prevention and Treatment: Literature-Based Data Analysis Study

Shuyan Jin[1], MPH; Haobin Liang[2], MSc; Wenxia Zhang[1], PhD; Huan Li[1], MM

1

2

Corresponding Author:
Shuyan Jin, MPH

Abstract

Background: The incidence of breast cancer has remained high and continues to rise since the 21st century. Consequently, there has been a significant increase in research efforts focused on breast cancer prevention and treatment. Despite the extensive body of literature available on this subject, systematic integration is lacking. To address this issue, knowledge graphs have emerged as a valuable tool. By harnessing their powerful knowledge integration capabilities, knowledge graphs offer a comprehensive and structured approach to understanding breast cancer prevention and treatment.

Objective: We aim to integrate literature data on breast cancer treatment and prevention, build a knowledge graph, and provide support for clinical decision-making.

Methods: We used Medical Subject Headings terms to search for clinical trial literature on breast cancer prevention and treatment published on PubMed between 2018 and 2022. We downloaded triplet data from the Semantic MEDLINE Database (SemMedDB) and matched them with the retrieved literature to obtain triplet data for the target articles. We visualized the triplet information using NetworkX for knowledge discovery.

Results: Within the scope of literature research in the past 5 years, malignant neoplasms appeared most frequently (587/1387, 42.3%). Pharmacotherapy (267/1387, 19.3%) was the primary treatment method, with trastuzumab (209/1805, 11.6%) being the most commonly used therapeutic drug. Through the analysis of the knowledge graph, we have discovered a complex network of relationships between treatment methods, therapeutic drugs, and preventive measures for different types of breast cancer.

Conclusions: This study constructed a knowledge graph for breast cancer prevention and treatment, which enabled the integration and knowledge discovery of relevant literature in the past 5 years. Researchers can gain insights into treatment methods, drugs, preventive knowledge regarding adverse reactions to treatment, and the associations between different knowledge domains from the graph.

(JMIR Med Inform 2024;12:e52210) doi:10.2196/52210

KEYWORDS

knowledge graph; breast cancer; treatment; prevention; adverse reaction

Introduction

Breast cancer is the most common malignant tumor in women worldwide, with a reported death toll exceeding 600,000 in 2018 alone [1]. Breast cancer has emerged as the most prevalent cancer and a primary cause of mortality among women. The global incidence of new cases of female breast cancer witnessed a sharp increase from 1.05 million in 2000 to 2.09 million in 2018 [2]. In 2020, global cancer burden data revealed that new breast cancer cases reached 2.26 million, constituting 11.7% of all newly diagnosed cancer cases worldwide. The newly reported mortality cases numbered 0.68 million, representing 6.9% of global newly reported deaths [3]. Factors such as old age, young age at menarche, family history of breast cancer, smoking, and drinking alcohol increase the risk of breast cancer [4-6]. On the contrary, regular physical exercise; breastfeeding; regular work and rest; and intake of fruits, vegetables, whole grains, and dietary fiber can appropriately reduce the risk of breast cancer [7]. Various treatment methods are used for patients with breast cancer, including surgery, radiation therapy, endocrine therapy, chemotherapy, and targeted therapy. So far, most countries have primarily focused on population education for breast cancer prevention, including encouraging increased physical activity, controlling BMI, and limiting alcohol intake [8]. Despite the increasing number of research literature, a large amount of literature on breast cancer prevention and treatment has not been systematically integrated. Knowledge graph technology allows for the independent connection and integration of disparate literature, resulting in a more comprehensive and cohesive knowledge framework.

Knowledge Graph is a knowledge repository proposed by Google in 2012 to enhance the functionality of search engines.

It describes concepts and their relationships in the real world using triplets in the form of entity-relation-entity [9]. Knowledge graphs can integrate information from diverse sources and domains, including text, databases, and web pages, and intricately interlink them. These integrations serve to mitigate information silos, fostering the establishment of a more comprehensive knowledge framework. Knowledge graphs have been widely used in various fields, such as medicine, network security, journalism, finance, and education [10]. Knowledge graphs in the biomedical domain have applications in studies related to disease associations [11], genomics [12], drug interactions [13], and support for physicians in formulating individualized treatment regimens [14]. At present, there are well-established knowledge graphs, including DisGeNET [15], which integrate information on the associations between genes and diseases; DrugBank [16], a comprehensive bioinformatics and cheminformatics knowledge base; and ClinVar [17], a compilation of genetic variation information from diverse laboratories worldwide. One study extracted breast cancer–related features from Chinese breast cancer mammography reports and built a knowledge graph for diagnosing breast cancer by combining diagnosis and treatment guidelines and insights from clinical experts [18]. Another study integrated triples from clinical guidelines, medical encyclopedias, and electronic medical records to build a breast cancer knowledge graph [19]. Despite a small number of scholars having constructed knowledge graphs for breast cancer, the varied emphases and diverse data sources employed render their applicability limited. A knowledge graph specifically focused on the prevention and treatment of breast cancer has not been constructed at present. Therefore, this study primarily collects information related to the prevention and treatment of breast cancer to construct a knowledge graph.

In the biomedical field, there are already mature tools (eg, SemRep) for extracting knowledge from medical texts. SemRep is a natural language processing program based on the Unified Medical Language System (UMLS), which performs operations such as text tokenization, syntactic analysis, part-of-speech disambiguation, phrase mapping, semantic predicate normalization, and syntactic constraints [20]. It extracts entities and relationships from biomedical texts and outputs triplets stored in the Semantic MEDLINE Database (SemMedDB) [21]. SemMedDB currently encompasses details on approximately 96.3 million predications derived from all PubMed citations (around 29.1 million citations) and serves as the foundation for the Semantic MEDLINE application [22]. We downloaded the entity and relationship data provided by SemMedDB. NetworkX is an open-source library for Python, primarily designed for creating, analyzing, and visualizing complex network structures. NetworkX plays a significant role in knowledge visualization, facilitating users in intuitively presenting and comprehending intricate knowledge graphs or network data.

Methods

Ethics Approval

This study was approved by the Board of Medical Ethics Committee of Shenzhen Maternal and Child Health Hospital (SFYLS[2022]003).

Data Source

We conducted a search on PubMed using Medical Subject Headings terms “breast cancer,” “prevention,” and “treatment,” covering the period from January 1, 2018, to December 31, 2022, and the study type was clinical trials. A total of 3589 articles were retrieved. We obtained the entity and relationship data from SemMedDB.

Data Processing and Construction of Knowledge Graph

We matched the PMIDs of the retrieved articles with the database and extracted the corresponding triplet information. We initially obtained 33,060 Subject-Predicate-Object (SPO) triplets of data.

Next, we made improvements according to the SPO cleaning principles proposed by Fiszman et al [9] (ie, relevance, connectivity, novelty, and significance). We combined them with expert manual screening to ensure that the selected SPO triplets have a higher relevance. In the improved process, we did not predefine semantic patterns. Instead, we used a series of cleaning operations to select core SPO triplets and connected SPO triplets, eliminating SPO triplets lacking specific information and those that appeared only once in the frequency. The specific process is as follows:

1. In the same article, there may be repeated occurrences of identical SPO triplets. To maintain equal contribution from each article, we counted the repeated SPO triplets once within the same article.
2. To ensure statistical reliability, we calculated the occurrence frequency of each SPO triplet across different articles. SPO triplets with low occurrence frequencies may lack statistical significance. Therefore, we filtered SPO triplets with frequencies greater than or equal to 2.
3. Based on expert domain knowledge, we manually screened the selected SPO triplets with frequencies greater than or equal to 2 to identify those of research value.

Finally, we obtained 25,449 SPO triplets data. We imported the filtered SPO triplets information into the NetworkX for visual analysis to explore knowledge and information related to breast cancer prevention and treatment.

All analyses were conducted in a Python program (version 3.11.3; Python Software Foundation), primarily using Pandas, Matplotlib, WordCloud, and NetworkX packages [23-26].

Results

Summary of Included Literatures

A total of 3589 articles were published in 618 different journals. Among them, 191 articles were published in the same journal, while 293 journals had only 1 article published. The journals

were ranked based on the number of publications, and the top 100 journals accounted for 2631 articles, which is 73.30% of the total.

Semantic Relationships and Semantic Patterns

We mainly summarize semantic associations into 3 types: treatment and prevention, influencing or associated factors, and related diseases (Table S1 in Multimedia Appendix 1). Regarding treatment and prevention, the relationships include TREATS, ADMINISTERED_TO, USES, and PREVENTS, representing treatment drugs, surgeries, and preventive measures for breast cancer. Regarding influencing or associated factors, the relationships include ASSOCIATED_WITH, AFFECTS, and CAUSES, which represent diseases' impact and etiological factors. Regarding related diseases, the relationship COEXISTS_WITH represents the coexistence between different diseases. In the semantic patterns involving treatment (TREATS), the topp-TREATS-neop and topp-TREATS-podg have appeared over 1000 times.

Summary of SPO Triples

In terms of breast tumors, malignant neoplasms had the highest frequency, accounting for 42.3% (587/1387) of the total, followed by triple-negative breast neoplasms (56/1387, 4%) and human epidermal growth factor receptor 2 (*HER*2)–positive carcinoma of breast (54/1387, 4%; Table 1 and Multimedia Appendix 2).

Table . Summary of breast cancer subtypes and stages, treatment methods, and treatment drugs. The top 30 subtypes, treatment methods, and treatment drugs with higher frequencies in all data are presented for each group.

Group		Values, n (%)
Breast cancer subtypes and stages (n=1387)		
	Malignant neoplasm of breast	587 (42.3)
	Triple-negative breast neoplasms	56 (4)
	HER2[a]-positive carcinoma of breast	54 (3.9)
	Carcinoma breast stage IV	48 (3.5)
	Breast cancer metastatic	47 (3.4)
	Early-stage breast carcinoma	42 (3)
	Malignant neoplasms	31 (2.2)
	Neoplasm	30 (2.2)
	Metastatic triple-negative breast carcinoma	26 (1.9)
	High-risk cancer	24 (1.7)
	Neoplasm metastasis	21 (1.5)
	Advanced cancer	19 (1.4)
	Advanced breast carcinoma	19 (1.4)
	HER2-negative breast cancer	18 (1.3)
	Locally advanced malignant neoplasm	17 (1.2)
	Advanced malignant neoplasm	15 (1.1)
	Nonsmall cell lung carcinoma	15 (1.1)
	Noninfiltrating intraductal carcinoma	14 (1)
	Locally advanced breast cancer	13 (0.9)
	Breast cancer stage III	11 (0.8)
Treatment of breast cancer (n=1387)		
	Pharmacotherapy	267 (19.3)
	Neoadjuvant therapy	88 (6.3)
	Hormone therapy	68 (4.9)
	Chemotherapy (adjuvant)	54 (3.9)
	Therapeutic procedure	53 (3.8)
	Radiation therapy	48 (3.5)
	Treatment protocols	43 (3.1)
	Adjuvant therapy	36 (2.6)
	Breast-conserving surgery	35 (2.5)
	First-line treatment	31 (2.2)
	Single-agent therapy	27 (1.9)
	Mastectomy	27 (1.9)
	Operative surgical procedures	20 (1.4)
	Interventional procedure	16 (1.2)
	Radiotherapy (adjuvant)	14 (1)
	Excision of axillary lymph nodes group	13 (0.9)
	Combined modality therapy	12 (0.9)
	Excision	11 (0.8)
	Targeted therapy	11 (0.8)

Group		Values, n (%)
	Placebos	10 (0.7)
Drugs for breast cancer (n=1805)		
	Trastuzumab	209 (11.6)
	Capecitabine	88 (4.9)
	Paclitaxel	81 (4.5)
	Aromatase inhibitors	64 (3.5)
	Immunologic adjuvants	62 (3.4)
	Letrozole	58 (3.2)
	Bevacizumab	48 (2.7)
	Tamoxifen	40 (2.2)
	Gemcitabine	36 (2)
	Pertuzumab	36 (2)
	Fulvestrant	36 (2)
	Cyclophosphamide	32 (1.8)
	Pembrolizumab	30 (1.7)
	Docetaxel	27 (1.5)
	Taxane	27 (1.5)
	Ado-trastuzumab emtansine	22 (1.2)
	130-nm albumin-bound paclitaxel	22 (1.2)
	Carboplatin	22 (1.2)
	Eribulin	21 (1.2)
	Palbociclib	19 (1.1)
	Exemestane	19 (1.1)
	Everolimus	19 (1.1)
	Olaparib	18 (1)
	Talazoparib	17 (0.9)
	Pharmaceutical preparations	16 (0.9)
	Protein-tyrosine kinase inhibitor	15 (0.8)
	Cisplatin	14 (0.8)
	Lapatinib	14 (0.8)
	Fluorouracil	13 (0.7)
	Preservative free ingredient	13 (0.7)

[a]*HER*2: human epidermal growth factor receptor 2.

Pharmacotherapy is the most common treatment method, accounting for 19.2% (267/1387) of the overall frequency. Additionally, other high-frequency treatment modalities include neoadjuvant therapy (88/1387, 6%), hormone therapy (68/1387, 5%), adjuvant chemotherapy (54/1387, 4%), and radiation therapy (48/1387, 3%; Table 1 and Multimedia Appendix 3). In breast cancer treatment drugs, trastuzumab (209/1805, 11.6%), capecitabine (88/1805, 5%), paclitaxel (81/1805, 4%), aromatase inhibitors (64/1805, 4%), and immunologic adjuvants (62/1805, 3%) have a relatively high frequency of occurrence (Table 1 and Multimedia Appendix 4).

Breast Cancer Knowledge Graph

We visualized the SPO triples and displayed 3 subgroups: breast cancer treatment methods, therapeutic drugs, and relevant preventive measures. Figure 1 shows the relationship between different subtypes and stages of breast cancer and treatment methods. In different subtypes of breast cancer, the highest frequency is observed in malignant neoplasm of the breast, with pharmacotherapy having the highest frequency among various treatment modalities. Different subtypes simultaneously correspond to multiple treatment modalities; likewise, a single treatment modality corresponds to multiple breast cancer subtypes.

Figure 1. Relationship between different subtypes and stages of breast cancer and treatment methods. *HER2*: human epidermal growth factor receptor 2.

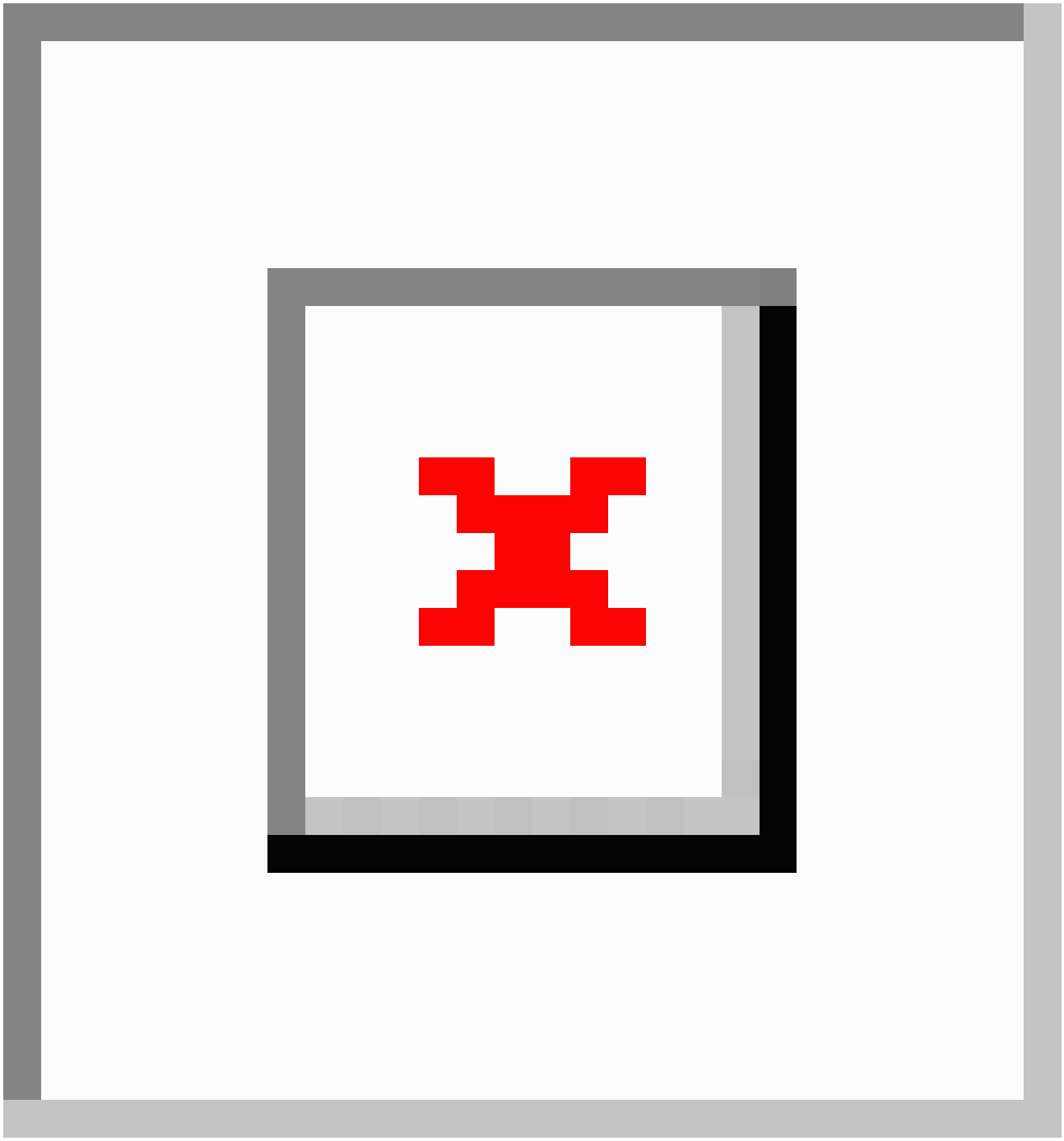


Figure 2 shows the relationship between different subtypes and stages of breast cancer and drugs. Among the therapeutic drugs for breast cancer, trastuzumab has the highest frequency and corresponds to the most types of breast cancer. Capecitabine, paclitaxel, aromatase inhibitors, and immunologic adjuvants also have relatively high frequencies. In comparison, immunologic adjuvants have the fewest connections with different types of breast cancer.

Figure 2. Relationship between different subtypes and stages of breast cancer and drugs. *HER2*: human epidermal growth factor receptor 2.

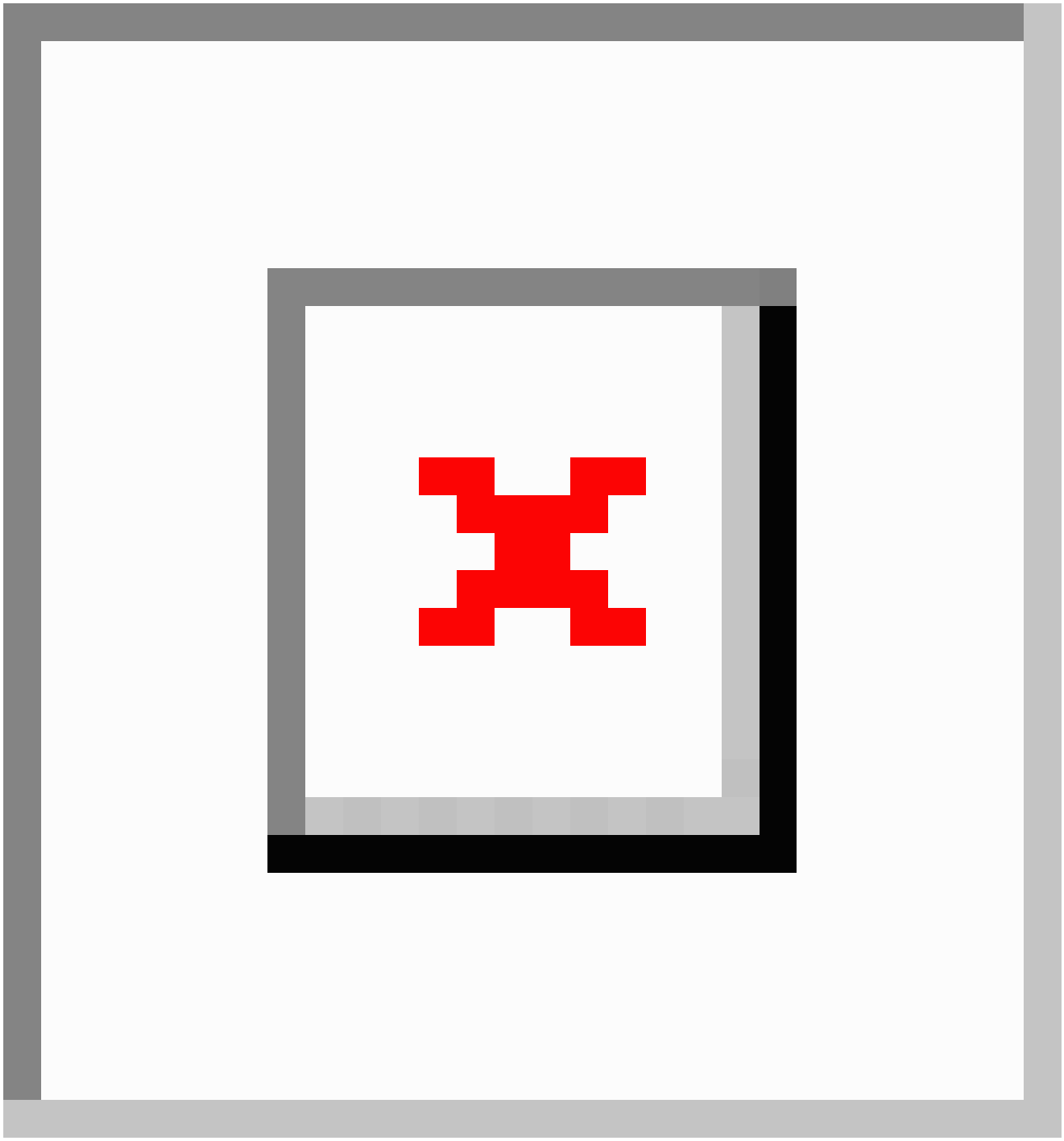


Figure 3 shows the relationship between breast cancer treatment and adverse reactions. Pharmacotherapy is associated with neuropathy, onycholysis, heart neutropenia failure, alopecia, febrile neutropenia, anemia, stomatitis, leukopenia, thrombocytopenia, premature menopause, and gastrointestinal dysfunction. Additionally, multiple nodes are connected, forming multiple pathways, such as pharmacotherapy-febrile neutropenia-adjuvant chemotherapy and pharmacotherapy-leukopenia-breast cancer therapeutic procedure-osteoporosis.

Figure 3. Relationship between breast cancer treatment and adverse reactions.

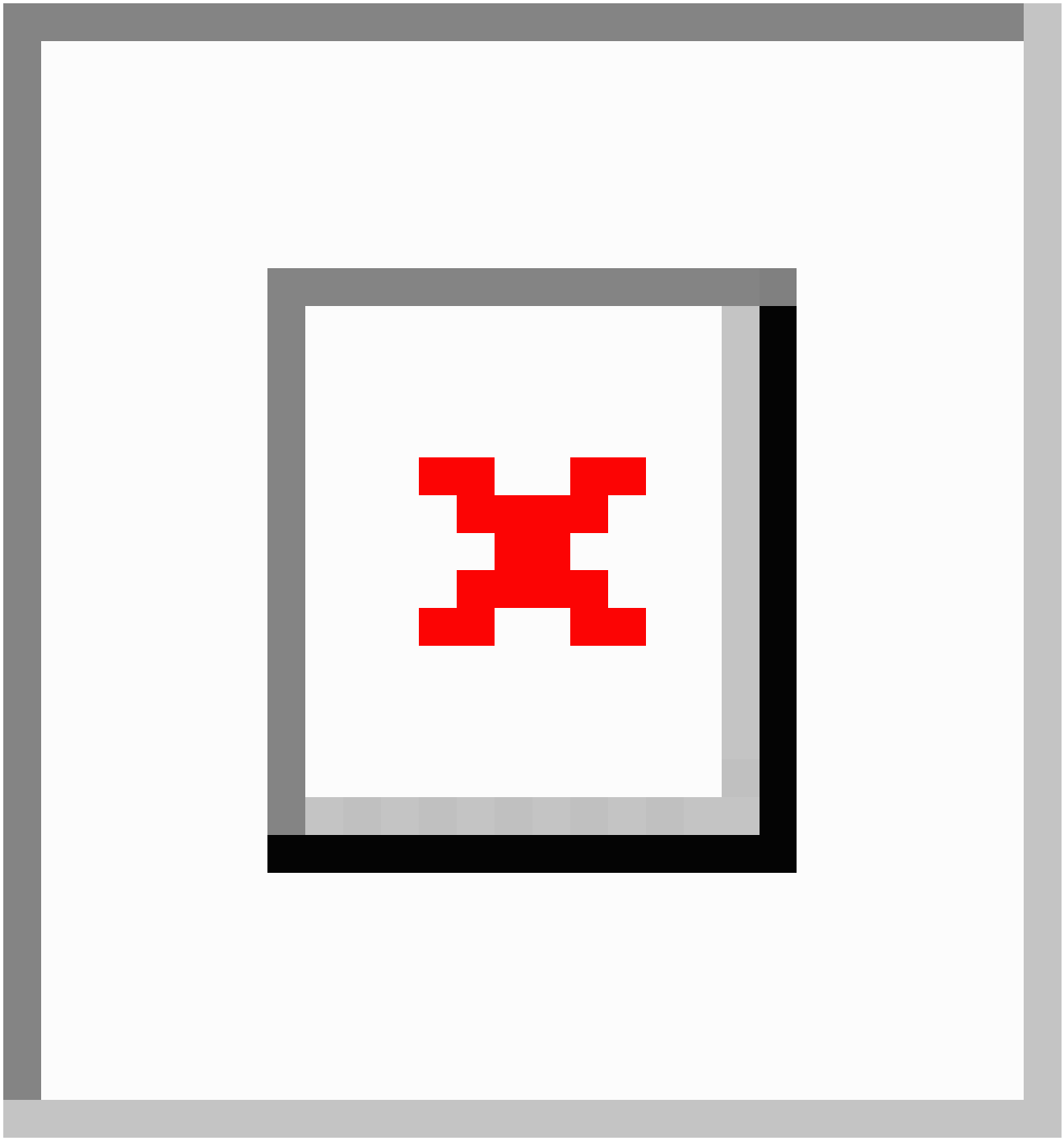
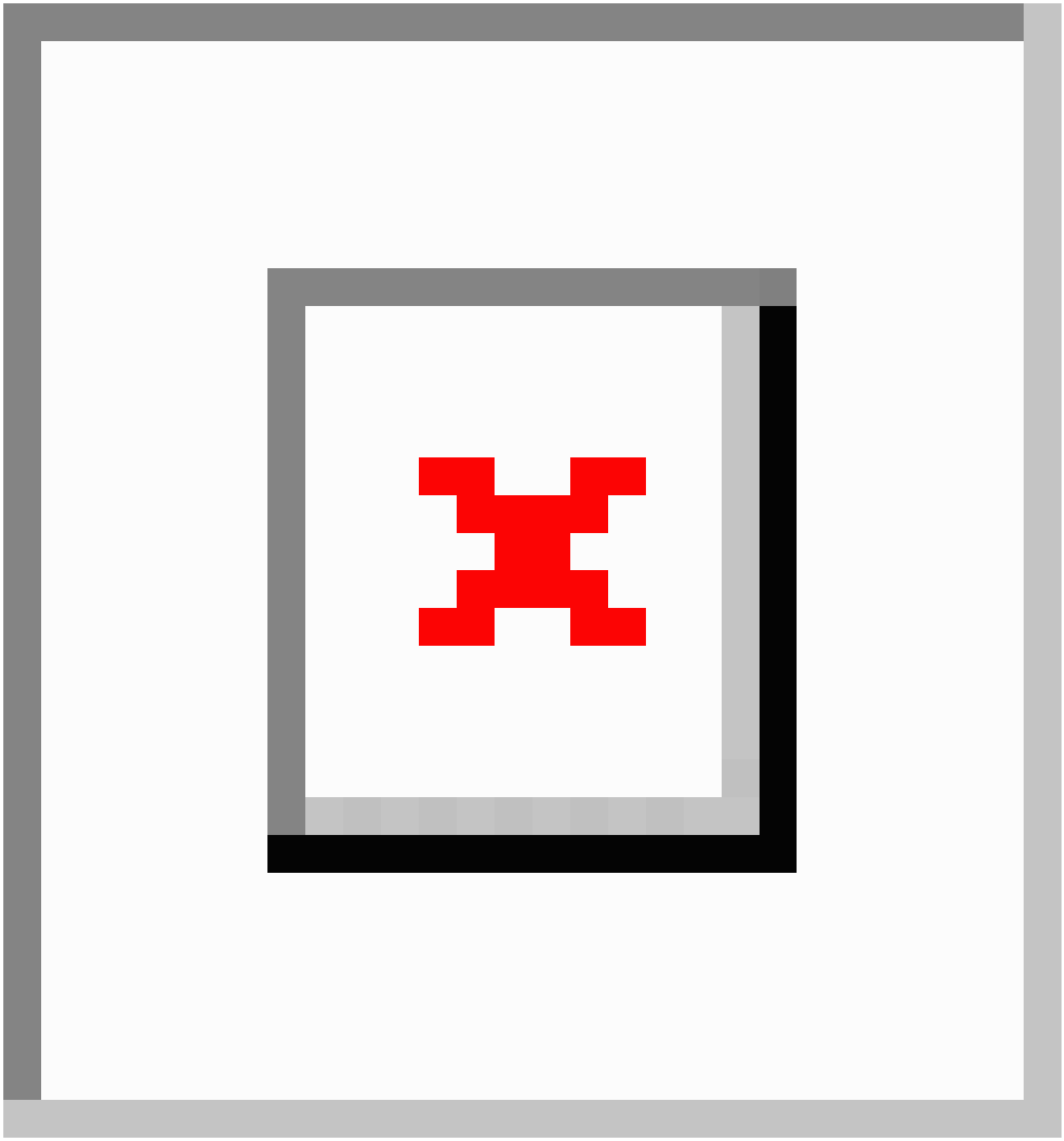


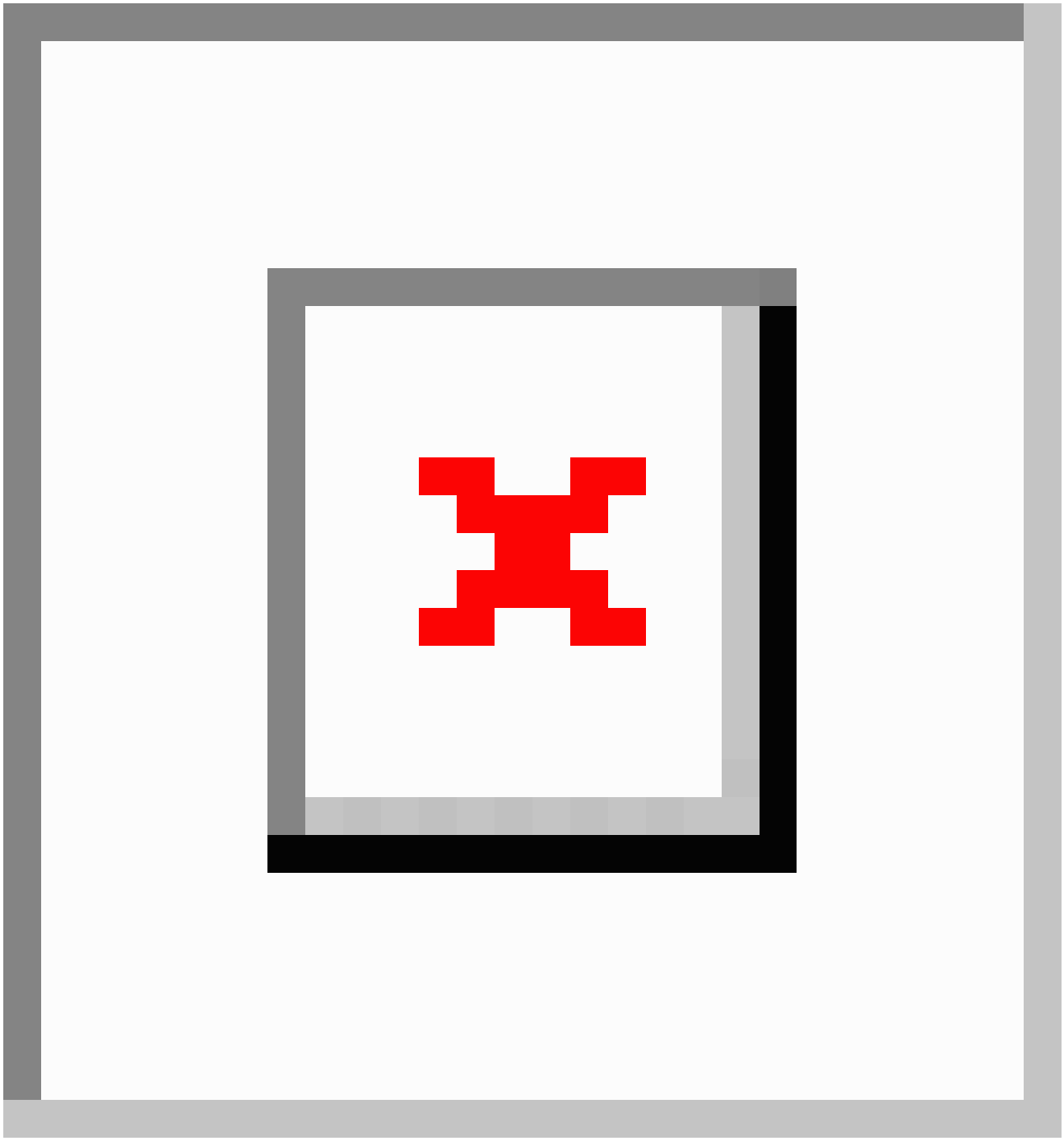
Figure 4 shows the relationship between adverse events after breast cancer treatment and preventive measures. Peripheral neuropathy is associated with cryotherapy, low-level laser therapy, compression procedure, acupuncture procedure, pharmacotherapy, and massage. Lymphedema is associated with resistance education, axillary lymph node dissection, physical therapy, excision of axillary lymph nodes group, and drainage of lymphatics. Early radiation dermatitis is associated with topical administration and bleomycin, cisplatin, or methotrexate protocol. In addition, there are some adverse reactions with relatively few treatment measures, such as stomatitis-diet, alopecia-scalp cooling.

Figure 4. Relationship between adverse reactions after breast cancer treatment and preventive measures.



We performed a relationship visualization to gain a better understanding of the association between types of breast cancer, treatments, drugs, and genes. Figure 5 intuitively reflects the high frequency of malignant neoplasm of the breast, pharmacotherapy, and trastuzumab. In addition, breast malignant tumors are associated with multiple genes, such as the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*) gene, platelet-derived growth factor receptor beta (*PDGFRB*) gene, phosphatase and tensin homolog (*PTEN*) gene, and erb-B2 receptor tyrosine kinase 2 (*ERBB2*) gene.

Figure 5. Relationship between types of breast cancer, treatments, drugs, and genes.



Discussion

Principal Findings

The knowledge graphs constructed in this study help researchers understand the research hot spots in breast cancer over the past 5 years. The complex network involving treatment methods, drugs, adverse reactions, preventive measures, and genes in breast cancer can assist clinicians in making decisions that comprehensively consider multiple aspects, ultimately aiding in decisions that are the most beneficial to patients. Additionally, the knowledge graph allows for personalized considerations based on specific genes for individualized patients.

This study found that from 2018 to 2022, breast malignancies appeared most frequently in the literature and were the primary concern for researchers. Research interest in triple-negative breast neoplasms is higher than in other subtypes. This phenomenon may be due to the higher risk of recurrence and poor prognosis in patients with early-stage triple-negative breast neoplasms [10], making it a subject of greater concern to clinicians and researchers. Among treatment modalities, pharmacotherapy receives the highest attention. Pharmacotherapy for breast cancer primarily involves chemotherapy, endocrine therapy, and targeted therapy [27]. Compared to traditional surgery and radiotherapy, pharmacotherapy can more precisely intervene in the growth and division of cancer cells by targeting specific molecules or

cellular structures, which reduces damage to normal cells and allows for the formulation of personalized treatment plans based on the patient's genotype and molecular characteristics [28]. Medications circulating through the bloodstream can also act on cancer cells throughout the body, preventing cancer cell metastasis. These advantages of pharmacotherapy may be related to the heightened emphasis on pharmacotherapy over the past 5 years. Trastuzumab receives the highest attention in breast cancer pharmacotherapy; it is a specific cancer-targeting medication used in the treatment of cancers characterized by elevated levels of HER2 protein [29].

Pharmacotherapy is associated with various adverse reactions, including neutropenia, neuropathy, onycholysis, heart failure, alopecia, and febrile neutropenia. Among these adverse reactions, peripheral neuropathy and lymphedema have the most corresponding preventive and treatment measures, with lymphedema being a common complication after surgery [30]. However, there is limited research on how to prevent and treat the potential adverse reactions of pharmacotherapy, and further studies are needed. Various adverse effects of breast cancer treatment may reduce patients' adherence to treatment. Therefore, when clinicians choose different treatments and drugs, they should pay close attention to their potential adverse reactions and how to prevent or mitigate them.

In existing knowledge graphs related to breast cancer, one study from China constructed a knowledge graph using electronic medical records, clinical guidelines, and expert opinions, primarily focusing on breast cancer diagnosis [18]. Another study by Chinese scholars also used data from various sources, including clinical guidelines, medical encyclopedias, and electronic medical records, to construct a knowledge graph primarily applied to medical knowledge question-answering and medical record retrieval [19]. These studies used data from multiple sources, including structured, unstructured, and semistructured data. Data extraction and accuracy face challenges. Therefore, they used neural network models for training and calculated a series of metrics to ensure data accuracy. For instance, they utilized BERT + Bi-LSTM+ CRF for textual data to achieve named entity recognition. In this study, SemMedDB was used as the data source, and the database was constructed by extracting semantic information from PubMed using SemRep, which demonstrated good performance in a biomedical text [21].

In summary, the knowledge graph constructed in this study for breast cancer treatment and prevention encompasses information on different stages, subtypes of breast cancer, treatment modalities, medications, adverse reactions, and preventive measures. This knowledge forms a complex network, providing clinical practitioners with a comprehensive and referenced knowledge base. We recommend that clinical practitioners apply our research findings in several aspects. First, clinicians can gain insights into the current state of breast cancer treatment and prevention research through our study. Additionally, there is a relative lack of preventive measures and strategies for mitigating postoperative and postmedication adverse reactions compared to breast cancer treatment, and more efforts are needed in these areas. Furthermore, our research can assist clinicians in making comprehensive decisions. For instance, when selecting a treatment approach for patients, the knowledge graph facilitates linking to available medications, associated adverse reactions, and measures to mitigate or prevent adverse effects.

Our research still has several limitations. First, SemRep, as a natural language processing program based on the UMLS, still exhibits shortcomings. Despite the extensive coverage and scale of the UMLS Metathesaurus, it has a relatively limited ability to recognize entities. There are still areas for improvement in processing natural language texts [20]. Second, clinical researchers often prefer causal relationships rather than pure correlations; however, our study can only reveal the connections between pieces of information and cannot determine the magnitude and direction of their effects. Third, with the release of new literature, the knowledge graph also needs to be updated promptly, increasing the burden on researchers. Future improvements should focus on automating the mining of literature data to ensure timely updates to the knowledge graph for breast cancer prevention and treatment, thereby alleviating the burden on researchers.

Conclusions

This study successfully constructed a knowledge graph for breast cancer prevention and treatment by integrating relevant literature from the past 5 years and conducting knowledge discovery. Through this knowledge graph, researchers can learn about breast cancer treatment methods, medications, and adverse reactions to preventive treatments and gain insights into the relationships between different pieces of knowledge.

Acknowledgments

The authors would like to thank Feng Xixi, associate chief physician and member of the Chronic Disease Special Committee of the Chengdu City Preventive Medicine Association, for her suggestions at the initial stage of the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table depicting the semantic relationship and semantic schema of breast cancer.
[DOCX File, 19 KB - medinform_v12i1e52210_app1.docx]

Multimedia Appendix 2
Different subtypes and stages of breast cancer.
[PNG File, 158 KB - medinform_v12i1e52210_app2.png]

Multimedia Appendix 3
Treatments of breast cancer.
[PNG File, 214 KB - medinform_v12i1e52210_app3.png]

Multimedia Appendix 4
Drugs for breast cancer.
[PNG File, 160 KB - medinform_v12i1e52210_app4.png]

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424. [doi: 10.3322/caac.21492] [Medline: 30207593]
2. Xiao Y, Xia J, Li L, et al. Associations between dietary patterns and the risk of breast cancer: a systematic review and meta-analysis of observational studies. Breast Cancer Res 2019 Jan 29;21(1):16. [doi: 10.1186/s13058-019-1096-1] [Medline: 30696460]
3. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249. [doi: 10.3322/caac.21660] [Medline: 33538338]
4. Thakur P, Seam RK, Gupta MK, Gupta M, Sharma M, Fotedar V. Breast cancer risk factor evaluation in a Western Himalayan state: a case-control study and comparison with the Western World. South Asian J Cancer 2017;6(3):106-109. [doi: 10.4103/sajc.sajc_157_16] [Medline: 28975116]
5. Badr LK, Bourdeanu L, Alatrash M, Bekarian G. Breast cancer risk factors: a cross- cultural comparison between the west and the east. Asian Pac J Cancer Prev 2018 Aug 24;19(8):2109-2116. [doi: 10.22034/APJCP.2018.19.8.2109] [Medline: 30139209]
6. Zhang X, Dong XP, Guan YZ, Me R, Guo DL, He YT, et al. Research progress on epidemiological trend and risk factors of female breast cancer. Cancer Res Prev Treat 2021;48(1):87-92.
7. Tan MM, Ho WK, Yoon SY, et al. A case-control study of breast cancer risk factors in 7,663 women in Malaysia. PLoS One 2018;13(9):e0203469. [doi: 10.1371/journal.pone.0203469] [Medline: 30216346]
8. Britt KL, Cuzick J, Phillips KA. Key steps for effective breast cancer prevention. Nat Rev Cancer 2020 Aug;20(8):417-436. [doi: 10.1038/s41568-020-0266-x] [Medline: 32528185]
9. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. In: Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004: Association for Computational Linguistics; 2004:76-83. [doi: 10.5555/1596431.1596442]
10. For the progress of adjuvant treatment of triple-negative breast cancer, just look at these 8 key clinical studies! [Article in Chinese]. Sohu. 2021 Dec 14. URL: https://www.sohu.com/a/508222106_121118854 [accessed 2023-06-25]
11. Feng B, Gao J. AnthraxKP: a knowledge graph-based, anthrax knowledge portal mined from biomedical literature. Database (Oxford) 2022 Jun 2;2022:baac037. [doi: 10.1093/database/baac037] [Medline: 35653350]
12. Feng F, Tang F, Gao Y, et al. GenomicKB: a knowledge graph for the human genome. Nucleic Acids Res 2023 Jan 6;51(D1):D950-D956. [doi: 10.1093/nar/gkac957] [Medline: 36318240]
13. James T, Hennig H. Knowledge graphs and their applications in drug discovery. Methods Mol Biol 2024;2716:203-221. [doi: 10.1007/978-1-0716-3449-3_9] [Medline: 37702941]
14. Lyu K, Tian Y, Shang Y, et al. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. J Biomed Inform 2023 Mar;139:104298. [doi: 10.1016/j.jbi.2023.104298] [Medline: 36731730]
15. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017 Jan 4;45(D1):D833-D839. [doi: 10.1093/nar/gkw943] [Medline: 27924018]
16. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018 Jan 4;46(D1):D1074-D1082. [doi: 10.1093/nar/gkx1037] [Medline: 29126136]
17. ClinVar. National Library of Medicine. URL: https://www.ncbi.nlm.nih.gov/clinvar [accessed 2023-11-18]
18. Li X, Sun S, Tang T, et al. Construction of a knowledge graph for breast cancer diagnosis based on Chinese electronic medical records: development and usability study. BMC Med Inform Decis Mak 2023 Oct 10;23(1):210. [doi: 10.1186/s12911-023-02322-0] [Medline: 37817193]
19. An B. Construction and application of Chinese breast cancer knowledge graph based on multi-source heterogeneous data. Math Biosci Eng 2023 Feb 6;20(4):6776-6799. [doi: 10.3934/mbe.2023292] [Medline: 37161128]

20. Li XY, Li JL, Li ZY. Integrated medical language system and its application in knowledge discovery. Digital Library Forum 2019;9:24-29.
21. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics 2020 May 14;21(1):188. [doi: 10.1186/s12859-020-3517-7] [Medline: 32410573]
22. Access to SemRep/SemMedDB/SKR resources. National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR.html [accessed 2023-11-18]
23. McKinney W. Pandas: a foundational Python library for data analysis and statistics. In: Python for High Performance and Scientific Computing: Deutsches Zentrum für Luft-und Raumfahrt; 2010:293-296.
24. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3):90-95. [doi: 10.1109/MCSE.2007.55]
25. WordCloud for Python documentation. Andreas C. Müller - Machine Learning Scientist. URL: https://amueller.github.io/word_cloud/ [accessed 2023-12-25]
26. Hagberg A, Swart PJ, Schult DA. Exploring Network Structure, Dynamics, and Function Using NetworkX: Los Alamos National Lab (LANL); 2008.
27. The difference between breast cancer radiotherapy, targeted therapy and chemotherapy! [Article in Chinese]. Sohu. 2018 Dec 7. URL: https://www.sohu.com/a/280208482_790163 [accessed 2023-11-18]
28. Nagini S. Breast cancer: current molecular therapeutic targets and new players. Anticancer Agents Med Chem 2017;17(2):152-163. [doi: 10.2174/1871520616666160502122724] [Medline: 27137076]
29. Trastuzumab. Cancer Research UK. URL: https://www.cancerresearchuk.org/about-cancer/treatment/drugs/trastuzumab [accessed 2023-11-18]
30. Bernas M, Thiadens SRJ, Smoot B, Armer JM, Stewart P, Granzow J. Lymphedema following cancer therapy: overview and options. Clin Exp Metastasis 2018 Aug;35(5-6):547-551. [doi: 10.1007/s10585-018-9899-5] [Medline: 29774452]

Abbreviations

ERBB2: erb-B2 receptor tyrosine kinase 2
HER2: human epidermal growth factor receptor 2
PDGFRB: platelet-derived growth factor receptor beta
PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PTEN: phosphatase and tensin homolog
SemMedDB: Semantic MEDLINE Database
SPO: Subject-Predicate-Object
UMLS: Unified Medical Language System

Edited by A Benis; submitted 26.08.23; peer-reviewed by C Gaudet-Blavignac, S Yang, Y Chu; revised version received 02.01.24; accepted 06.01.24; published 22.02.24.

Please cite as:
Jin S, Liang H, Zhang W, Li H
Knowledge Graph for Breast Cancer Prevention and Treatment: Literature-Based Data Analysis Study
JMIR Med Inform 2024;12:e52210
URL: https://medinform.jmir.org/2024/1/e52210
doi:10.2196/52210

Original Paper

Recognition of Daily Activities in Adults With Wearable Inertial Sensors: Deep Learning Methods Study

Alberto De Ramón Fernández[1], PhD; Daniel Ruiz Fernández[1], PhD; Miguel García Jaén[2], PhD; Juan M. Cortell-Tormo[2], PhD

[1]Department of Computer Technology, University of Alicante, San Vicente del Raspeig, Spain

[2]Department of General Didactics and Specific Didactics, University of Alicante, San Vicente del Raspeig, Spain

Corresponding Author:
Daniel Ruiz Fernández, PhD
Department of Computer Technology
University of Alicante
Carretera San Vicente del Raspeig s/n
San Vicente del Raspeig, 03690
Spain
Phone: 34 965 90 9656 ext 3331
Email: druiz@dtic.ua.es

Abstract

Background: Activities of daily living (ADL) are essential for independence and personal well-being, reflecting an individual's functional status. Impairment in executing these tasks can limit autonomy and negatively affect quality of life. The assessment of physical function during ADL is crucial for the prevention and rehabilitation of movement limitations. Still, its traditional evaluation based on subjective observation has limitations in precision and objectivity.

Objective: The primary objective of this study is to use innovative technology, specifically wearable inertial sensors combined with artificial intelligence techniques, to objectively and accurately evaluate human performance in ADL. It is proposed to overcome the limitations of traditional methods by implementing systems that allow dynamic and noninvasive monitoring of movements during daily activities. The approach seeks to provide an effective tool for the early detection of dysfunctions and the personalization of treatment and rehabilitation plans, thus promoting an improvement in the quality of life of individuals.

Methods: To monitor movements, wearable inertial sensors were developed, which include accelerometers and triaxial gyroscopes. The developed sensors were used to create a proprietary database with 6 movements related to the shoulder and 3 related to the back. We registered 53,165 activity records in the database (consisting of accelerometer and gyroscope measurements), which were reduced to 52,600 after processing to remove null or abnormal values. Finally, 4 deep learning (DL) models were created by combining various processing layers to explore different approaches in ADL recognition.

Results: The results revealed high performance of the 4 proposed models, with levels of accuracy, precision, recall, and F_1-score ranging between 95% and 97% for all classes and an average loss of 0.10. These results indicate the great capacity of the models to accurately identify a variety of activities, with a good balance between precision and recall. Both the convolutional and bidirectional approaches achieved slightly superior results, although the bidirectional model reached convergence in a smaller number of epochs.

Conclusions: The DL models implemented have demonstrated solid performance, indicating an effective ability to identify and classify various daily activities related to the shoulder and lumbar region. These results were achieved with minimal sensorization—being noninvasive and practically imperceptible to the user—which does not affect their daily routine and promotes acceptance and adherence to continuous monitoring, thus improving the reliability of the data collected. This research has the potential to have a significant impact on the clinical evaluation and rehabilitation of patients with movement limitations, by providing an objective and advanced tool to detect key movement patterns and joint dysfunctions.

(JMIR Med Inform 2024;12:e57097) doi:10.2196/57097

KEYWORDS

activities of daily living; ADL; ADLs; deep learning; deep learning models; wearable inertial sensors; clinical evaluation; patient's rehabilitation; rehabilitation; movement; accelerometers; accelerometer; accelerometry; wearable; wearables; sensor; sensors; gyroscopes; gyroscope; monitor; monitoring

Introduction

Activities of daily living (ADL) are the most basic tasks of the person, as they enable them to function with a minimum of autonomy. ADL are crucial for maintaining quality of life and personal well-being, serving as indicators of functional status [1-3]. ADL are an indicator of a person's functional status and include basic physical tasks such as moving, eating, dressing, maintaining personal hygiene, and grooming, as well as more complex and instrumental activities such as working, shopping, cleaning, exercising, and participating in recreational activities [2-4]. Impaired physical function can limit the execution of these tasks, affecting personal goals and independent living. This condition can affect the individual's ability to achieve personal goals and maintain an independent quality of life [2,5,6]. Therefore, it is necessary to assess this deterioration during the execution of ADL in different preventive, clinical, or rehabilitation contexts [6-8].

The functional assessment of ADL is complex, so it is advisable to approach it based on the evaluation of fundamental movement patterns on which these ADL are developed [9-11]. The shoulder and lumbar region are key joint complexes in this regard. Specifically, the shoulder joint is essential in many basic ADL, providing the mobility and stability necessary to perform actions in all planes of movement. It is essential to position the hand in space in a way that allows one to reach objects, eat, button a shirt, unbutton a bra, or comb one's hair [9,12-14]. The movement patterns most used in its assessment are scapula-humeral elevation in the sagittal and frontal plane and rotations at different elevation angles [9,13,14]. Similarly, the lumbar region is a joint complex that has a close relationship with basic movement patterns such as flexion and extension of the trunk in the sagittal plane but also in extremely important actions such as sitting and standing up [10,15-18]. Various ADL derive from this fundamental movement pattern, the most studied being the gestures of sitting and getting up from a chair, bending or crouching, and lifting an object or weight [15-17,19].

The precise evaluation, control, and monitoring of ADL performance are fundamental tasks, although not simple, in the development of effective intervention tools in these clinical and rehabilitation contexts. Traditionally, the assessment of ADL has been based on direct observation and subjective evaluation by therapists, which entails biases, errors, and lack of precision in the results [6,20-22]. In contrast, recent advancements in technology, including wearable health monitoring devices, smart clothing sensors, and mobility assistance devices, enable the objective assessment and quantification of personal performance during ADL [23-27]. This technology includes wearable devices, motion sensors, and 2D or 3D motion capture systems, which allow complex movements and functionality of key joints, such as the shoulder or lumbar region, to be accurately recorded and analyzed during the performance of ADL [4,9,15]. However, limitations such as its high acquisition and implementation cost, its specialized technical knowledge, its lack of transparency and complexity, or its lack of validation and reliability hinder its applicability in the specific clinical or rehabilitation context [4,9,24,25].

A promising solution to overcome the aforementioned limitations is the use of wearable inertial sensors [28-34]. These have been gaining substantial scientific interest due to their potential to provide real-time information on kinematic aspects of human movement through continuous, dynamic, and minimally invasive monitoring. In the clinical and rehabilitation field, this technology has emerged as a simple and low-cost alternative to obtain precise information on accelerations, angular velocities, and trajectories in the different planes of movement during the execution of different basic ADL. This technology offers several advantages. It allows for a more accurate and objective assessment of the functionality of key joint complexes, identifying specific areas of weakness or limitation in movement during ADL and providing quantitative data on the person's progress over time [28,35,36]. On the other hand, it favors the motivation of patients, by being able to visualize their evolution, thus improving treatment adherence [28,31,37].

However, inertial sensors have some limitations. Despite being light and small, these devices may not be entirely transparent for users, especially due to the high number of sensors that, in many cases, must be used to obtain data that accurately interpret human movement [30,38,39].

Compared with the traditional approach of most studies that only use wearable inertial sensors to monitor kinematic aspects of human movement, the use of artificial intelligence (AI) techniques has been gaining popularity, by helping to improve the process of assessing and supervising different body movements using inertial sensors, in addition to reducing the number of sensors necessary for this [40-42].

In Yen at al [40], a wearable device consisting of a microcontroller and an inertial sensor placed on the participant's waist is presented. The signals collected by the accelerometer and gyroscope were used to train a 1D convolutional neural network–based feature learning model, enabling the identification of 6 ADL. The results demonstrated high accuracy in both external and study data, validating the effectiveness of the proposed method.

The study by Huynh-The et al [41] introduces an innovative method for recognizing ADL- and sports-related activities using wearable sensors. This method involves converting inertial data into color images, facilitating the learning of highly discriminative features using convolutional neural networks. Experimental results showed recognition accuracy of over 95%, outperforming other deep learning (DL)–based approaches for human activity recognition (HAR).

In Ronald et al [43], a novel DL model inspired by the Inception-ResNet architecture is presented for HAR tasks. The proposed model, trained on data collected from smartphones and inertial sensors capturing accelerometer, gyroscope, magnetometer, GPS, temperature, and heart rate signals, achieved remarkable performance across different data sets, demonstrating its flexibility and adaptability to varying signal types and quantities.

Meanwhile, Poulose et al [44] address the challenges of HAR in health care systems by proposing an approach based on a human image threshing machine using smartphone camera images. The human image threshing system uses mask region–based convolutional neural networks for human detection and a DL model for activity classification, achieving a precision of 98.53% and surpassing conventional sensor-based HAR approaches.

This study is based on the combination of accelerometer and gyroscope signals with AI techniques for the assessment of the shoulder and lumbar spine. AI algorithms can process the data captured by inertial sensors and perform sophisticated analyses to detect patterns, identify alterations in movement, and provide relevant clinical information. This facilitates a more complete and accurate evaluation of the joint movement of the shoulder or lower back, allowing a better understanding of dysfunctions and personalization of treatment and rehabilitation plans. The key contributions made by this study are summarized as follows:

1. Accelerometer and gyroscope signals with AI integration for enhanced ADL assessment: This combination shows great potential for the assessment of shoulder and lumbar region motion in basic ADL performance, providing an objective and advanced perspective in clinical evaluation and rehabilitation. However, validly and reliably demonstrating its use as a control and evaluation tool for ADL performance, in gestures such as eating, combing hair, dressing, sitting, or standing, still appears as an unresolved research challenge. Therefore, in this study, we aim to address the automatic detection and monitoring, using AI techniques, of the patient's basic ADL related to the shoulder and back.
2. Enhanced activity recognition precision: Our study relies on direct capture of inertial sensor signals, potentially offering a more precise and less image quality–dependent solution.
3. Efficient sensors use: For signal capture, only 2 sensors are used. Furthermore, it is intended to achieve this objective through minimal, noninvasive, and practically transparent sensorization for the user, improving adherence to the monitoring process and facilitating the integration of technology into the individual's daily life at a low cost.
4. Direct inertial data approach: Our study focuses on the direct use of accelerometer and gyroscope data without requiring additional conversion for model training.
5. Broad scope and versatility: It covers a wide range of activities, showcasing its versatility and adaptability.

We believe that this novel approach will make a significant contribution to this field of research, as it can be used in the prevention, clinical, or rehabilitation contexts of the shoulder and lumbar region.

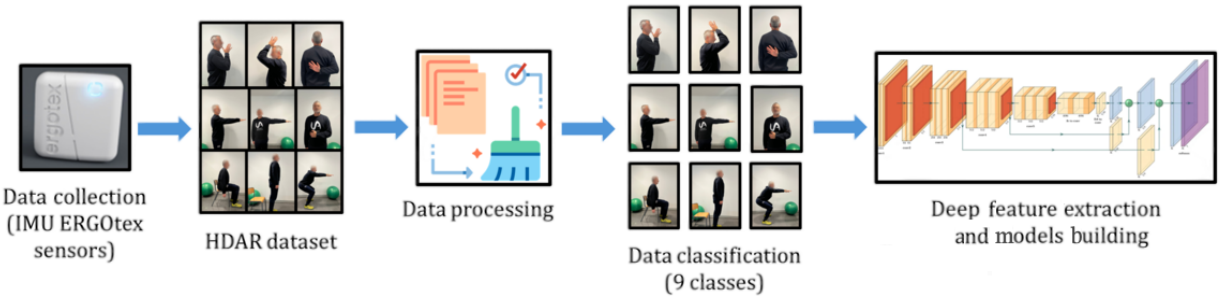
The remainder of the paper is organized as follows: the *Methods* section addresses how the database was generated, the processing layers used, and the architecture of the 4 developed DL models, as well as the parameters selected for their training and optimization. In the *Results* section, the evaluation outcomes obtained by the 4 DL models are presented, analyzed, and compared. Finally, the *Discussion* section presents a discussion of the principal findings and conclusions regarding our study.

Methods

Overview

This research work focuses on the detection and automatic monitoring of ADL using AI models (DL models) and wearable inertial sensors to prevent or diagnose injuries, as well as supervise rehabilitation processes. Figure 1 presents an overview of the methodology proposed. In the following subsections, each step is explained in depth.

Figure 1. System overview. HDAR: Human Daily Activities Recognition ; IMU: inertial measurement unit.



Ethical Considerations

This study has been conducted in strict accordance with the ethical principles outlined in the Declaration of Helsinki. Approval for this research was obtained from the Ethics Committee of the University of Alicante (protocol code UA-2023-11-16).

Prior to commencement, participants provided written informed consent. Respect for participants, including their autonomy, confidentiality, and well-being, has been ensured.

All collected data have undergone a rigorous anonymization process, safeguarding the privacy of the individuals involved in the research. Protective measures were implemented in

accordance with institutional guidelines to ensure the security of participant information throughout the study.

Participants involved in human subjects research were not provided with any form of compensation. This decision was made to uphold transparency and fairness in the research process and to minimize potential biases associated with compensation.

Data Collection

A total of 9 ADLs were included in the study, 6 of them related to the shoulder (eating [E], combing hair [CH], fastening the bra [FB], opening the door [OD], reaching for an object [RO], and buttoning up [BU]) and 3 related to the back (sitting [S], standing up [SU], and half squat [HS]). Figure 2 graphically shows these movements.

Figure 2. Graphic description of activities of daily living movements. In the top row (from left to right): eating, combing hair, and fastening the bra. In the middle row (from left to right): opening the door, reaching for an object, and buttoning up. Bottom row (left to right): sitting, standing up, and half squat.

To monitor movements, we used 2 self-developed inertial measurement unit ERGOtex model sensors [45,46]. This inertial measurement unit ERGOtex sensor comprises 3 triaxial accelerometers (±2 g, controlled noise at 100 µg/√Hz), triaxial gyroscopes (±1000 deg/s, sensitivity error within ±1%, and low noise level, at ±4 mdeg/s/√Hz), and magnetometers, encapsulated in a device (weight=8 g, dimensions=23×21×10 mm). The ICM-20602 MEMS MotionTracking (TDK Corp) device was selected for its high-performance specifications, critical for the reliability of the device. The incorporation of a 1K-byte FIFO buffer reduces serial bus congestion, enhancing measurement consistency and optimizing device power use. It operates at a sampling rate of 20 Hz, has an autonomy of 8 hours, and can be attached to the skin using double-sided tape or secured elsewhere using an elastic strap. These enhancements guarantee reliable response times and sensitivity levels, crucial for maintaining data accuracy (Figure 3).

Figure 3. ERGOtex inertial measurement unit sensors were developed for movement identification.

The inertial sensors were attached to the skin over the sacrum (S1) and the distal part of the upper extremity (close to the wrist). Primarily designed for monitoring spine posture and arm, this device records acceleration data across all 3 axes. Internal integration of the acceleration signal occurs within the device, transmitting data instantly via Bluetooth (frequency=2.4 GHz) to a smartphone or tablet equipped with the preinstalled app. This application enables immediate data visualization and facilitates export to a spreadsheet in comma-separated text format (CSV).

The database generated initially had 53,165 records of all activities. The records were grouped into batches of time series (of different lengths) that represented the different movements. Each record was made up of 12 attributes or numerical variables, corresponding to the value obtained by the accelerometer and gyroscope of each sensor during the execution of the movement according to its 3 axes (Acx, Acy, Acz, Gyx, Gyy, Gyz). After the processing stage, where null, missing, and abnormal values were eliminated, the database was reduced to 52,600 records (RO: n=6423, FB: n=6956, E: n=6216, OD: n=6472, SU: n=3678, CH: n=6010, BU: n=5915, HS: n=6630, and S: n=4300).

DL Models

Processing Layers

To create the DL models, different processing layers that perform the transformation, regularization, feature extraction, regularization, and dependency capture operations were combined. The basics of each of them are presented below.

1D Convolution Layer for Feature Extraction

A 1D convolutional layer is specifically designed to process data that follows a 1D structure, such as time series or text sequences. In the case of a 1D time series, the 1D convolution operation follows a similar process as a standard convolutional layer but is performed along 1 dimension instead of 2 [47]. The convolution operation is the key component of this type of layer. During the 1D convolution operation, a filter (kernel) of defined size slides along the time series, multiplying the filter values by the corresponding values in the time series and summing them to produce a single value at the output. This process is repeated for each filter position throughout the time series, thus generating a feature map that highlights relevant patterns in the data sequences. The 1D convolutional layer is essential for the automatic identification of patterns in time series, allowing efficient extraction of important features during the training process. By reducing the number of parameters and avoiding overfitting, 1D convolution helps capture the temporal structure of data and improve model performance in time series prediction or classification tasks [48]. Given an input 1D time series X and a set of filters F, the convolution operation is performed as follows (equation 1):

where Y_i is the output value at the feature map position i, X_{i+j} is the time series value at position $i+j$, * denotes the convolution operation, and b_i is the bias associated with the output F_j, and m is the filter size.

Long Short-Term Memory Layer for Modeling Temporal Dependencies

Long short-term memory (LSTM) layers are a type of recurrent layer designed to overcome the limitations of traditional recurrent neural networks in capturing long-term dependencies in temporal sequences [49]. Its design is based on the idea of using internal memory structures controlled by gates to manage information over time and make decisions about what information to retain and discard. In an LSTM, 3 main gates are introduced: the forget gate, which decides what information should be discarded from the previous memory; the input gate, which decides what new information should be stored in memory; and the output gate, which determines what memory information should be used to generate the output of the layer. These gates are controlled by activation functions and adjustable weights during training.

An overview of the fundamental equations of an LSTM cell is presented below, which describe how an LSTM cell manages information and gates to process and retain relevant information over time in a temporal sequence, given one input at a time step t, denoted as x_t, and the outputs of the previous time step [h_{t-1}] (LSTM cell output) and C_{t-1} (LSTM cell state).

Forget gate (f_t): decides what information should be discarded or forgotten from the cell state (equation 2)

Input gate (i_t): decide what new information to store in the cell state (equations 3 and 4)

The forgotten information and new information are then combined to update the state of the cell (equation 5).

Output gate (o_t): finally, the final activation at the current position (h_t) is calculated with the output gate (o_t), which regulates the amount of information to be output (equation 6)

Where σ is the sigmoid function; *tanh* is the hyperbolic tangent function; W_i,W_C and W_o are weight matrices that are learned during training; and b_f, b_i, b_c, and b_o are biases. [h_{t-1},x_t] denotes the concatenation of h_{t-1} and x_t before applying the linear operation.

Dropout Regularization Layer

The Dropout layer is a regularization strategy that prevents overfitting by introducing variability into the network during training [50]. This technique randomly turns off a percentage of units in each iteration, temporarily removing them and forcing the network to learn more robust representations. Based on the assembly concept, it simulates the presence or absence of units, improving effectiveness and reducing dependence on specific units. In addition to its impact on generalization, the Dropout layer acts as an effective regularization mechanism, improving modeling efficiency and performance by preventing overoptimization and facilitating generalization to unseen data [51,52].

Flatten and Fully Connected (Dense) Transformation Layers

The Flatten layer aims to transform 2D or 3D data into a 1D format, allowing for a more manageable representation and facilitating the transition from convolutional layers to dense layers [53]. Given a 3D input matrix where m, n, and p are the spatial dimensions, the Flatten layer converts this matrix into a 1D vector X' of size $m * n * p$.

The fully connected (FC) or Dense layer connects all neurons in 1 layer to all neurons in the next layer [48]. It performs linear transformations on the data followed by nonlinear activation functions, allowing complex representations to be learned. If X is the input of the Dense layer, W is the weight matrix, and b is the bias vector, the output Y is calculated as (equation 8):

where σ is the activation function.

1D MaxPooling Layer for Feature Reduction

The 1D MaxPooling layer is a technique used in neural networks to reduce the spatial dimensionality of data by retaining only the maximum values in specific regions [54,55]. In the context of 1D time series, 1D MaxPooling is used to summarize the most relevant information and reduce the computational cost by decreasing the number of parameters in the network. Given a 1D input data set X with elements and a pooling window of size p, the output Y is calculated by taking the maximum value in each window. Mathematically, this can be expressed as (equation 9):

where i is the index of the pooling window. This process is repeated until the entire length of the entry is covered.

Proposed Architecture

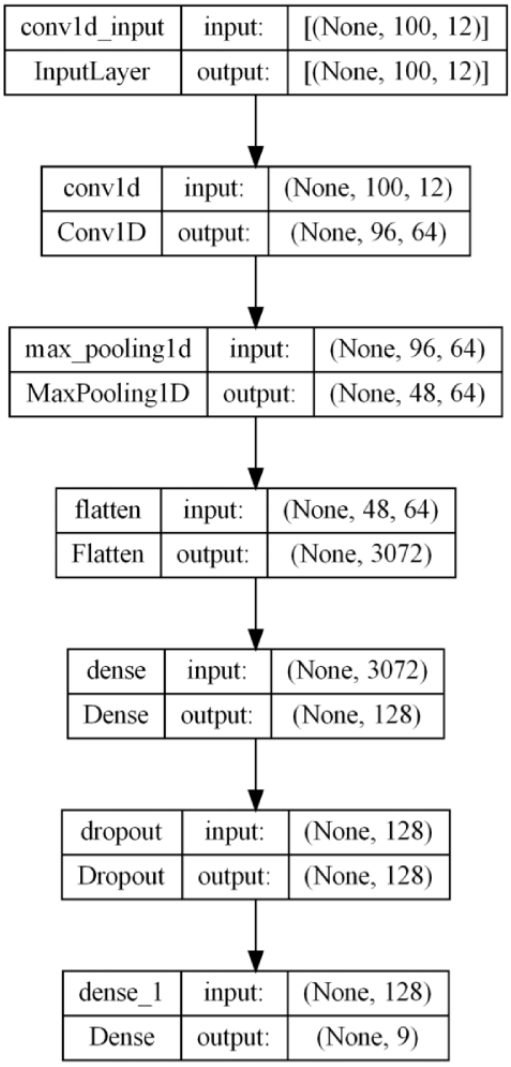
Overview

The processing layers described above were combined to create 4 DL models of different complexity. Each model was designed to explore and exploit different approaches in data processing for the ADL recognition task. The architectures and distinctive features of each of these models are detailed below.

Convolutional Approach

The first proposed architecture uses a convolutional approach. It is composed of 3 main layers: a 1D convolutional layer, a pooling layer, and an FC layer (Figure 4). The convolutional layer, with 64 filters and a kernel size of 5, performs local feature extraction. Next, the pooling layer with pool size 2 is applied to reduce the dimensionality and preserve the most relevant features. Subsequently, a Flatten layer is used to convert the output into a 1D vector before connecting it to an FC layer with 128 neurons and a rectified lineal unit (ReLU) activation function. ReLU is a nonlinear activation function commonly used in neural networks to introduce nonlinearities and aid in model convergence [56]. Finally, a Dropout layer with a rate of 40% is incorporated to prevent overfitting. The output layer uses the Softmax function and is designed for multiclass classification. The output layer uses the Softmax function, which is commonly used in multiclass classification tasks to compute the probabilities of each class outcome and facilitate decision-making based on the highest probability class [57].

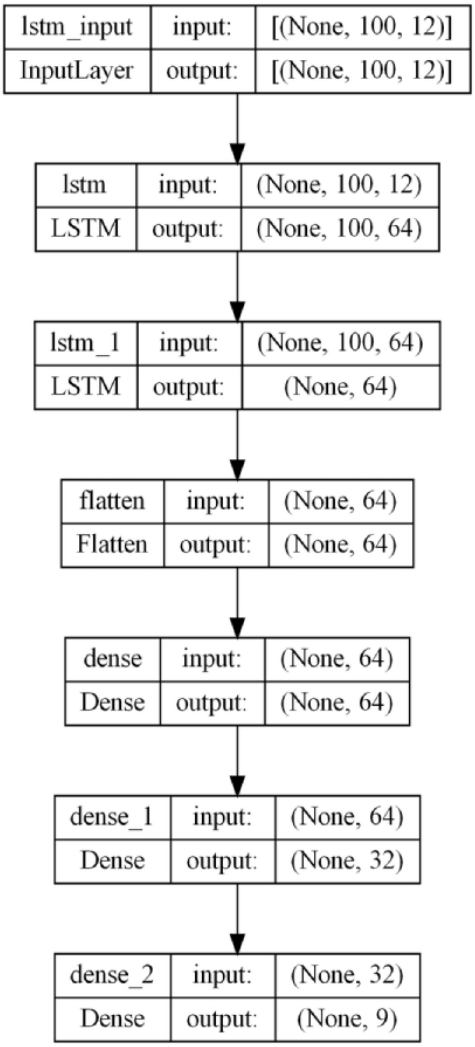
Figure 4. Model architecture based on a convolutional approach.



Deep LSTM Approach

The second architecture is based on a deep LSTM networking approach. It includes 2 LSTM layers, both with 64 units, followed by a Flatten layer. Then, 2 FC layers, with 64 and 32 neurons, respectively, and ReLU activation function are incorporated. The output layer uses the Softmax function for multiclass classification (Figure 5). This architecture deepens into the LSTM network with multiple layers, allowing more complex temporal patterns to be learned. The complexity increases compared with the convolutional model due to the deepening of the LSTM layers and the increase in FC connections. This approach seeks to capture more elaborate temporal dependencies in time series data.

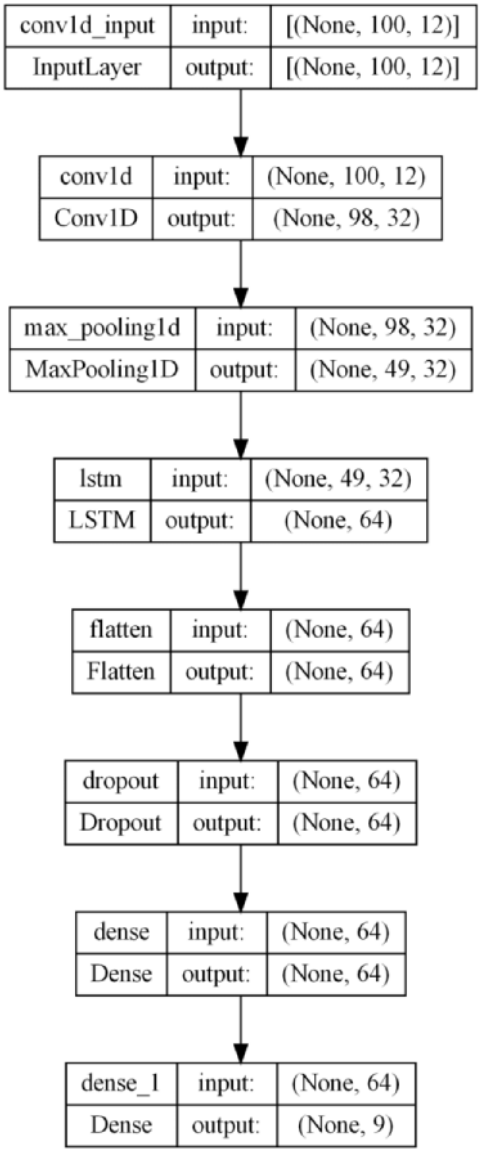
Figure 5. Model architecture based on a deep LSTM approach. LSTM: long short-term memory.



Hybrid Approach: 1D Convolutional + LSTM

The third architecture adopts a hybrid approach combining convolutional layers and LSTM networks (Figure 6). It starts with a 1D convolutional layer with 32 filters and kernel size 3, followed by an LSTM layer with 64 units. Subsequently, a pooling layer and a Flatten layer are applied. A Dropout layer (30%) is introduced to prevent overfitting before connecting to an FC layer with 64 neurons and ReLU activation. The output layer uses Softmax for multiclass classification. This architecture seeks to take advantage of the ability of convolutional layers to extract local features and the ability of LSTMs to model long-term temporal dependencies, offering a combination of both capabilities. Its complexity lies in the integration of 2 different approaches to improve the representation and understanding of time series data.

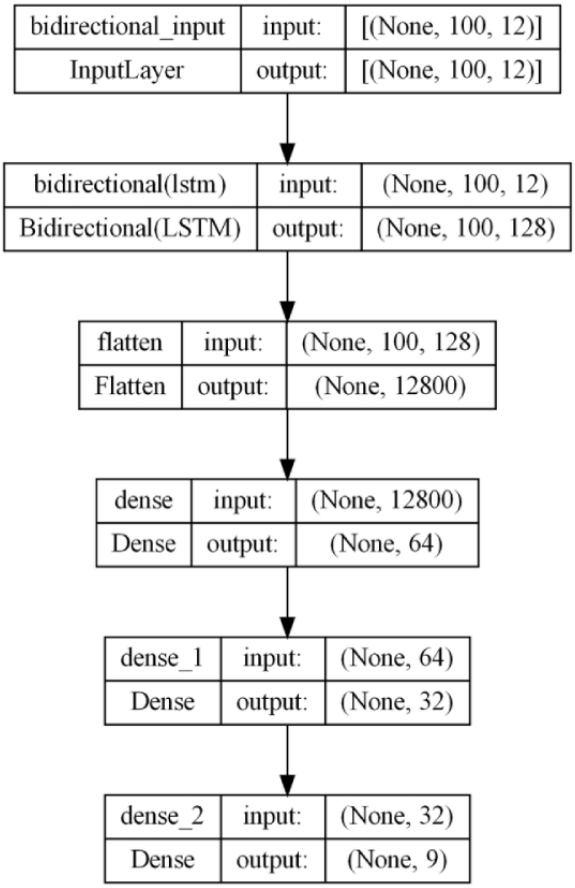
Figure 6. Model architecture based on a hybrid approach (convolutional + LSTM). LSTM: long short-term memory.



Bidirectional LSTM Approach

The fourth architecture adopts a bidirectional approach using LSTM layers (Figure 7). It starts with a bidirectional LSTM layer with 64 units to capture temporal patterns in both directions. Then, a Flatten layer is applied before connecting with 2 FC layers of 64 and 32 neurons, respectively, with ReLU activation function. The output layer uses Softmax for multiclass classification. This architecture represents a more sophisticated and complex model by taking advantage of the ability of bidirectional LSTMs to capture both forward and backward temporal dependencies.

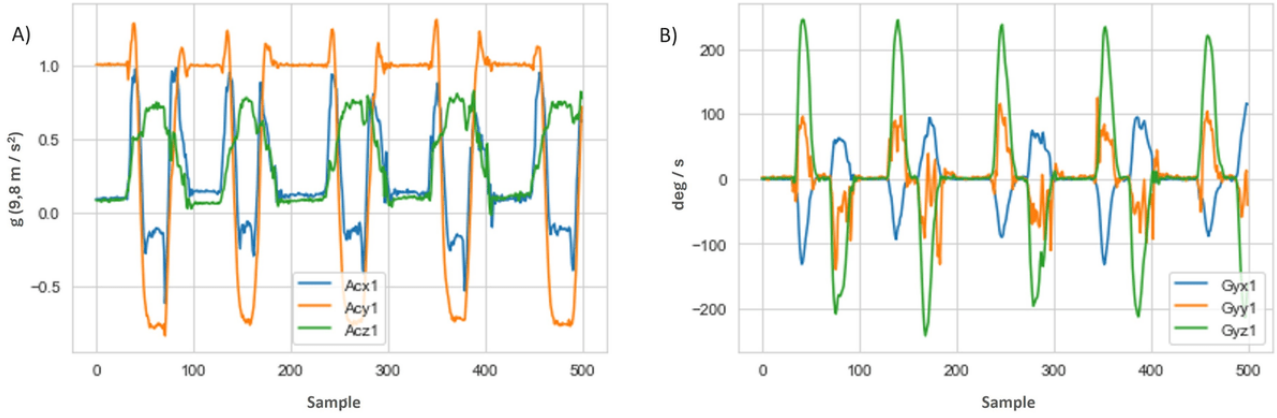
Figure 7. Model architecture based on a bidirectional LSTM approach. LSTM: long short-term memory.



Selection of Parameters for Training and Optimization of Models

For a better understanding of the data and selection of the hyperparameters of the AI model, the accelerometry and gyroscope values of each movement were analyzed separately (Figure 8). Based on this, the temporal sequences were divided into windows of 100 records with a 10-record overlap between adjacent windows.

Figure 8. Time series of eating activity: (A) accelerometer and (B) gyroscope.



Each model was trained over 150 epochs, representing a complete iteration through the training data. The model weights were updated every 1024 records (batch size), and training was stopped if the validation accuracy did not improve for 15 consecutive epochs (early stopping) to prevent overfitting.

As the optimization algorithm during training, Adam was used. Its primary goal is to adjust the network's weights and biases so that the model's loss function is minimized. Adam enhances the standard gradient descent technique by adjusting the learning rate for each parameter individually, potentially leading to faster convergence and better model performance.

Additionally, an L2 regularization term with a strength of 0.0015 was also applied to mitigate overfitting. This term controls the excessive growth of weights during training by adding a penalty term to the model's loss function. The regularization strength determines how much large weights are penalized. By penalizing large weights, L2 regularization helps smooth out the model's

decisions and prevents it from fitting too closely to the training data.

Categorical cross-entropy was used as the loss function. This function measures the discrepancy between the probability distributions predicted by the model and the actual distributions. The primary evaluation metric was accuracy, indicating the proportion of the model's predictions in the test set that were correct.

Results

Evaluation Metrics

The experiment was performed on a personal computer with Microsoft Windows 10, an Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz-2.40 GHz, 6 GB RAM, and no GPU. All software was implemented using the Python programming language and the TensorFlow library in the Spyder development environment. After preparing the data, the DL models were trained using 70% of the data, and the remaining 30% was used to evaluate their performance. For this, popular evaluation metrics were used in classification problems, including precision, recall, F_1-score, and accuracy.

Accuracy (equation 10) refers to the proportion of correct predictions, true positives (TPs) and true negatives (TNs) in relation to the total predictions made by the model, which include false positives (FPs) and false negatives (FNs).

Precision (equation 11) represents the proportion of positive predictions that were correct. It is calculated as the number of TPs divided by the sum of TPs and FPs.

Recall (equation 12) refers to the proportion of TP cases that were correctly identified by the model, calculated as the number of TPs divided by the sum of TPs and FNs.

F_1-score (equation 13) is a measure that combines precision and recall. It is calculated as the harmonic mean between precision and recall and provides a more balanced assessment of model performance, particularly useful when there is an imbalance in the class distribution in the data.

Evaluation Outcomes

The obtained results show the high performance of the 4 proposed models, with accuracy, precision, recall, and F_1-score ranging between 95% and 97% for all cases (Table 1), while the loss function indicates an error rate of approximately 0.10 for the models. The high accuracy, precision, and recall suggest an ability to accurately identify multiple classes of activities, while the high F_1-score indicates a good balance between precision and recall. These results suggest that the models have effectively learned the relationships in the training, enabling them to identify patterns and generalize effectively to data they have not encountered during training, demonstrating strong and reliable predictive capabilities.

Table 1. Evaluation metrics of the designed deep learning models.

Models	Accuracy (%)	Precision (%)	Recall (%)	F_1-score (%)
CNN[a]	97.11	97.19	97.14	97.14
Deep LSTM[d]	95.52	95.64	95.58	95.54
CNN+LSTM	96.19	96.19	96.14	96.14
Bidirectional LSTM	97.56	97.51	97.52	97.51

[a]CNN: convolutional neural network.

[b]LSTM: long short-term memory.

When comparing different modeling approaches, it is evident that both the convolutional and bidirectional methods yield similar results across all evaluated metrics. This suggests that, despite the bidirectional approach's inherent complexity in processing sequences in both directions, it does not offer a significant improvement over the simpler convolutional method. The convolutional model may have struck an optimal balance between learnability and generalization, enabling it to match or even surpass more complex models in terms of accuracy. However, it is worth noting that the bidirectional model achieved convergence in a smaller number of epochs (n=30; Figure 9), which is particularly valuable when rapid training and model responsiveness are required.

It is also noteworthy that more complex models, such as deep LSTM and the hybrid approach, exhibit slightly inferior results compared with the convolutional approach. This observation may stem from several factors. First, the generalization ability of these models may be compromised due to the inherent complexity of their architectures and sensitivity to weight initialization. Additionally, the nature of the data and the suitability of different modeling approaches to capture the relevant characteristics of the time series should be considered. The activities represented in the data may benefit more from a simpler, more straightforward approach, such as convolutional, rather than more complex methods that may be prone to capturing irrelevant features or noise in the data.

At the activity or class level, the confusion matrix provides a detailed breakdown of the model predictions for each class compared with the real class. Referring to the confusion matrix of the model with the best performance (Figure 10), it is observed that the majority of the predictions align with the main diagonal of the matrix, indicating that, for the most part, the classes are classified correctly. However, the activity of eating exhibits the most erroneous predictions, primarily being confused with the activities of opening a door and combing one's hair. This confusion may arise due to overlapping movements and shared characteristics, such as acceleration and rotation patterns, making it challenging for the model to distinguish between them. Moreover, variations in the sequence of movements and the context in which these activities are performed may lead to different interpretations by the model. Variability in the execution of activities and differences in movements between individuals can also contribute to confusion among these classes.

Figure 9. Training sessions progress over iterations.

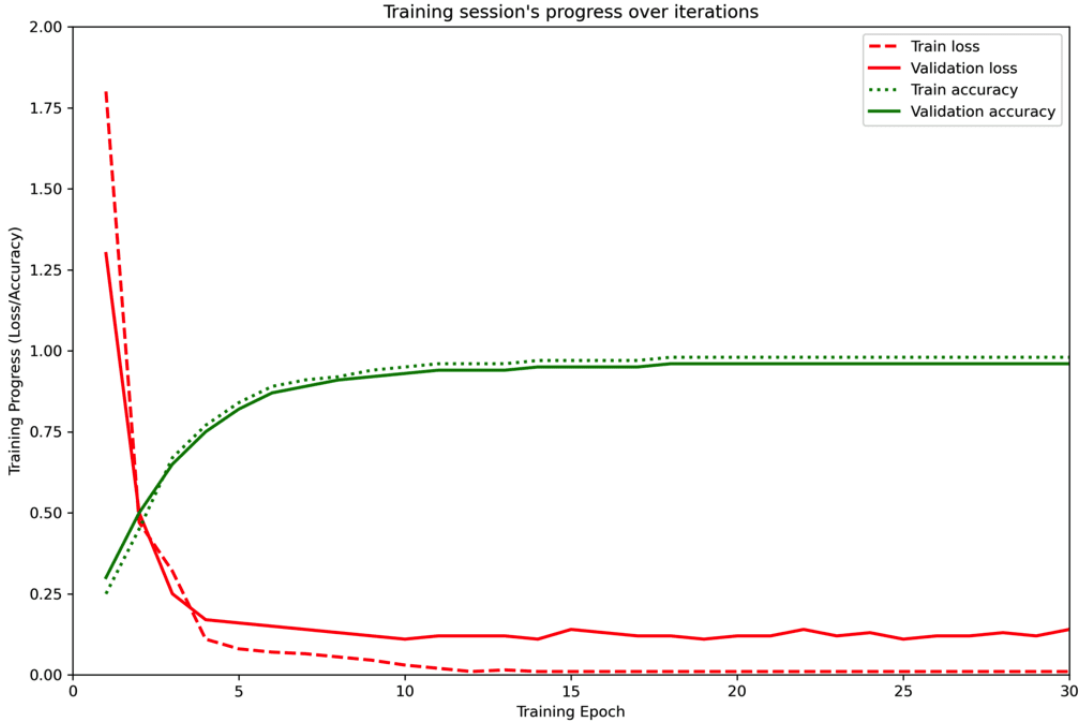
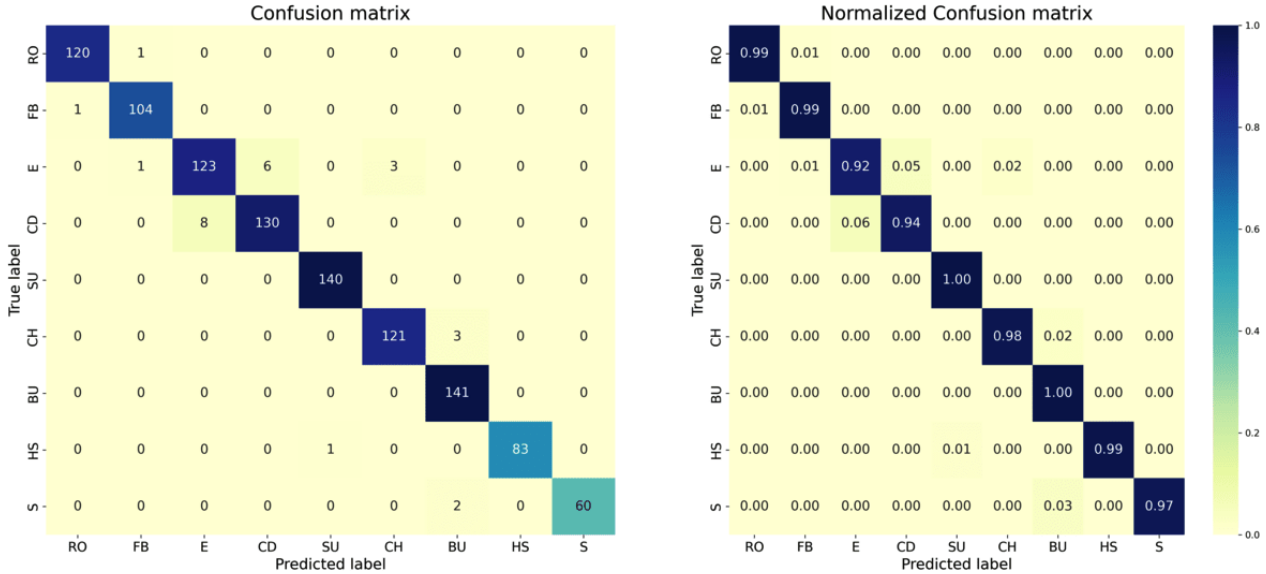


Figure 10. Confusion matrix of the winning model based on the bidirectional approach: (left) standard and (right) normalized. BU: buttoning up; CH: combing hair; E: eating; FB: fastening the bra; HS: half squat; OD: opening the door; RO: reaching for an object; S: sitting SU: standing up.



Discussion

Background

ADL are fundamental tasks that enable individuals to function with a minimum of autonomy and maintain their quality of life. Precise evaluation of ADL, especially in clinical and rehabilitation contexts, is crucial for understanding individuals' functional status and designing effective interventions. Traditionally, the assessment of ADL has relied on direct observation and subjective evaluation by therapists, which can lead to biases and errors. Innovative technology, including wearable inertial sensors and AI, offers new opportunities for objective and quantitative evaluation of ADL performance.

Principal Findings

This study presents an innovative initiative by combining wearable inertial sensors with AI techniques to evaluate human movement in ADL. The implemented AI models have demonstrated solid performance, exhibiting high accuracy, precision, recall, and F_1-score (ranging between 95% and 97%), indicating an effective ability to identify and classify a variety of daily activities related to the shoulder and lumbar region. Furthermore, these results have been achieved through minimal sensorization, which is noninvasive and practically imperceptible to the user, thus minimizing interference with their daily life. This feature is crucial as it promotes user acceptance and adherence to continuous monitoring, contributing to the reliability of the collected data.

Comparison to Prior Work

This study presents significant improvements in the identification and monitoring of activities of ADL compared with other existing methods. Unlike most previous approaches that primarily focus on activities involving the lumbar region (sitting, lying down, standing up, etc), our proposal allows for the precise identification of complex movements involving both the lumbar region and the shoulder. This is achieved using only 2 low-cost inertial sensors, contrasting with other solutions that require a higher degree of sensorization or bulkier devices. This minimally invasive monitoring enables individuals to perform daily activities naturally, promoting a more authentic representation of movement.

The information provided by the sensors is used by DL algorithms for movement identification, without requiring additional processing. This enables immediate analysis of movement patterns during the performance of everyday activities, avoiding the delay associated with data processing needed in image-based motion capture systems, which tend to be more expensive and complex to set up and maintain.

Furthermore, the use of inertial sensors offers versatility and adaptability, making them suitable for monitoring a wide range of ADL in different environments and contexts. They provide valuable information on movement patterns and functional abilities that may not be effectively captured or may be more difficult to capture by traditional 2D and 3D motion capture systems, which are more limited by factors such as image quality, potential obstructions in the line of sight between the camera and the person, or the need to use a greater number of cameras or sensors to capture all movement details.

Limitations and Strengths of This Study

This study demonstrates notable strengths in its methodology and approach. It uses the integration of inertial sensors and AI to improve the assessment of shoulder and lumbar motion during basic ADL performance, providing an objective and advanced perspective for clinical evaluation and rehabilitation. Although challenges persist in validating its use across various ADL gestures, such as eating or dressing, our focus on automatic detection and monitoring using AI techniques addresses this gap. Furthermore, by directly capturing inertial sensor signals and using only 2 sensors, our approach ensures enhanced activity recognition precision and efficiency. This strategy facilitates seamless integration into individuals' daily lives at a low cost, promoting improved adherence to monitoring. Additionally, our study's direct use of accelerometer and gyroscope data without conversion for model training emphasizes its versatility and broad scope, highlighting its adaptability across a wide range of activities.

However, it is essential to acknowledge potential limitations to encourage further research and refinement. One limitation lies in the scope of activities monitored, which primarily focuses on specific muscle groups. Future research should aim to expand the scope of using AI and wearable inertial sensors beyond the assessment of shoulder and lumbar motion, broadening the range of monitored ADL. Given this limitation, it would be interesting to conduct in the future more extensive studies that encompass a broader range of ADL and other more distal body segments. For instance, investigations could explore the application of these technologies in assessing motion patterns related to limb motion (ie, elbow and wrist, or knee and ankle movements), offering valuable insights into biomechanical segmentary dynamics and enhancing our understanding of musculoskeletal movement patterns through AI approaches. Despite this limitation, the study sets a solid foundation for future endeavors in this field, showcasing its potential for advancement and application in clinical and rehabilitative settings.

Conclusions

This research has the potential to significantly impact the clinical evaluation and rehabilitation of patients with movement limitations, offering an objective and advanced tool to detect key movement patterns and joint dysfunctions. Such information can assist professionals in tailoring treatment plans to be more precise and personalized, addressing specific areas of weakness, and designing interventions to improve the patient's functionality and quality of life.

Acknowledgments

This research was funded by the Valencian Innovation Agency of Spain (grant INNVA1/2020/81).

Data Availability

The datasets generated during this study are not publicly available due to ethical restrictions but are available from the corresponding author on reasonable request.

Authors' Contributions

DRF and JCT designed the research. JCT and MGJ collected and supervised the data. ADR analyzed the data and developed and evaluated the deep learning models. All authors drafted the manuscript. DRF and JCT critically reviewed the manuscript. DRF had primary responsibility for the final content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Edemekong PF, Bomgaars D, Sukumaran S, Levy SB. Activities of Daily Living. Treasure Island, FL: StatPearls Publishing LLC; 2019.
2. Merrilees J. Activities of daily living. Encyclopedia of the Neurological Sciences 2014. [Medline: 29261878]
3. Katz S. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. J Am Geriatr Soc 1983;31(12):721-727. [doi: 10.1111/j.1532-5415.1983.tb03391.x] [Medline: 6418786]
4. Kang H, Lee C, Kang SJ. A smart device for non-invasive ADL estimation through multi-environmental sensor fusion. Sci Rep 2023;13(1):17246 [FREE Full text] [doi: 10.1038/s41598-023-44436-5] [Medline: 37821665]
5. Chan CS, Slaughter SE, Jones CA, Wagg AS. Greater independence in activities of daily living is associated with higher health-related quality of life scores in nursing home residents with dementia. Healthcare (Basel) 2015;3(3):503-518 [FREE Full text] [doi: 10.3390/healthcare3030503] [Medline: 27417776]
6. Giebel CM, Sutcliffe C, Challis D. Activities of daily living and quality of life across different stages of dementia: a UK study. Aging Ment Health 2015;19(1):63-71. [doi: 10.1080/13607863.2014.915920] [Medline: 24831511]
7. Herero VG, Extremera N. Daily life activities as mediators of the relationship between personality variables and subjective well-being among older adults. Pers Individ Differ 2010;49(2):124-129. [doi: 10.1016/j.paid.2010.03.019]
8. Osborne M, Rizzo J. Chapter 106 - neurorehabilitation. Neurol Clin Neurosci 2007:1423-1432. [doi: 10.1016/b978-0-323-03354-1.50110-3]
9. Klemt C, Prinold JA, Morgans S, Smith SHL, Nolte D, Reilly P, et al. Analysis of shoulder compressive and shear forces during functional activities of daily life. Clin Biomech (Bristol, Avon) 2018;54:34-41 [FREE Full text] [doi: 10.1016/j.clinbiomech.2018.03.006] [Medline: 29550641]
10. Vaisy M, Gizzi L, Petzke F, Consmüller T, Pfingsten M, Falla D. Measurement of lumbar spine functional movement in low back pain. Clin J Pain 2015;31(10):876-885. [doi: 10.1097/AJP.0000000000000190] [Medline: 25503596]
11. Kaljić E, Pašalić A, Katana B, Mačak Hadžiomerović A, Bojičić S, Jaganjac A, et al. Influence of motion therapy on daily life activities of people with lumbar pain syndrome. J Health Sci 2022;12:213-222. [doi: 10.17532/jhsci.2022.1975]
12. Poppen NK, Walker PS. Normal and abnormal motion of the shoulder. J Bone Joint Surg Am 1976 Mar;58(2):195-201. [Medline: 1254624]
13. Magda A, Cáceres L. Doctoral Thesis. University of Valencia. 2019. URL: http://hdl.handle.net/10251/133994 [accessed 2024-07-26]
14. Michiels I, Grevenstein J. Kinematics of shoulder abduction in the scapular plane. on the influence of abduction velocity and external load. Clin Biomech (Bristol, Avon) 1995;10(3):137-143. [doi: 10.1016/0268-0033(95)93703-v] [Medline: 11415544]
15. Sánchez-Zuriaga D, López-Pascual J, Garrido-Jaén D, de Moya MFP, Prat-Pastor J. Reliability and validity of a new objective tool for low back pain functional assessment. Spine (Phila Pa 1976) 2011;36(16):1279-1288. [doi: 10.1097/BRS.0b013e3181f471d8] [Medline: 21240051]
16. Artacho PCA, Andrea C. Biomechanical assessment of the spine based on functional analysis of various activities of daily living. University of Valencia. 2018. URL: https://roderic.uv.es/handle/10550/68281#.ZA1QDVfRrrs.mendeley [accessed 2018-12-14]
17. Fuster Ortí MA. Effects of a manual spinal traction technique on the lumbo-pelvic movement pattern and activation of the erector spinae during trunk flexion-extension in patients with low back pain. University of Valencia. 2021. URL: https://dialnet.unirioja.es/servlet/tesis?codigo=311244 [accessed 2024-07-09]
18. Lehman GJ. Biomechanical assessments of lumbar spinal function. how low back pain sufferers differ from normals. implications for outcome measures research. part i: kinematic assessments of lumbar function. J Manipulative Physiol Ther 2004;27(1):57-62. [doi: 10.1016/j.jmpt.2003.11.007] [Medline: 14739876]
19. Arguisuelas MD, Lisón JF, Doménech-Fernández J, Martínez-Hurtado I, Salvador Coloma P, Sánchez-Zuriaga D. Effects of myofascial release in erector spinae myoelectric activity and lumbar spine kinematics in non-specific chronic low back

pain: randomized controlled trial. Clin Biomech (Bristol, Avon) 2019;63:27-33. [doi: 10.1016/j.clinbiomech.2019.02.009] [Medline: 30784788]
20. Katz S, Ford AB, Moskowitz RB, Jackson BA, Jaffe MW. Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function. JAMA J Am Med Assoc 1963;185:914-919. [doi: 10.1001/jama.1963.03060120024016] [Medline: 14044222]
21. Morris JN, Fries BE, Morris SA. Scaling ADLs within the MDS. J Gerontol A Biol Sci Med Sci 1999;54(11):M546-M553. [doi: 10.1093/gerona/54.11.m546] [Medline: 10619316]
22. Graf C, Hartford Institute for Geriatric Nursing. The Lawton instrumental activities of daily living (IADL) scale. Medsurg Nurs 2008;17(5):343-344. [Medline: 19051984]
23. Sánchez-Zuriaga D, Artacho-Pérez C, Biviá-Roig G. Lumbopelvic flexibility modulates neuromuscular responses during trunk flexion-extension. J Electromyogr Kinesiol 2016;28:152-157. [doi: 10.1016/j.jelekin.2016.04.007] [Medline: 27155332]
24. Pashmdarfard M, Azad A. Assessment tools to evaluate activities of daily living (ADL) and instrumental activities of daily living (IADL) in older adults: a systematic review. Med J Islam Repub Iran 2020;34:33 [FREE Full text] [doi: 10.34171/mjiri.34.33] [Medline: 32617272]
25. Jekel K, Damian M, Storf H, Hausner L, Frölich L. Development of a proxy-free objective assessment tool of instrumental activities of daily living in mild cognitive impairment using smart home technologies. J Alzheimers Dis 2016;52(2):509-517 [FREE Full text] [doi: 10.3233/JAD-151054] [Medline: 27031479]
26. Amaral Gomes ES, Ramsey KA, Rojer AGM, Reijnierse EM, Maier AB. The association of objectively measured physical activity and sedentary behavior with (instrumental) activities of daily living in community-dwelling older adults: a systematic review. Clin Interv Aging 2021;16:1877-1915 [FREE Full text] [doi: 10.2147/CIA.S326686] [Medline: 34737555]
27. Goverover Y, Kalmar J, Gaudino-Goering E, Shawaryn M, Moore NB, Halper J, et al. The relation between subjective and objective measures of everyday life activities in persons with multiple sclerosis. Arch Phys Med Rehabil 2005;86(12):2303-2308. [doi: 10.1016/j.apmr.2005.05.016] [Medline: 16344027]
28. Bonato P. Advances in wearable technology and applications in physical medicine and rehabilitation. J Neuroeng Rehabil 2005;2(1):2 [FREE Full text] [doi: 10.1186/1743-0003-2-2] [Medline: 15733322]
29. Kim J, Campbell AS, de Ávila BEF, Wang J. Wearable biosensors for healthcare monitoring. Nat Biotechnol 2019;37(4):389-406 [FREE Full text] [doi: 10.1038/s41587-019-0045-y] [Medline: 30804534]
30. Rodgers MM, Alon G, Pai VM, Conroy RS. Wearable technologies for active living and rehabilitation: current research challenges and future opportunities. J Rehabil Assist Technol Eng 2019;6:2055668319839607 [FREE Full text] [doi: 10.1177/2055668319839607] [Medline: 31245033]
31. Lang CE, Barth J, Holleran CL, Konrad JD, Bland MD. Implementation of wearable sensing technology for movement: pushing forward into the routine physical rehabilitation care field. Sensors (Basel) 2020;20(20):5744 [FREE Full text] [doi: 10.3390/s20205744] [Medline: 33050368]
32. Porciuncula F, Roto AV, Kumar D, Davis I, Roy S, Walsh CJ, et al. Wearable movement sensors for rehabilitation: a focused review of technological and clinical advances. PM R 2018;10(9 Suppl 2):S220-S232 [FREE Full text] [doi: 10.1016/j.pmrj.2018.06.013] [Medline: 30269807]
33. Jalloul N. Wearable sensors for the monitoring of movement disorders. Biomed J 2018;41(4):249-253 [FREE Full text] [doi: 10.1016/j.bj.2018.06.003] [Medline: 30348268]
34. Wu W, Dasgupta S, Ramirez EE, Peterson C, Norman GJ. Classification accuracies of physical activities using smartphone motion sensors. J Med Internet Res 2012;14(5):e130 [FREE Full text] [doi: 10.2196/jmir.2208] [Medline: 23041431]
35. Rast FM, Labruyère R. Systematic review on the application of wearable inertial sensors to quantify everyday life motor activity in people with mobility impairments. J Neuroeng Rehabil 2020;17(1):148 [FREE Full text] [doi: 10.1186/s12984-020-00779-y] [Medline: 33148315]
36. Kristoffersson A, Lindén M. A systematic review of wearable sensors for monitoring physical activity. Sensors (Basel) 2022;22(2):573 [FREE Full text] [doi: 10.3390/s22020573] [Medline: 35062531]
37. Camomilla V, Bergamini E, Fantozzi S, Vannozzi G. Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: a systematic review. Sensors (Basel) 2018;18(3):873 [FREE Full text] [doi: 10.3390/s18030873] [Medline: 29543747]
38. Picerno P, Iosa M, D'Souza C, Benedetti MG, Paolucci S, Morone G. Wearable inertial sensors for human movement analysis: a five-year update. Expert Rev Med Devices 2021;18(sup1):79-94. [doi: 10.1080/17434440.2021.1988849] [Medline: 34601995]
39. Iosa M, Picerno P, Paolucci S, Morone G. Wearable inertial sensors for human movement analysis. Expert Rev Med Devices 2016;13(7):641-659. [doi: 10.1080/17434440.2016.1198694] [Medline: 27309490]
40. Yen CT, Liao JX, Huang YK. Human daily activity recognition performed using wearable inertial sensors combined with deep learning algorithms. IEEE Access 2020;8:174105-174114. [doi: 10.1109/access.2020.3025938]
41. Huynh-The T, Hua CH, Kim DS. Visualizing inertial data for wearable sensor based daily life activity recognition using convolutional neural network. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:2478-2481. [doi: 10.1109/EMBC.2019.8857366] [Medline: 31946400]

42. Mustafa Z. A study of machine learning techniques based on human daily living activities via inertial sensors. 2023 Presented at: International Conference on IT Innovation and Knowledge Discovery, ITIKD; 08 March 2023; Manama, Bahrain. [doi: 10.1109/itikd56332.2023.10099820]
43. Ronald M, Poulose A, Han DS. iSPLInception: an Inception-ResNet deep learning architecture for human activity recognition. IEEE Access 2021;9:68985-69001. [doi: 10.1109/access.2021.3078184]
44. Poulose A, Kim JH, Han DS. HIT HAR: human image threshing machine for human activity recognition using deep learning models. Comput Intell Neurosci 2022;2022:1808990 [FREE Full text] [doi: 10.1155/2022/1808990] [Medline: 36248917]
45. García-Luna MA, Jimenez-Olmedo JM, Pueo B, Manchado C, Cortell-Tormo JM. Concurrent validity of the ergotex device for measuring low back posture. Bioengineering (Basel) 2024;11(1):98 [FREE Full text] [doi: 10.3390/bioengineering11010098] [Medline: 38275578]
46. Jimenez-Olmedo JM, Tortosa-Martínez J, Cortell-Tormo JM, Pueo B. Assessing the validity of the ergotex IMU in joint angle measurement: a comparative study with optical tracking systems. Sensors (Basel) 2024;24(6):1903 [FREE Full text] [doi: 10.3390/s24061903] [Medline: 38544165]
47. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: a survey. Mech Syst Signal Process 2021;151:107398. [doi: 10.1016/j.ymssp.2020.107398]
48. Heaton J. Ian goodfellow, yoshua bengio, and aaron courville: deep learning. Genet Program Evolvable Mach 2018;19(1-2):305-307. [doi: 10.1007/s10710-017-9314-z]
49. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
50. Park S, Kwak N. Analysis on the dropout effect in convolutional neural networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017 Presented at: Asian Conference on Computer Vision; 10 March 2017; Springer, Cham p. 189-204. [doi: 10.1007/978-3-319-54184-6_12]
51. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. JMLR 2014;15(56):1929-1958. [doi: 10.5555/2627435.2670313]
52. Salehin I, Kang DK. A review on dropout regularization approaches for deep neural networks within the scholarly domain. Electronics (Switzerland) 2023;12(14):3106. [doi: 10.3390/electronics12143106]
53. Jin J, Dundar A, Culurciello E. Flattened convolutional neural networks for feedforward acceleration. 2015 Presented at: 3rd International Conference on Learning Representations, ICLR - Workshop Track Proceedings; 20 November 2015; USA.
54. Christlein V, Spranger L, Seuret M, Nicolaou A, Kral P, Maier A. Deep generalized max pooling. 2019 Presented at: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR; 26 February 2024; Australia. [doi: 10.1109/icdar.2019.00177]
55. Lee CY, Gallagher P, Tu Z. Generalizing pooling functions in CNNs: mixed, gated, and tree. IEEE Trans Pattern Anal Mach Intell 2018;40(4):863-875. [doi: 10.1109/TPAMI.2017.2703082] [Medline: 28504932]
56. Banerjee C, Mukherjee T, Pasiliao E. An empirical study on generalizations of the RelU activation function. 2019 Presented at: ACMSE; 18 April 2019; New York, NY. [doi: 10.1145/3299815.3314450]
57. Zhu D, Lu S, Wang M, Lin J, Wang Z. Efficient precision-adjustable architecture for softmax function in deep learning. IEEE Trans Circuits Syst II Express Briefs 2020;67(12):3382-3386. [doi: 10.1109/tcsii.2020.3002564]

Abbreviations

ADL: activities of daily living
AI: artificial intelligence
BU: buttoning up
CH: combing hair
DL: deep learning
E: eating
FB: fastening the bra
FC: fully connected
FN: false negative
FP: false positive
HAR: human activity recognition
HS: half squat
OD: opening the door
LSTM: long short-term memory
RO: reaching for an object
ReLU: rectified lineal unit
S: sitting
SU: standing up

TN: true negative
TP: true positive

Edited by C Lovis; submitted 05.02.24; peer-reviewed by A Poulose, M Maximiano, J Colado Sanchez; comments to author 01.03.24; revised version received 27.03.24; accepted 30.06.24; published 09.08.24.

Please cite as:
De Ramón Fernández A, Ruiz Fernández D, García Jaén M, Cortell-Tormo JM
Recognition of Daily Activities in Adults With Wearable Inertial Sensors: Deep Learning Methods Study
JMIR Med Inform 2024;12:e57097
URL: https://medinform.jmir.org/2024/1/e57097
doi:10.2196/57097
PMID:39121473

Evaluation of SNOMED CT Grouper Accuracy and Coverage in Organizing the Electronic Health Record Problem List by Clinical System: Observational Study

Rashaud Senior[1], MMCi, MD; Timothy Tsai[2], MMCi, DO; William Ratliff[3], MBA; Lisa Nadler[1], MD; Suresh Balu[3], MS, MBA; Elizabeth Malcolm[4], MSHS, MD; Eugenia McPeek Hinz[1], MS, MD

1

2

3

4

Corresponding Author:
Rashaud Senior, MMCi, MD

Abstract

Background: The problem list (PL) is a repository of diagnoses for patients' medical conditions and health-related issues. Unfortunately, over time, our PLs have become overloaded with duplications, conflicting entries, and no-longer-valid diagnoses. The lack of a standardized structure for review adds to the challenges of clinical use. Previously, our default electronic health record (EHR) organized the PL primarily via alphabetization, with other options available, for example, organization by clinical systems or priority settings. The system's PL was built with limited groupers, resulting in many diagnoses that were inconsistent with the expected clinical systems or not associated with any clinical systems at all. As a consequence of these limited EHR configuration options, our PL organization has poorly supported clinical use over time, particularly as the number of diagnoses on the PL has increased.

Objective: We aimed to measure the accuracy of sorting PL diagnoses into PL system groupers based on Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) concept groupers implemented in our EHR.

Methods: We transformed and developed 21 system- or condition-based groupers, using 1211 SNOMED CT hierarchal concepts refined with Boolean logic, to reorganize the PL in our EHR. To evaluate the clinical utility of our new groupers, we extracted all diagnoses on the PLs from a convenience sample of 50 patients with 3 or more encounters in the previous year. To provide a spectrum of clinical diagnoses, we included patients from all ages and divided them by sex in a deidentified format. Two physicians independently determined whether each diagnosis was correctly attributed to the expected clinical system grouper. Discrepancies were discussed, and if no consensus was reached, they were adjudicated by a third physician. Descriptive statistics and Cohen κ statistics for interrater reliability were calculated.

Results: Our 50-patient sample had a total of 869 diagnoses (range 4-59; median 12, IQR 9-24). The reviewers initially agreed on 821 system attributions. Of the remaining 48 items, 16 required adjudication with the tie-breaking third physician. The calculated κ statistic was 0.7. The PL groupers appropriately associated diagnoses to the expected clinical system with a sensitivity of 97.6%, a specificity of 58.7%, a positive predictive value of 96.8%, and an F_1-score of 0.972.

Conclusions: We found that PL organization by clinical specialty or condition using SNOMED CT concept groupers accurately reflects clinical systems. Our system groupers were subsequently adopted by our vendor EHR in their foundation system for PL organization.

(JMIR Med Inform 2024;12:e51274) doi:10.2196/51274

KEYWORDS

electronic health record; problem List; problem list organization; problem list management; SNOMED CT; SNOMED CT Groupers; Systematized Nomenclature of Medicine; clinical term; ICD-10; International Classification of Diseases

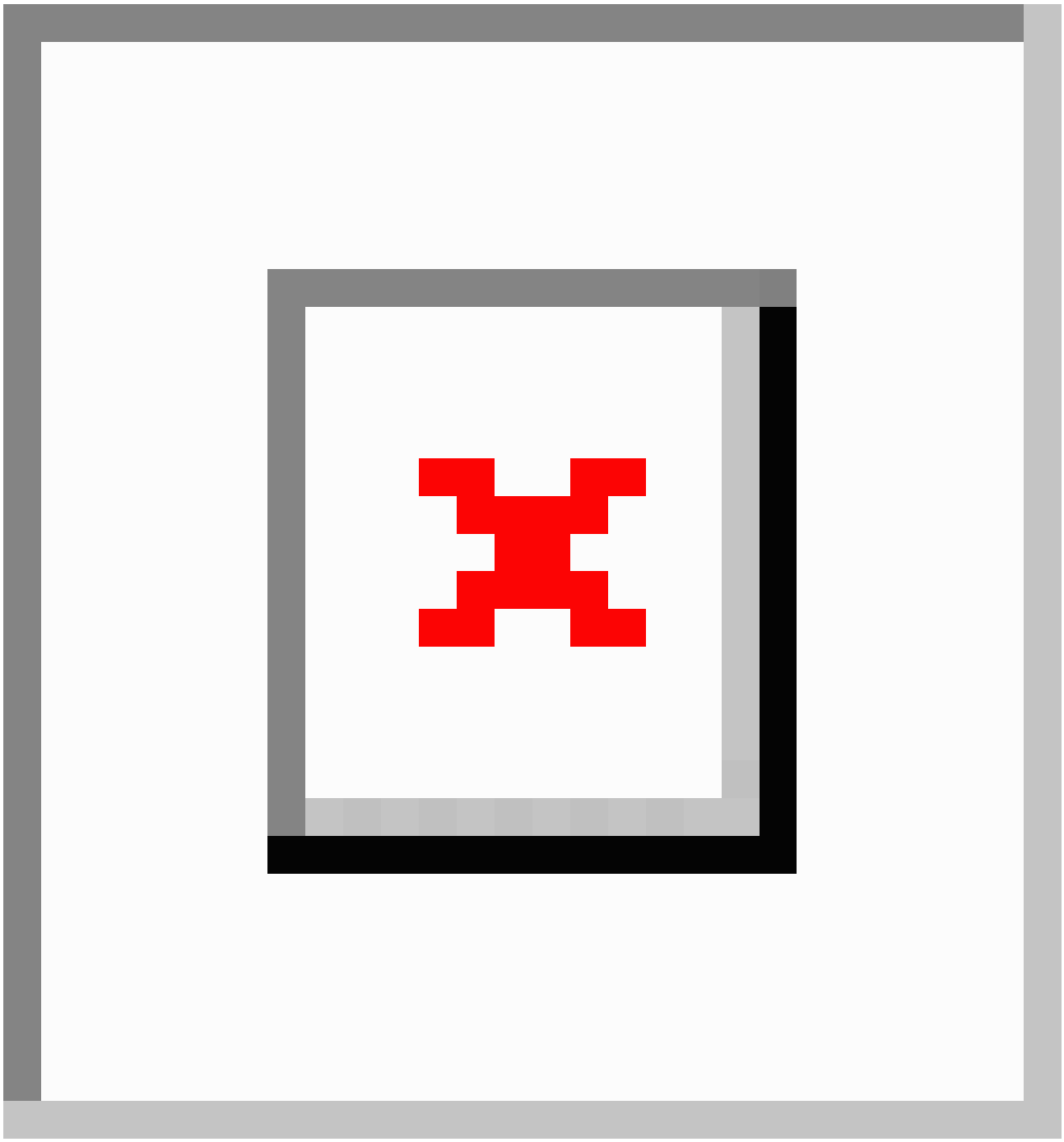
Introduction

The electronic health record (EHR) problem list (PL) is a dynamic repository of a patient's current and historical conditions as well as other health-related issues. As such, it supports communication across a wide range of potential caregivers and clinical environments. An accurate PL serves as a foundation for clinical care and population health management, with multiple derivative secondary processes, including phenotype extraction and disease prediction.

Understanding the history of the PL helps to illustrate why this construct has become the default format for summarizing patients' clinical history. In the 1960s, Lawrence L Weed, MD, proposed the concepts of the problem-oriented medical record; the PL; and the Subjective, Objective, Assessment, and Plan (SOAP) notes for documentation [1]. The idea was to colocate clinical problems with clinical results to focus on systematically addressing all of a patient's diagnoses [2]. Although the SOAP note became the standard format for clinical notes, the PL has encountered more inconsistent use, struggling with problems of inaccuracy, missing diagnoses, not being updated, and bloating [3]. In 2009, the HITECH (Health Information Technology Economic and Clinical Health) Act codified the requirement for an up-to-date PL for meaningful use [4]. Until recently, our vendor EHR had relied on relatively ineffective organization strategies for the PL.

With no one owner, the PLs have become disorganized and cluttered with duplications, conflicting entries, and no-longer-valid diagnoses that contribute to information overload and bloat, obscuring the patient's clinical picture [5]. In its former state, our EHR PL was organized primarily alphabetically, with other options based on primary specialty or priority, all of which have limited clinical utility, especially as the number of diagnoses on the PL increases (Figure 1) . For example, for one patient, we found active diagnoses of lung nodule (Respiratory System), then lung cancer (Oncology System), and then lung cancer with brain metastases (Oncology System). These diagnoses were all related to the same problem but were added sequentially with previous diagnoses that were no longer clinically relevant and were not removed.

Figure 1. Appearance of a problem list before and after grouping algorithm application. Items were reorganized into 21 system groupers using Boolean logic with the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) codes; they were then translated into the *International Classification of Diseases, Tenth Version* (*ICD-10*) codes. Groupers were based on a combination of traditional medical specialty categories, clinically relevant care coordination, and procedure-based groupings, some of which were themselves combined due to overlapping diagnostic coverage. The final order of the problem list items was determined by Epic System's base hierarchy. CMS: The Centers for Medicare and Medicaid Services; HCC: Hierarchical Condition Category; HHS: US Department of Health and Human Services; FEN/GI: Fluids, Electrolytes, Nutrition/Gastrointestinal; GFR=Glomerular Filtration Rate.



There are several major terminology standards that capture patient diagnoses, symptoms, and other health-related conditions, two of which are the *International Classification of Diseases* (*ICD*) [6] and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [7]. The World Health Organization maintains *ICD* codes, which are designed to classify diseases, conditions, and other health-related issues [8,9]. Organized into 21 chapters, they use an alphanumeric classification format to identify diseases, injuries, and factors influencing health. The United States uses an additional Clinical Modifier for further specificity [10]. The *ICD* codes are used in various clinical and nonclinical settings, including disease description, treatment selection, billing, and research applications [11]. There are 78,044 total codes in the 2024 code set, according to the Centers for Medicare and Medicaid Services [12].

Currently managed by Systematized Nomenclature of Medicine (SNOMED) International (previously known as the International

Health Terminology Standards Development Organization [IHTSDO]), SNOMED CT was designed to be a US standard for health information exchange. It functions as a highly granular ontology used to describe clinical observations and findings [13]. SNOMED CT uses a polyhierarchal (parent-child) format organized around a general root concept (eg, a clinical finding, procedure, situation with explicit context, or event) with increased granularity achieved by differentiating more specific descriptions of that root concept. This process allows for the representation of specific clinical content in a machine-readable format [14,15]. Updated monthly, the total number of concepts is 512,087 (as of April 1, 2024 [16]) and continues to increase over time.

Though required for billing, the *ICD, Tenth Revision* (*ICD-10)* terminology is not used directly in clinical care, as many code names are not consistent with clinical vernacular. For example, code Z91.038 "Allergy status to unspecified drugs, medicaments and biological substances" is not as intuitive as "Allergy to insect stings." For this example, our third-party vendor, Intelligent Medical Objects (IMO), transforms the *ICD-10* codes into clinically relevant human-readable concepts. IMO additionally maps at least 1 SNOMED CT concept for each diagnosis, attached as metadata, upon which the PL groupers can be organized. Although SNOMED CT concepts are mapped to *ICD-10* codes [17,18], as with most ontologies, there are gaps in clinical concept coverage.

PL organization and cleanup is challenging for many reasons, including there being no single owner of a patient's PL [19] and its maintenance and cleanup being secondary to other direct clinical care priorities. The tools in the EHR for cleanup are limited to a single patient and do not allow for automated processing or opportunities to categorize or define the state of the PL or large-scale maintenance at the population level. Multiple interventions, including reconfiguration of the EHR PL and re-education, have been met with limited success [20]. Despite these attempts, PL bloat and inaccuracy are widely recognized as issues affecting clinical care and secondary downstream uses of the data [3]. We have come to recognize that curating a clinically relevant and updated PL is a difficult challenge; our primary option for improving its organization was to extend and improve SNOMED CT groupers.

In this paper, we present a PL reorganization developed around clinical specialty groupings using SNOMED CT codes and Boolean logic. We describe the evaluation of the new PL groupers for clinical accuracy and efficiency using a convenience set of patients and their diagnoses. This system allows for future characterization of the PLs at the patient and population levels; it also provides potential for automated cleanup options in the future.

Methods

SNOMED CT Grouper Development and Evaluation

Author EMH extended and extensively modified 19 previously defined groupers initially developed by Heidi Twedt, MD, and added 2 newly defined system- or condition-based groupers, one for pediatric and one for transplant-specific conditions (Table 1). System groupers included traditional medical specialty categories as well as clinically relevant care coordination and procedure-based groupings. Some specialties were combined due to overlapping diagnosis domains (eg, "Respiratory and Allergy" and "Orthopedic and Musculoskeletal" domains). The primary focus was for the system grouper diagnoses to be organized around clinical use. For example, "acute myocardial infarction" and "venous thromboembolism" were sorted into the "Cardiovascular and Peripheral Vascular" grouper, while addiction issues, such as "alcohol use disorder," were sorted into the "Behavioral Health" grouper.

Table . List of system groupers with example diagnoses.

Condition or specialty grouper	Example diagnosis	Notable deviations
1. Care Coordination	Physical deconditioning, food insecurity, risk for falls	Includes health-related social needs
2. Oncology	Malignancies and radiation therapy diagnoses	Excludes dermatology cancers and includes treatment complications
3. Cardiovascular and Peripheral Vascular	Atrial fibrillation and deep vein thrombosis	Excludes cerebral vascular diagnoses
4. Respiratory and Allergy	Asthma and peanut allergy	—[a]
5. Endocrine	Diabetes mellitus, gout, and hypothyroidism	—
6. Behavioral Health	Schizophrenia and opioid use disorder	—
7. Transplant	Living-related kidney transplant and graft versus host disease	Includes transplant complications
8. Infectious Disease, Immune, or Lymphatic	Pneumonia and immune deficiencies	—
9. Blood	Anemia	—
10. Neurology or Sleep	Seizure and sleep disorders	Excludes chronic pain
11. Ears, Nose, Throat (ENT)	Nasal polyps, cleft palate, and hearing loss	—
12. Fluids, Electrolytes, Nutrition, and Gastrointestinal	Hyponatremia and Crohns disease	—
13. Obstetrics and Gynecology	Ovarian cysts; hemolysis, elevated liver enzymes, low platelet count (HELLP) syndrome; and dense breast tissue	Female-specific diagnoses
14. Genitourinary and Nephrology	Ureteral calculus and prostatitis	Includes male-specific genitourinary issues
15. Dermatology	Atopic dermatitis and melanoma	Includes all dermatology-specific cancers (eg, squamous or basal cell carcinoma)
16. Rheumatology	Rheumatoid arthritis	—
17. Orthopedic and Musculoskeletal	Hip fracture	—
18. Ophthalmology or Eye	Uveitis	Includes complications of eye from other diseases
19. Genetics	Trisomy 21	Includes all nonspecific system genomic issues
20. Pediatrics	28-week prematurity	Includes developmental disorders
21. Surgery, Trauma, Wound, and Pain	Gunshot wound and complex regional pain syndrome	—
22. Other	Edema and medication management	Includes any diagnosis that does not fit into another grouper

[a]Not applicable.

Due to its polyhierarchal framework, all child-related SNOMED CT concepts include all related downstream concepts, unless excluded by the Boolean logic. In this format, fewer SNOMED CT concepts can represent many derivative *ICD-10* codes more comprehensively than could be achieved by directly curating *ICD-10* codes. For example, 167 SNOMED CT concepts within the Neurology grouper were mapped to 9243 IMO *ICD-10* diagnoses. Our default EHR PLs were reorganized according to this system-based methodology in the order presented in Table 1.

Using our EHR vendor's built-in tools for grouper build, author EMH iteratively refined the groupers to be consistent with clinical systems using 1211 SNOMED CT concepts. System groupers included the highest parent concept that was appropriate with logic to exclude child-related SNOMED CT concepts not clinically appropriate for a system. For example, squamous cell carcinoma and skin cancers in general are managed clinically by dermatology. In the build for the Oncology grouper, therefore, all dermatologic cancers were excluded and instead added to the Dermatology grouper. As another example, our Cardiovascular grouper includes peripheral vascular diseases like "deep venous thrombosis" but excludes cerebral vascular concepts. This allows diagnoses like "cerebral avascular malformation" to be presented within our Neurology grouper. Both examples highlight the focus of this grouper organization to support clinical specialty coordination of diagnoses.

Multisystem disorders were grouped according to the specialty that would typically manage each disease entity. For example, systemic lupus erythematosus was grouped under "Rheumatology." If a diagnosis's SNOMED CT concept was too broad to be captured by one of the 21 groupers, it defaulted

into the "Other" category. For example, "edema" is a clinical finding that can be reasonably attributed to multiple diseases. As such, it does not have a specific condition or specialty and instead falls into the "Other" category.

To evaluate the effectiveness of specialty sorting, we used a convenience sample of 50 patients randomly identified in January 2022. These patients had at least 3 encounters in the previous year and were selected across all age groups, ranging from newborn to geriatric patients, with an equal ratio of sexes (Table 2). The encounter criteria ensured that identified patients had multiple recent opportunities to have their PLs updated. The PLs for these patients were extracted through screen capture software by author EMH to develop a cohort with no patient identifiers. Standard EHR PL functionality included system grouper name, time frame since the problem was added to the PL, and a limited free-text overview if included with the entry. These study PL entries were reconfigured into a study document with labels indicating sequential patient number, patient age, and patient sex.

Two of the authors, both family medicine physicians (TT and RS), independently examined each patient's PL to determine the clinical accuracy of system groupings for all diagnoses (Table 3). For any items whose system attribution they questioned, the reviewers identified the SNOMED CT code attached to the *ICD-10* code. Diagnoses that were deemed correctly grouped into the appropriate system grouper were considered true positives, while those that were incorrectly grouped were considered false positives. All diagnoses in the dropout "Other" category were examined by their associated SNOMED CT code for options for attribution to a defined system grouper. A diagnosis for which the SNOMED CT code was too vague or not specific enough to be grouped was considered a true negative. Any diagnosis that had a SNOMED code that could have been placed in a relevant system grouper but was not was considered a false negative.

Table . Patient demographics and baseline descriptive statistics. A total of 50 patients, subdivided by age and sex, with descriptive statistics, were reported for each age range.

Age ranges (years)	Gender			Problems			
	Total, n (%)	Male, n (%)	Female, n (%)	Total, n	Mean	Median	Min-Max
<1	6 (12)	3 (50)	3 (50)	72	12.0	10	6-20
1-17	7 (14)	3 (42.9)	4 (57.1)	157	22.4	26	4-35
18-64	24 (48)	11 (45.8)	13 (54.2)	342	14.3	10	4-43
≥65	13 (26)	8 (61.5)	5 (38.5)	298	22.9	21	4-59
All ages	50 (100)	25 (50)	25 (50)	869	17.4	12	4-59

Table . Description of metrics used to determine the effectiveness of automated system grouping. Two reviewers examined individual problem list items and their assigned grouping, placing each into a category.

Assessment category	Definition	Example
True positive (correct system association)	Diagnosis falls into the right disease system—the SNOMED[a] grouper is specific and attributable.	"Community-acquired pneumonia" in the Infectious Disease system
False positive (incorrect system association)	Diagnosis falls in the wrong system grouper.	"Diaphragmatic stimulation by cardiac pacemaker" grouped under "Central Hypoventilation Syndrome".
True negative (Other—correct system association)	The SNOMED grouper associated with a diagnosis is not specific enough to be in anything but the Other category.	"Anticoagulated" placed with the SNOMED grouper "Drug therapy finding". This is not specific enough to be attributed to just anticoagulation status.
False negative (Other—incorrect system association)	Diagnosis belongs to a specified system grouper but falls into the Other category due to logic deficits in the grouper.	"Genetic disorder" falling into the "Other" category until the VCG Grouper is corrected.

[a]SNOMED: Systematized Nomenclature of Medicine.

Each diagnosis was independently categorized according to the scheme in Table 3; then the reviewers compared their determinations. A third independent clinician (author LN) served as a tie-breaker for those PL items for which an agreement was not reached. We calculated descriptive statistics to summarize the volume of diagnoses for the 50 test patients and performance metrics to assess the accuracy and validity of the groupers. The correlation coefficient (κ statistic) was calculated for the degree of agreement between reviewers and for SNOMED CT grouper attributions.

This work was performed using Epic Systems (version May 2021; Verona, WI) initially deployed with ambulatory applications in July 2012 and inpatient applications in June 2013 within the Duke University Health System.

Ethical Considerations

All patient data were anonymized with all demographic identifiers removed except for age. This study was approved by the Duke University Internal Review Board for exempt status (IRB #PRO-00108903).

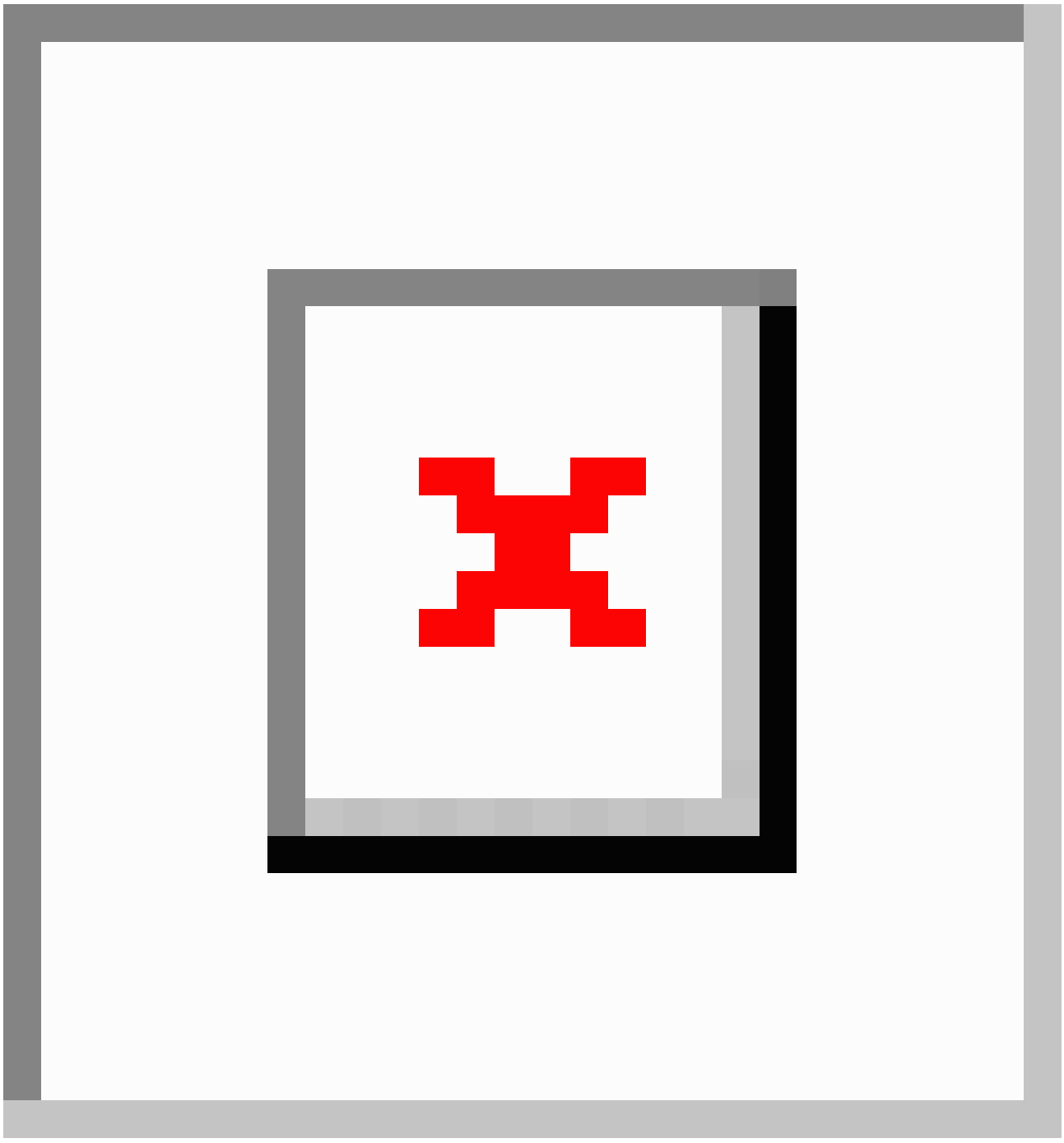
Results

Across the 50 patients, aged 14 days to 93 years, there were a total of 869 (range 4-59) diagnoses identified, with a median of 12 diagnoses per patient. Table 2 includes the breakdown of the volume of PL entries across age and sex.

After their independent evaluations of the PLs, the reviewers initially agreed on 821 (94.4%) of the 869 total problems (Cohen κ coefficient of 0.7, indicating moderate agreement [21]). Of the remaining 48 diagnoses, they subsequently agreed on 32 for a revised agreement rate of 98.2%. The remaining 16 were adjudicated by author LN for attribution.

Based on the definitions presented in Table 3, Figure 2 describes our results. Our final attribution evaluation found that the diagnoses were correctly attributed to a system grouper (ie, sensitivity) in 97.6% of cases, and the nonspecific diagnoses were correctly placed in the "Other" category (ie, specificity) in 58.7% of cases. The positive predictive value, or the correct grouper accuracy rate, was 96.8%. We found 37 (4.3%) true negatives, representing concepts without a SNOMED CT code or diagnoses too general to be attributed to a clinical system. The calculated F_1-Score was 0.972.

Figure 2. Two clinicians' review of problem list sorting algorithm. FN: false negative; FP: false positive; PPV: positive predictive value; NPV: negative predictive value; TN: true negative; TP: true positive; Sn: sensitivity; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms; Sp: specificity.



Discussion

Overview

The PL is the repository of medical diagnoses intended to reflect the patient's clinical conditions. Without groupings reflective of the larger specialty formats of clinical care, the PL can become overloaded and difficult to use as a tool to communicate a patient's clinical status across encounters. We developed 21 SNOMED CT groupers for system concepts to standardize the organization of our EHR PL based on 1211 concepts (Table 1). We chose to evaluate these PL groupers across all ages and sexes to provide a more representative sample of diagnoses across our EHR patient population, recognizing that this is only a subset of the total potential diagnoses. Taking advantage of the hierarchal logic of the SNOMED CT concepts refined with Boolean logic allowed for more than 95% of the diagnoses to be attributed to a system grouper [22,23].

We established the effectiveness of these SNOMED CT groupers in organizing the PL by clinical system related to clinical specialty or condition. We propose that this standardized format for PL organization permits the sharing and reproduction of concepts across other health systems and EHRs.

Comparison to Prior Work

Other groups have used conceptually similar methods with SNOMED CT codes for clinical phenotyping [22]. However, those code sets are typically more narrow in scope for a more specific clinical description. The United Medical Language System Clinical Observations Recordings and Encoding Problem List Subset is meant to "facilitate the use of SNOMED CT as the primary coding terminology for PLs or other summary level clinical documentation" [24]. Compared to these other SNOMED CT code sets, our implementation includes broader clinical coordination groupings (eg, surgical, transplant, care coordination, and infectious disease) that are more reflective of the PL clinical care needs within our institution. Our work here builds upon those efforts and applies them at the system level, which is more accessible for clinical use.

Limitations

We noted some limitations and challenges in using SNOMED CT concepts for this build. Despite ongoing international mapping efforts [17,18], SNOMED CT concepts are not fully representative of all the *ICD-10* codes because of differences in original intended uses [8,13] and baseline granularity [25,26]. For example, the *ICD-10* code "Encounter for pre-transplant evaluation for chronic liver disease" is mapped to the SNOMED CT concept "patient encounter status," as there is no other comparable SNOMED CT coding option. Estimates for the proportions of completely mapped concepts or codes are found in studies reviewing the automation of mapping SNOMED CT and *ICD-10* codes; one study estimated the proportion of complete mappings to *ICD-10-Clinical Modification* (*ICD-10-CM*) at 74% in 2012 [25], and another one estimated the proportion of complete mappings to *ICD-10-Procedure Coding System* (*ICD-10-PCS*) (used to capture inpatient procedures) to be about 86% in 2017 [27].

There were many *ICD-10* diagnoses that were too broad to easily match a SNOMED CT system grouper. "Fatigue" is a good example of an inherently vague constitutional or multisystemic symptom that does not have a clearly identifiable system-level grouper in our schema. For these diagnoses, the "Other" category was used to capture the remaining nonspecific diagnoses. It is important to note that this category is not the same as the *ICD-10* options for "Not Otherwise Specified" (NOS) or "Not Elsewhere Classifiable" (NEC) codes for lesser defined diagnoses. For example, "Pneumonia due to other infectious organisms, NEC" still falls into our "Infectious Disease, Immune, Lymphatic" grouper.

We also note that the mappings are not completely represented across all specialties in terms of the breadth of coverage of concepts. For example, we found more SNOMED CT cardiology-specific concepts and fewer pediatric-specific concepts. These differences may reflect the relative volume of cardiology diagnoses in the general population. The more specific diagnosis of "Encounter for assessment of implantable cardioverter-defibrillator" was mapped to an appropriate SNOMED CT concept and was correctly placed into our cardiovascular system grouper. However, the pediatric diagnosis "Concern about growth" was only mapped to the SNOMED CT code "Finding reported by subject or history provider," which was too broad to be added to the Pediatric grouper only, consequently falling into the "Other" category. Specialties such as pediatrics also require greater levels of specificity for their diagnoses than is always possible with the SNOMED CT concepts currently available.

There were also multiple *ICD-10* codes mapped to the same SNOMED CT code that made attribution to a system grouper challenging. For example, the diagnoses "Diaphragmatic stimulation by pacemaker" and "Disorder of cardiac pacemaker system" mapped to the same SNOMED CT code of "Disorder of cardiac pacemaker system," placing them into the Cardiovascular grouper, although the former would ideally be attributed to the Pulmonary grouper.

As we consider the future challenges of algorithm-based PL sorting, it will be important to investigate the implications of updating ontologies as the World Health Organization has already published the 11th edition of *ICD* (*ICD-11)* with 35 countries now implementing it [28]. We do not suspect that *ICD-11* will replace SNOMED CT as an ontology organization method, as SNOMED CT maintains greater flexibility for clinical use. Health systems are always evolving, and it will be important to consider how such algorithms and their applications will evolve within them.

Conclusions

We leveraged a PL sorting algorithm based on the clinical system–based SNOMED CT groupers to create a standardized PL format in our EHR, reorganizing the diagnoses, symptoms, and medical problems for better clinical utility. We found subjective positive outcomes for our clinical users who reported streamlining their clinical review processes and easier ability to identify similar and duplicate diagnoses. This may be especially helpful for patients with complex issues and many associated diagnoses. A structured PL also enables a shift from patient-level evaluation to potentially population-level assessments and cleanup automation.

As with improvements in the provider experience, automated PL maintenance may also impact researchers leveraging PL diagnoses for machine learning and other similar research. Such possibilities underscore the need for accurate and updated PL diagnoses to achieve and maintain high-fidelity outputs. It will be important to further evaluate methods to automate the maintenance of accurate Pls and best influence care delivery.

Acknowledgments

The authors would also like to acknowledge the support of Tres Brown from the Duke Health Technology System's Maestro Care Electronic Health Records (EHR) Application Team. This work was done while authors RS and TT were Clinical Informatics Fellows at Duke University.

This work was partially funded by a grant from the Duke Institute for Health Innovation.

Data Availability

The authors have uploaded the groupers to GitHub and will share if emailed directly.

Authors' Contributions

All the authors of this manuscript participated in and contributed equally to the conceptualization, design, and evaluation of the program list (PL) grouper categorization. EHM used the previous work of Heidi Twedt, MD, at Stanford (172 concepts across 19 systems and conditions) to develop the final set by extending 8 systems' coverage (eg, "Pulmonary" to "Pulmonary and Allergy"), adding 2 system groupers, and extending them to a total of 1211 concepts with the Boolean logic. Authors RS and EMH primarily authored the manuscript with other authors contributing to editing. Epic Systems incorporated these groupers into their standard development for PL organization by system as default in November of 2022. EMH will continue to make yearly updates to the groupers.

Conflicts of Interest

None declared.

References

1. Weed LL. Medical records that guide and teach. N Engl J Med 1968 Mar 14;278(11):593-600. [doi: 10.1056/NEJM196803142781105] [Medline: 5637758]
2. Wright A, Sittig DF, McGowan J, Ash JS, Weed LL. Bringing science to medicine: an interview with Larry Weed, inventor of the problem-oriented medical record. J Am Med Inform Assoc 2014 Dec;21(6):964-968. [doi: 10.1136/amiajnl-2014-002776] [Medline: 24872343]
3. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: an audit of problem list completeness during the COVID-19 pandemic. Int J Med Inform 2021 Jun;150:104452. [doi: 10.1016/j.ijmedinf.2021.104452] [Medline: 33864979]
4. Henricks WH. “Meaningful use" of electronic health records and its relevance to laboratories and pathologists”. J Pathol Inform 2011 Feb 11;2:7. [doi: 10.4103/2153-3539.76733] [Medline: 21383931]
5. Wright A, McCoy AB, Hickman TT, et al. Problem list completeness in electronic health records: a multi-site study and assessment of success factors. Int J Med Inform 2015 Oct;84(10):784-790. [doi: 10.1016/j.ijmedinf.2015.06.011] [Medline: 26228650]
6. International statistical classification of diseases and related health problems (ICD). World Health Organization. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2023-02-23]
7. SNOMED international recognizes entity linking challenge winners. SNOMED. URL: https://www.snomed.org/ [accessed 2023-02-23]
8. Moriyama IM, Loy RM, Robb-smith AHT, Rosenberg HM, Hoyert DL. History of the Statistical Classification of Diseases and Causes of Death: National Center for Health Statistics; 2011.
9. International Classification of Diseases, Tenth Revision (ICD-10). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/icd/icd10.htm [accessed 2023-02-23]
10. 2.05: ICD-10-CM. MedicalBillingandCoding.org. URL: https://www.medicalbillingandcoding.org/icd-10-cm/ [accessed 2023-03-14]
11. Alharbi MA, Isouard G, Tolchard B. Historical development of the statistical classification of causes of death and diseases. Cogent Med 2021 Jan 1;8(1):1893422. [doi: 10.1080/2331205X.2021.1893422]
12. 2024 ICD-10-CM. Centers for Medicare and Medicaid Services. URL: https://www.cms.gov/medicare/coding-billing/icd-10-codes/2024-icd-10-cm [accessed 2024-04-20]
13. Overview of SNOMED CT. National Library of Medicine. URL: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html [accessed 2023-03-14]
14. Davidso D, Rawson M. SNOMED CT: why it matters to you. Wolters Kluwer. URL: https://www.wolterskluwer.com/en/expert-insights/snomed-ct-why-it-matters-to-you [accessed 2023-02-23]
15. 5-step briefing. SNOMED International. URL: https://www.snomed.org/five-step-briefing [accessed 2023-02-24]
16. Release summary. SNOMED CT Release Statistics 2024-04-01. URL: https://browser.ihtsdotools.org/qa/#/SNOMEDCT/release-summary [accessed 2024-04-20]
17. SNOMED CT to ICD-10-CM map. US National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html [accessed 2023-03-02]
18. March 2021 - SNOMED CT managed service - US edition (US). SNOMED Confluence. 2021. URL: https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=121867053 [accessed 2023-03-02]

19. Klappe ES, de Keizer NF, Cornet R. Factors influencing problem list use in electronic health records-application of the unified theory of acceptance and use of technology. Appl Clin Inform 2020 May;11(3):415-426. [doi: 10.1055/s-0040-1712466] [Medline: 32521555]
20. Kreuzthaler M, Pfeifer B, Vera Ramos JA, et al. EHR problem list clustering for improved topic-space navigation. BMC Med Inform Decis Mak 2019 Apr 4;19(Suppl 3):72. [doi: 10.1186/s12911-019-0789-9] [Medline: 30943968]
21. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012 Oct;22(3):276-282. [doi: 10.11613/BM.2012.031] [Medline: 23092060]
22. Willett DL, Kannan V, Chu L, et al. SNOMED CT concept hierarchies for sharing definitions of clinical conditions using electronic health record data. Appl Clin Inform 2018 Jul;9(3):667-682. [doi: 10.1055/s-0038-1668090] [Medline: 30157499]
23. Chu L, Kannan V, Basit MA, et al. SNOMED CT concept hierarchies for computable clinical phenotypes from electronic health record data: comparison of Intensional versus extensional value sets. JMIR Med Inform 2019 Jan 16;7(1):e11487. [doi: 10.2196/11487] [Medline: 30664458]
24. The CORE Problem List Subset of SNOMED CT®. US National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html [accessed 2023-03-18]
25. Fung KW, Xu J. Synergism between the mapping projects from SNOMED CT to ICD-10 and ICD-10-CM. AMIA Annu Symp Proc 2012;2012:218-227. [Medline: 23304291]
26. McGlothlin S. SNOMED and ICD: aligning to standards. J2 Interactive. 2021. URL: https://www.j2interactive.com/blog/snomed-and-icd/ [accessed 2023-03-02]
27. Fung KW, Xu J, Ameye F, Gutierrez AR, D'Have A. Achieving logical equivalence between SNOMED CT and ICD-10-PCS surgical procedures. AMIA Annu Symp Proc 2018 Apr 16;2017:724-733. [Medline: 29854138]
28. ICD-11 2022 release. World Health Organization. 2022. URL: https://www.who.int/news/item/11-02-2022-icd-11-2022-release [accessed 2023-03-15]

Abbreviations

EHR: electronic health record
HITECH: Health Information Technology Economic and Clinical Health
***ICD*:** *International Classification of Diseases*
IHTSDO: International Health Terminology Standards Development Organization
IMO: Intelligent Medical Objects
NEC: Not Elsewhere Classifiable
NOS: Not Otherwise Specified
PL: problem list
SNOMED: Systematized Nomenclature of Medicine
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SOAP: Subjective, Objective, Assessment, and Plan

Edited by C Lovis; submitted 27.07.23; peer-reviewed by C Gaudet-Blavignac, L Chu, T Karen; revised version received 01.12.23; accepted 22.02.24; published 09.05.24.

Please cite as:
Senior R, Tsai T, Ratliff W, Nadler L, Balu S, Malcolm E, McPeek Hinz E
Evaluation of SNOMED CT Grouper Accuracy and Coverage in Organizing the Electronic Health Record Problem List by Clinical System: Observational Study
JMIR Med Inform 2024;12:e51274
URL: https://medinform.jmir.org/2024/1/e51274
doi:10.2196/51274

Original Paper

Dermoscopy Differential Diagnosis Explorer (D3X) Ontology to Aggregate and Link Dermoscopic Patterns to Differential Diagnoses: Development and Usability Study

Rebecca Z Lin[1], MD; Muhammad Tuan Amith[2,3,4], PhD; Cynthia X Wang[5], MPHS, MD; John Strickley[6], MD; Cui Tao[7], PhD

[1]Division of Dermatology, Washington University School of Medicine, St. Louis, MO, United States
[2]Department of Information Science, University of North Texas, Denton, TX, United States
[3]Department of Biostatistics and Data Science, The University of Texas Medical Branch, Galveston, TX, United States
[4]Department of Internal Medicine, The University of Texas Medical Branch, Galveston, TX, United States
[5]Department of Dermatology, Kaiser Permanente Redwood City Medical Center, Redwood City, CA, United States
[6]Division of Dermatology, University of Louisville, Louisville, KY, United States
[7]Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL, United States

Corresponding Author:
Cui Tao, PhD
Department of Artificial Intelligence and Informatics
Mayo Clinic
4500 San Pablo Road
Jacksonville, FL, 32224
United States
Phone: 1 9049530255
Email: Tao.Cui@mayo.edu

Abstract

Background: Dermoscopy is a growing field that uses microscopy to allow dermatologists and primary care physicians to identify skin lesions. For a given skin lesion, a wide variety of differential diagnoses exist, which may be challenging for inexperienced users to name and understand.

Objective: In this study, we describe the creation of the dermoscopy differential diagnosis explorer (D3X), an ontology linking dermoscopic patterns to differential diagnoses.

Methods: Existing ontologies that were incorporated into D3X include the elements of visuals ontology and dermoscopy elements of visuals ontology, which connect visual features to dermoscopic patterns. A list of differential diagnoses for each pattern was generated from the literature and in consultation with domain experts. Open-source images were incorporated from DermNet, Dermoscopedia, and open-access research papers.

Results: D3X was encoded in the OWL 2 web ontology language and includes 3041 logical axioms, 1519 classes, 103 object properties, and 20 data properties. We compared D3X with publicly available ontologies in the dermatology domain using a semiotic theory–driven metric to measure the innate qualities of D3X with others. The results indicate that D3X is adequately comparable with other ontologies of the dermatology domain.

Conclusions: The D3X ontology is a resource that can link and integrate dermoscopic differential diagnoses and supplementary information with existing ontology-based resources. Future directions include developing a web application based on D3X for dermoscopy education and clinical practice.

(JMIR Med Inform 2024;12:e49613) doi:10.2196/49613

KEYWORDS

medical informatics; biomedical ontology; ontology; ontologies; vocabulary; OWL; web ontology language; skin; semiotic; web app; web application; visual; visualization; dermoscopic; diagnosis; diagnoses; diagnostic; information storage; information retrieval; skin lesion; skin diseases; dermoscopy differential diagnosis explorer; dermatology; dermoscopy; differential diagnosis; information storage and retrieval

Introduction

Dermoscopy is a noninvasive, in vivo microscopic technique used to examine skin lesions by detecting morphological features that may not be seen by the naked eye [1-4]. Studies have demonstrated that dermoscopy improves the diagnosis of both pigmented skin lesions [3,5-7] and nonpigmented skin lesions [8], including neoplasms [9] and infectious and inflammatory skin diseases [4,10]. Notably, the diagnostic accuracy of dermoscopy is dependent on the examiner's experience, as dermoscopy by untrained or less experienced examiners was found to be no better than clinical inspection without dermoscopy [6]. Learning dermoscopy is not just relevant to dermatologists, but also for physicians in other medical specialties. Patients with new or changing skin lesions often first consult their primary care physician (PCP) rather than a dermatologist. Dermoscopy has shown to be an effective tool for the assessment and triage of pigmented skin lesions in primary care, with improved diagnostic accuracy and referral accuracy to dermatologists [11-13]. However, dermoscopy training for PCPs is currently highly variable, with many PCPs citing a lack of training as a key barrier to the use of dermoscopy [14-16]. Furthermore, short dermoscopy training programs [14] may be insufficient to establish long-term competency in dermoscopy, with poor continuing use of dermoscopy and the need for refresher sessions [17]. The need for dermoscopy training among plastic surgeons has recently been documented as well [18]. Thus, the development of machine-based tools for dermoscopy may enhance clinical practice for dermatology providers and other medical professionals.

The use of standard terminologies organized through taxonomies has a long history with the life sciences, starting with Carl Linnaeus' taxonomy [19]: a classification system to name and group species according to their shared characteristics. Centuries later and with advances in computing infrastructure, these types of classification systems have continued to be of interest to the science community. An ontology is "a representational artifact comprising a taxonomy as proper part, whose representational units are intended to designate some combination of universals, defined classes, and certain relations between them" [20]. Essentially, an ontology is a graphical representation of linked concepts to formalize a schema (Tbox) for data (Abox). The formalization leverages semantic links (Rbox) between the concepts to give data more meaning and to aggregate related data of any heterogeneous format. This ensures the normalization of heterogeneous data. Furthermore, with semantics, ontologies could support machine reasoning to generate references via deductive reasoning. As related to the medical field, ontologies can extend the computability of standard controlled terminologies to provide descriptive and composite representations of medical information (such as features related to various diagnoses). Ontologies represent the data in a machine-readable format to give computing tools more context, making them highly valuable for artificial intelligence.

Within the dermatology domain, some existing ontologies aim to describe cutaneous disorders. For example, the dermatology lexicon (DERMLEX) was created with the American Academy of Dermatology with a nosology, anatomical distributions, classical signs, and therapeutic procedures; however, maintenance was discontinued in 2009 [21,22]. More recently, the human dermatological disease ontology (DERMO) was developed to classify cutaneous diseases by etiology, anatomical location or cell type, and phenotype consistent with current clinical practice [23,24]. Some other dermatology-specific ontologies exist, including the skin physiology ontology (SPO; last updated in 2008) [25], but notably, none of these ontologies connect cutaneous disorders to metaphoric terms like "strawberry pattern" which may be difficult for a machine to understand. Similarly, none of the aforementioned ontologies specifically address dermoscopy, which is a specialized technique that may have special considerations when used in diagnosis. For instance, the colors of certain lesions are best seen under polarized light [26]. As such, there is a need to develop an ontology that adequately addresses the field of dermoscopy, with the capability of processing both descriptive and metaphoric terminology.

In our previous work, we developed the elements of visuals ontology (EVO) to decompose the fundamental features of visualizations, such as shapes, colors, and textures. The dermoscopy elements of visuals ontology (DEVO) then applied the visual features described in EVO to dermoscopic terminology [27]. For instance, DEVO characterizes dermoscopic metaphoric terms such as "shiny white streaks" and "leaflike areas" by shapes, colors, and textures, along with other features involved. Discussion with domain experts revealed that while DEVO is capable of responding to queries to find visual features associated with metaphoric terms and vice versa, linking the dermoscopic terms to differential diagnoses would significantly enhance its clinical utility. A list of differential diagnoses indicates many possible diagnoses that share similar features to the patient's symptoms and signs. These differential diagnoses can then be narrowed down to aid the clinician in identifying the final diagnosis. As dermatology is a technical field, the landscape of differential diagnoses is wide and difficult to parse [28]. In this study, we describe the extension of EVO and DEVO to create the dermoscopy differential diagnosis explorer (D3X), an ontology linking metaphoric terms to differential diagnoses. We further propose a use case integrating D3X into a web application in dermoscopy education and clinical practice.

Methods

Ethical Considerations

This article adheres to the Committee on Publication Ethics guidelines. This research did not involve human subjects.

Integration of Existing Ontologies

Overview

A common practice in the development of ontologies [29] is to reuse existing ontologies' components to ensure semantic interoperability. We used the following ontologies to build the D3X ontology.

About EVO

EVO is a foundational ontology model that describes the basic constituents of visualizations: shapes, colors, strokes (lines), size, perceived texture, etc. It also represents the dimensional extended 9-intersection model, a mathematical model for spatial relationships between elements [30]. Further, EVO imports and reuses controlled terminologies and standards from the W3C scalable vector graphics, Wikidata, phenotype and trait ontology, and the simple knowledge organization system to supplement our core representational model of visualizations. EVO is hosted on GitHub for public release and is coded in the OWL 2 web ontology language.

About DEVO

DEVO is an extension of EVO that reuses the foundational understanding of visualizations for the dermoscopy domain. DEVO incorporates some of the controlled terminologies—"metaphoric" and "descriptive"—that are used in practice by dermatologists, with a focus on the metaphoric terminologies. With DEVO, we developed a core model that encodes and describes the "visual language" of the dermoscopic terms' definitions. Further, one important outcome of this work was a computable representational model of an agreed understanding of visual elements of dermoscopic patterns, which we used as a framework to generate differential diagnoses. Similarly, DEVO was coded in OWL 2 and is hosted on GitHub for public consumption.

Miscellaneous Ontologies and Vocabularies

We also aligned D3X with commonly used top-level ontologies. The information artifact ontology (IAO) [31] is part of the open biological and biomedical ontology (OBO) foundry. IAO represents a general abstraction of informational objects (like documents and components within those documents—eg, figures, images). Like many OBO foundry ontologies, IAO uses the basic formal ontology and relation ontology as part of its architecture model. We minimally reused some of the term entities and properties like IAO:image and "denoted by." We also reused the software ontology (SWO) [32] for its licensing entity terms—SWO:license and "has license"—to describe the licensing information for any imaging resource of skin lesions. Lastly, we used Schema.org's [33] schema::image to link image resources.

Development of D3X

To generate a list of differential diagnoses, we started with the metaphoric terms defined in DEVO from the third consensus conference conducted by the International Society of Dermoscopy [34]. We then searched the literature [34-36] for corresponding differential diagnoses for each term and consulted 2 domain experts to independently edit the list of diagnoses for accuracy. These differential diagnoses were later encoded using Protégé [37] in our ontology. Following this, we reviewed open-source resources (DermNet, Dermoscopedia, and open-access research papers) for a collection of hosted images that matched individual differential diagnoses. We tracked the provenance information and associated data (caption, description, etc) in a spreadsheet as a central organized resource that mapped the images for each diagnosis to the concept diagnosis used in D3X. To streamline the data transfer process, we developed a management code to transfer data from the spreadsheet to the ontology. The source code is available on our GitHub repository, using the OWL API to facilitate efficient custom import. This approach allowed centralized data collection and also enabled an ad-hoc import and data creation pipeline.

Semiotic Evaluation

Semiotic theory is the study of signs and symbols, and considering ontologies are symbolic representations of a specific domain, we used a metric suite grounded in that theoretical framework [38]. Semiotic theory is composed of 3 basic qualities: *syntactic*, *semantic*, and *pragmatic*. Essentially, in the context of ontologies, the metric suite components refer to aspects of the ontology artifact—*syntactic* concerning encoding adherence and standards; *semantic* concerning the effective use of human-friendly labels for entities and concepts; and *pragmatic* concerning function. Each of these qualities is quantified based on a computation of representative quantifiable features of an ontology file (eg, the number of classes, the average number of word senses for labels, etc). This is described in detail in previously published works [38]. This suite helps to measure some of the intrinsic qualities of our ontology concerning other ontologies in the same domain. We used publicly available ontologies from the skin and dermatology domain—DERMLEX, DERMO, and SPO—that are found in the National Center for Biomedical Ontology (NCBO) BioPortal. We used a command line version of our tool OntoKeeper [39] to quickly generate scores from the metric suite and then calculated *z* scores to determine how D3X fares in terms of intrinsic quality with other ontologies of the dermatology domain.

Results

Development of D3X

The D3X ontology was encoded in the OWL 2 web ontology language. In terms of the size of the ontology, there are 3041 logical axioms, 1519 classes, 103 object properties, and 20 data properties. Imported image data are encoded as 387 instances. Figure 1 displays a sample series of screenshots showing Kaposi sarcoma, as an example entity, linked to DEVO's rainbow pattern, standard medical terminologies (eg, Systematized Nomenclature of Medicine–Clinical Terms [SNOMED CT], National Cancer Institute [NCI] Thesaurus), and the open-sourced image example. For ongoing data management, we host the spreadsheet with image data (n=364 images) and the OWL API software code to allow for an automated process of adding new image data. The software will pull the data from the spreadsheet and will add and export a version of our ontology that has the image instance data. Both the spreadsheet and the software are available on our GitHub repository [32]. As more dermoscopy images become available for the public domain, we will include them in our spreadsheet and generate an encoded export with the new instance data. D3X uses our pre-existing work of DEVO and also leverages terminology from the IAO, SWO, and Schema.org. Figure 2 shows a global

overview of the D3X ontology and the various linked terminologies that compose the entire model.

Figure 1. Sample screenshot of the D3X ontology through Protégé showing related metadata and information about Kaposi sarcoma. D3X: dermoscopy differential diagnosis explorer; DDX: differential diagnosis.

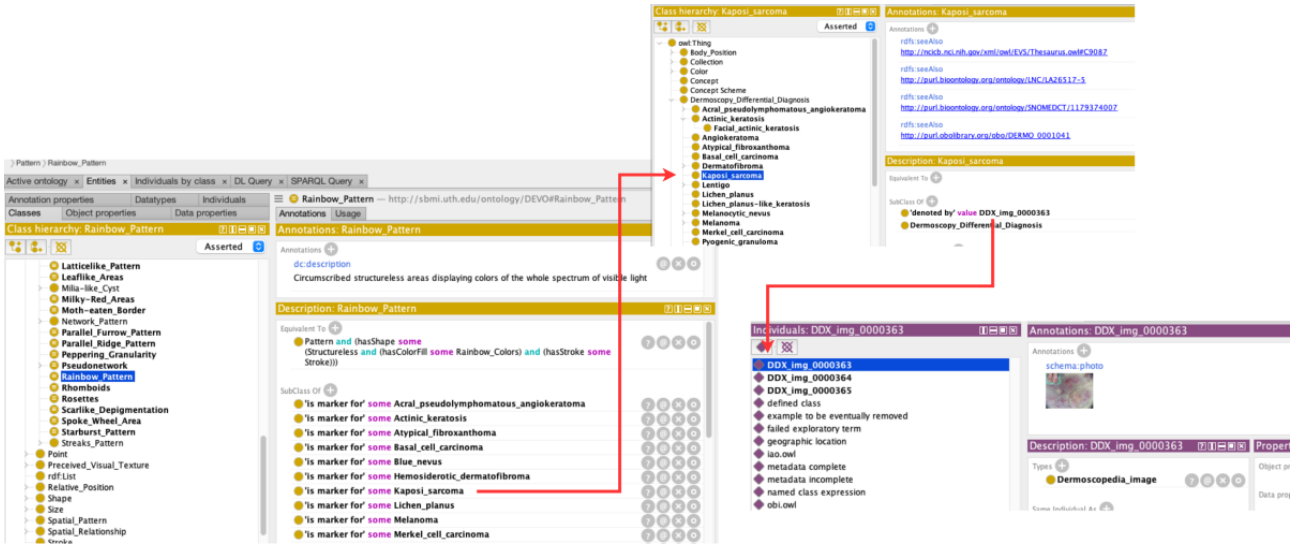
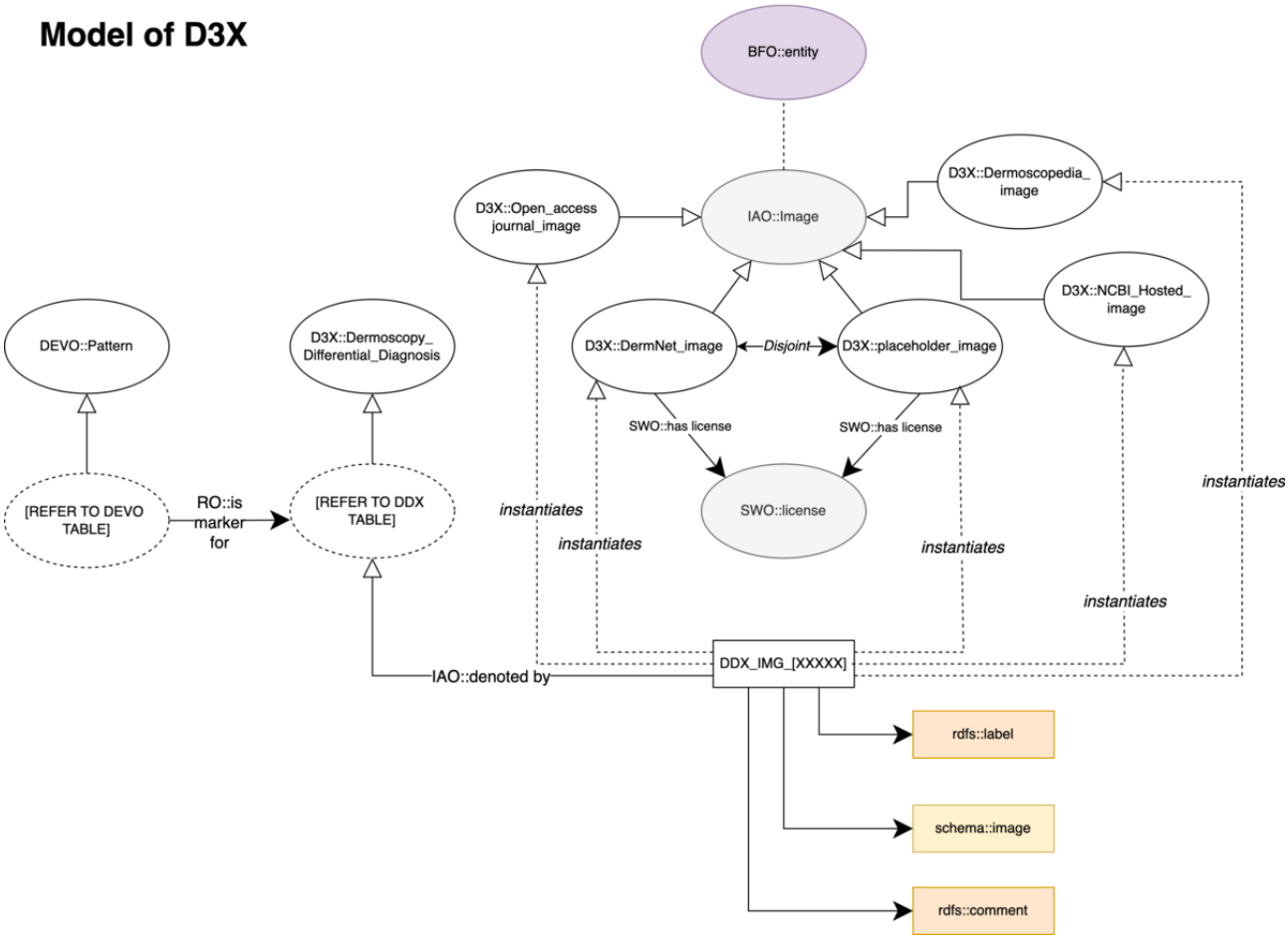


Figure 2. Global overview of the D3X ontology and linked ontologies, including the DEVO, BFO, IAO, and SWO. BFO: basic formal ontology; D3X: dermoscopy differential diagnosis explorer; DEVO: dermoscopy elements of visuals ontology; IAO: information artifact ontology; rdfs: resource description framework schema; RO: relation ontology; SWO: software ontology.



Each dermoscopy sample image is represented as a single instance data value with a unique ID (DDX_IMG_[DIGITS]). As an instance data value representing the digital image, it links to the exact file on the web using schema:image from Schema.org. Caption information is used as an annotation for RDF:comment (RDF: resource description framework) and rdf:label (rdf: resource description framework). The instance data value is an instantiation of a specific image class from the

following sources: IAO—open access journal images, DermNet images, National Center for Biotechnology Information–hosted images, and Dermoscopedia images. For some of them, the "has licenses" predicate links to a license, signifying that any instance of this image class has some license agreement (eg, Creative Commons). The licensing terminology is derived from the SWO and is hosted on our GitHub repository as an external import.

With D3X, we declared a new dermoscopy differential diagnosis class. This class provides a list of associated diagnoses for skin lesions. Each of the dermoscopy differential diagnosis classes is linked to a pattern from DEVO. The pattern in DEVO ontologically describes each dermoscopic pattern (metaphoric term) using visual elements, such as lines, shapes, colors, and spatial relationships. Table S1 in Multimedia Appendix 1 provides a comprehensive list of the metaphoric patterns listed in DEVO and their corresponding differential diagnoses in D3X. Each pattern in DEVO is linked to its differential diagnoses using OBO's "is marker for" (eg, angular lines > is a marker for > Lentigo_maligna). Additionally, the instance images described above are linked to the differential diagnoses using "denoted by," such that each diagnosis is provided with at least one visual example. Lastly, for each of the dermoscopy differential diagnosis classes, there are associated annotations that link the class to the other standardized ontologies like the Medical Dictionary for Regulatory Activities (MedDRA), SNOMED CT, NCI Thesaurus, and LOINC (logical observation identifier names and codes). MedRA covered 63% (n=25) of the classes, while SNOMED CT and NCI Thesaurus covered 53% (n=21) and 55% (n=22) of the classes, respectively. The remaining, like DERMO and LOINC, covered 15% (n=6) and 3% (n=1) of the classes.

Semiotic Evaluation

Semiotic theory is composed of 3 basic qualities: *syntactic*, *semantic*, and *pragmatic*. Table 1 displays the *z* scores for each of the qualities and subqualities of D3X compared to other publicly available ontologies in the dermatology domain. Examining the *syntactic* quality of D3X (z=0.17), while it lacks diverse syntactic *richness* (z=–0.74) in comparison with its other domain counterparts, D3X does adhere to syntactic *lawfulness* (z=0.49). D3X compares satisfactorily with other ontologies in the *semantic* quality (z=0.77). Although the semantic *clarity* subquality was below average than its peers (z=–0.91; the ambiguity of labels), D3X does better with semantic *consistency* (z=0.56; the number of essentially unique labels) and semantic *interpretability* (z=0.65; whether the label has meaning). The *pragmatic* quality is composed of 1 score: *comprehensiveness*, a measure of the coverage of the domain scope of the ontology based on the number of entities encoded, which was nearly below average for D3X (z=–0.66). Lastly, the overall score of D3X (z=0.58) points to a somewhat better overall quality score than DERMLEX and SPO (z=–1.41 and 0.00, respectively). Although DERMO had a slightly higher overall quality than D3X (z=0.83), its score is still within 1 SD of the D3X ontology score, so the quality of D3X appears at least comparable to that of the other ontologies within its own domain.

Table 1. Semiotic comparison of D3X[a] to other dermatology ontologies: the DERMLEX[b], DERMO[c], and SPO[d] using *z* scores.

Quality and subquality	Mean (SD)	D3X-z	DERMLEX-z	DERMO-z	SPO-z
Syntactic	0.57 (0.11)	0.17[e]	–1.33	0.08	1.09[e]
Richness	0.26 (0.11)	–0.74	0.62[e]	–0.96	1.08[e]
Lawfulness	0.87 (0.25)	0.49	–1.50	0.51[e]	0.51 [e]
Semantic	0.85 (0.13)	0.77[e]	–1.38	0.70[e]	–0.08
Clarity	0.99 (0.01)	–0.91	0.91[e]	0.82[e]	–0.82
Consistency	0.73 (0.49)	0.56[e]	–1.50	0.43	0.51[e]
Interpretability	0.87 (0.21)	0.65[e]	0.65[e]	0.16	–1.46
Pragmatic	0.11 (0.15)	–0.66	1.44[e]	–0.09[e]	–0.69
Comprehensiveness	0.11 (0.15)	–0.66	1.44[e]	–0.09[e]	–0.69
Overall score	0.52 (0.03)	0.58 [e]	–1.41	0.83[e]	0.00

[a]D3X: dermoscopy differential diagnosis explorer.

[b]DERMLEX: dermatology lexicon.

[c]DERMO: human dermatological disease ontology.

[d]SPO: skin physiology ontology.

[e]These values indicate the 2 highest values for each quality and subquality.

Discussion

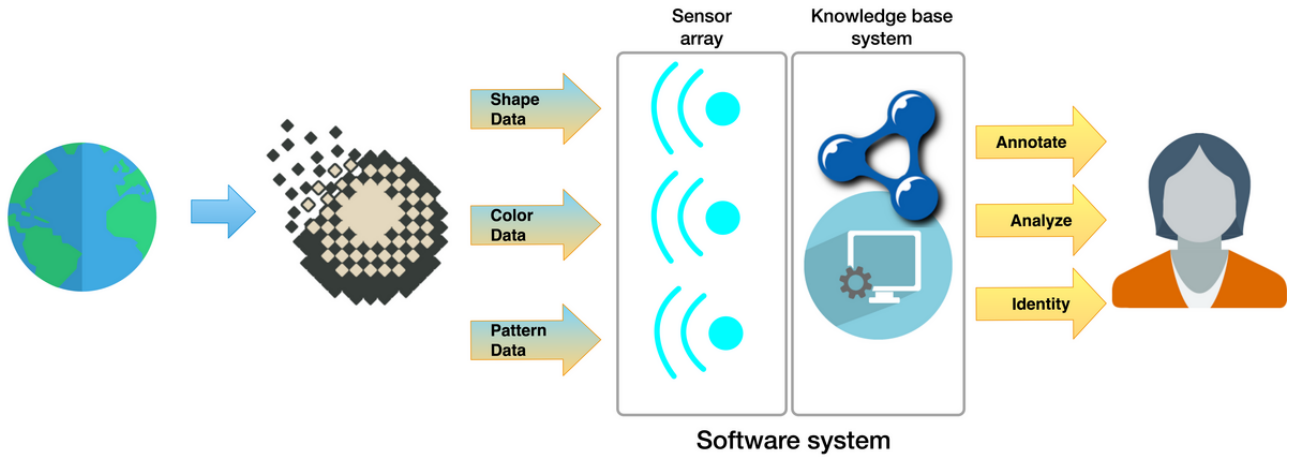
Principal Results and Limitations

D3X is an ontology that connects dermoscopic patterns (metaphoric terms) with differential diagnoses. It is an extension of the DEVO to describe patterns based on their visual elements, which is in turn an extension of the EVO. D3X also leverages terminology from IAO, SWO, Relation Ontology, and Schema.org, and its differential diagnoses are linked to MedDRA, SNOMED CT, and NCI Thesaurus. Using the semiotic theory framework proposed by Burton-Jones et al [38], we measured D3X in comparison with similar publicly available ontologies to assess its intrinsic quality. Our assessment indicates that while comparably better to the other ontologies of the same domain in its overall score, D3X does lack diverse syntactic *richness* and could improve its semantic *clarity* (despite a better overall semantic quality score than its ilk) and pragmatic *comprehensiveness*. Leveraging additional OWL 2 syntactic features could improve the syntactic richness. However, since the purpose of our ontology is to retrieve and aggregate information and metadata about dermoscopic features, some of the more sophisticated OWL 2 features like symmetry, inverse, etc, may not be necessary for our use case. As for the pragmatic score, it might improve over time as we collect more instance data of images to link to our ontology. Further, our assessment was limited to 3 ontologies as there are no other publicly available ontologies that deal solely with a dermatology subject. Additionally, OntoKeeper uses a subset of scores as the social quality (composed of *authority* and *history*), and the pragmatic subscores of *accuracy* and *relevancy* are difficult to compute, so they are not listed in our semiotic analysis [39]. Despite this, the quality scores are sufficient for an application ontology, since the role of this artifact is to aggregate and consolidate skin diagnostic information—an area where it is likely to shine.

The aforementioned evaluation included DERMLEX, DERMO, and the SPO. DERMLEX was originally created by the American Academy of Dermatology to describe dermatological diagnosis and related domain vocabularies, aligned to *International Classification of Diseases, Ninth Revision* (*ICD-9*). However, the upkeep ended in 2009 [22]. DERMO is another ontology that also aims to describe dermatological diseases, but unlike DERMLEX, it is aligned to *International Classification of Diseases, Tenth Revision* (*ICD-10*). The latest version was last released in 2015, according to the NCBO BioPortal record [23]. Not much is known about SPO, other than a presence on NCBO BioPortal and the latest release dating back to 2008 [25]. Compared to these existing works, D3X yields richer semantics and applicability by the OWL2 encoding in EVO and DEVO that describes lesions using primitive visualization elements. Another advantage of this work is the use of semantic web properties of our work, namely the linking of heterogeneous resources (external entities, images, metadata, etc). This allows D3X to be an application-driven artifact that can be integrated into software tools, and other analytical and educational tools. According to researchers, terminologies enriched with semantics will yield opportunities to develop innovative tools and applications [40]. We further discuss our vision in the subsequent sections (see Proposed Web Application and Use Case). Overall, we presume this work provides a richer ontological artifact compared to similar ontologies of the same domain.

Aside from our aforementioned application use case, this work can advance machine learning models for dermoscopy diagnosis support. There has been some preliminary evidence that machine learning models can be supported or improved by ontologies [41-43]. Potentially, the combined stack of EVO, DEVO, and D3X could augment tools that analyze real-world entities (eg, lesions). In Figure 3, we illustrate a hypothetical example where a software application segments signals from an entity using machine learning in a sensor array to detect shape, color, and pattern data. The structured information from the sensor array could then be linked to an ontological knowledge base system that expresses meaning and context.

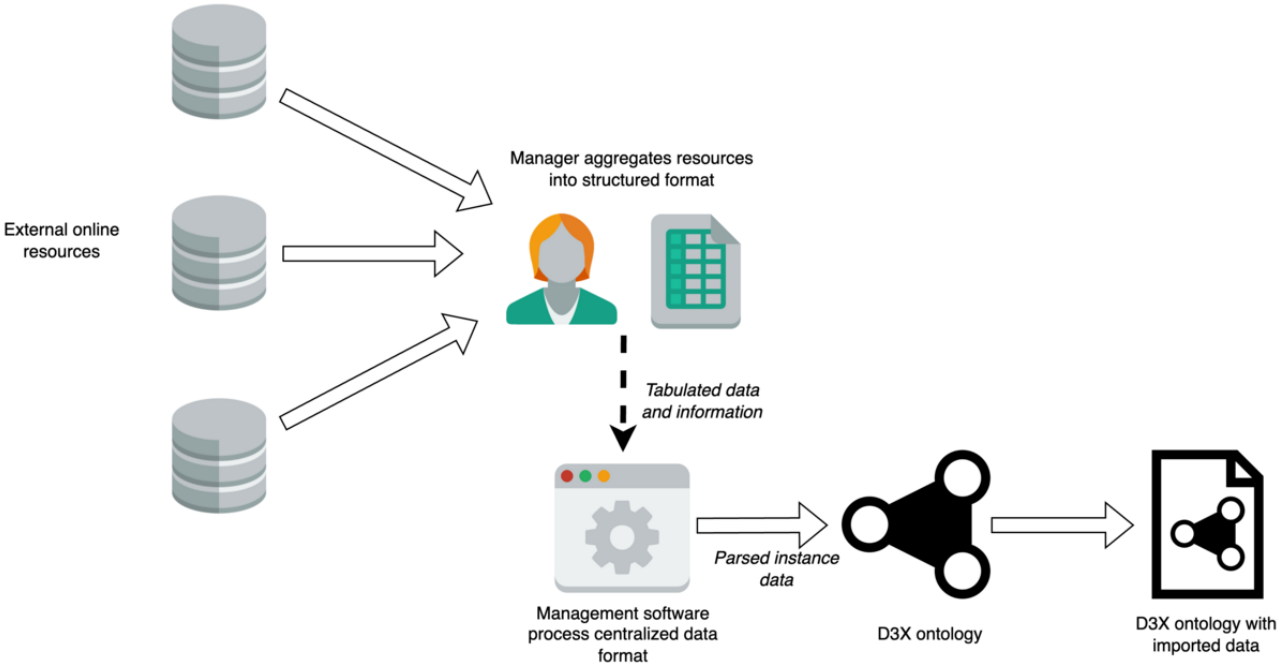
Figure 3. Diagram of a software system using segmented machine learning with a sensor array linked to an ontological knowledge base system. Pixel art icon by DesignContest is licensed under CC BY 4.0. Earth icon by Treetog ArtWork. Globe icon and user female icon by paomedia.



Data Upkeep and Management Plan

Noted earlier, we produced a basic management system to allow for continued integration of data and information from external sources to be added to D3X. Continued data management is an issue with some ontology and controlled terminology resources. By having this basic management system, we can ensure that D3X will be up to date with little resources and time needed to integrate new diagnostic information and metadata. Figure 4 shows the basic management pipeline, with the tools needed, hosted on our GitHub repository under the ddx_data_management folder. In the aforementioned figure, any new or updated digital resources (images, web page text, knowledge graph, and ontology resources) will be added to a centralized spreadsheet for the human-friendly organization of data for diagnosis information. The management software will import the spreadsheet and parse the data for the D3X ontology. The final output of the software is the D3X ontology with the updated linked information. Future plans could include using shapes constraint language (SHACL) to ensure the quality of the data is validated, and further development of data management software to facilitate ease of use.

Figure 4. Outline of the D3X ontology data upkeep and management plan. D3X: dermoscopy differential diagnosis explorer. OWL Lite icon and OWL Lite document icon by Picol Team are licensed under CC BY 4.0. File excel icon, database icon, window system icon, user female alt icon by paomedia.

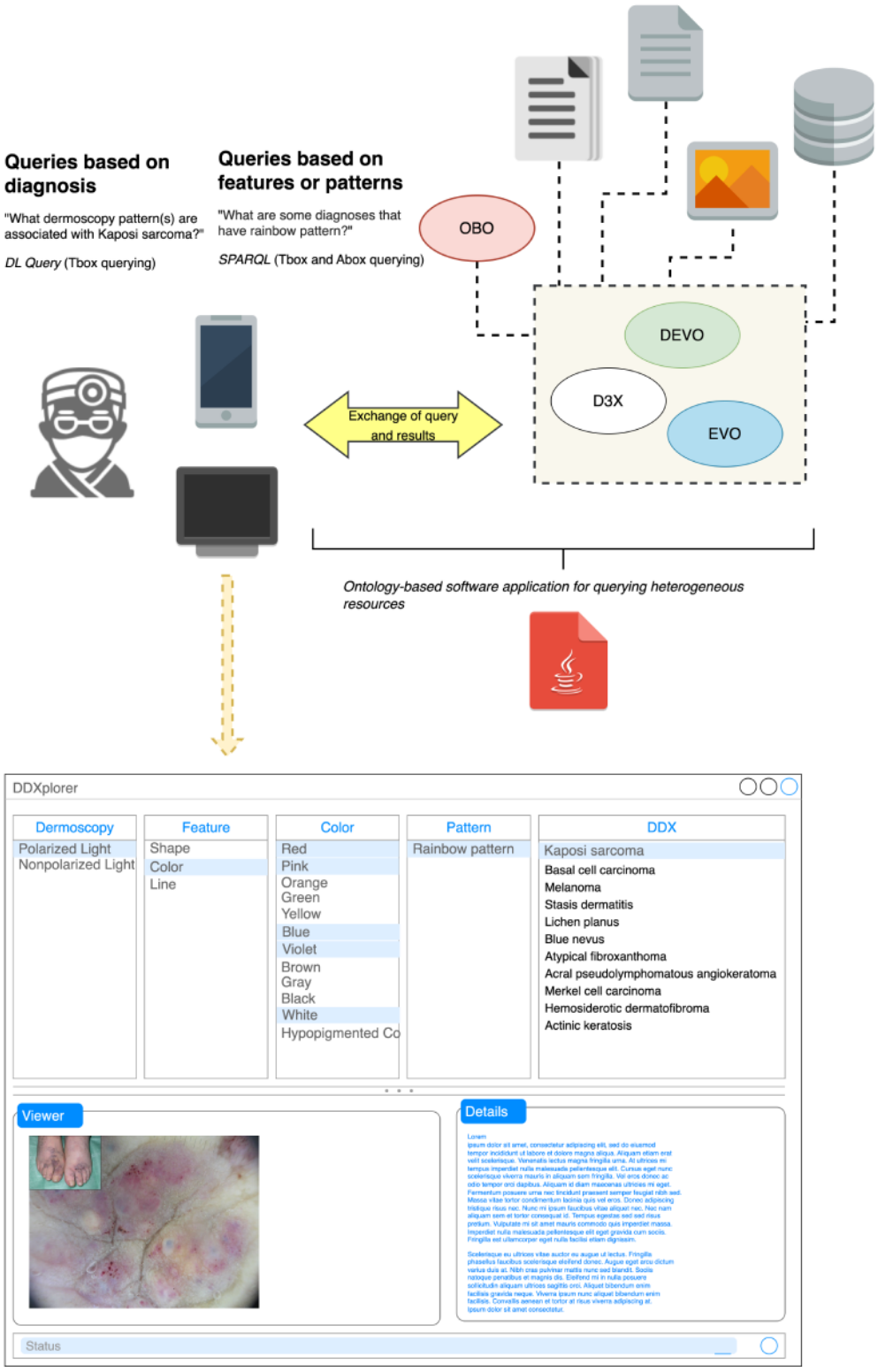


Proposed Web Application and Use Case

For use in clinical practice, an ontology-based software application could be designed using D3X to guide the identification of differential diagnoses. After the user performs dermoscopy on a skin lesion, they can open the web application and select various features about the lesion, which will then suggest differential diagnoses. A mock-up for the use case of Kaposi sarcoma is shown in Figure 5. The user can select options for the first 3 boxes ("dermoscopy," "feature," and the chosen feature [eg, "color"], with multi-select functionality available for the latter). The web application would then generate the relevant patterns and a list of differential diagnoses. Thus, if the user indicates that under polarized light, red, pink, blue, violet, and white colors were visualized, this corresponds to a "rainbow pattern," which is associated with Kaposi sarcoma, among other differential diagnoses. Clicking on each differential diagnosis will then display any relevant dermoscopic images in the "viewer," as well as a description under "details" with a link to learn more about the condition. By reviewing images and descriptions of these differential diagnoses, this would ideally help the user narrow down the list and identify the most likely diagnosis. This web application was independently reviewed by 2 domain experts who agreed that the format was understandable to the user; they also stated that the information provided would be useful in clinical practice as a quick search for differential diagnoses. Of note, the aforementioned example illustrates a query based on features or patterns, but the web application would also be capable of querying based on diagnoses (eg, "what dermoscopy patterns are associated with Kaposi sarcoma?"). This would be more useful in an educational setting for those who want to gain an understanding of dermoscopic patterns and the features comprising each pattern. Furthermore, we proposed the development of a web application harnessing D3X capable of carrying out the following queries: (1) given dermoscopic features or patterns, output a list of differential diagnoses; and (2) given a differential diagnosis, output associated features and patterns. Along with a description of each differential diagnosis, the application would also display images from DermNet, Dermoscopedia, National Center for Biotechnology Information Hosted, or open-access journals for ease of understanding the relationship between each differential diagnosis and its visual elements. There is a growing body of literature on machine learning models for automated diagnosis of dermoscopic images, such as convolutional neural networks (CNNs) [9]. Both ontologies and CNNs fall under the artificial intelligence umbrella, but ontologies relate to knowledge representation, while CNN is statistical machine learning. These

are fundamentally different approaches to power artificial intelligence that are difficult to compare directly. While automated diagnosis via CNNs is a very promising area of study, research has largely focused on the diagnosis of melanoma [44-46], with few studies including pigmented nonmelanocytic lesions [47,48] and largely ignoring nonpigmented lesions. D3X labels dermoscopic patterns of pigmented and nonpigmented lesions, so it may apply to a broader range of patient visits. Additionally, the likelihood of provider acceptance of automated diagnosis systems is unclear. With our proposed web application, providers would be able to input search criteria themselves and see a list of differential diagnoses, rather than a binary output for 1 diagnosis (eg, melanoma) suggested by the machine, which may not be as likely to be accepted by physicians.

Figure 5. Mock-up of a web application harnessing the D3X ontology to perform queries for differential diagnoses associated with dermoscopic features and patterns, with Kaposi sarcoma as a use case. Doctor Icon by MedicalWP is licensed under CC BY 4.0. Computer icon and text x java icon by Papirus Dev Team are licensed under GNU GPL (version 3.0). Database icon, file text icon, file picture icon, device mobile phone icon by paomedia. Abox: assertion component of a knowledge base; D3X: dermoscopy differential diagnosis explorer; DEVO: dermoscopy elements of visuals ontology; DL: description logic; EVO: elements of visuals ontology; OBO: open biological and biomedical ontology; SPARQL: SPARQL protocol and RDF query language (recursive acronym); Tbox: terminology component of a knowledge base.



In discussion with domain experts, we chose not to integrate diagnostic rules into D3X, as experienced dermoscopy users could assess the list of differential diagnoses fairly quickly and decide on the most likely diagnosis using their clinical expertise. If the user is relatively inexperienced, they may gain more understanding by reading the description of each differential diagnosis, viewing the images, and accessing additional information by clicking “learn more” in the web application. Nevertheless, in the future, it may be useful to integrate diagnostic rules into D3X for more sophisticated suggestions

of differential diagnoses (eg, ranking the most to least likely differential diagnoses in a prioritized list). Another limitation is that there are dermoscopic terms and differential diagnoses not included in D3X, given that we built D3X from the terms mentioned in the International Society of Dermoscopy's third consensus conference [34]. Similarly, while we aimed to include images from a variety of external sources, we acknowledge that they may not be fully representative of all patient skin tones. Our work is only a starting point, as we plan to continue updating D3X and anticipate that its library will become more comprehensive with time.

To conclude, the web application based on D3X has great potential for use in several areas. First, it could be included alongside formal dermoscopy training as a supplementary educational tool for dermatology trainees and providers in other specialties (PCPs, plastic surgeons). Given that providers may require ongoing dermoscopy refresher sessions to feel fully comfortable even after completing an initial training program [17], this web application could be a helpful reference to deepen understanding of dermoscopic patterns associated with different skin conditions. Furthermore, in a clinical setting, providers could quickly query the web application for a list of differential diagnoses after dermoscopic examination of a lesion, which would aid in the identification of their patient's diagnosis. This may be useful for inexperienced and experienced dermoscopy users alike, as it is intended to augment, not replace, the provider's clinical reasoning. The next steps include using our proposed interface to build a functional web application. Creating the web application could reveal additional flaws in the design that require clarification, and we would continue to improve aspects of D3X and the web application in an iterative process. After a beta version of the web application is finalized, we would aim to conduct user testing to evaluate the user experience as well as clinical or educational utility among physicians.

Conclusions

We introduce and discuss the design and development of the D3X ontology as a resource that can link and integrate dermoscopic differential diagnoses and supplementary information with existing ontology-based resources (MedDRA, SNOMED CT, and NCI). We repurposed a previous work of the EVO and DEVO to construct and support D3X, along with other supplementary standardized ontologies, like IAO and SWO. Using the semiotic theoretical framework to compare D3X with other dermatology-related ontologies, its overall quality score was similar to existing ontologies' scores. One of the outcomes of this work is providing a means to aggregate and link dermoscopic patterns to differential diagnoses, thereby enhancing understanding of dermoscopy for educational and clinical use. This outcome has fueled our next objective in developing a web-based application that can query D3X and fetch the linked information for the user. Currently, D3X and its resources are available on GitHub for public release and use.

Acknowledgments

This work was partially supported by the National Institutes of Health (R01AI130460 and U24CA194215) and the Cancer Prevention Research Institute of Texas (RP220244).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patterns in the dermoscopy elements of visuals ontology and their associated differential diagnoses in the dermoscopy differential diagnosis explorer.
[DOCX File , 19 KB - medinform_v12i1e49613_app1.docx]

References

1. Argenziano G, Soyer HP. Dermoscopy of pigmented skin lesions—a valuable tool for early diagnosis of melanoma. Lancet Oncol 2001;2(7):443-449. [doi: 10.1016/s1470-2045(00)00422-8] [Medline: 11905739]
2. Vázquez-López F, Manjón-Haces JA, Maldonado-Seral C, Raya-Aguado C, Pérez-Oliva N, Marghoob AA. Dermoscopic features of plaque psoriasis and lichen planus: new observations. Dermatology 2003;207(2):151-156. [doi: 10.1159/000071785] [Medline: 12920364]
3. Rosendahl C, Tschandl P, Cameron A, Kittler H. Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. J Am Acad Dermatol 2011;64(6):1068-1073. [doi: 10.1016/j.jaad.2010.03.039] [Medline: 21440329]
4. Lallas A, Argenziano G, Apalla Z, Gourhant JY, Zaballos P, Di Lernia V, et al. Dermoscopic patterns of common facial inflammatory skin diseases. J Eur Acad Dermatol Venereol 2014;28(5):609-614. [doi: 10.1111/jdv.12146] [Medline: 23489377]
5. Bafounta ML, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. Arch Dermatol 2001;137(10):1343-1350. [doi: 10.1001/archderm.137.10.1343] [Medline: 11594860]
6. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002;3(3):159-165. [doi: 10.1016/s1470-2045(02)00679-4] [Medline: 11902502]

7. Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br J Dermatol 2008;159(3):669-676. [doi: 10.1111/j.1365-2133.2008.08713.x] [Medline: 18616769]
8. Sinz C, Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, et al. Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. J Am Acad Dermatol 2017;77(6):1100-1109. [doi: 10.1016/j.jaad.2017.07.022] [Medline: 28941871]
9. Weber P, Tschandl P, Sinz C, Kittler H. Dermatoscopy of neoplastic skin lesions: recent advances, updates, and revisions. Curr Treat Options Oncol 2018;19(11):56 [FREE Full text] [doi: 10.1007/s11864-018-0573-6] [Medline: 30238167]
10. Haliasos EC, Kerner M, Jaimes-Lopez N, Rudnicka L, Zalaudek I, Malvehy J, et al. Dermoscopy for the pediatric dermatologist part I: dermoscopy of pediatric infectious and inflammatory skin lesions and hair disorders. Pediatr Dermatol 2013;30(2):163-171. [doi: 10.1111/pde.12097] [Medline: 23405886]
11. Westerhoff K, McCarthy WH, Menzies SW. Increase in the sensitivity for melanoma diagnosis by primary care physicians using skin surface microscopy. Br J Dermatol 2000;143(5):1016-1020. [doi: 10.1046/j.1365-2133.2000.03836.x] [Medline: 11069512]
12. Argenziano G, Puig S, Zalaudek I, Sera F, Corona R, Alsina M, et al. Dermoscopy improves accuracy of primary care physicians to triage lesions suggestive of skin cancer. J Clin Oncol 2006;24(12):1877-1882. [doi: 10.1200/JCO.2005.05.0864] [Medline: 16622262]
13. Herschorn A. Dermoscopy for melanoma detection in family practice. Can Fam Physician 2012;58(7):740-745 [FREE Full text] [Medline: 22859635]
14. Chappuis P, Duru G, Marchal O, Girier P, Dalle S, Thomas L. Dermoscopy, a useful tool for general practitioners in melanoma screening: a nationwide survey. Br J Dermatol 2016;175(4):744-750. [doi: 10.1111/bjd.14495] [Medline: 26914613]
15. Morris JB, Alfonso SV, Hernandez N, Fernández MI. Examining the factors associated with past and present dermoscopy use among family physicians. Dermatol Pract Concept 2017;7(4):63-70 [FREE Full text] [doi: 10.5826/dpc.0704a13] [Medline: 29214111]
16. Fee JA, McGrady FP, Rosendahl C, Hart ND. Training primary care physicians in dermoscopy for skin cancer detection: a scoping review. J Cancer Educ 2020;35(4):643-650 [FREE Full text] [doi: 10.1007/s13187-019-01647-7] [Medline: 31792723]
17. Robinson JK, MacLean M, Reavy R, Turrisi R, Mallett K, Martin GJ. Dermoscopy of concerning pigmented lesions and primary care providers' referrals at intervals after randomized trial of mastery learning. J Gen Intern Med 2018;33(6):799-800 [FREE Full text] [doi: 10.1007/s11606-018-4419-5] [Medline: 29637481]
18. Brennan MC, Kabuli MN, Dargan MD, Pinder MR. A short correspondence piece to the editor in chief: the need for increased training in the technique of dermoscopy amongst plastic surgeons and the under recognised value of dermoscopy in the assessment of non-pigmented cutaneous lesions. J Plast Reconstr Aesthet Surg 2022;75(1):496-498. [doi: 10.1016/j.bjps.2021.11.014] [Medline: 34852970]
19. Paterlini M. There shall be order. The legacy of Linnaeus in the age of molecular biology. EMBO Rep 2007;8(9):814-816 [FREE Full text] [doi: 10.1038/sj.embor.7401061] [Medline: 17767191]
20. Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a reference terminology for ontology research and development in the biomedical domain. 2006 Presented at: Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2006) ;222?65; November 8, 2006; Baltimore, MA p. 57.
21. Papier A, Chalmers RJG, Byrnes JA, Goldsmith LA, Dermatology Lexicon Project. Framework for improved communication: the Dermatology Lexicon Project. J Am Acad Dermatol 2004;50(4):630-634. [doi: 10.1016/s0190-9622(03)01571-8] [Medline: 15034516]
22. Dermatology lexicon. NCBO BioPortal. 2009. URL: https://bioportal.bioontology.org/ontologies/DERMLEX/?p=summary [accessed 2024-01-05]
23. Fisher HM, Hoehndorf R, Bazelato BS, Dadras SS, King LE, Gkoutos GV, et al. DermO; an ontology for the description of dermatologic disease. J Biomed Semantics 2016;7:38 [FREE Full text] [doi: 10.1186/s13326-016-0085-x] [Medline: 27296450]
24. Human dermatological disease ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/DERMO [accessed 2024-01-05]
25. Skin physiology ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/SPO [accessed 2024-01-05]
26. Benvenuto-Andrade C, Dusza SW, Agero ALC, Scope A, Rajadhyaksha M, Halpern AC, et al. Differences between polarized light dermoscopy and immersion contact dermoscopy for the evaluation of skin lesions. Arch Dermatol 2007;143(3):329-338. [doi: 10.1001/archderm.143.3.329] [Medline: 17372097]
27. Lin R, Amith M, Zhang X, Wang C, Light J, Strickley J, et al. Developing ontologies to standardize descriptions of visual and dermoscopic elements. 2021 Presented at: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 09-12, 2021; Houston, TX p. 1813. [doi: 10.1109/bibm52615.2021.9669477]
28. Ashton R, Leppard B. Differential Diagnosis in Dermatology. In: CRC Press. London: CRC Press; 2014:978-1001.

29. Noy NF, McGuinness DL. Ontology development 101: A guide to creating your first ontology. Corais. 2021. URL: https://corais.org/sites/default/files/ontology_development_101_aguide_to_creating_your_first_ontology.pdf [accessed 2024-05-23]
30. Egenhofer M, Herring J. A mathematical framework for the definition of topological relations. 1990 Presented at: Proceedings of the Fourth International Symposium on Spatial Data Handing; September 1990; California p. 803.
31. Ceusters W. An information artifact ontology perspective on data collections and associated representational artifacts. In: Studies in Health Technology and Informatics. Amsterdam, the Netherlands: IOS Press; 2012:68-72.
32. Malone J, Brown A, Lister AL, Ison J, Hull D, Parkinson H, et al. The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation. J Biomed Semantics 2014;5:25 [FREE Full text] [doi: 10.1186/2041-1480-5-25] [Medline: 25068035]
33. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Commun ACM 2016;59(2):44-51. [doi: 10.1145/2844544]
34. Kittler H, Marghoob AA, Argenziano G, Carrera C, Curiel-Lewandrowski C, Hofmann-Wellenhof R, et al. Standardization of terminology in dermoscopy/dermatoscopy: results of the third consensus conference of the International Society of Dermoscopy. J Am Acad Dermatol 2016;74(6):1093-1106 [FREE Full text] [doi: 10.1016/j.jaad.2015.12.038] [Medline: 26896294]
35. Yélamos O, Braun RP, Liopyris K, Wolner ZJ, Kerl K, Gerami P, et al. Dermoscopy and dermatopathology correlates of cutaneous neoplasms. J Am Acad Dermatol 2019;80(2):341-363 [FREE Full text] [doi: 10.1016/j.jaad.2018.07.073] [Medline: 30321581]
36. Draghici C, Vajaitu C, Solomon I, Voiculescu VM, Popa MI, Lupu M. The dermoscopic rainbow pattern—a review of the literature. Acta Dermatovenerol Croat 2019;27(2):111-115. [Medline: 31351506]
37. Musen MA, Protégé Team. The Protégé project: a look back and a look forward. AI Matters 2015;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
38. Burton-Jones A, Storey VC, Sugumaran V, Ahluwalia P. A semiotic metrics suite for assessing the quality of ontologies. Data Knowl Eng 2005;55(1):84-102. [doi: 10.1016/j.datak.2004.11.010]
39. Amith M, Manion F, Liang C, Harris M, Wang D, He Y, et al. Architecture and usability of OntoKeeper, an ontology evaluation tool. BMC Med Inform Decis Mak 2019;19(Suppl 4):152 [FREE Full text] [doi: 10.1186/s12911-019-0859-z] [Medline: 31391056]
40. Obrst L. Ontologies for semantically interoperable systems. 2003 Presented at: Proceedings of the Twelfth International Conference on Information and Knowledge Management; November 3, 2003; New York, NY p. 366-369. [doi: 10.1145/956863.956932]
41. Mullin S, Zola J, Lee R, Hu J, MacKenzie B, Brickman A, et al. Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. J Biomed Inform 2021;122:103889. [doi: 10.1016/j.jbi.2021.103889] [Medline: 34411708]
42. Sahoo SS, Kobow K, Zhang J, Buchhalter J, Dayyani M, Upadhyaya DP, et al. Ontology-based feature engineering in machine learning workflows for heterogeneous epilepsy patient records. Sci Rep 2022;12(1):19430 [FREE Full text] [doi: 10.1038/s41598-022-23101-3] [Medline: 36371527]
43. Zemmouchi-Ghomari L. Ontology and machine learning: a two-way street to improved knowledge representation and algorithm accuracy. : Springer Nature; 2023 Presented at: Proceedings of International Conference on Paradigms of Communication, Computing and Data Analytics; October 11, 2023; Singapore p. 181-189. [doi: 10.1007/978-981-99-4626-6_15]
44. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. Br J Dermatol 2009;161(3):591-604. [doi: 10.1111/j.1365-2133.2009.09093.x] [Medline: 19302072]
45. Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A, International Skin Imaging Collaboration. Results of the 2016 International Skin Imaging Collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78(2):270-277. [doi: 10.1016/j.jaad.2017.08.016] [Medline: 28969863]
46. Pham TC, Luong CM, Hoang VD, Doucet A. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. Sci Rep 2021;11(1):17485 [FREE Full text] [doi: 10.1038/s41598-021-96707-8] [Medline: 34471174]
47. Tschandl P, Kittler H, Argenziano G. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. Br J Dermatol 2017;177(3):867-869. [doi: 10.1111/bjd.15695] [Medline: 28569993]
48. Shetty B, Fernandes R, Rodrigues AP, Chengoden R, Bhattacharya S, Lakshmanna K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Sci Rep 2022;12(1):18134. [doi: 10.1038/s41598-022-22644-9] [Medline: 36307467]

Abbreviations

CNN: convolutional neural network
D3X: dermoscopy differential diagnosis explorer
DERMLEX: dermatology lexicon
DERMO: human dermatological disease ontology
DEVO: dermoscopy elements of visuals ontology
EVO: elements of visuals ontology
IAO: information artifact ontology
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
LOINC: logical observation identifier names and codes
MedDRA: Medical Dictionary for Regulatory Activities
NCBO: National Center for Biomedical Ontology
NCI: National Cancer Institute
OBO: Open Biological and Biomedical Ontology
OWL: web ontology language
PCP: primary care physician
rdf: resource description framework
RDF: resource description framework
SHACL: shapes constraint language
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
SPO: skin physiology ontology
SWO: software ontology

Edited by C Lovis, G Eysenbach; submitted 12.07.23; peer-reviewed by C Gaudet-Blavignac; comments to author 09.12.23; revised version received 18.04.24; accepted 04.05.24; published 21.06.24.

Please cite as:
Lin RZ, Amith MT, Wang CX, Strickley J, Tao C
Dermoscopy Differential Diagnosis Explorer (D3X) Ontology to Aggregate and Link Dermoscopic Patterns to Differential Diagnoses: Development and Usability Study
JMIR Med Inform 2024;12:e49613
URL: https://medinform.jmir.org/2024/1/e49613
doi:10.2196/49613
PMID:38904996

Original Paper

An Ontology-Based Decision Support System for Tailored Clinical Nutrition Recommendations for Patients With Chronic Obstructive Pulmonary Disease: Development and Acceptability Study

Daniele Spoladore[1,2*], DPhil; Vera Colombo[1*], DPhil; Alessia Fumagalli[3*], MD, DPhil; Martina Tosi[4,5*], RD; Erna Cecilia Lorenzini[4,6*], MD; Marco Sacco[1*]

[1]Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, National Research Council of Italy, Lecco, Italy

[2]Department of Pure and Applied Sciences, Computer Science Division, Insubria University, Varese, Italy

[3]Unit of Pulmonary Rehabilitation, IRCCS, Italian National Research Center on Aging, Casatenovo, Italy

[4]Institute of Agricultural Biology and Biotechnology, National Research Council of Italy, Milan, Italy

[5]Department of Health Science, University of Milan, Milan, Italy

[6]Department of Biomedical Sciences for Health, Chair of Clinical Pathology, University of Milan, Milan, Italy

[*]all authors contributed equally

Corresponding Author:
Daniele Spoladore, DPhil
Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing
National Research Council of Italy
Via G. Previati 1E
Lecco, 23900
Italy
Phone: 39 03412350202
Email: daniele.spoladore@stiima.cnr.it

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is a chronic condition among the main causes of morbidity and mortality worldwide, representing a burden on health care systems. Scientific literature highlights that nutrition is pivotal in respiratory inflammatory processes connected to COPD, including exacerbations. Patients with COPD have an increased risk of developing nutrition-related comorbidities, such as diabetes, cardiovascular diseases, and malnutrition. Moreover, these patients often manifest sarcopenia and cachexia. Therefore, an adequate nutritional assessment and therapy are essential to help individuals with COPD in managing the progress of the disease. However, the role of nutrition in pulmonary rehabilitation (PR) programs is often underestimated due to a lack of resources and dedicated services, mostly because pneumologists may lack the specialized training for such a discipline.

Objective: This work proposes a novel knowledge-based decision support system to support pneumologists in considering nutritional aspects in PR. The system provides clinicians with patient-tailored dietary recommendations leveraging expert knowledge.

Methods: The expert knowledge—acquired from experts and clinical literature—was formalized in domain ontologies and rules, which were developed leveraging the support of Italian clinicians with expertise in the rehabilitation of patients with COPD. Thus, by following an agile ontology engineering methodology, the relevant formal ontologies were developed to act as a backbone for an application targeted at pneumologists. The recommendations provided by the decision support system were validated by a group of nutrition experts, whereas the acceptability of such an application in the context of PR was evaluated by pneumologists.

Results: A total of 7 dieticians (mean age 46.60, SD 13.35 years) were interviewed to assess their level of agreement with the decision support system's recommendations by evaluating 5 patients' health conditions. The preliminary results indicate that the system performed more than adequately (with an overall average score of 4.23, SD 0.52 out of 5 points), providing meaningful and safe recommendations in compliance with clinical practice. With regard to the acceptability of the system by lung specialists (mean age 44.71, SD 11.94 years), the usefulness and relevance of the proposed solution were extremely positive—the scores on each of the perceived usefulness subscales of the technology acceptance model 3 were 4.86 (SD 0.38) out of 5 points, whereas the score on the intention to use subscale was 4.14 (SD 0.38) out of 5 points.

Conclusions: Although designed for the Italian clinical context, the proposed system can be adapted for any other national clinical context by modifying the domain ontologies, thus providing a multidisciplinary approach to the management of patients with COPD.

(JMIR Med Inform 2024;12:e50980) doi:10.2196/50980

KEYWORDS

ontology-based decision support system; nutritional recommendation; chronic obstructive pulmonary disease; clinical decision support system; pulmonary rehabilitation

Introduction

Background

Chronic obstructive pulmonary disease (COPD) is one of the leading causes of morbidity and mortality worldwide. In Italy, 3.2% of the population in 2019 had a diagnosis of COPD, and such numbers are expected to increase in the next years due to the worsening of risk factors [1]. COPD is characterized by chronic inflammation in the lungs and airflow obstruction. Starting from the respiratory system, COPD has several systemic effects, including skeletal muscle wasting (which limits exercise capacity), increased risk of cardiovascular disease, osteoporosis, depression, and anxiety [2]. Pulmonary rehabilitation (PR) is an evidence-based, nonpharmacological intervention that helps people with COPD improve their health condition and quality of life by increasing exercise capacity and reducing symptoms of dyspnea and fatigue [3]. PR is based on a multidisciplinary approach and includes several components, among which nutritional status evaluation is considered relevant.

However, as highlighted recently by the American Thoracic Society, the scarcity of resources often limits the nutritional component, which, in some settings, can be accessible only outside the PR program [4]. Moreover, even though the influence of nutrition on COPD has been recognized and a relationship between some nutrients and pulmonary functions has been identified, PR programs often do not formally include patient-specific dietary recommendations. Therefore, more studies are required to drive the implementation of tailored nutritional strategies [5].

The Role of Nutrition in Patients With COPD: Evidence

The relationship between COPD and nutrition has been investigated by many authors, who have put together a set of evidence that underlines the role of nutrients in PR and the disease exacerbations. The results of the systematic review and meta-analysis by Collins et al [6] demonstrated that nutritional support, primarily in the form of oral nutritional supplements, improves total energy intake, anthropometric measurements, and grip strength in patients with COPD. This work highlighted how dietary changes supported protein and energy intake in patients with COPD, resulting in weight gain and a moderate increase in muscle strength—which can significantly impact respiratory muscle strength, walking distance, and quality of life. Another systematic review [7] found that nutritional supplementation improves body weight gain in patients with COPD, especially when malnourished, in terms of body composition, respiratory muscle strength, and quality of life.

Inadequate nutritional intake, higher daily energy expenditure, weight loss, and fat-free mass (FFM) depletion affect the functional capacity of patients with COPD. The progression of the disease also contributes to an increased respiratory muscle load, hypoxemia, physical inactivity, and release of inflammation mediators [8]. In addition, Liu et al [9] explored the associations among COPD, diet, and inflammation, identifying a correlation between diet and COPD inflammatory status. By calculating the Diet Inflammatory Index in COPD, they observed that COPD was more prevalent in patients with a worse Diet Inflammatory Index.

A systematic review of factors influencing the risk of developing COPD [10] concluded that the Western diet, characterized by the consumption of processed meat and alcohol and a low intake of fruit and vegetables, is associated with the prevalence of COPD. A cross-sectional study [11] assessed the nutritional status of outpatients with COPD and found a significant association between malnutrition and COPD—most of the patients did not have an adequate daily food intake and showed a lower quality of life, highlighting the need for a more tailored nutritional intervention for managing malnutrition in patients with COPD with cachexia. In COPD, low FFM and sarcopenia are predictive factors of mortality [12]. Considering that literature data demonstrate that the presence of malnutrition and, above all, a low FFM index is associated with an increased risk of mortality, Schols et al [13] identified different nutritional risk profiles based on nutritional phenotypes that appear to be predictors of outcome independent of impaired respiratory function. These nutritional phenotypes included obesity, sarcopenic obesity, sarcopenia, and cachexia and required at least three items for their identification: (1) BMI and body circumference, (2) bioelectrical impedance analysis, and (3) history of unintentional weight loss. A recent study [14] demonstrated that patients with COPD with sarcopenia and sarcopenic obesity had worse muscle strength than patients with healthy body weight, claiming that body composition is associated with physical functions. Patients with COPD with obesity have a higher risk of developing comorbidities such as diabetes and metabolic and cardiovascular diseases in addition to typical COPD symptoms.

Research Challenge: Nutritional Status to Prevent COPD Exacerbation

The evidence reported previously indicates that clinicians and dietitians must evaluate the nutritional status and body composition of patients when giving nutritional advice or diet recommendations to patients with COPD with obesity to foster an adequate intake of macro- and micronutrients and control obesity as well [15].

Malnutrition plays a pivotal role among the challenges that must be faced in treating COPD. The study by Mete et al [16] affirms that malnutrition and risk of malnutrition are very frequent conditions among patients with COPD and, together with a significantly lower BMI, are associated with disease severity, as highlighted by pulmonary function tests. In clinical practice, sarcopenia assessment is not part of the standard of care, implying a loss of personalized treatment that should be essential to improve health outcomes related to the disease [17]. The current literature illustrates that screening for sarcopenia and tailored nutrition intervention in patients with COPD could cost-effectively achieve better health outcomes [18,19]. Functional capacity, respiratory muscle strength, and quality of life can all be improved through nutritional supplementation and a better assessment of nutritional status.

The future nutrition challenges require the identification of specific targets of intervention, considering body composition, nutritional status, and inflammation in addition to the strictly clinical aspects. Nowadays, diet is not always an integral part of COPD therapeutic strategy [20]. Moreover, different studies have highlighted the necessity of new methods to assess malnutrition and evaluate nutritional status in patients with COPD, not considering the BMI alone [21] as it misses important changes in body composition.

This new vision places nutritional intervention as an integral part of COPD management, not only in the advanced and early stages but also to prevent the evolution toward respiratory failure. Creating a decision support system (DSS) that brings together data and knowledge in the nutritional field to identify the different nutritional phenotypes and suggest specific dietary recommendations could be useful to meet these challenges. In particular, this DSS could become a valuable tool for pulmonologists to develop nutritional interventions as an integral part of the COPD therapeutic strategy even in the absence of specific resources.

Objective of This Study

Digital health care applications are already used to improve PR models, mainly for self-management; modification of lifestyle factors; and modification of risk factors, such as smoking cessation and fostering physical activity [22]. However, to the best of our knowledge, there are only a few applications focused on nutritional aspects associated with COPD. This work leveraged clinical expert knowledge—formalized into domain ontologies—from the Italian health care context to develop a DSS to support pneumologists in considering nutritional aspects in PR to avoid the disease's exacerbation and provide patients with tailored dietary guidelines. The application exploiting the DSS fits the context of a preprototype according to the World Health Organization (WHO) guidelines for the monitoring and evaluation of digital health interventions [23].

Related Work

Overview

Ontology-based technologies have been adopted for both clinical DSSs and patient-centered DSSs. Regarding the ontological formalization of requisites and tailored suggestions for patients with COPD in DSSs, no work has tackled this issue in PR—except for an early version of the system we propose in this work [24]. Nonetheless, the use of ontologies to formalize COPD has been established in some works. Moreover, specific digital applications for patients with COPD have been developed in the past years. This section presents works relevant to the fields of ontology-based clinical and patient-centered DSSs and digital applications specifically developed for patients with COPD.

Ontologies and Ontology-Based Systems for COPD Management

COPD has been formalized in ontologies or knowledge bases since the early 2010s. The COPD ontology by Greenberg et al [25] was developed to support COPD longitudinal research and clinical trials, leveraging a preexisting model dedicated to representing subpopulations of patients. Cano et al [26] developed a knowledge base devoted to collecting clinical experimental data; the COPD Knowledge Base can semantically map data to physiological and molecular data to support clinical decision-making through a predictive mathematical model.

Ontologies can play a pivotal role in diagnostic systems. In the case of COPD, Rayner et al [27] developed an ontology for the early diagnosis of the illness. The ontological model takes advantage of clinical tests and patients' data (eg, spirometry, forced expiratory volume in the first second [FEV1], age, and smoking status) and tests its robustness against a large data set of patients. The model proved able to categorize patients as "Unlikely COPD," "Probable COPD," and "Definite COPD" cases.

As ontology-enabled reasoning processes are appreciated in the context of clinical decision-making, the adoption of formal models in health care systems aimed at managing chronic conditions—including systems for patients with COPD—has also been established. CHRONIOUS [28] is an open and ubiquitous system that exploits ontology-based inferential reasoning to adapt the platform's services, including monitoring patients' conditions—which is also performed leveraging wearable sensors. Lasierra et al [29] developed an ontology-based telemonitoring system aimed at monitoring patients with chronic diseases at home. The ontology layer (Home Ontology for Integrated Management in Home-Based Scenarios) represents the patient profile, vital signs, and chronic conditions (including COPD) to observe the patient's evolution and plan activities. Similarly, Ajami and McHeick [30] developed a domain ontology encompassing environmental features, patient data, clinical status and diseases (including COPD), and devices to monitor and identify patients' conditions. This DSS aimed to foster patients' adherence to a healthy lifestyle and identify and avoid possibly dangerous situations.

Digital Applications for Patients With COPD

Digital applications may offer solutions to both clinicians—to ease the process of assessment and support clinical decisions—and patients—mainly to support them in the management of the disease. Digital applications currently available for patients with COPD are mainly telemedicine and telerehabilitation solutions. Most of them are focused on educational programs, symptom tracking, behavior change,

support for medication or treatment, and activity report [31]. In some cases, as both research prototypes [32] and commercial products [33], the educational programs include specific sections on nutrition, in which helpful tips and recommendations are provided to manage symptoms during meal consumption and help maintain a balanced diet. However, such applications do not provide personalized information tailored to the patient's nutritional status. In such cases, the applications targeted at health care professionals comprise the "clinician side" of telemedicine platforms (ie, monitoring dashboard allowing for remote visualization of patients' data and applications for remote teleconsulting).

Differently from the aforementioned solutions, the DSS proposed in this work tackles the role of nutritional therapy in the PR of patients with COPD—an aspect neglected in existing DSSs.

Methods

A DSS was developed to tackle the aims described in the Introduction section and the research challenge described in the Objective of This Study section, thus supporting pneumologists in suggesting specific dietary recommendations for patients with COPD. The DSS leverages expert knowledge in the form of ontologies and, through automatic inference processes, is expected to provide guidelines to prepare a tailored dietary plan.

The development of the ontology underlying the DSS leveraged expert clinical knowledge to maximize its acceptability. Contrary to purely data-driven approaches, ontologies formalize information to enable a system to perform inferences (based on the knowledge formalized in the ontology). The inference process simulates human inference capabilities [34] so that ontology-based approaches are perceived as transparent. Thus, ontologies are widely adopted in several artificial intelligence and health-related applications [35].

However, the development of a domain ontology is not a trivial task—it is a process that may involve several activities (eg, the acquisition of knowledge, its conceptualization, the survey of existing models that can be reused, the development of the model in a formal language, and the testing of the developed ontology [36]). The ontology engineered for the proposed DSS was developed following the Agile, Simplified, and Collaborative Ontology Engineering Methodology (AgiSCOnt) [37] engineering methodology, which involves knowledge elicitation techniques and domain experts in the development phase of the ontology. This collaborative ontology engineering methodology adopts unstructured interviews, scientific literature surveys, and discussions to elicit the necessary knowledge to minimize the impact of the "knowledge elicitation bottleneck" (ie, it takes longer to gather knowledge from experts and documentation than to write the software [38]). Its collaborative and agile features and validation [37] were the reasons behind the adoption of AgiSCOnt in this work.

The methodology involves three phases:

1. *Domain analysis and conceptualization*, which includes the identification of the knowledge to be included and the preparation of competency questions (CQs) [39] that the ontology is expected to answer; it enables the conceptualization of the domain (which results in a conceptual map) and the preliminary identification of existing ontological resources that can be reused.
2. *Development and testing*, which involves the selection of the ontological languages to formalize the conceptualization developed in the previous step and the identification of ontology design patterns (ie, "micro-ontologies" that can be reused to model recurrent problems in ontology engineering [40]). This step results in the prototypization of the ontology, which undergoes a preliminary test to assess the validity of its inferences.
3. *Ontology use and updating*, which includes activities such as use of the developed ontology in an application, extended validation, and feedback collection. The following subsections delve into the engineering process.

The development of this ontology involved the following team members: 1 ontologist with experience in modeling using agile ontology engineering methodologies, 1 biomedical engineer with previous experience in ontology engineering and knowledge of COPD, 2 senior dieticians with clinical experience with patients with COPD, and 1 senior pneumologist. The team was composed of clinical personnel from universities (dieticians) and a research and cure center (Scientific Institute for Research, Hospitalisation and Health Care) with a specialization in COPD (pneumologist) and yearly experience treating such patients. The team delved into nutrition and diet's role in tackling this disease, with examples from the literature and clinical trials in which the clinical personnel was involved. The discussion was then oriented to identify some of the issues that the ontology was expected to answer (ie, the CQs).

Ethical Considerations

This study does not include human subjects research (no human subjects experimentation or intervention was conducted) and so does not require institutional review board approval.

Results

This section describes the development of the COPD and Nutrition domain ontology for the DSS and the ontology-based application for clinical personnel.

The COPD and Nutrition Domain Ontology

Domain Analysis and Conceptualization

The considerations reported in The Role of Nutrition in Patients With COPD: Evidence and Research Challenge: Nutritional Status to Prevent COPD Exacerbation sections were gathered by the team in this phase. Leveraging the objective (ie, providing clinical personnel with support in identifying tailored nutritional recommendations for patients with COPD), the team decided that the ontology should focus on representing the patients' health condition and the stage of their COPD. On the basis of their expertise, clinicians pointed out that the purpose of the ontology should be to illustrate, for each patient, a tailored percentage of macro- and micronutrients they are advised to consume on a daily basis to avoid exacerbations, as well as to provide nutritional guidance. The summary of the entities and

expected outputs of the ontology is provided in Textbox 1, listing the CQs and their answers.

Textbox 1. The list of competency questions and answers for the Chronic Obstructive Pulmonary Disease (COPD) and Nutrition ontology engineering process.

What information identifies a patient? What basic information is used to identify the patient? What clinical information is used to identify the patient?

- A patient is identified by an ID and gender. Each patient is associated with 1 health condition and 1 anthropometric phenotype (defined via BMI cutoffs). Each patient can be classified as a patient with sarcopenia or cachexia or as a patient without sarcopenia or cachexia.

How is COPD evaluated?

- COPD is evaluated according to the criteria defined in the gold standard—it can be mild, moderate, severe, and very severe. The criterion to be analyzed is the forced expiratory volume in the first second.

How is sarcopenia evaluated?

- The status of sarcopenia is evaluated according to clinical standards (operational definition of sarcopenia). The first criterion comprises low muscle strength, the second criterion comprises a low muscle quantity or quality, and the third criterion comprises low physical performance. The presence of the first criterion alone indicates probable sarcopenia, the presence of both the first and second criteria indicates diagnosed sarcopenia, and the presence of all 3 criteria indicates severe sarcopenia.

How is cachexia evaluated?

- Cachexia is evaluated by means of biochemical indicators according to the study by Evans et al [41]. Albuminemia, iron transport, and polymerase chain reaction (PCR) criteria must be copresent to indicate a cachexia diagnosis.

Which data characterize the patient's health condition? How is the nutritional risk index assessed?

- Patients' health condition must indicate the stage of COPD and the nutritional risk index profile characterizing the patient. Each health condition must illustrate anthropometric measures (current weight, usual weight, height in meters, and BMI), physical performance indicators (hand grip and gait speed), and biochemical indicators (albuminemia, PCR, resistance, reactance, and iron transport). The nutritional risk index assessment is performed following clinical standards.

What recommendations are given to clinical personnel?

- The recommendations provided to clinical personnel indicate (for each patient) the basal metabolic rate and the corrected caloric intake, the daily macronutrient shares (protein, minimum and maximum share of carbohydrates, minimum and maximum share of fats, minimum and maximum share of fiber, maximum share of sugar, and maximum share of saturated fats), the amount of cholesterol and sodium, and whether the patient should increase their caloric intake by means of branched-chain amino acid or energy-protein supplementations.

How are recommendation values calculated?

- The indications provided in the patient's recommendation are calculated according to clinical standards and differentiated according to the patient's gender, stage of COPD, and anthropometric phenotype.

In this phase, the pneumologist specialized in clinical nutrition and a clinical dietitian defined the phenotypes and their nutritional requirements based on anthropometric and clinical parameters and comorbidities according to national and international guidelines. For the definition of each anthropometric phenotype, COPD stage (defined via FEV1 and forced vital capacity and in particular FEV1–to–forced vital capacity ratio [42,43]; Table 1) and the presence of sarcopenia or cachexia were considered. A total of 5 metabolic phenotypes for patients with COPD were developed (underweight, normal weight, overweight, first-degree obesity, and second-degree obesity), and each can be characterized by the presence of sarcopenia, cachexia, or none of the 2 conditions (Table 2).

Moreover, nutritional risk was assessed by means of the nutritional risk index (NRI) formula [44] as follows:

(1)

Thus, patients were classified according to this formula (Table 3).

The diagnosis of sarcopenia was based on the analysis of specific patients' value indicators. The first one was the appendicular skeletal muscular mass, which is calculated according to the work by Sergi et al [45]:

(2)

Appendicular skeletal muscular mass and other indicators allow for the classification of patients' sarcopenic condition according to the 3 criteria in Table 4 [46].

If the first criterion applied, then the patient was *probable* sarcopenic; if the first and second criteria applied, then the patient was *diagnosed* sarcopenic; if all 3 criteria applied, then the patient was *severe* sarcopenic.

A similar approach was adopted to identify whether a patient was cachectic. If a patient's polymerase chain reaction was >10, the level of iron transport was <150, the level of albuminemia was <3.5, and the patient was sarcopenic, then they were also cachectic [41]. As such, it is safe to infer that cachexia is a particular case of sarcopenia.

Calculation of nutritional recommendations was performed using the "if-then" type rules produced by clinical personnel considering reference values reported in scientific literature and national and international guidelines. The COPD DSS provides the following nutritional information and recommendations: (1) basal metabolic rate (BMR), (2) total daily energy requirement (kcal), (3) meal frequency, (4) indication for energy-protein supplementation (yes or no), (5) indication for branched-chain amino acid (BCAA) supplementation (yes or no), (6) protein intake (percentage and grams), (7) carbohydrate intake (percentage and grams), (8) lipid intake (percentage and grams), (9) sugar intake (maximum percentage and grams), (10) saturated fat intake (maximum percentage and grams), (11) cholesterol intake (maximum milligrams), (12) fiber intake (minimum and maximum grams), (13) sodium intake (maximum milligrams), and (14) calcium intake (milligrams).

For BMR calculation, Harris-Benedict or Mifflin predictive equations based on sex, age, weight, and height were considered. The Harris-Benedict equation was preferred for patients who were underweight or had normal weight, whereas the Mifflin equation was used for patients with COPD with overweight or obesity [47-49]. According to the COPD stage, different correction factors to BMR were applied (Multimedia Appendix 1)—for patients who were underweight, had a normal weight, were overweight, or had first- or second-class obesity presenting with COPD at the first or second stage and who were nonsarcopenic and noncachectic or sarcopenic, a 1.5 correction factor was used to calculate total daily energy requirement; for the same categories of patients presenting with COPD at the third or fourth stage, a correction factor of 1.8 was preferred to counteract the important energy expenditure due to the respiratory work of these patients. Only for patients with cachexia a 1.8 correction factor was always applied regardless of BMI and COPD stage.

The impact of physical activity was deemed marginal for the correction factor's definition as patients with COPD are a vulnerable population presenting with comorbidities that limit their capacity to perform physical activity. As seen in the pathological lung mechanics of patients with COPD, dynamic hyperinflation influences the proper operation of the chest's horizontal and vertical diameters that expand during inspiration due to the activation of the external intercostal muscles and the diaphragmatic contraction. This impairment plays a role in how much exercise a patient with COPD can tolerate. As shown in individuals with severe-stage COPD and weight loss related to COPD, respiratory muscle weakness exacerbates the breathing mechanism. Increased dyspnea and decreased exercise tolerance are directly related to this respiratory muscle weakening. Moreover, it has been observed that patients with COPD are characterized by higher levels of physical inactivity [50]. As far as protein requirement was concerned, clinical experts decided to provide different recommendations according to real or ideal weight or BMI or focusing on FFM considering specific body composition or COPD stage to give personalized recommendations to not compromise health and nutritional status as well as prevent further decrease in metabolically active lean mass [12]. For this reason, cachectic phenotype, regardless of BMI and COPD stage, was always given a high percentage of protein intake. BCAA was suggested when energy requirements but not protein requirements were met through diet. Differently, protein-caloric supplements were indicated when neither energy nor protein requirements were satisfied through food intake [51]. Phenotypes characterized by cachexia, regardless of BMI, underweight sarcopenic and normal-weight phenotype, and COPD stage, were always recommended BCAA and energy-protein supplementation in consideration of their health status. For the same reason, BCAA supplementation was always suggested for sarcopenic phenotypes [52]. As far as carbohydrate metabolism was concerned, and in line with the lower levels reported in the reference values of nutrients and energy for the Italian population [13,47], a carbohydrate intake of 45% to 50% of total daily calories and a maximum sugar intake of 15% of total calories were indicated except for individuals with type 2 diabetes mellitus, for whom a maximum of 10% of total calories was indicated to be reached through sugar intake. A lower recommendation for carbohydrate intake permits the promotion of protein and lipid intake, which are functional for respiratory work and to prevent further weight or lean mass loss. Regarding dietary fiber, it was decided that providing a lower indication than that provided by the Livelli di Assunzione di Riferimento di Nutrienti ed energia (LARN; National Recommended Energy and Nutrient Intake Levels) was preferred to encourage the intake of energy and protein foods, especially considering the difficulties met by patients with COPD in feeding and the early sense of satiety as a result of high-fiber food intake. Compared to the LARN, a higher intake of lipids was recommended, especially for phenotypes presenting with a partial pressure of carbon dioxide of >50 mm Hg, a measure of carbon dioxide in arterial or venous blood [53]. For saturated fats, a maximum intake of 10% of total daily calories was suggested according to LARN guidelines. Regarding cholesterol intake, a LARN nutritional goal for prevention, a maximum of 300 mg per day was recommended except for patients with high cholesterol levels, for whom the target was lowered to 200 mg per day. Hypercholesterolemia was defined as elevated total or low-density lipoprotein cholesterol levels or low levels of high-density lipoprotein cholesterol. Regarding micronutrients, sodium and calcium intake was considered. According to the recent WHO report on sodium intake [54], a maximum intake of <2000 mg per day of sodium (<5 g per day of salt) was recommended. A slightly higher recommendation than that of the LARN was given for calcium to prevent or counteract osteoporosis [55]. Finally, given the difficulties in feeding and early satiety observed, an indication for fractioned meals was given to meet energy and nutritional requirements throughout the day with small and frequent meals.

All the parameters involved in the evaluations presented previously (ie, FEV1; partial pressure of carbon dioxide; resistance; reactance; iron transport; albuminemia; polymerase chain reaction; hand grip; gait speed; and general patient

information such as age, gender, height in meters, current weight, and usual weight) are usually acquired during patient assessment (such as spirometry and blood tests). AgiSCOnt's outputs for the domain analysis phase consisted of the conceptual map reported in Figure 1 and a list of CQs (Textbox 1).

Table 1. The cutoffs identifying the chronic obstructive pulmonary disease (COPD) stages based on the forced expiratory volume in the first second (FEV1) values.

COPD stage number	COPD stage name	FEV1 (%)
I	Mild	≥80
II	Moderate	≥50 to <80
III	Severe	≥30 to <50
IV	Very severe	<30

Table 2. World Health Organization cutoffs for nutritional status categories based on BMI.

Nutritional status	BMI (kg/m^2)
Underweight	<18.5
Healthy weight	≥18.5 to ≤24.9
Overweight (preobesity)	≥25.0 to <29.9
Obesity degree I	≥30.0 to <34.9
Obesity degree II	≥35.0 to <39.9
Obesity degree III	≥40

Table 3. The cutoffs identifying the 4 levels of nutritional risk based on the nutritional risk index (NRI).

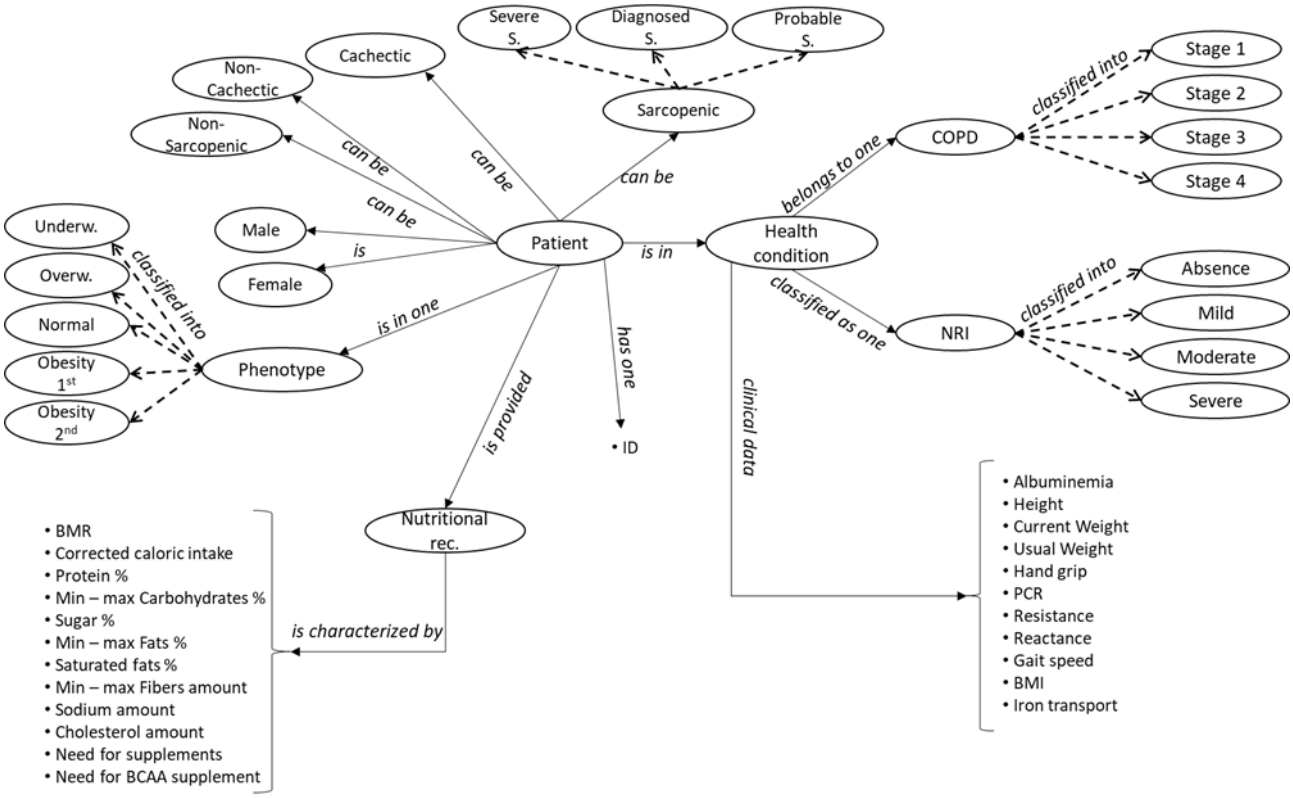
NRI	Index
Absence of risk	>100
Mild risk	≥97.5 to ≤100
Moderate risk	≥83.5 to <97.5
Severe risk	<83.5

Table 4. Criteria for the classification of sarcopenia in patients.

Criteria	Cutoff for male patients	Cutoff for female patients
Low muscular strength—hand grip	<27 kg	<16 kg
Low muscular quantity	ASMM[a]<20 kg; ASMM/height^2<7 kg/m^2	ASMM<15 kg; ASMM/height^2<5.5 kg/m^2
Poor physical performance	≤0.8 m/s	≤0.8 m/s

[a]ASMM: appendicular skeletal muscular mass.

Figure 1. An excerpt of the conceptual map developed by the team involved in the ontology engineering process for the clinical chronic obstructive pulmonary disease (COPD) and nutrition ontology. BCAA: branched-chain amino acid; BMR: basal metabolic rate; NRI: nutritional risk index; PCR: polymerase chain reaction.



Development and Testing

The development phase adopted the conceptual map and CQs produced in the previous phase to guide the entire development process and discuss whether to model some concepts pertaining to the patients or their health conditions. The ontology editor Protégé (Stanford Center for Biomedical Informatics Research) [56] that supports Resource Description Framework [57] and Web Ontology Language [58] with a DL (description logic) profile [59] was adopted. Clinicians explicitly asked any significant advancement in the development of TBox and Abox (eg, patient modeling, health condition characterization, and recommendation modeling) to be illustrated to ensure that undesired entailments were not modeled in the ontology. From a reuse perspective, no ontology able to describe the conceptualization reached in this study was found; the only ontology design pattern reused in this ontology was the one that relates a copd:Patient to their copd:Health_Condition via the copd:isInHealthCondition object property [60]. The developed ontology (prefixed with copd:), discussed in this subsection, is accessible in Multimedia Appendix 2.

The development started with the identification of concepts that could be translated into owl:Classes. The concept of copd:Patient is pivotal in this ontology. Each patient is defined by exactly 1 patient ID, is given at least 1 nutritional recommendation individual, and is characterized by a health condition individual.

Each patient needs to be classified as copd:Female or copd:Male—disjoint classes—and as copd:Cachectic (or its complement copd:non-Cachectic) or copd:Sarcopenic (or its complement copd:non-Sarcopenic). Sarcopenia and cachexia are modeled as attributes of the patient and not of their health condition. The clinical personnel deemed essential to state that these 2 conditions have systemic status; therefore, they characterize the individual as a whole. The class copd:Sarcopenic is further detailed into the subclasses copd:Probable_Sarcopenic, copd:Diagnosed_Sarcopenic, and copd:Severe_Sarcopenic to reflect the operational definition standard provided by clinicians (presented in the previous subsection).

In the same way, the copd:Anthropometric_Phenotypes are characteristics of the copd:Patients, and this class lists 5 subclasses for the representation of the phenotypes.

Similarly to copd:Patient, the development of the TBox pertaining to the patient's health condition was discussed among the team members—each health condition is characterized by an NRI profile, but in general, a copd:Health_Condition is not necessarily characterized by COPD. The terms adopted in the conceptual map to sketch the relationships among copd:Health_Condition, copd:Nutritional_Risk_Index_Profile, and copd:COPD_HC were found indicative of the clinicians' perspective—both NRI and COPD are considered particular attributes of a health condition (ie, there could be health conditions characterized only by an NRI profile but lacking COPD). Therefore, the classes copd:Nutritional_Risk_Index_Profile and copd:COPD_HC were modeled as rdfs:subclassOf copd:Health_Condition. This decision was also encouraged by the fact that the datatype properties copd:FEV1 and copd:nutritionalRiskIndex have copd:Health_Condition as the domain.

The copd:Nutritional_Risk_Index_Profile and copd:COPD_HC subclasses are characterized by restrictions that allow for the classification of individual health conditions whose copd:nutritionalRiskIndex and copd:FEV1 object values fall under specific restrictions.

Each copd:Health_Condition is described by a set of datatype properties, which represent the clinical data elicited in the previous phase and are required to enable the patient's classification and recommendations. Each owl:Individual belonging to this class also materializes inferred triples related to the copd:AppendicularSkeletalMuscleMass, the copd:ResistiveIndex, and the copd:nutritionRiskIndex. While the copd:nutritionalRiskIndex is calculated using semantic web rule language (SWRL) rules and used to classify each copd:Health_Condition into one of the NRI's subclasses, the copd:ResistiveIndex is a piece of information necessary to calculate the copd:AppendicularSkeletalMuscleMass (both are inferred as the result of 2 different SWRL rules). Figure 2 illustrates an example of copd:Health_Condition completed with all its datatype properties (both asserted and inferred).

The ontology makes use of 39 datatype properties and 2 object properties (copd:isInHealthCondition and copd:hasRecommendation)—almost all datatype properties were used to provide values for the patient's health condition and nutritional recommendation.

The ontology also contains 79 SWRL rules, which are largely used to represent the tuples in Multimedia Appendix 1, depicting the conditions that determine the shares and amounts of nutrients characterizing a patient's diet (see the full ontology in Multimedia Appendix 2). The equations adopted to calculate the BMR and the corrected caloric intake were adapted with SWRL using mathematical built-ins [61]. Taking as an example a copd:Overweight, copd:non-Sarcopenic, and copd:non-Cachectic male patient characterized by copd:Stage2 disease, the BMR is inferred through the following rule (for each male patient not characterized by sarcopenia or cachexia, calculate the BMR using the Harris-Benedict equation and round the result):

Male(?p), Overweight(?p), (not (Cachectic))(?p), (not (Sarcopenic))(?p), hasRecommendation(?p, ?rec), isInHealthCondition(?p, ?hc), age(?hc, ?age), currentWeight(?hc, ?kg), height_meters(?hc, ?m), multiply(?a, ?kg, 13.75), multiply(?b, ?m, 5, ?100), multiply(?c, 6.78, ?age), add(?d, 66.5, ?a, ?b), subtract (?e, ?d, ?c), round(?f, ?e) -> regularRecommendedCaloricIntake(?rec, ?f)

Then, the correction is applied (for each male patient not characterized by sarcopenia or cachexia with stage-1 or stage-2 COPD, correct the BMR calculated using the Harris-Benedict equation by multiplying it by 1.5):

(Normal_Weight or Obesity_1st_Degree or Overweight or Underweight)(?p), hasRecommendation(?p, ?rec), isInHealthCondition(?p, ?hc), (Stage1 or Stage2)(?hc), (not (Cachectic))(?p), (not (Sarcopenic))(?p), regularRecommendedCaloricIntake(?rec, ?reg), multiply(?corin, ?reg, 1.5), round(?f, ?corin) -> correctedRecommendedCaloricIntake(?rec, ?f)

The definition of the share of protein that a patient with COPD needs is calculated by identifying the amount (in grams) of protein. With the sole exception of patients with cachexia—who are given a 25% protein share for clinical reasons—for each patient, the daily quantity of protein is calculated according to their weight (for each patient with normal or overweight status and not characterized by sarcopenia or cachexia, obtain the amount of protein in grams by multiplying their current weight by 1.2):

(Normal_Weight or Overweight)(?p), (not (Cachectic))(?p), (not (Sarcopenic))(?p), hasRecommendation(?p, ?rec), isInHealthCondition (?p, ?hc), currentWeight (?hc, ?w), multiply (?pgra, ?w, 1.2) -> proteinsGrams (?rec, ?pgra)

This amount is then converted into calories, bearing in mind that 1 protein is equal to 4 kcal, and then the daily protein share is calculated. This approach also enables the possibility to correct the amount of protein for particular classes of patients; for example, dieticians indicated that patients classified as copd:non-Sarcopenic, copd:non-Cachectic, and copd:Underweight should have their protein share calculated considering a different BMI (which is set to a higher value to fight their underweight condition and is established at 22.5 kg/m^2).

As mentioned previously, the ontology provides enough SWRL rules to model all the information identified by the domain experts and elicited in Multimedia Appendix 1. As established by the development step in AgiSCOnt, the ontology underwent a test with data from 6 patients provided by clinicians. The test was divided into 2 steps. The first was dedicated to assessing whether the ontology provided a correct classification of the patients (ie, whether it identified copd:Sarcopenic and copd:Cachectic status for each patient and whether the stage of COPD and the copd:Nutritional_Risk_Index_Profile were correctly inferred). Thus, by querying the ontology using SPARQL (World Wide Web Consortium) [62], it was possible to assess the accuracy of the inferences (reported in Table 5).

The pneumologist and dieticians verified the correctness of the classification for each patient. All 6 individuals representing patients were found to be correctly classified. The second phase dealt with the retrieval of nutritional suggestions and their evaluation by the clinical personnel with the aim of assessing the validity of the SWRL rules modeled in the COPD and Nutrition ontology. The ontology was queried using SPARQL to retrieve all the nutrient minimum and maximum shares and quantities deemed important for patients with COPD (the results of the query are reported in Multimedia Appendix 3). Each copd:Nutritional_Recommendation and its inferred nutrient values were evaluated by clinical personnel and were found to be correct—although, for some values such as the copd:proteinShare and copd:fiberMINamount, a rounding of the decimal was suggested by dieticians.

Figure 2. The complete datatype property set for a copd:Health_Condition. The datatype properties with a yellow background represent inferred values.

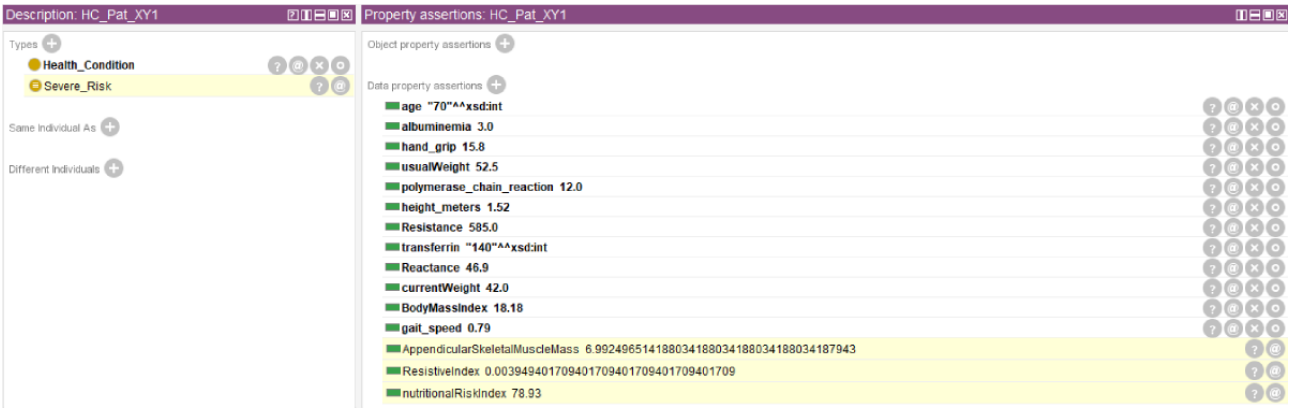


Table 5. An excerpt of the results retrieved for the query investigating, for each patient, their ID, their status (whether they had sarcopenia or cachexia), the stage of chronic obstructive pulmonary disease (COPD), and their nutritional risk index profile. For patients characterized by sarcopenia, all the subclasses of copd:Sarcopenic are illustrated so that clinical personnel can easily see the importance of this condition.

?id	?status	?antrPhen	?copdstage	?nri
001	copd:Cachectic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Diagnosed_Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Female	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Probable_Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Severe_Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
...[a]	...	...	...	...
BB	copd:Female	copd:Normal_Weight	copd:Stage4	copd:Absence_of_Risk
BB	copd:non-Cachetic	copd:Normal_Weight	copd:Stage4	copd:Absence_of_Risk
BB	copd:non-Sarcopenic	copd:Normal_Weight	copd:Stage4	copd:Absence_of_Risk
CV	copd:Male	copd:Obesity_1st_Degree	copd:Stage3	copd:Mild_Risk
CV	copd:non-Cachetic	copd:Obesity_1st_Degree	copd:Stage3	copd:Mild_Risk
CV	copd:non-Sarcopenic	copd:Obesity_1st_Degree	copd:Stage3	copd:Mild_Risk
...	...	...	...	...

[a]Data not reported for conciseness.

Ontology Use and Updating

The COPD and Nutrition ontology described in the previous subsection is a tested prototype able to classify patients properly and provide clinicians with nutrition-related recommendations. To support clinical personnel, the ontology-based DSS needs to be integrated into digital applications, enabling clinicians to access its functionalities intuitively and easily (as described in the following section). This would also enable the assessment of the usefulness and accuracy of the inferences by leveraging on clinicians outside the development team.

The Application for Clinical Personnel

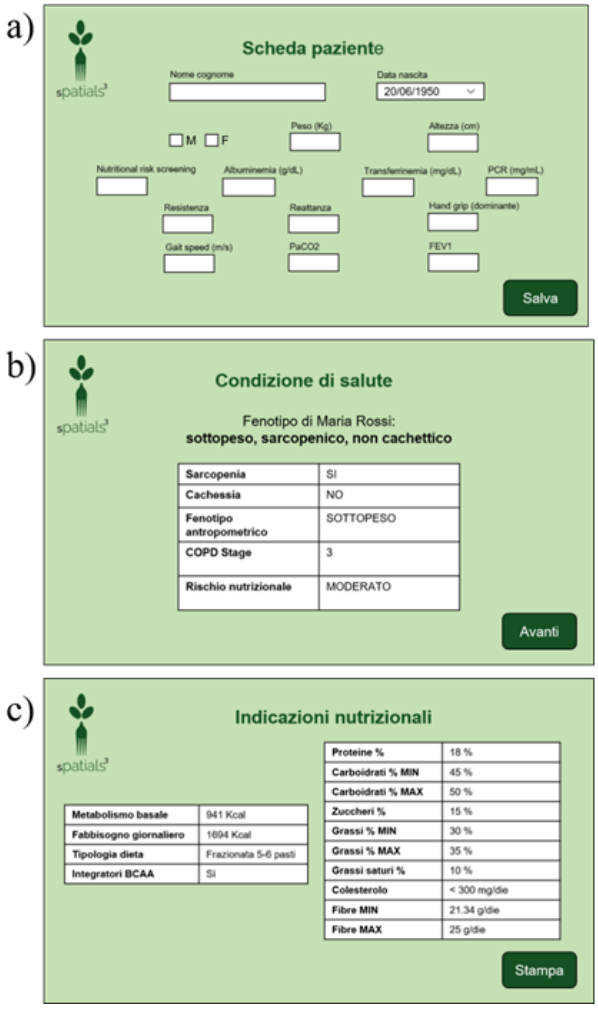
The clinician application was developed to run on a Windows PC or laptop and allows the lung specialist to obtain an overview of the patient's nutritional condition and generate tailored dietary recommendations. To achieve this, the application is connected to the DSS (hosted on a semantic repository) with permission to modify the ontology, reason over new input data, and obtain a new set of dietary recommendations. The application and DSS communication are based on SPARQL queries running over the Stardog reasoner. The entire architecture has already been tested in previous work [24,63].

The application is based on a simple and intuitive graphical user interface (GUI). Its flow is as follows:

1. The clinician logs in to the application either to create a new patient profile or to modify an existing one. The Patient Profile (Scheda Paziente) panel (Figure 3A) has several fields corresponding to the input information needed by the DSS to represent the user's health condition. These include personal data and the results of the patient assessment.
2. Once the new patient profile is saved, the application sends a SPARQL query INSERT to upload the new information into the ontology. The DSS reasons over the new data to obtain the patient's classification and the nutritional recommendations. The output is sent back to the application in the form of a JSON file to populate the GUI panels.

3. The Health Condition (Condizione di Salute) panel, shown in Figure 3B, shows the classification results: metabolic phenotype, presence of sarcopenia and cachexia, anthropometric phenotype, COPD stage, and NRI.
4. The subsequent panel, the Nutritional Recommendations (Indicazioni Nutrizionali), shown in Figure 3C, summarizes the nutritional recommendations, indicating basal metabolism, daily intake, type of diet, suggested BCAA supplement, percentage, and grams of micro- and macronutrients.
5. The clinician can save the recommendations in a PDF file or print them to share them with the dietician to support them in preparing a diet for the patient.

Figure 3. The 3 main graphical user interface panels of the clinician application: (A) patient profile, (B) health condition, and (C) nutritional recommendations.



Preliminary Validation With Clinical Personnel

Overview

Before the implementation in a real use-case scenario, we performed an expert validation of the DSS as a preliminary but necessary step. In total, 2 experiments were performed—considering the multidisciplinary nature of the tool, it was necessary to validate 2 aspects of the DSS with 2 different groups of clinicians.

Therefore, each experiment was carried out by a specific group of clinical experts: (1) a group of nutrition experts validated the recommendations generated by the DSS, and (2) a group of specialists in respiratory diseases evaluated the acceptability of the digital application in clinical practice based on the COPD DSS.

Procedure

Participants were recruited through email invitations among the national experts in nutrition and pulmonology, identified by searching the literature and professional networks. Once they expressed willingness to participate in the study, they signed a written informed consent form and agreed to the data treatment according to the General Data Protection Regulation. In total, 2 experimenters scheduled the web-based video calls—one for each participant—between November 2022 and January 2023. During the test, the experimenters briefly introduced the aim of the system and the main steps of the validation and recorded the participants' answers to brief ad hoc questionnaires and spontaneous comments. Descriptive statistics were calculated for each variable, and the spontaneous comments were analyzed and categorized based on their content.

Experiment 1: Validation of the DSS Recommendations

The first experiment focused on validating the nutritional recommendations generated by the DSS. A total of 7 nutrition experts participated in the validation. The experimenters showed the participants the profiles of 5 real patients and the inferred recommendations. The patient profiles were obtained from real clinical cases provided by the clinicians involved in the project. Each profile included the patient ID, age, gender, and health condition containing all the clinical parameters needed as input to the system.

After presenting the patient's condition, the experimenter showed the recommendations generated by the DSS, which included the patient's inferred classification (metabolic phenotype, presence of sarcopenia or cachexia, anthropometric phenotype, COPD stage, and NRI) and the nutritional recommendations with information on metabolism and suggested quantities of macro- and micronutrients. An example of a patient profile used during the experiment is presented in Figure 4, whereas all the patient profiles used during the evaluation are available in Multimedia Appendix 3. Participants were granted up to 15 minutes to observe the presented patient's health condition. During this time, participants were free to perform calculations using the data shown, ask questions (if necessary) to the experimenter, and consult external sources (eg, books and papers). After the 15-minute period, for each patient, we asked the participants to rate on a scale from 1 to 5 how much they agreed with the recommendation; in case of a score of <5, we asked the participant to provide a brief explanation. We also collected spontaneous comments that emerged during the experiment. The maximum duration of the experiment for each participant was 1 hour and 15 minutes.

Figure 4. An example of a patient profile provided to clinicians (on the left) and the inferences drawn by the ontology-based decision support system (on the right).

Records	
ID	BB
Age	85
Gender	F

Clinical data	
Height (m)	1,5
Current weight (Kg)	43
Usual weight (Kg)	43
BMI	19,11

PaCO2 (Hgmm)	40,4
FEV1 (%)	100

CRP (mg/ml)	0,1
Albuminemia (g/dl)	3,86
Transferrin (mg/dl)	264
Resistance	431
Reactance	17,3
Hand grip (dominant) (kg)	16
Gait speed (m/s)	1,31

General inferences	
	BB
Appendicular Skeletal Muscle Mass	5,193
Resistive Index	0,005
Nutritional Risk Index	100,333
Anthropometric phenotype	Normal weight
Sarcopenia	no
Cachexia	no
Nutritional risk	Absent
COPD Stage	4

Nutritional recommendations inferred

Number of meals	Kcal BMR	Kcal Intake
5 or 6 meals per day	946	1703

% min carbohydrates	% max carbohydrates	% simple sugars	% proteins	proteins (g)
45.0	50.0	15	12.12	51.6

% min fats	% max fats	% saturated fats	cholesterol (mg)
30	35	10	300

min fibers (g)	max fibers (g)	sodium	calcium (mg)	required BCAA?
21.457	25.0	2.4 grams per day, 6 grams of salt per day	1500	no

Experiment 2: Acceptability Evaluation

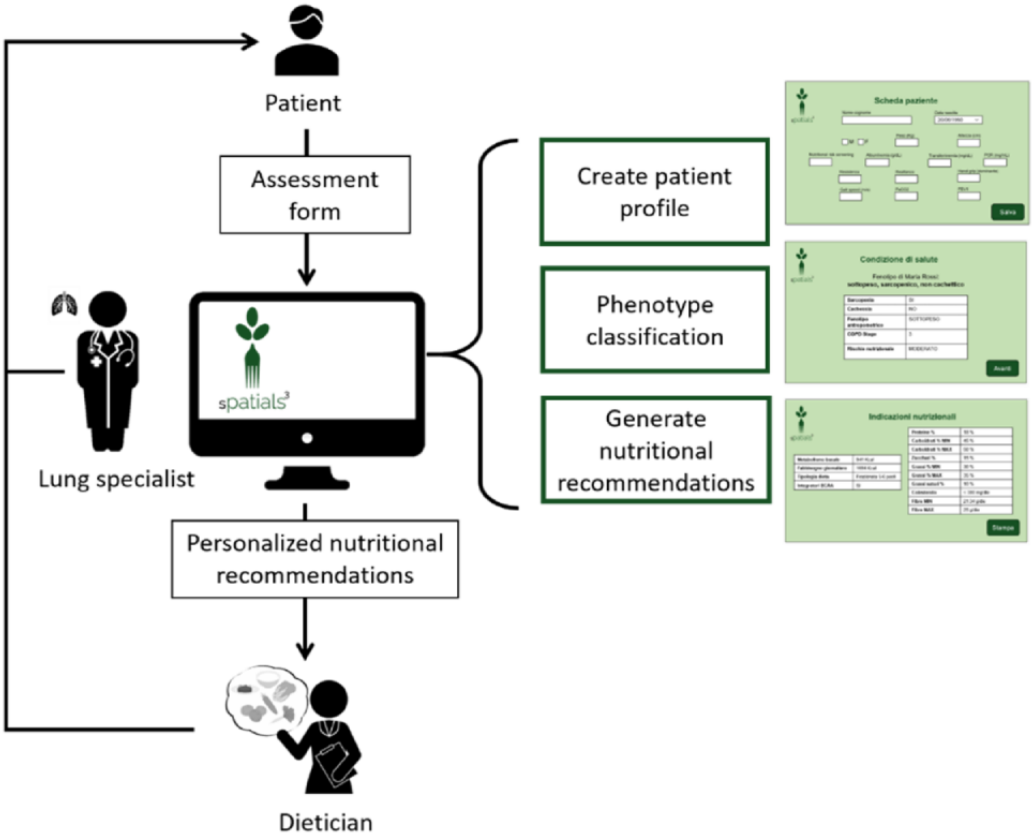
The second evaluation focused on assessing the overall system acceptability, including the DSS and a digital application working as a GUI for inserting patient data and retrieving personalized recommendations. A total of 7 lung specialists agreed to participate in the experiment. The experimenter explained to each participant the expected process of use of the system and the application data flow in a daily clinical routine, as summarized in Figure 5. The lung specialists performed the patient assessment; inserted the clinical data to generate the patient profile on the application GUI, as described in the Application for Clinical Personnel section; and generated the nutritional recommendations**.**

Participants were granted 10 minutes to assess the application, ask the experimenters questions, or ask to be presented with the data flow again.

After this time, the experimenter administered an ad hoc questionnaire based on the subscales of the technology acceptance model by Davis [64] and its subsequent amendments [65] focused on perceived usefulness and intention to use. The participants had to rate on a scale from 1 to 5 the level of agreement with the following statements: (1) "I think that the proposed system is useful for clinicians"; (2) "I think that by using this application could enhance the treatment of individuals with COPD"; and (3) "if I had this application at work, I would use it."

We also asked them to specify at least one reason why the application could be useful or not.

Figure 5. The process of use of the chronic obstructive pulmonary disease decision support system allowing the professionals to generate nutritional recommendations for a specific patient and some screenshots of the main graphical user interface panels (patient profile, patient classification, and nutritional recommendations).



Results

Experiment 1 Results: Validation of the DSS Recommendations

A total of 7 experts (mean age 46.60, SD 13.35 years; n=7, 100% female) participated in the first validation phase. They were all dieticians with an average of 25.57 (SD 11.49) years of professional experience. The level of agreement with each recommendation is reported in Table 6 in terms of mean, SD, and minimum and maximum score.

The comments provided by each participant were analyzed and categorized according to their content. The categorization and frequency of the comments and the patient profiles that originated them are reported in Table 7.

Table 6. Validation—level of agreement scores (mean and SD) for each patient profile (BB, FG, LA, TM, and XY2 are the patient IDs) expressed by each participant; mean and SD for each patient profile.

	Scores (1-5)							Values, mean (SD)
BB	5	4	5	5	3	4	5	4.43 (0.79)
FG	5	1	4	2	3	4	4	3.29 (1.38)
LA	5	4	5	5	4	4	4	4.43 (0.53)
TM	5	4	5	5	3	4	5	4.43 (0.79)
XY2	5	4	5	5	4	4	5	4.57 (0.53)

Table 7. The list of comments provided by clinicians categorized according to their content (category and comment) and linked to the patients that originated them (patient IDs); the frequency of each comment is reported in parentheses.

Category	Comment	Patient ID (frequency)
Quantity	"Too high-calorie uptake"	BB (1), FG (6), TM (1), and ANM (1)
Quantity	"Too high protein uptake"	FG (2) and LA (2)
Quantity	"Consider reducing simple sugars from 15 to 10%"	FG (1), LA (1), TM (2), and ANM (1)
Assessment	"I suggest including the physical activity level during the assessment and adjust correction factors accordingly"	All (2)

Experiment 2 Results: Acceptability Evaluation

A total of 7 experts (mean age 44.71, SD 11.94 years; n=7, 100% female) participated in the second validation phase. Of the 7 experts, 5 (71%) were lung specialists, 1 (14%) was a specialized lung physician, and 1 (14%) was a surgeon of the respiratory system. They had, on average, 14.73 (SD 9.67) years of professional experience, ranging from a minimum of 1 year for the specialized physician to 31 years. The acceptability score for the 3 subscales of perceived usefulness and intention to use was >4 points out of 5, as reported in Table 8.

Our system was considered useful for clinical practice because (1) it promotes the importance and facilitates the inclusion of the nutritional aspect in PR, as stated by 43% (3/7) of the participants; (2) it quickly provides a complete overview of the patient's condition, according to 57% (4/7) of the participants; and (3) it fosters the multidisciplinary collaboration between lung specialists and dieticians. In addition, 86% (6/7) of the participants spontaneously commented on the ease of use of the application GUI that allowed the clinician to insert the patient assessment and obtain the nutritional recommendations.

Table 8. Acceptability—technology acceptance model subscale scores reported by each participant for perceived usefulness (PU1 and PU2) and intention to use (INT); mean and SD are reported for each subscale.

	Scores (1-5)							Values, mean (SD)
INT	4	4	4	5	4	4	4	4.86 (0.38)
PU1	5	5	5	5	4	5	5	4.86 (0.38)
PU2	5	5	5	5	5	5	4	4.14 (0.38)

Discussion

Principal Findings

We performed 2 types of evaluation with 2 different experiments aimed at validating the nutritional recommendations generated by our system by nutrition experts and assessing the system's acceptability by lung specialists. The first validation—performed by 7 nutrition experts—demonstrated that our system is able to provide meaningful and safe recommendations overall in compliance with clinical practice. In 80% (4/5) of the cases, the level of agreement between the human and the "digital" expert was approximately 4.5 points out of 5. In one case (patient FG), the experts did not completely agree with the recommendations, reporting a score of 3.29, which is not considered a disagreement. However, such a score was associated with the case of a critical patient who, in addition to COPD, had second-degree obesity. This is because our system is highly specialized in treating patients with COPD, who need a higher energy intake to cope with impaired respiratory functionality. In such critical cases, the active involvement of the clinician in the process becomes essential. For instance, the clinician may adjust the nutritional recommendations for the patient to lose weight while monitoring them. It is crucial that such a patient does not lose muscular mass instead of fat mass.

The comments were positive overall and could be considered more as suggestions than criticisms. The presence of small disagreements among experts (eg, regarding the calculation of the quantity of macro- and micronutrients) confirms the need for maintaining "the clinician in the loop," as prescribed by the AgiSCOnt methodology adopted for the development of our DSS. In fact, although our system provides a useful and easy way of generating recommendations, each clinical case should be carefully considered, and slight modifications should be made by the clinician themselves in person.

The same considerations apply to the spontaneous comments, summarized in Table 7. The main concerns were about the percentage of simple sugars, which, for 4 patients, could be reduced from 15% to 10%. The guidelines indicate 15% as the maximum value, which should be adjusted by the clinician based on the percentage of other macronutrients. The other comments revealed a slight disagreement on the overall energy and protein intake, which was sometimes considered too high. However, our system is specifically focused on COPD; therefore, a higher intake was justified by the need to compensate for impaired respiratory function. Our group of experts was representative of the Italian clinical scenario, in which the influence of COPD on a patient's nutrition is sometimes neglected. Therefore, such a result strengthens the rationale of our work, which provides a system able to help professionals and clinical care facilities, which often lack specialized services, identify particular needs toward more personalized and effective care.

The second evaluation demonstrated the acceptability of our system by a group of final users (ie, lung specialists involved daily in the assessment and therapy of patients with COPD). The usefulness of our system was confirmed and was especially

related to the possibility of strengthening the consideration of nutritional aspects as part of PR standard practice. This is considered crucial by most specialists and the scientific community; however, due to organizational issues, it is not always considered [4]. Another crucial aspect that emerged was related to the importance of a multidisciplinary approach, and our system could especially help ease the cooperation between lung and nutrition specialists. At the same time, it could help lung professionals in extending their knowledge by considering aspects not strictly related to their expertise. Finally, all participants expressed their willingness to have such a system available in their daily clinical routine. They also considered it easy to use as the GUI was clear and the process was intuitive.

Limitations

This work is not without limitations. First, our system was designed for the Italian context. The DSS is based on the national nutritional guidelines—it was necessary to follow a recognized standard, which may be different from one country to another. Similarly, the dietary recommendations are based on the Italian diet. This was necessary to provide a tool that can be effectively used by our target users (ie, lung specialists treating patients with COPD in the Italian health care system). The DSS's modularity allows it to overcome such a limitation easily—the DSS could be adapted to include nutritional recommendations for other countries. The second limitation is about the participants of our validation experiments. The first experiment was based on the evaluation of 5 patient profiles by a group of nutrition experts. The number of patients examined was identified as the best compromise between a comprehensive representation of the clinical context and organizational aspects. Despite being few, the proposed patient profiles covered most of the potential real clinical cases. Regarding the acceptability evaluation, the main limitation resides in the fact that participants were homogeneous in terms of age (most of them were aged 45-50 years), culture, geographical location, and language (all of them were Italian). Such sociocultural factors are known to impact digital health technology use [66]; however, as previously stated, our work at this stage is focused on the Italian scenario, and therefore, our sample can be considered representative of the final population of target users.

Future Work

As recently noted, most digital health applications remain limited to pilot studies—mainly because they fail in the proposed aims or face significant implementational barriers [67]. From a digital health application perspective, our experiments aimed to verify the stability of the developed solution—in line with the WHO's guidelines [23]. In particular, the validations verified the performance consistency, the proposed solution's overall feasibility, and the digital tool's efficacy. Considering the early stage of the DSS and its application, we need to further investigate the implementation protocols and to acquire long-term proof of the efficacy of the tool among pneumologists and patients with COPD (ie, the acceptability of the tool needs to be verified with a larger sample of end users and in a real clinical setting so that more pneumologists can provide feedback regarding the tool's perceived usefulness, ease of use, and satisfaction; moreover, the effectiveness of the proposed diets should be tested with patients with COPD). To achieve these objectives, more extensive experimentation with larger samples of participants (clinical nutritionists, dieticians, and pneumologists) is necessary. Moreover, the involvement of clinical personnel can support the identification of implementation protocols suitable for the adoption of the digital tool in clinical practice. In this way, the application's level of maturity could move from early to mild (according to the WHO [23]), where its effectiveness can be tested in a nonresearch (uncontrolled) setting.

To support the prompt identification of barriers and implementational challenges, reporting the development of the COPD DSS within a framework for the definition of digital health application implementation can be useful. In particular, the Guidelines and Checklist for the Reporting on Digital Health Implementations [67], by providing a list of 20 items to be monitored, can foster the identification of issues in the Technical design phase in the Interoperability and Data management areas.

In this regard, the availability of data and the implementation of the application within the health system are another relevant node to be investigated. Although the current version of the digital application is still in its prototypical phase, scaling up the application to the regional or national level (ie, *coverage* in the Guidelines and Checklist for the Reporting on Digital Health Implementations) is essential to ensure its use in clinical practice. Therefore, toward this aim, strategies for collecting the outputs and making them available in patients' data (or electronic health records) need to be investigated. In this regard, scientific literature offers some interesting approaches grounded in the Italian health care system that could be considered to make the COPD DSS interoperable with existing tools [68-70]. As the COPD DSS leverages ontologies to represent its data, this technology can be used to achieve semantic interoperability of the information [68], moving a step toward the longitudinal collection of health data about patients [70] while ensuring data protection according to the national and European laws [69].

Finally, from an ontological perspective, the domain ontology regarding COPD presented in this work could benefit from mapping with existing (and larger) biomedical standard ontologies to increase its shareability (eg, the WHO's International Classification of Diseases and International Classification of Functioning, Disability and Health, as well as upper biomedical ontologies)—a best practice of ontology engineering [71]. Moreover, considering that the proposed system can be adapted to any other national clinical context by modifying the domain ontologies, a possible future research direction consists of involving international clinicians to increase the knowledge formalized in the ontologies so that it is possible for the DSS to cover the specific nutritional indications of different countries.

Conclusions

The role of nutrition in the management of patients with COPD is often underestimated, although scientific evidence points toward the important role that diet plays in PR. The nutritional status of patients with COPD is essential to prevent exacerbations and avoid comorbidities, but attention to the patient's body composition and nutritional status is often

secondary in clinical practice. This may be partially because lung specialists may lack specialized training in clinical nutrition, although they recognize the relevance of dietary recommendations in PR.

An ontology-based DSS was developed to support pneumologists in considering nutritional aspects. The DSS formalizes expert knowledge in computable models able to infer patient-tailored nutritional recommendations, leveraging a set of information to capture the nutritional and physical status of the patient; therefore, by applying rules, it can support the classification of patients with COPD and provide tailored recommendations indicating the percentages and amounts of micro- and macronutrients that should make up their diet. The domain ontologies act as the backbone of a clinician-dedicated application.

The application was validated to assess the clinical compliance of the DSS's recommendations and the acceptability of such an application in clinical practice by lung specialists. For both validations, the proposed system performed more than adequately—in particular, pneumologists underlined the role that such an application may play in achieving a multidisciplinary approach in PR. This paper concludes by investigating future research directions to implement the COPD DSS further into a fully-fledged digital health application.

Acknowledgments

The sPATIALS3 (Miglioramento delle Produzioni Agroalimentari e Tecnologie Innovative per un'Alimentazione più Sana, Sicura e Sostenibile) project is financed by the European Regional Development Fund under the ROP (Regional Operational Programme) of the Lombardy Region European Regional Development Fund 2014 to 2020 (axis I, "Strengthen technological research, development and innovation"; action 1.b.1.3, "Support for cooperative R&D activities to develop new sustainable technologies, products and services"; Call Hub).

Data Availability

All data generated or analyzed during this study are included in this published paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A table representing the set of rules for providing phenotype-tailored recommendations.
[DOCX File , 46 KB - medinform_v12i1e50980_app1.docx]

Multimedia Appendix 2
The chronic obstructive pulmonary disease decision support system domain ontology.
[ZIP File (Zip Archive), 24 KB - medinform_v12i1e50980_app2.zip]

Multimedia Appendix 3
A table reporting the nutritional recommendations for each patient used to test the validity of the ontology rule set.
[DOCX File , 31 KB - medinform_v12i1e50980_app3.docx]

References

1. Benjafield A, Tellez D, Barrett M, Gondalia R, Nunez C, Wedzicha J, et al. An estimate of the European prevalence of COPD in 2050. Eur Respir J 2021;58(suppl 65):OA2866 [FREE Full text] [doi: 10.1183/13993003.congress-2021.oa2866]
2. MacNee W. Pathology, pathogenesis, and pathophysiology. BMJ 2006 May 18;332(7551):1202-1204. [doi: 10.1136/bmj.332.7551.1202]
3. Spruit MA, Singh SJ, Garvey C, Zu Wallack R, Nici L, Rochester C, ATS/ERS Task Force on Pulmonary Rehabilitation. An official American thoracic society/European respiratory society statement: key concepts and advances in pulmonary rehabilitation. Am J Respir Crit Care Med 2013 Oct 15;188(8):e13-e64. [doi: 10.1164/rccm.201309-1634ST] [Medline: 24127811]
4. Holland AE, Cox NS, Houchen-Wolloff L, Rochester CL, Garvey C, ZuWallack R, et al. Defining modern pulmonary rehabilitation: an official American thoracic society workshop report. Ann Am Thorac Soc 2021 May;18(5):e12-e29 [FREE Full text] [doi: 10.1513/AnnalsATS.202102-146ST] [Medline: 33929307]
5. Scoditti E, Massaro M, Garbarino S, Toraldo DM. Role of diet in chronic obstructive pulmonary disease prevention and treatment. Nutrients 2019 Jun 16;11(6):1357 [FREE Full text] [doi: 10.3390/nu11061357] [Medline: 31208151]
6. Collins PF, Stratton RJ, Elia M. Nutritional support in chronic obstructive pulmonary disease: a systematic review and meta-analysis. Am J Clin Nutr 2012 Jun;95(6):1385-1395 [FREE Full text] [doi: 10.3945/ajcn.111.023499] [Medline: 22513295]

7. Ferreira I, Brooks D, White J, Goldstein R. Nutritional supplementation for stable chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2012 Dec 12;12:CD000998. [doi: 10.1002/14651858.CD000998.pub3] [Medline: 23235577]
8. Collins PF, Yang IA, Chang YC, Vaughan A. Nutritional support in chronic obstructive pulmonary disease (COPD): an evidence update. J Thorac Dis 2019 Oct;11(Suppl 17):S2230-S2237 [FREE Full text] [doi: 10.21037/jtd.2019.10.41] [Medline: 31737350]
9. Liu H, Tan X, Liu Z, Ma X, Zheng Y, Zhu B, et al. Association between diet-related inflammation and COPD: findings from NHANES III. Front Nutr 2021 Oct 18;8:732099 [FREE Full text] [doi: 10.3389/fnut.2021.732099] [Medline: 34733875]
10. van Iersel LE, Beijers RJ, Gosker HR, Schols AM. Nutrition as a modifiable factor in the onset and progression of pulmonary function impairment in COPD: a systematic review. Nutr Rev 2022 May 09;80(6):1434-1444 [FREE Full text] [doi: 10.1093/nutrit/nuab077] [Medline: 34537848]
11. Nguyen HT, Collins PF, Pavey TG, Nguyen NV, Pham TD, Gallegos DL. Nutritional status, dietary intake, and health-related quality of life in outpatients with COPD. Int J Chron Obstruct Pulmon Dis 2019;14:215-226 [FREE Full text] [doi: 10.2147/COPD.S181322] [Medline: 30666102]
12. van Bakel SI, Gosker HR, Langen RC, Schols AM. Towards personalized management of sarcopenia in COPD. Int J Chron Obstruct Pulmon Dis 2021;16:25-40 [FREE Full text] [doi: 10.2147/COPD.S280540] [Medline: 33442246]
13. Schols AM, Ferreira IM, Franssen FM, Gosker HR, Janssens W, Muscaritoli M, et al. Nutritional assessment and therapy in COPD: a European respiratory society statement. Eur Respir J 2014 Dec 18;44(6):1504-1520 [FREE Full text] [doi: 10.1183/09031936.00070914] [Medline: 25234804]
14. Machado FV, Schneider LP, Fonseca J, Belo LF, Bonomo C, Morita AA, et al. Clinical impact of body composition phenotypes in patients with COPD: a retrospective analysis. Eur J Clin Nutr 2019 Nov 14;73(11):1512-1519. [doi: 10.1038/s41430-019-0390-4] [Medline: 30643222]
15. Hanson C, Rutten EP, Wouters EF, Rennard S. Influence of diet and obesity on COPD development and outcomes. Int J Chron Obstruct Pulmon Dis 2014;9:723-733 [FREE Full text] [doi: 10.2147/COPD.S50111] [Medline: 25125974]
16. Mete B, Pehlivan E, Gülbaş G, Günen H. Prevalence of malnutrition in COPD and its relationship with the parameters related to disease severity. Int J Chron Obstruct Pulmon Dis 2018;13:3307-3312 [FREE Full text] [doi: 10.2147/COPD.S179609] [Medline: 30349235]
17. Sepúlveda-Loyola W, Osadnik C, Phu S, Morita AA, Duque G, Probst VS. Diagnosis, prevalence, and clinical impact of sarcopenia in COPD: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle 2020 Oct 30;11(5):1164-1176 [FREE Full text] [doi: 10.1002/jcsm.12600] [Medline: 32862514]
18. Hoong JM, Ferguson M, Hukins C, Collins PF. Economic and operational burden associated with malnutrition in chronic obstructive pulmonary disease. Clin Nutr 2017 Aug;36(4):1105-1109. [doi: 10.1016/j.clnu.2016.07.008] [Medline: 27496063]
19. Iheanacho I, Zhang S, King D, Rizzo M, Ismaila AS. Economic burden of chronic obstructive pulmonary disease (COPD): a systematic literature review. Int J Chron Obstruct Pulmon Dis 2020;15:439-460 [FREE Full text] [doi: 10.2147/COPD.S234942] [Medline: 32161455]
20. Beijers RJ, Steiner MC, Schols AM. The role of diet and nutrition in the management of COPD. Eur Respir Rev 2023 Jun 30;32(168):230003 [FREE Full text] [doi: 10.1183/16000617.0003-2023] [Medline: 37286221]
21. Raad S, Smith C, Allen K. Nutrition status and chronic obstructive pulmonary disease: can we move beyond the body mass index? Nutr Clin Pract 2019 Jun;34(3):330-339. [doi: 10.1002/ncp.10306] [Medline: 30989731]
22. Watson A, Wilkinson TM. Digital healthcare in COPD management: a narrative review on the advantages, pitfalls, and need for further research. Ther Adv Respir Dis 2022 Mar 02;16:17534666221075493 [FREE Full text] [doi: 10.1177/17534666221075493] [Medline: 35234090]
23. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-04-29]
24. Spoladore D, Colombo V, Arlati S, Mahroo A, Trombetta A, Sacco M. An ontology-based framework for a telehealthcare system to foster healthy nutrition and active lifestyle in older adults. Electronics 2021 Sep 01;10(17):2129. [doi: 10.3390/electronics10172129]
25. Greenberg J, Deshmukh R, Huang L, Mostafa J, La Vange L, Carretta E, et al. The COPD ontology and toward empowering clinical scientists as ontology engineers. J Libr Metadata 2010 Aug 31;10(2-3):173-187. [doi: 10.1080/19386389.2010.520604]
26. Cano I, Tényi Á, Schueller C, Wolff M, Huertas Migueláñez MM, Gomez-Cabrero D, et al. The COPD knowledge base: enabling data analysis and computational simulation in translational COPD research. J Transl Med 2014;12(Suppl 2):S6. [doi: 10.1186/1479-5876-12-s2-s6]
27. Rayner L, Sherlock J, Creagh-Brown B, Williams J, deLusignan S. The prevalence of COPD in England: an ontological approach to case detection in primary care. Respir Med 2017 Nov;132:217-225 [FREE Full text] [doi: 10.1016/j.rmed.2017.10.024] [Medline: 29229101]
28. Rosso R, Munaro G, Salvetti O, Colantonio S, Ciancitto F. CHRONIOUS: an open, ubiquitous and adaptive chronic disease management platform for chronic obstructive pulmonary disease (COPD), chronic kidney disease (CKD) and renal insufficiency. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:6850-6853. [doi: 10.1109/IEMBS.2010.5626451] [Medline: 21096301]

29. Lasierra N, Alesanco A, Guillén S, García J. A three stage ontology-driven solution to provide personalized care to chronic patients at home. J Biomed Inform 2013 Jun;46(3):516-529 [FREE Full text] [doi: 10.1016/j.jbi.2013.03.006] [Medline: 23567539]
30. Ajami H, McHeick H. Ontology-based model to support ubiquitous healthcare systems for COPD patients. Electronics 2018 Dec 02;7(12):371. [doi: 10.3390/electronics7120371]
31. McCabe C, McCann M, Brady AM. Computer and mobile technology interventions for self-management in chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2017 May 23;5(5):CD011425 [FREE Full text] [doi: 10.1002/14651858.CD011425.pub2] [Medline: 28535331]
32. Colombo V, Mondellini M, Gandolfo A, Fumagalli A, Sacco M. A mobile diary app to support rehabilitation at home for elderly with COPD: a preliminary feasibility study. In: Proceedings of the 17th International Conference on Computers Helping People with Special Needs. 2020 Presented at: ICCHP '20; September 9-11, 2020; Lecco, Italy p. 224-232 URL: https://dl.acm.org/doi/abs/10.1007/978-3-030-58805-2_27 [doi: 10.1007/978-3-030-58805-2_27]
33. myCOPD - empowering patients to manage their COPD for a lifetime. my mhealth Limited. URL: https://mymhealth.com/mycopd [accessed 2024-04-29]
34. The problem with AI. Earley Information Science. 2017. URL: https://www.earley.com/insights/problem-with-ai [accessed 2024-04-29]
35. Spoladore D, Sacco M, Trombetta A. A review of domain ontologies for disability representation. Expert Syst Appl 2023 Oct;228:120467. [doi: 10.1016/j.eswa.2023.120467]
36. Kotis KI, Vouros GA, Spiliotopoulos D. Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations. Knowl Eng Rev 2020 Jan 31;35:e4 [FREE Full text] [doi: 10.1017/s0269888920000065]
37. Spoladore D, Pessot E, Trombetta A. A novel agile ontology engineering methodology for supporting organizations in collaborative ontology development. Comput Ind 2023 Oct;151:103979. [doi: 10.1016/j.compind.2023.103979]
38. Owen T. Building expert systems, edited by Frederick Hayes-Roth, Donald A. Waterman and Douglas B. Lenat Addison-Wesley Publishing Company, Massachusetts, USA, 1983 (£32.95). Robotica 2009 Mar 09;6(2):165. [doi: 10.1017/s0263574700004069]
39. Grüninger M, Fox MS. The role of competency questions in enterprise engineering. In: Rolstadås A, editor. Benchmarking — Theory and Practice. Cham, Switzerland: Springer; 1995:22-31.
40. Gangemi A. Ontology design patterns for semantic web content. In: Proceedings of the 4th International Semantic Web Conference on Semantic Web. 2005 Presented at: ISWC '05; November 6-10, 2005; Galway, Ireland p. 262-276 URL: https://link.springer.com/chapter/10.1007/11574620_21 [doi: 10.1007/11574620_21]
41. Evans WJ, Morley JE, Argilés J, Bales C, Baracos V, Guttridge D, et al. Cachexia: a new definition. Clin Nutr 2008 Dec;27(6):793-799. [doi: 10.1016/j.clnu.2008.06.013] [Medline: 18718696]
42. Agustí A, Celli BR, Criner GJ, Halpin D, Anzueto A, Barnes P, et al. Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. Eur Respir J 2023 Apr;61(4):2300239 [FREE Full text] [doi: 10.1183/13993003.00239-2023] [Medline: 36858443]
43. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2013 Feb 15;187(4):347-365. [doi: 10.1164/rccm.201204-0596pp]
44. Prendergast JM, Coe RM, Chavez MN, Romeis JC, Miller DK, Wolinsky FD. Clinical validation of a nutritional risk index. J Community Health 1989;14(3):125-135. [doi: 10.1007/BF01324362] [Medline: 2600200]
45. Sergi G, De Rui M, Veronese N, Bolzetta F, Berton L, Carraro S, et al. Assessing appendicular skeletal muscle mass with bioelectrical impedance analysis in free-living Caucasian older adults. Clin Nutr 2015 Aug;34(4):667-673. [doi: 10.1016/j.clnu.2014.07.010] [Medline: 25103151]
46. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, Writing Group for the European Working Group on Sarcopenia in Older People 2 (EWGSOP2), the Extended Group for EWGSOP2. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jan 01;48(1):16-31 [FREE Full text] [doi: 10.1093/ageing/afy169] [Medline: 30312372]
47. LARN - Livelli di Assunzione di Riferimento di Nutrienti ed energia per la popolazione italiana. Società Italiana di Nutrizione Umana (SINU). 2021. URL: https://sinu.it/tabelle-larn-2014/ [accessed 2024-06-03]
48. Frankenfield D, Roth-Yousey L, Compher C. Comparison of predictive equations for resting metabolic rate in healthy nonobese and obese adults: a systematic review. J Am Diet Assoc 2005 May;105(5):775-789. [doi: 10.1016/j.jada.2005.02.005] [Medline: 15883556]
49. Bendavid I, Lobo DN, Barazzoni R, Cederholm T, Coëffier M, de van der Schueren M, et al. The centenary of the Harris-Benedict equations: how to assess energy requirements best? Recommendations from the ESPEN expert group. Clin Nutr 2021 Mar;40(3):690-701. [doi: 10.1016/j.clnu.2020.11.012] [Medline: 33279311]
50. Pitta F, Troosters T, Spruit MA, Probst VS, Decramer M, Gosselink R. Characteristics of physical activities in daily life in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2005 May 01;171(9):972-977. [doi: 10.1164/rccm.200407-855OC] [Medline: 15665324]

51. Matheson EM, Nelson JL, Baggs GE, Luo M, Deutz NE. Specialized oral nutritional supplement (ONS) improves handgrip strength in hospitalized, malnourished older patients with cardiovascular and pulmonary disease: a randomized clinical trial. Clin Nutr 2021 Mar;40(3):844-849. [doi: 10.1016/j.clnu.2020.08.035] [Medline: 32943241]
52. Bai GH, Tsai MC, Tsai HW, Chang CC, Hou WH. Effects of branched-chain amino acid-rich supplementation on EWGSOP2 criteria for sarcopenia in older adults: a systematic review and meta-analysis. Eur J Nutr 2022 Mar 27;61(2):637-651. [doi: 10.1007/s00394-021-02710-0] [Medline: 34705076]
53. Guerra BA, Pereira TG, Eckert IC, Bernardes S, Silva FM. Markers of respiratory function response to high-carbohydrate and high-fat intake in patients with lung diseases: a systematic review with meta-analysis of randomized clinical trials. JPEN J Parenter Enteral Nutr 2022 Sep 31;46(7):1522-1534. [doi: 10.1002/jpen.2385] [Medline: 35437762]
54. WHO global report on sodium intake reduction. World Health Organization.: World Health Organization; 2023. URL: https://www.who.int/publications/i/item/9789240069985 [accessed 2024-04-29]
55. Chen Y, Ramsook AH, Coxson HO, Bon J, Reid WD. Prevalence and risk factors for osteoporosis in individuals with COPD: a systematic review and meta-analysis. Chest 2019 Dec;156(6):1092-1110. [doi: 10.1016/j.chest.2019.06.036] [Medline: 31352034]
56. Musen MA, Protégé Team. The Protégé project: a look back and a look forward. AI Matters 2015 Jun;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
57. Pan JZ. Resource description framework. In: Staab S, Studer R, editors. Handbook on Ontologies. Berlin, Germany: Springer; 2009:71-90.
58. Antoniou G, van Harmelen F. Web ontology language: OWL. In: Antoniou G, van Harmelen F, editors. Handbook on Ontologies. Cham, Switzerland: Springer; 2009:91-110.
59. Alamri A, Bertok P. Distributed store for ontology data management. In: Lee R, editor. Computer and Information Science. Cham, Switzerland: Springer; 2012:15-35.
60. Spoladore D, Mahroo A, Trombetta A, Sacco M. DOMUS: a domestic ontology managed ubiquitous system. J Ambient Intell Human Comput 2021 Mar 31;13(6):3037-3052. [doi: 10.1007/S12652-021-03138-4]
61. O'Connor M, Tu S, Nyulas C, Das A, Musen M. Querying the semantic web with SWRL. In: Proceedings of the 2007 International Symposium on Advances in Rule Interchange and Applications.: Springer; 2007 Presented at: RuleML '07; October 25-26, 2007; Orlando, FL p. 155-159 URL: https://link.springer.com/chapter/10.1007/978-3-540-75975-1_13 [doi: 10.1007/978-3-540-75975-1_13]
62. Hommeaux EP, Seaborne A. SPARQL query language for RDF. W3C Recommendation. 2008. URL: https://www.w3.org/TR/rdf-sparql-query/ [accessed 2024-04-29]
63. Spoladore D, Mahroo A, Sacco M. Leveraging ontology to enable indoor comfort customization in the smart home. In: Proceedings of the 13th International Conference on Flexible Query Answering Systems. 2019 Presented at: FQAS '19; July 2-5, 2019; Amantea, Italy p. 63-74 URL: https://link.springer.com/chapter/10.1007/978-3-030-27629-4_9 [doi: 10.1007/978-3-030-27629-4_9]
64. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340 [FREE Full text] [doi: 10.2307/249008]
65. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
66. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
67. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
68. Ciampi M, Esposito A, Guarasci R, De Pietro G. Towards interoperability of EHR systems: the case of Italy. In: Proceedings of the 2nd International Conference on Information and Communication Technologies for Ageing Well and e-Health. 2016 Presented at: ICT4AGEINGWELL '16; April 21-22, 2016; Rome, Italy p. 138 URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0005916401330138 [doi: 10.5220/0005916401330138]
69. Bologna S, Bellavista A, Corso PP, Zangara G. Electronic health record in Italy and personal data protection. Eur J Health Law 2016 Jun 14;23(3):265-277. [doi: 10.1163/15718093-12341403] [Medline: 27491249]
70. Ciampi M, Sicuranza M, Esposito A, Guarasci R, De Pietro G. A technological framework for EHR interoperability: experiences from Italy. In: Proceedings of the 2nd International Conference on Information and Communication Technologies for Ageing Well and e-Health. 2016 Presented at: ICT4AWE '16; April 21-22, 2016; Rome, Italy p. 80-99 URL: https://link.springer.com/chapter/10.1007/978-3-319-62704-5_6 [doi: 10.1007/978-3-319-62704-5_6]
71. Spoladore D, Pessot E. Collaborative ontology engineering methodologies for the development of decision support systems: case studies in the healthcare domain. Electronics 2021 Apr 29;10(9):1060. [doi: 10.3390/electronics10091060]

Abbreviations

AgiSCOnt: Agile, Simplified, and Collaborative Ontology Engineering Methodology

BCAA: branched-chain amino acid
BMR: basal metabolic rate
COPD: chronic obstructive pulmonary disease
CQ: competency question
DSS: decision support system
FEV1: forced expiratory volume in the first second
FFM: fat-free mass
GUI: graphical user interface
LARN: Livelli di Assunzione di Riferimento di Nutrienti ed energia
NRI: nutritional risk index
PR: pulmonary rehabilitation
SWRL: semantic web rule language
WHO: World Health Organization

Edited by C Lovis; submitted 18.07.23; peer-reviewed by A AL-Asadi, T Salzmann; comments to author 19.01.24; revised version received 01.02.24; accepted 23.04.24; published 26.06.24.

Please cite as:
Spoladore D, Colombo V, Fumagalli A, Tosi M, Lorenzini EC, Sacco M
An Ontology-Based Decision Support System for Tailored Clinical Nutrition Recommendations for Patients With Chronic Obstructive Pulmonary Disease: Development and Acceptability Study
JMIR Med Inform 2024;12:e50980
URL: https://medinform.jmir.org/2024/1/e50980
doi:10.2196/50980
PMID:38922666

Use of Video in Telephone Triage in Out-of-Hours Primary Care: Register-Based Study

Mette Amalie Nebsbjerg[1], MD; Claus Høstrup Vestergaard[1], MSc; Katrine Bjørnshave Bomholt[1], MD; Morten Bondo Christensen[1,2], MD, PhD; Linda Huibers[1], MD, PhD

1

2

Corresponding Author:
Mette Amalie Nebsbjerg, MD

Abstract

Background: Out-of-hours primary care (OOH-PC) is challenging due to high workloads, workforce shortages, and long waiting and transportation times for patients. Use of video enables triage professionals to visually assess patients, potentially ending more contacts in a telephone triage contact instead of referring patients to more resource-demanding clinic consultations or home visits. Thus, video use may help reduce use of health care resources in OOH-PC.

Objective: This study aimed to investigate video use in telephone triage contacts to OOH-PC in Denmark by studying rate of use and potential associations between video use and patient- and contact-related characteristics and between video use and triage outcomes and follow-up contacts. We hypothesized that video use could serve to reduce use of health care resources in OOH-PC.

Methods: This register-based study included all telephone triage contacts to OOH-PC in 4 of the 5 Danish regions from March 15, 2020, to December 1, 2021. We linked data from the OOH-PC electronic registration systems to national registers and identified telephone triage contacts with video use (video contact) and without video use (telephone contact). Calculating crude incidence rate ratios and adjusted incidence rate ratios (aIRRs), we investigated the association between patient- and contact-related characteristics and video contacts and measured the frequency of different triage outcomes and follow-up contacts after video contact compared to telephone contact.

Results: Of 2,900,566 identified telephone triage contacts to OOH-PC, 9.5% (n=275,203) were conducted as video contacts. The frequency of video contact was unevenly distributed across patient- and contact-related characteristics; it was used more often for employed young patients without comorbidities who contacted OOH-PC more than 4 hours before the opening hours of daytime general practice. Compared to telephone contacts, notably more video contacts ended with advice and self-care (aIRR 1.21, 95% CI 1.21-1.21) and no follow-up contact (aIRR 1.08, 95% CI 1.08-1.09).

Conclusions: This study supports our hypothesis that video contacts could reduce use of health care resources in OOH-PC. Video use lowered the frequency of referrals to a clinic consultation or a home visit and also lowered the frequency of follow-up contacts. However, the results could be biased due to confounding by indication, reflecting that triage GPs use video for a specific set of reasons for encounters.

(JMIR Med Inform 2024;12:e47039) doi:10.2196/47039

KEYWORDS

primary health care; after-hours care; referral and consultation; general practitioner; GP; triage; remote consultation; telemedicine

Introduction

General practice serves as a gatekeeper to secondary care in many countries [1]. However, the services in out-of-hours primary care (OOH-PC) are challenging due to high workloads, workforce shortages, and long waiting and transportation times for patients. This development has received much political attention and has caused public debate and reorganization [2,3].

Existing health care systems are currently undergoing a digital transformation, which was pushed by the COVID-19 pandemic [4-9]. As a central part of this digitization, video consultations have been implemented broadly in general practice [5,8-12]. Many countries have introduced video as part of telephone triage in OOH-PC [12-14]. Video use enables triage professionals to visually assess patients, which may imply that more contacts can be ended in a telephone triage contact instead of referring patients to clinic consultations or home visits, which demand more resources. Thereby, video use might reduce use of health care resources related to clinic consultations and home visits.

Research has shown that patients welcome the use of video in general practice in the daytime and also after hours [4,14-16]. However, in daytime general practice, general practitioners (GPs) experience both benefits of (eg, care delivery) and barriers to (eg, technical difficulties, varying suitability for different

health problems and patient groups) video use [6,10,16-19]. Two qualitative studies indicated that video use in OOH-PC is beneficial to both triage professionals (eg, it improved patient assessment and reassurance) [13,14] and patients (eg, it led to better reassurance and higher satisfaction) [14]. Two register-based studies found that video use in OOH-PC increased during the COVID-19 pandemic [12,20]. However, little is still known about video use and its effects. This study aimed to investigate video use in telephone triage contacts to OOH-PC in Denmark by studying rate of use and potential associations between video use and patient- and contact-related characteristics and between video use and triage outcomes and follow-up contacts.

Methods

Design and Population

We conducted a register-based study of video use in telephone triage contacts to OOH-PC in 4 of the 5 Danish regions (North Denmark Region, Central Denmark Region, Region of Southern Denmark, and Region Zealand). As the Capital Region of Denmark runs a different OOH-PC system than the other 4 regions, this region was not included in this study. We included all telephone contacts from March 15, 2020, to December 1, 2021, and followed each patient for 7 days to record the outcomes. In Region Zealand, telephone contacts were included from March 1, 2021, because this region started using video from this date.

Setting

Denmark has free public health care for its residents. The health care system is centrally regulated, but most services are provided by the local governments of the 5 regions. Outside office hours, Danish GPs and GP trainees cover shifts in the regional OOH-PC service, which is open on weekdays from 4 PM to 8 AM and 24 hours during weekends and holidays. GPs and GP trainees in their last year of specialist training (hereinafter referred to jointly as triage GPs) perform telephone triage and determine the triage outcome: telephone triage with video use (video contacts) or telephone triage without video use (telephone contacts), clinic consultation, home visit, or hospital admission. The triage GPs assesses whether the problem is suitable for a video contact. If so and if the patient approves, a video link is sent to the patient via text message. When the link is activated, the triage GP can see the patient, but the patient cannot see the triage GP. Triage GPs are paid a fee for service using remuneration codes.

Outcome Measures

The following outcome measures were defined: the proportion of video contacts (number of video contacts per 100 telephone contacts); the association between video contact and patient- and contact-related characteristics (sex and age of the patient, cohabitation status, comorbidity, educational level, ethnicity, income, urbanization, employment status, region, and time of contact); the frequency of triage outcomes (advice and self-care, referral to clinic consultation, home visit, or hospital admission) and their association with video contact; and the frequency of follow-up contacts in daytime general practice or OOH-PC within 7 days or a hospital admission within 1 day and their association with video contact.

Data Collection

We used data from the OOH-PC electronic registration system, which provided information on date, time, region, type of contact (telephone contact or video contact), and triage outcome (advice and self-care, referral to clinic consultation, home visit, or hospital admission). We constructed a "time of contact" variable, which was defined by its relation to the next opening time of daytime general practice and dichotomized into >4 hours or ≤4 hours, as the option to refer a patient to their regular GP may influence the triage decision.

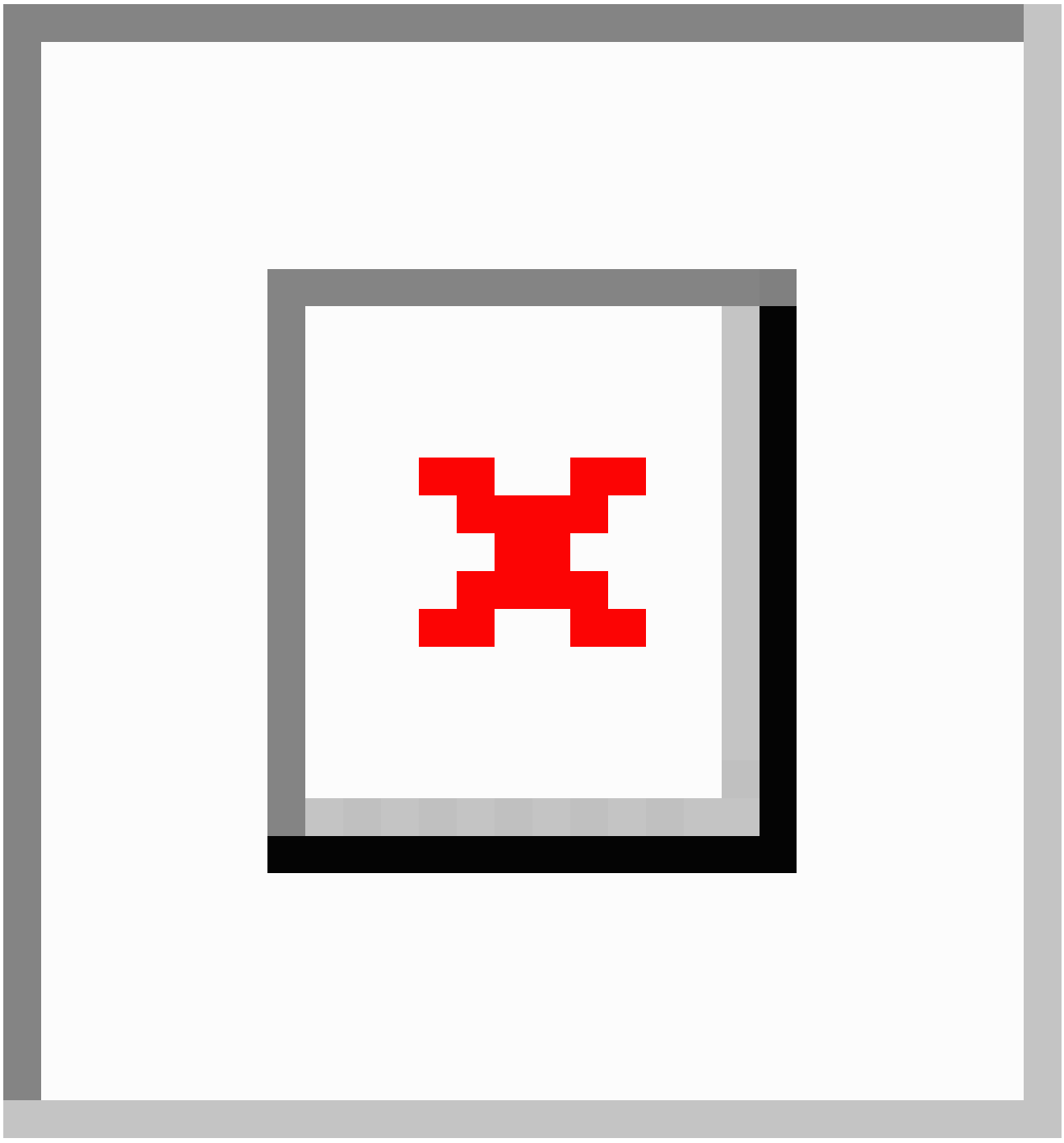
To investigate follow-up contacts, we linked data from the OOH-PC registration system to 2 Danish national registers using each patient's unique personal identification number [21]. The Danish National Health Service Register [22] provided information on date and type of contact to daytime general practice (telephone contacts, video contacts, clinic consultations, or home visits). The Danish National Patient Registry [23] provided information on date of contact to the hospital (emergency department visits and unscheduled hospital admissions) and comorbidity. Comorbidity was defined as the number of diagnoses from the Charlson Comorbidity Index that were recorded as diagnosis codes in hospital charts. Data on socioeconomic characteristics of the patients (sex, age, cohabitation status, educational level, ethnicity, income, urbanization, and employment status) were obtained from Statistics Denmark [24]. All covariates (except for age, sex, and comorbidity) were reported at the household level. For example, household educational level was determined by the member with the longest education. Hence, it was possible to avoid excluding contacts involving children because of missing values. We included only persons with registered socioeconomic characteristics.

Data Analyses

People with more than 25 contacts to OOH-PC during the study period (comprising 98,126/2,900,566 contacts, 3.4%) were excluded from the data analyses since they were considered outliers. Likewise, people aged >104 years (162/2,900,566 contacts, 0%) and patients with missing covariates (18,740/2,900,566 contacts, 0.7%) were excluded.

Descriptive analyses were used to describe the study population. To ensure convergence of the regressions, we used Poisson regression models to measure the association between patient- and contact-related characteristics and video contacts, and we calculated incidence rate ratios (IRRs) and 95% CIs [25]. Results are presented as a forest plot (Figure 1). Using a Poisson regression model, we also calculated crude and adjusted IRRs (aIRRs) of triage outcomes and follow-up contacts after a video contact compared to after a telephone contact. IRRs were adjusted for patient- and contact-related characteristics. Stata (version 17; StataCorp) was used to analyze all data. Reporting of results was conducted in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.

Figure 1. Forest plot presenting the association between patient- and contact-related characteristics and the likelihood of having a video contact (incidence rate ratios [IRRs] with 95% CIs). An IRR >1 indicates a higher use of video contacts compared to the reference group, marked in the right column of the figure. Conversely, an IRR <1 indicates a lower use of video contacts.



Ethical Considerations

The Committee on Health Research Ethics in the Central Denmark Region approved the data collection from the electronic patient records in the OOH-PC registration system (1-45-70-22-22) without informed consent from participants or any provision for them to opt out. The study was listed in the record of processing activities at the Research Unit for General Practice in Aarhus in accordance with the provisions of the General Data Protection Regulation (GDPR). All data on participants were deidentified. Finally, conduct of the study was endorsed by the regional association of GPs.

Results

Study Population

During the study period, 2,900,566 telephone triage contacts to OOH-PC were identified (Table 1). Patient- and contact-related characteristics varied between telephone and video contacts; the largest variation was seen for patient age, comorbidity, employment status, region, and time of contact.

Table . Distribution of patient- and contact-related characteristics (N=2,900,566).

Characteristics		Telephone contacts (n=2,625,363, 90.5%), n (%)	Video contacts (n=275,203, 9.5%), n (%)	Total (n=2,900,566, 100%), n (%)
Sex				
	Female	1,427,918 (54.4)	140,684 (51.1)	1,568,602 (54.1)
	Male	1,197,445 (45.6)	134,519 (48.9)	1,331,964 (45.9)
Age (years)				
	0-4	311,749 (11.9)	83,672 (30.4)	395,421 (13.6)
	5-10	138,745 (5.3)	25,196 (9.2)	163,941 (5.7)
	11-20	310,179 (11.8)	40,830 (14.8)	351,009 (12.1)
	21-40	688,182 (26.2)	64,780 (23.5)	752,962 (26)
	41-60	508,558 (19.4)	40,955 (14.9)	549,513 (18.9)
	61-80	433,400 (16.5)	16,042 (5.8)	449,442 (15.5)
	≥81	234,550 (8.9)	3728 (1.4)	238,278 (8.2)
Cohabitation status				
	Single	942,003 (35.9)	70,266 (25.5)	1,012,269 (34.9)
	Cohabiting	488,499 (18.6)	69,928 (25.4)	558,427 (19.3)
	Married	1,194,861 (45.5)	135,009 (49.1)	1,329,870 (45.8)
Comorbidities (n)				
	None	1,946,569 (74.1)	241,173 (87.6)	2,187,742 (75.4)
	1	426,677 (16.3)	27,329 (9.9)	454,006 (15.7)
	2	155,654 (5.9)	4647 (1.7)	160,301 (5.5)
	≥3	96,463 (3.7)	2054 (0.8)	98,517 (3.4)
Education (years)				
	<10	601,787 (22.9)	42,890 (15.6)	644,677 (22.2)
	10-15	1,175,364 (44.8)	124,310 (45.2)	1,299,674 (44.8)
	>15	807,635 (30.8)	105,160 (38.2)	912,795 (31.5)
	Unknown	40,577 (1.5)	2843 (1)	43,420 (1.5)
Ethnicity				
	Non-Western	195,638 (7.4)	21,790 (7.9)	217,428 (7.5)
	Western, not born in Denmark	54,761 (2.1)	5773 (2.1)	60,534 (2.1)
	Native, born in Denmark	2,325,363 (90.5)	247,640 (90)	2,622,604 (90.4)
Income (quintiles)				
	1	487,441 (18.6)	50,997 (18.5)	538,438 (18.6)
	2	598,552 (22.8)	45,088 (16.4)	643,640 (22.2)
	3	586,618 (22.3)	62,783 (22.8)	649,401 (22.4)
	4	540,238 (20.6)	65,139 (23.7)	605,377 (20.9)
	5	405,990 (15.5)	50,674 (18.4)	456,664 (15.7)
	Negative or zero	6524 (0.2)	522 (0.2)	7046 (0.2)
Urbanization (population)				
	>100,000	507,926 (19.4)	54,166 (19.7)	562,092 (19.4)
	20,000-100,000	655,472 (25)	67,516 (24.5)	722,988 (24.9)
	1,000-20,000	867,341 (33)	85,347 (31)	952,688 (32.9)
	<1,000	593,691 (22.6)	68,121 (24.8)	661,812 (22.8)

Characteristics		Telephone contacts (n=2,625,363, 90.5%), n (%)	Video contacts (n=275,203, 9.5%), n (%)	Total (n=2,900,566, 100%), n (%)
	Unplaceable	933 (0)	53 (0)	986 (0)
Employment status				
	Unemployed	332,428 (12.7)	27,305 (9.9)	359,733 (12.4)
	Retired	252,098 (20)	11,941 (4.3)	537,039 (18.5)
	Employed	1,767,837 (67.3)	235,957 (85.8)	2,003,794 (69.1)
Time of contact (hours until opening time of daytime general practice)				
	>4 before office hours	2,527,937 (96.3)	268,668 (97.6)	2,796,605 (96.4)
	<4 before office hours	97,426 (3.7)	6535 (2.4)	103,961 (3.6)
Region				
	North Denmark Region	405,869 (15.5)	34,556 (12.5)	440,415 (15.2)
	Central Denmark Region	989,549 (37.7)	88,821 (32.3)	1,078,370 (37.2)
	Region of Southern Denmark	946,931 (36.1)	134,029 (48.7)	1,080,960 (37.3)
	Region Zealand	283,014 (10.7)	17,807 (6.5)	300,821 (10.4)

Proportion of Video Contacts

During the study period, 9.5% (275,203/2,900,566) of telephone triage contacts to OOH-PC were video contacts. After the introduction of video, a range of 5%-15% video contacts was achieved within weeks across all regions. This level remained stable throughout the study period (data not shown).

Association Between Video Contact and Patient- and Contact-Related Characteristics

The frequency of video contacts was unevenly distributed across patient- and contact-related characteristics. The strongest associations were seen for age, comorbidity, employment status, region, and time of contact (Figure 1). Patients aged <20 years had a notably higher frequency of video contacts than patients aged 21 to 40 years (aIRR range: 2.39-1.31). This was also the case for employed compared to unemployed patients (aIRR 1.21). The frequency of video contacts was significantly higher for contacts to OOH-PC at more than 4 hours before the opening of daytime general practice (aIRR 1.40; reference: ≤4 hours), and was more frequent in the Region of Southern Denmark (aIRR 1.55; reference: North Denmark Region). In contrast, the frequency of video contacts was significantly lower for patients >40 years (aIRR range: 0.40-0.90; reference: 21-40 years), patients with comorbidities (aIRR range 0.59-0.90; reference: no comorbidities), and retired patients (aIRR 0.66; reference: unemployed). The frequency of video contacts was also significantly lower in Region Zealand (aIRR 0.73; reference: North Denmark Region).

Triage Outcomes

Patients receiving a video contact had a significantly higher frequency of ending the contact with advice and self-care compared to patients receiving a telephone contact (aIRR 1.21, 95% CI 1.21-1.21) (Table 2). Conversely, patients receiving a video contact had a significant lower frequency of being referred to a clinic consultation (aIRR 0.59, 95% CI 0.59-0.60) or a home visit compared to patients receiving a telephone contact (aIRR 0.31, 95% CI 0.29-0.32). The frequency of being admitted to a hospital was significantly higher after a video contact compared to a telephone contact (aIRR 1.20, 95% CI 1.17-1.23).

Table . Frequency of triage outcomes and their association with video contacts (incidence rate ratio).

Outcome	Telephone contacts (n=2,625,363), n (%)	Video contacts (n=275,203), n (%)	Total (n=2,900,566), n (%)	Incidence rate ratio (95% CI)	
				Crude	Adjusted[a]
Advice and self-care	1,663,681 (63.4)	215,484 (78.3)	1,879,567 (64.8)	1.24 (1.23-1.24)	1.21 (1.21-1.21)
Clinic consultation	712,255 (27.1)	49,262 (17.9)	759,948 (26.2)	0.66 (0.66-0.67)	0.59 (0.59-0.60)
Home visit	165,052 (6.3)	1926 (0.7)	168,233 (5.8)	0.12 (0.11-0.12)	0.31 (0.29-0.32)
Hospital admission	84,375 (3.2)	8531 (3.1)	92,818 (3.2)	0.95 (0.93-0.97)	1.20 (1.17-1.23)

[a]Adjusted for patient sex, age, cohabitation status, comorbidity, educational level, ethnicity, income, urbanization, employment status, region, and time of contact.

Follow-Up Contacts

In general, patients receiving a video contact had a significantly higher frequency of no follow-up contact compared to patients receiving a telephone contact (aIRR 1.09, 95% CI 1.08-1.09) (Table 3). For those who had a follow-up contact, the patients who received a video contact had a significantly higher frequency of having a follow-up contact with their regular GP compared to those receiving a telephone contact (aIRR 1.02, 95% CI 1.01-1.03). Conversely, patients receiving a video contact had a significant lower frequency of a follow-up contact in OOH-PC (aIRR 0.96, 95% CI 0.95-0.97) or at the hospital (aIRR 0.75, 95% CI 0.74-0.76) compared to patients receiving a telephone contact.

Table . Frequency of follow-up contacts and association between use of video contacts and subsequent follow-up contacts (incidence rate ratio).

Type of follow-up contact	Telephone contacts (n=2,625,363), n (%)	Video contacts (n=275,203), n (%)	Total (n=2,900,566)	Incidence rate ratio (95% CI)	
				Crude	Adjusted[a]
No follow-up	1,097,402 (41.8)	137,601 (50)	1,232,741 (42.5)	1.20 (1.19-1.20)	1.09 (1.08-1.09)
Daytime general practice[b]	719,349 (27.4)	70,728 (25.7)	791,854 (27.3)	0.94 (0.93-0.94)	1.02 (1.01-1.03)
OOH-PC[c,d]	396,430 (15.1)	37,703 (13.7)	435,085 (15)	0.90 (0.90-0.91)	0.96 (0.95-0.97)
Hospital[e]	412,182 (15.7)	29,171 (10.6)	440,886 (15.2)	0.68 (0.67-0.69)	0.75 (0.74-0.76)

[a]Adjusted for patient's sex and age, cohabitation status, comorbidity, educational level, ethnicity, income, urbanization, employment status, region, and time of contact.

[b]Contacts (telephone contacts, video contacts, clinic consultations, or home visits) to daytime general practice within 7 days from the index contact to OOH-PC.

[c]OOH-PC: out-of-hours primary care.

[d]All telephone triage contacts to OOH-PC within 7 days from the index contact to OOH-PC.

[e]All nonscheduled hospital contacts (emergency department visits and hospital admissions) within 1 day from the index contact to OOH-PC.

Discussion

Principal Results

Video was used in 9.5% (275,203/2,900,566) of all telephone triage contacts to OOH-PC. Video contacts were unevenly distributed across patient- and contact-related characteristics; video contacts were more often used for patients who were employed, young, without comorbidities, and contacting OOH-PC more than 4 hours before the opening hours of daytime general practice. Compared to telephone contacts, significantly more video contacts ended with advice and self-care and significantly fewer had follow-up contacts.

Strengths and Limitations

This study was based on a large data set, including codes for remuneration by GPs. The economic incentive for GPs to register all services provided contributed to the completeness of the data, though validity has not been studied [22].

Our study also had some limitations. First, we had no information on the reasons for encounters (RFEs), as this is not systematically registered in OOH-PC contacts. In each telephone triage contact, the triage GP assessed the relevance of video use based on the current RFE balanced against the specific patient- and contact-related characteristics. Therefore, telephone contacts and video contacts had different diagnostic scope, which could have influenced the differences found in triage outcome and follow-up contacts through confounding by indication. Second, we followed each patient for 7 days to record follow-up contacts to OOH-PC and to daytime general practice, as previously described in the literature [26]. This led to an overestimation of follow-up contacts, as we could not link these follow-up contacts to the index contact in OOH-PC using the RFE. However, any overestimation would be independent of type of contact. Finally, we used the Charlson Comorbidity Index to define comorbidity based on hospital diagnosis codes. This approach might have led to an underestimation of comorbidity [27], as patients with mild chronic diseases are often treated solely in general practice.

Several factors must be considered when generalizing the results of this study. First, the study period was defined according to the date of initiation of video contact in each of the regions. Therefore, the regions were included in different periods of the COVID-19 pandemic, and they had different contact patterns to primary care both inside and outside office hours [8,12,20,28] and probably also different distributions of triage outcomes. Second, triage GPs perform telephone triage with no decision support tool in Danish OOH-PC; this is unlike most countries with comparable OOH-PC services, which often use other health care professionals with decision support systems [3]. Compared to other triage professionals, GPs may be able to triage more patients via video contact. Lastly, Danish triage GPs were remunerated on a fee-for-service basis. As the fee for a video contact was higher than the fee for a telephone contact, this could have been an incentive to aim for a higher share of video contacts in this setting compared to countries with other payment structures.

Comparison With Prior Work

We found a 9.5% rate of use of video contacts to OOH-PC. To our knowledge, no previous studies used a data collection period of this length to report on video use in OOH-PC. Studies on

changing contact patterns in OOH-PC during the COVID-19 pandemic have found an overall increase in telehealth consultations (email, video, or telephone) [8,12,28]. Video use in daytime general practice has previously been reported to range from 1% to 6.4% [15,29-32]. However, as patient populations and RFEs are known to differ between daytime general practice and OOH-PC [33], these results cannot be compared with our findings. Furthermore, video contacts in OOH-PC guide triage professionals in the assessment of patients and in improving patient reassurance [13,14]. In contrast, video contact has often been used as a substitute for clinic consultations in daytime general practice for practical reasons, for example, to reduce travel time or limit the risk of contamination, but both patients and GPs seem to prefer in-person consultations in the postpandemic era [10,13,18].

Our study showed that video contacts were used more often for employed young patients without comorbidities. To the best of our knowledge, this is the first study to report on associations between patient- and contact-related characteristics and video contacts in OOH-PC. Studies conducted in daytime general practice have found higher video rates of use during COVID-19 lockdown periods [34] and among people from socioeconomically advantaged areas [34,35]. Previous studies have reported inconsistent results on the association between patient age and video use, as higher use has been reported for both younger [32,35] and older patients [30]. Moreover, daytime video use seems to be associated with patients with high morbidity [36] compared to patients with low morbidity. These findings are not in line with our study results, which could be due to differences in patient populations between daytime general practice and OOH-PC [33]. Furthermore, some previous studies were conducted during the peak of the COVID-19 pandemic, and different countries have different health care systems and had different approaches to tackling the pandemic.

We found that video contacts more often ended with advice and self-care and no follow-up contact compared to telephone contacts. Two qualitative studies investigating the effect of video contacts on the patient flow in daytime general practice found that GPs experience uncertainties when referring patients to secondary care after a video contact [9,37]. A UK study on follow-up contact after using a video contact service (used by hospitals, daytime general practices, and other services) found no significant difference in the number of subsequent referrals compared to telephone contacts [38]. However, these studies focused on video use in the daytime rather than on telephone triage in OOH-PC.

Implications for Practice and Future Research

Our study suggests that video contacts could help reduce the use of health care resources in the OOH-PC setting by lowering the number of subsequent clinic consultations and home visits. More studies are needed on the effect of video contact on patient flow. First, further research is needed to investigate the impact of video contact in relation to different RFEs. Second, future studies should explore if the findings of this study are maintained in the postpandemic period and across different OOH-PC organizations. Third, future studies should investigate if the video option might generate more contacts to OOH-PC overall. Fourth, our study indicates an association between video contacts and specific patient characteristics: video was more often used for employed young patients without comorbidities. This finding contrasts with most studies in daytime general practice and should be further investigated. Finally, it is important to note that we did not study costs associated with video use and its effects on resource use. Therefore, future studies should investigate the costs as well.

Conclusion

This study supports our hypothesis that video contacts could reduce use of health care resources in OOH-PC. Video use lowered the frequency of referrals to a clinic consultation or home visit and also lowered the frequency of follow-up contacts. However, the results could be biased due to confounding by indication, reflecting that triage GPs use video for a specific set of RFEs.

Acknowledgments

We gratefully acknowledge the financial support for this study provided by the Danish Health Insurance Foundation (Sygeforsikringen danmark), the General Practice Research Foundation of the Central Denmark Region (Praksisforskningsfonden), and the Department of Public Health, Aarhus University. The funding bodies had no role in the study design, data collection, data analysis, data interpretation, writing of the manuscript, or submission of the final article.

Authors' Contributions

All authors contributed to the study design, interpretation of results, and drafting and revising of the manuscript. MAN and CHV conducted the data management and the statistical analysis. All authors have agreed to the final submitted version of the manuscript.

Conflicts of Interest

None declared.

References

1. Pedersen KM, Andersen JS, Søndergaard J. General practice and primary health care in Denmark. J Am Board Fam Med 2012 Mar;25 Suppl 1:S34-S38. [doi: 10.3122/jabfm.2012.02.110216] [Medline: 22403249]

2. Smits M, Rutten M, Keizer E, Wensing M, Westert G, Giesen P. The development and performance of after-hours primary care in the Netherlands: a narrative review. Ann Intern Med 2017 May 16;166(10):737-742. [doi: 10.7326/M16-2776] [Medline: 28418455]
3. Steeman L, Uijen M, Plat E, Huibers L, Smits M, Giesen P. Out-of-hours primary care in 26 European countries: an overview of organizational models. Fam Pract 2020 Nov 28;37(6):744-750. [doi: 10.1093/fampra/cmaa064] [Medline: 32597962]
4. Drerup B, Espenschied J, Wiedemer J, Hamilton L. Reduced no-show rates and sustained patient satisfaction of telehealth during the COVID-19 pandemic. Telemed J E Health 2021 Dec;27(12):1409-1415. [doi: 10.1089/tmj.2021.0002] [Medline: 33661708]
5. Green MA, McKee M, Katikireddi SV. Remote general practitioner consultations during COVID-19. Lancet Digit Health 2022 Jan;4(1):e7. [doi: 10.1016/S2589-7500(21)00279-X] [Medline: 34952678]
6. Johnsen TM, Norberg BL, Kristiansen E, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 8;23(2):e26433. [doi: 10.2196/26433] [Medline: 33465037]
7. Saint-Lary O, Gautier S, Le Breton J, et al. How GPs adapted their practices and organisations at the beginning of COVID-19 outbreak: a French national observational survey. BMJ Open 2020 Dec 2;10(12):e042119. [doi: 10.1136/bmjopen-2020-042119] [Medline: 33268433]
8. Sigurdsson EL, Blondal AB, Jonsson JS, et al. How primary healthcare in Iceland swiftly changed its strategy in response to the COVID-19 pandemic. BMJ Open 2020 Dec 7;10(12):e043151. [doi: 10.1136/bmjopen-2020-043151] [Medline: 33293329]
9. Wherton J, Greenhalgh T, Shaw SE. Expanding video consultation services at pace and scale in Scotland during the COVID-19 pandemic: national mixed methods case study. J Med Internet Res 2021 Oct 7;23(10):e31374. [doi: 10.2196/31374] [Medline: 34516389]
10. Due TD, Thorsen T, Andersen JH. Use of alternative consultation forms in Danish general practice in the initial phase of the COVID-19 pandemic - a qualitative study. BMC Fam Pract 2021 Jun 2;22(1):108. [doi: 10.1186/s12875-021-01468-y] [Medline: 34078281]
11. Joy M, McGagh D, Jones N, et al. Reorganisation of primary care for older adults during COVID-19: a cross-sectional database study in the UK. Br J Gen Pract 2020 Aug;70(697):e540-e547. [doi: 10.3399/bjgp20X710933] [Medline: 32661009]
12. Ramerman L, Rijpkema C, Bos N, Flinterman LE, Verheij RA. The use of out-of-hours primary care during the first year of the COVID-19 pandemic. BMC Health Serv Res 2022 May 21;22(1):679. [doi: 10.1186/s12913-022-08096-x] [Medline: 35597939]
13. Greenhalgh T, Ladds E, Hughes G, et al. Why do GPs rarely do video consultations? Qualitative study in UK general practice. Br J Gen Pract 2022 May;72(718):e351-e360. [doi: 10.3399/BJGP.2021.0658] [Medline: 35256385]
14. Gren C, Egerod I, Linderoth G, et al. “We can't do without it”: parent and call-handler experiences of video triage of children at a medical helpline. PLoS One 2022;17(4):e0266007. [doi: 10.1371/journal.pone.0266007] [Medline: 35421109]
15. Assing Hvidt E, Christensen NP, Grønning A, Jepsen C, Lüchau EC. What are patients’ first-time experiences with video consulting? A qualitative interview study in Danish general practice in times of COVID-19. BMJ Open 2022 Apr 15;12(4):e054415. [doi: 10.1136/bmjopen-2021-054415] [Medline: 35428624]
16. Mold F, Hendy J, Lai YL, de Lusignan S. Electronic consultation in primary care between providers and patients: systematic review. JMIR Med Inform 2019 Dec 3;7(4):e13042. [doi: 10.2196/13042] [Medline: 31793888]
17. Koch S, Guhres M. Physicians’ experiences of patient-initiated online consultations in primary care using direct-to-consumer technology. Stud Health Technol Inform 2020 Jun 16;270:643-647. [doi: 10.3233/SHTI200239] [Medline: 32570462]
18. Meurs M, Keuper J, Sankatsing V, Batenburg R, van Tuyl L. “Get used to the fact that some of the care is really going to take place in a different way”: general practitioners' experiences with e-health during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Apr 22;19(9):5120. [doi: 10.3390/ijerph19095120] [Medline: 35564519]
19. Nordtug M, Assing Hvidt E, Lüchau EC, Grønning A. General practitioners’ experiences of professional uncertainties emerging from the introduction of video consultations in general practice: qualitative study. JMIR Form Res 2022 Jun 14;6(6):e36289. [doi: 10.2196/36289] [Medline: 35653607]
20. Huibers L, Bech BH, Kirk UB, Kallestrup P, Vestergaard CH, Christensen MB. Contacts in general practice during the COVID-19 pandemic: a register-based study. Br J Gen Pract 2022 Nov;72(724):e799-e808. [doi: 10.3399/BJGP.2021.0703] [Medline: 36253113]
21. Pedersen CB. The Danish civil registration system. Scand J Public Health 2011 Jul;39(7 Suppl):22-25. [doi: 10.1177/1403494810387965] [Medline: 21775345]
22. Andersen JS, Olivarius NDF, Krasnik A. The Danish national health service register. Scand J Public Health 2011 Jul;39(7 Suppl):34-37. [doi: 10.1177/1403494810394718] [Medline: 21775348]
23. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish national patient registry: a review of content, data quality, and research potential. Clin Epidemiol 2015 Nov;7:449-490. [doi: 10.2147/CLEP.S91125] [Medline: 26604824]

24. Thygesen LC, Daasnes C, Thaulow I, Brønnum-Hansen H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. Scand J Public Health 2011 Jul;39(7 Suppl):12-16. [doi: 10.1177/1403494811399956] [Medline: 21898916]
25. Nguyen AD, Frensham LJ, Baysari MT, Carland JE, Day RO. Patients' use of mobile health applications: what general practitioners think. Fam Pract 2019 Mar 20;36(2):214-218. [doi: 10.1093/fampra/cmy052] [Medline: 29873708]
26. van Uden CJT, Zwietering PJ, Hobma SO, et al. Follow-up care by patient's own general practitioner after contact with out-of-hours care. A descriptive study. BMC Fam Pract 2005 Jun 9;6(1):23. [doi: 10.1186/1471-2296-6-23] [Medline: 15946382]
27. Prior A, Fenger-Grøn M, Larsen KK, et al. The association between perceived stress and mortality among people with multimorbidity: a prospective population-based cohort study. Am J Epidemiol 2016 Aug 1;184(3):199-210. [doi: 10.1093/aje/kwv324] [Medline: 27407085]
28. Morreel S, Philips H, Verhoeven V. Organisation and characteristics of out-of-hours primary care during a COVID-19 outbreak: a real-time observational study. PLoS One 2020;15(8):e0237629. [doi: 10.1371/journal.pone.0237629] [Medline: 32790804]
29. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857. [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
30. Murphy M, Scott LJ, Salisbury C, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021 Feb;71(704):e166-e177. [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
31. Chang JE, Lindenfeld Z, Albert SL, et al. Telephone vs. video visits during COVID-19: safety-net provider perspectives. J Am Board Fam Med 2021;34(6):1103-1114. [doi: 10.3122/jabfm.2021.06.210186] [Medline: 34772766]
32. Dai Z, Sezgin G, Hardie RA, et al. Sociodemographic determinants of telehealth utilisation in general practice during the COVID-19 pandemic in Australia. Intern Med J 2023 Mar;53(3):422-425. [doi: 10.1111/imj.16006] [Medline: 36624629]
33. Huibers L, Moth G, Bondevik GT, et al. Diagnostic scope in out-of-hours primary care services in eight European countries: an observational study. BMC Fam Pract 2011 May 13;12:30. [doi: 10.1186/1471-2296-12-30] [Medline: 21569483]
34. Savira F, Orellana L, Hensher M, et al. Use of general practitioner telehealth services during the COVID-19 pandemic in regional Victoria, Australia: retrospective analysis. J Med Internet Res 2023 Feb 7;25:e39384. [doi: 10.2196/39384] [Medline: 36649230]
35. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26. [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
36. Glazier RH, Green ME, Wu FC, Frymire E, Kopp A, Kiran T. Shifts in office and virtual primary care during the early COVID-19 pandemic in Ontario, Canada. CMAJ 2021 Feb 8;193(6):E200-E210. [doi: 10.1503/cmaj.202303] [Medline: 33558406]
37. Randhawa RS, Chandan JS, Thomas T, Singh S. An exploration of the attitudes and views of general practitioners on the use of video consultations in a primary healthcare setting: a qualitative pilot study. Prim Health Care Res Dev 2019 Jan;20:e5. [doi: 10.1017/S1463423618000361] [Medline: 29909798]
38. Smith C, Kubanova B, Ahmed F, Manickavasagam J. The effectiveness of remote consultations during the COVID-19 pandemic: a tool for modernising the national health service (NHS). Cureus 2022 Dec;14(12):e32301. [doi: 10.7759/cureus.32301] [Medline: 36627990]

Abbreviations

aIRR: adjusted incidence rate ratio
GDPR: General Data Protection Regulation
GP: general practitioner
ICPC-2: International Classification of Primary Care, 2nd Edition
IRR: incidence rate ratio
OOH-PC: out-of-hours primary care
RFE: reason for encounter
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by C Lovis; submitted 10.03.23; peer-reviewed by F Savira, T Koskela; revised version received 07.02.24; accepted 10.02.24; published 04.04.24.

Please cite as:
Nebsbjerg MA, Vestergaard CH, Bomholt KB, Christensen MB, Huibers L
Use of Video in Telephone Triage in Out-of-Hours Primary Care: Register-Based Study
JMIR Med Inform 2024;12:e47039
URL: https://medinform.jmir.org/2024/1/e47039
doi:10.2196/47039

Telehealth Uptake Among Hispanic People During COVID-19: Retrospective Observational Study

Di Shang[1,*], PhD; Cynthia Williams[2,*], PhD; Hera Culiqi[1,*]

[1]

[2]

[*]all authors contributed equally

Corresponding Author:
Cynthia Williams, PhD

Abstract

Background: The Hispanic community represents a sizeable community that experiences inequities in the US health care system. As the system has moved toward digital health platforms, evaluating the potential impact on Hispanic communities is critical.

Objective: The study aimed to investigate demographic, socioeconomic, and behavioral factors contributing to low telehealth use in Hispanic communities.

Methods: We used a retrospective observation study design to examine the study objectives. The COVID-19 Research Database Consortium provided the Analytics IQ PeopleCore consumer data and Office Alley claims data. The study period was from March 2020 to April 2021. Multiple logistic regression was used to determine the odds of using telehealth services.

Results: We examined 3,478,287 unique Hispanic patients, 16.6% (577,396) of whom used telehealth. Results suggested that patients aged between 18 and 44 years were more likely to use telehealth (odds ratio [OR] 1.07, 95% CI 1.05-1.1; *P*<.001) than patients aged older than 65 years. Across all age groups, patients with high incomes were at least 20% more likely to use telehealth than patients with lower incomes (*P*<.001); patients who had a primary care physician (*P*=.01), exhibited high medical usage (*P*<.001), or were interested in exercise (*P*=.03) were more likely to use telehealth; patients who had unhealthy behaviors such as smoking and alcohol consumption were less likely to use telehealth (*P*<.001). Male patients were less likely than female patients to use telehealth among patients aged 65 years and older (OR 0.94, 95% CI 0.93-0.95; *P*<.001), while male patients aged between 18 and 44 years were more likely to use telehealth (OR 1.05, 95% CI 1.03-1.07; *P*<.001). Among patients younger than 65 years, full-time employment was positively associated with telehealth use (*P*<.001). Patients aged between 18 and 44 years with high school or less education were 2% less likely to use telehealth (OR 0.98, 95% CI 0.97-0.99; *P*=.005). Results also revealed a positive association with using WebMD (WebMD LLC) among patients aged older than 44 years (*P*<.001), while there was a negative association with electronic prescriptions among those who were aged between 18 and 44 years (*P*=.009) and aged between 45 and 64 years (*P*=.004).

Conclusions: This study demonstrates that telehealth use among Hispanic communities is dependent upon factors such as age, gender, education, socioeconomic status, current health care engagement, and health behaviors. To address these challenges, we advocate for interdisciplinary approaches that involve medical professionals, insurance providers, and community-based services actively engaging with Hispanic communities and promoting telehealth use. We propose the following recommendations: enhance access to health insurance, improve access to primary care providers, and allocate fiscal and educational resources to support telehealth use. As telehealth increasingly shapes health care delivery, it is vital for professionals to facilitate the use of all available avenues for accessing care.

(JMIR Med Inform 2024;12:e57717) doi:10.2196/57717

KEYWORDS

telehealth; telemedicine; ICT; eHealth; e-health; Hispanic; health equity; health access; Hispanics; digital divide; usage; utilization; equity; inequity; inequities; access; accessibility; Spanish; observational; demographic; demographics; socioeconomic; socioeconomics; information and communication technology

Introduction

Health inequities among people of racial and ethnic minority groups are a significant concern across the United States of America. The Hispanic or Latino (hereafter “Hispanic”) population is a community that experiences inequities in access to care. However, health inequities in the United States of America are primarily compared between non-Hispanic White and non-Hispanic Black populations, with some positioning of

the Hispanic population. Therefore, we must highlight the needs of the Hispanic community as a matter of priority. A previous study suggested that health inequities between non-Hispanic Black and Hispanic people are similar in health risks and outcomes [1]. Boen and Hummer [1] attributed much of the similarities to the influence of socioeconomic status and stress in these communities. Social determinants of health significantly impact the health and wellness of people of color. In addition, the inequity in health outcomes is confounded by challenges in access to care. Gaining access to and using health care services is essential to mitigating adverse health outcomes and promoting equity in quality care [2]. Low access to health care services is due to a complexity of factors, which include a lack of insurance, education, and fiscal resources [2]. Non-Hispanic White women and men experienced greater usage of care, 83% and 74%, respectively; among Hispanic individuals, women had higher usage rates than men, 68% and 50%, respectively; however, they still fall short of their non-Hispanic White counterparts [2].

While there are considerable efforts to decrease inequities in health care access, recent events such as the proliferation of technology in health care and the global COVID-19 pandemic eclipse equity in access efforts. COVID-19 accelerated the use of telehealth, which was essential to access health care information and services [3]. Telehealth was highly encouraged to mitigate the ill effects of the contagion and support lockdown measures. While other racial and ethnic communities experienced telehealth growth, the Hispanic population experienced relatively lower amounts of telehealth increases and had a disproportioned rise in the use of health care resources [4-7]. When compared to non-Hispanic White individuals, Hispanic patients had lower usage of telehealth services (27% and 6.8%, respectively) [8]. Telehealth usage among Hispanic people fell below that of their non-Hispanic Black counterparts. Hispanic people showed lower adjusted odds of using telehealth when compared to non-Hispanic Black persons before and during the pandemic [9,10]. However, Hispanic patients and people in low-income positions had increased levels of telehealth use during the pandemic [11]. People who experience inequities in access are more likely to have unmet health care needs and have a higher prevalence of morbidity and mortality rates [12]. The lack of telehealth usage was a critical issue during this time. Qian et al [13] found higher rates of COVID-19 cases in regions with low telehealth use. Given the lack of telehealth usage and reduced access to in-person services during the pandemic, it is reasonable to assume that overall health outcomes were compromised.

Research suggested that telehealth use differed by demographics such as gender, age, and education [14]. Studies have noted gender differences in telehealth use, as male participants are less likely to use telehealth than female participants [3,9,15,16]. Among racial and ethnic groups, Hispanic patients often opt for in-person rather than telehealth visits [5,10]. Whether a patient has access to and knowledge to use telehealth depends on socioeconomic factors [5]. Ramirez et al [5] observed that the Hispanic community had lower telehealth use due to cultural perceptions, inadequate financial resources, and digital literacy barriers. Health inequities are concerning, as research suggests that health outcomes are favorable across racial and ethnic groups when services are used [17,18]. The persistence of inequity in health care suggests that access to care is complex and should consider economic and social factors. To effectively incorporate factors influencing care, we comprehensively assess the components contributing to lower usage among Hispanic people. This study aims to examine the socioeconomic, demographic, and behavioral factors that influence telehealth use among people who identify as Hispanic.

Methods

Data Source

The study period was from March 2020 to April 2021. The COVID-19 Research Database Consortium provided access to the Office Ally and Analytics IQ PeopleCore consumer databases. The Office Ally database provided access to US claims data from 100 million unique patients and 3.4 billion medical claims. The Analytics IQ PeopleCore consumer database provides individual-level data across demographics, behaviors, and economic indicators and is a national representation of 242.5 million US adults aged 19 years and older. We joined the Office Ally claims data with Analytics IQ PeopleCore consumer data through an identifier, which enabled us to retrieve patient claims data during the study period and examine telehealth usage from socioeconomic and behavioral perspectives, as shown in the “Results” section.

Ethical Considerations

The COVID-19 Research Database was established by complying with regulatory standards to protect patient privacy. The COVID-19 Research Database received a waiver of patient consent certified from the Western Institutional Review Board for using HIPAA (Health Insurance Portability and Accountability Act)–certified deidentified data on April 20, 2020 [19]. The Western Institutional Review Board granted exemption status for HIPAA-limited data sets and non-HIPAA–covered data on May 14, 2020. This exemption covers all research performed in the COVID-19 Research Database. In addition, researchers with approved study proposals are granted access only to specific data sets necessary to answer their research question or questions. Only deidentified and limited data sets are made available through the database and certified before access is granted. Individual project institutional board approval was optional.

Results

Overview

We retrieved 16.43 million Office Ally claim records of Hispanic patients during the study period, and 3,478,287 unique Hispanic patients were included to investigate telehealth usage. A descriptive summary of the patients included in the analyses is in Table 1. Telehealth claims were identified by screening for the procedure modifier codes 95, GT, and GQ. Among the patients, 16.6% (578,945/3,478,287) had one or more telehealth claims during the study period. Female patients had slightly higher telehealth usage (338,795/1,958,350, 17.3%) than male patients (240,150/1,519,937, 15.8%).

Table . Characteristics of participants in the study (N=3,478,287).

Characteristics		Value
Sex, n (%)		
	Female	1,958,350 (56.3)
	Male	1,519,937 (43.7)
Age group (years), n (%)		
	18-44	1,133,417 (32.6)
	45-64	1,326,391 (38.1)
	65+	1,018,479 (29.3)
Education qualification, n (%)		
	Bachelor or higher	922,435 (26.5)
	High school or less	2,555,852 (73.5)
Employment status, n (%)		
	Unemployed	1,227,541 (35.3)
	Part-time	634, 818 (18.2)
	Full-time	1,615,928 (46.5)
Household annual income (US $), n (%)		
	Low: <46,000 (1st quartile)	941,294 (27.1)
	Medium: 46,000-144,000	1,670,747 (48)
	High: >114,000 (4th quartile)	866,246 (24.9)
Having a primary care doctor, n (%)		
	Yes	2,782,591 (80)
	No	695,696 (20)
Exhibit high medical use, n (%)		
	Yes	671,349 (19.3)
	No	2,806,938 (80.7)
Interest in exercise, n (%)		
	Yes	700,691 (20.1)
	No	2,777,596 (79.9)
Measure of alcohol consumption, n (%)		
	No consumption	353,921 (10.2)
	Some consumption	2,855,276 (82.1)
	High consumption	269,090 (7.7)
Frequency of smoking, n (%)		
	Never	413,982 (11.9)
	Some	2,919,447 (83.9)
	Daily	144,858 (4.2)
Using electronic prescription services, n (%)		
	Yes	578,768 (16.6)
	No	2,899,519 (83.4)
Using WebMD, n (%)		
	Yes	828,260 (23.8)
	No	2,650,027 (76.2)
Total number of claims during the study period, mean (SD)		4.73 (4.71)

Logistic Regression Analysis

The data were aggregated at the patient level to investigate telehealth use and determine whether a patient used telehealth during the study period. A patient with one or more telehealth claims during the study period was assigned a value of 1 for the dependent variable; otherwise, the value was assigned as 0. Categorical variables were created to stratify patients into groups by their demographics and socioeconomic status, as listed in Table 1. We conducted a multiple logistic regression to determine the odds of patients using telehealth. Each patient's total number of claims during the study period was included as an offset variable in the logistic regression to control its potential impact on the dependent variable. Results suggested that compared to patients aged older than 65 years, patients aged between 18 and 44 years are 1.07 times (odds ratio [OR] 1.07, 95% CI 1.05-1.1; *P*<.001) likely to use telehealth, while patients aged between 45 and 64 years showed a nonsignificant difference (*P*=.49).

Results from our logistic regression analysis of the 3 age groups are shown in Table 2. Male patients in the older group (aged older than 65 years) are 6% less likely to use telehealth (OR 0.94, 95% CI 0.93-0.95; *P*<.001), while male patients in the young group (aged between 18 and 44 years) are 5% more likely to use telehealth (OR 1.05, 95% CI 1.03-1.07; *P*<.001). Patients with a primary care doctor (*P*=.01) or high medical usage (*P*<.001) are significantly more likely to use telehealth, especially in the patients aged older than 65 years group. Patients who use WebMD (WebMD LLC) are significantly (*P*<.001) more likely to use telehealth among those who are aged older than 44 years. In comparison, the negative association with electronic prescriptions is significant among those who are aged between 18 and 44 years (*P*=.009) and aged between 45 and 64 years (*P*=.004). Patients aged between 18 and 44 years with high school or less education are 2% less likely to use telehealth (OR 0.98, 95% CI 0.97-0.99; *P*=.005). Patients with high incomes across all age groups were more likely to use telehealth than patients with lower incomes, as follows: aged between 18 and 44 years (OR 1.25, 95% CI 1.23-1.28; *P*<.001), aged between 45 and 64 years (OR 1.33, 95% CI 1.3-1.35; *P*<.001), and aged older than 65 years (OR 1.2, 95% CI 1.18-1.22; *P*<.001). Patients younger than 65 years with full-time employment (*P*<.001) are significantly more likely to use telehealth.

Patients with unhealthy behaviors such as alcohol use and smoking are significantly less likely to use telehealth (*P*<.001). In the patients older than 65 years group, patients with high alcohol consumption are 39% less likely to use telehealth than patients with no alcohol consumption (OR 0.61, 95%CI 0.56-0.65; *P*<.001). Patients aged 65 years and older who smoke daily are 36% less likely to use telehealth than patients who never smoke (OR 0.64, 95% CI 0.6-0.69; *P*<.001). Meanwhile, patients interested in exercise were significantly more likely to use telehealth (*P*=.03).

Table . Odds ratios from logistic regression analysis by age groups.

Variables		Aged 65+ years	Aged between 45 and 64 years	Aged between 18 and 44 years
Sex: male (reference: female)				
	Odds ratio (95% CI)	0.94 (0.93-0.95)	0.99 (0.98-1.01)	1.05 (1.03-1.07)
	P value	<.001	.35	<.001
Primary care doctor (reference: no)				
	Odds ratio (95% CI)	1.19 (1.1-1.29)	1.02 (1-1.04)	1.06 (1.05-1.08)
	P value	.001	.01	<.001
Medical use (reference: no)				
	Odds ratio (95% CI)	1.09 (1.07-1.1)	1.07 (1.06-1.08)	1.03 (1.02-1.05)
	P value	<.001	<.001	<.001
WebMD (reference: no)				
	Odds ratio (95% CI)	1.07 (1.04-1.1)	1.05 (1.04-1.07)	1.01 (0.99-1.02)
	P value	<.001	<.001	.32
Electronic prescriptions (reference: no)				
	Odds ratio (95% CI)	1 (0.99-1.02)	0.97 (0.95-0.99)	0.93 (0.87-0.98)
	P value	.46	.004	.009
Education: high school or less (reference: bachelor or higher)				
	Odds ratio (95% CI)	1.01 (0.98-1.04)	1 (0.99-1.01)	0.98 (0.97-0.99)
	P value	.34	.61	.005
Employment: full-time (reference: unemployed)				
	Odds ratio (95% CI)	1.02 (0.97-1.07)	1.07 (1.06-1.09)	1.05 (1.02-1.08)
	P value	.40	<.001	<.001
Income: high (reference: low)				
	Odds ratio (95% CI)	1.25 (1.23-1.28)	1.33 (1.3-1.35)	1.2 (1.18-1.22)
	P value	<.001	<.001	<.001
Exercise fan: yes (reference: no)				
	Odds ratio (95% CI)	1.04 (1-1.08)	1.05 (1.03-1.07)	1.03 (1.02-1.05)
	P value	.03	<.001	<.001
Alcohol: high consumption (reference: no consumption)				
	Odds ratio (95% CI)	0.61 (0.56-0.65)	0.79 (0.77-0.82)	0.81 (0.78-0.85)
	P value	<.001	<.001	<.001
Smoking: daily (reference: never)				
	Odds ratio (95% CI)	0.64 (0.6-0.69)	0.85 (0.82-0.88)	0.93 (0.89-0.97)
	P value	<.001	<.001	<.001

Discussion

Overview

This study found that among Hispanic people, male participants aged between 18 and 44 years were more likely to use telehealth than female participants, but male participants aged older than 44 years were less likely to use telehealth than female participants. Across all age groups, people with high incomes, people with primary care physicians, current users of the health care system, people who used WebMD, and people who reported full-time employment were more likely to use telehealth. Patients aged between 18 and 44 years with high school or less education were 2% less likely to use telehealth. There was a negative association with electronic prescriptions among those aged between 18 and 44 years and aged between 45 and 64 years. In addition, regardless of age, people with unhealthy behaviors, such as smoking, alcohol consumption, and a lack of interest in exercise, were less likely to use telehealth services.

Telehealth and Demographic Factors

The presented analyses represent factors contributing to telehealth use among Hispanic people. Like other studies, people aged older than 65 years were less likely to use telehealth than people in younger age groups [3,9]. Male participants in the aged older than 65 years group were 6% less likely to use telehealth, while male participants in the youngest group were 5% more likely to use telehealth. Saeed and Masters [16] indicated that female participants have higher telehealth usage due to caregiver burdens that make attending in-person visits more challenging. Furthermore, our findings indicate that the positive influence of health care use factors (eg, having a primary care doctor, medical usage, and use of WebMD) and the detrimental effects of unhealthy behaviors (alcohol consumption and smoking) are more pronounced among patients aged older than 65 years compared to younger age groups. Our findings suggest the necessity of considering various age groups when examining usage differences between age and gender groups.

Telehealth and Socioeconomic Factors

This study's results on educational background align with other studies that suggested that people with more than a high school diploma have higher telehealth usage [18]. A lack of education contributes to low health and digital literacy and interferes with a person's inability to access and use health-related information [20,21]. This study suggests that people with a primary care physician and those using the health care system (medical usage) are more likely to use telehealth [9,22]. These characteristics describe people who already have access to the health care system and are perhaps representative of the segment of the Hispanic population that experiences more health-related equity [23]. This has implications for increasing convenient access to care rather than increasing access for people who do not already have access to care. Research suggested that a primary care physician was critical to accessing traditional, in-person services. Having access to a primary care physician provides an avenue to get telehealth services; therefore, we can reasonably speculate that people who have access to in-person health care will have access to telehealth [16,24]. People who live in low-income communities are particularly vulnerable due to a lack of resources [14]. Darrat et al [15] suggested that among people with incomes less than US $30,000 annually, 29% lacked a smartphone, 44% did not have home broadband access, and 46% did not own a computer. The lack of internet access exacerbates the inequity in service access [18]. Jain et al [25] suggested that 84% of telehealth users had broadband internet access [26]. A study by Chau et al [27] indicated that 30% of Hispanic people do not have a computer in their home and are 10 years behind non-Hispanic White people with regard to broadband internet access. Researchers suggested that cultural perspectives influence technology use even when Hispanic individuals have similar technologies [2,26,28]. The results of this study indicate that Hispanic persons who used the internet for health information, such as WebMD, were more likely to use telehealth services; however, there was no relationship with electronic prescription behavior. Haun et al [29] also suggested no statistically significant relationship exists between telehealth use and electronic prescription behavior, as the provider, not the patient, initiates this service. Among Hispanic patients who used telehealth, they had the highest rate of missed telehealth visits at 42% [3]. A study by Ghaddar et al [6] suggested that 60% of Hispanic individuals access the internet or send or receive emails. However, only 24% communicate electronically with health care providers, and 40% report having low digital literacy [6].

Telehealth and Health Behaviors

In this study, people with unhealthy behaviors, such as smoking, alcohol consumption, and a lack of interest in exercise, had significantly lower odds of using telehealth across all age groups. Researchers suggested that smoking disproportionately affects people of low income and educational status, and alcohol misuse is increasing among people of color and people older than 60 years [30,31]. Health behaviors often include factors such as smoking, drinking, and physical activity and contribute to health inequity [32,33]. In addition, health behaviors have a significant effect on health care usage [34,35]. The association between telehealth use and smoking and alcohol is not conclusive in the body of literature. Jaffe et al [36] suggested that smoking has no relationship with exercise behavior, alcohol use, smoking, or telehealth. They suggested that 12% of people who smoked had a telehealth visit, as compared to 61% of people who never smoked and had a telehealth visit; however, it was not statistically significant (P=.45) [36]. The authors also suggested that alcohol consumption was not a deterrent to telehealth use, as 69% of alcohol users experienced a telehealth visit. A study by Wegermann et al [3] indicated that there were minimal to no differences in telehealth use among people who reported alcohol use or smoking. However, another study by Kim [37] indicated that exercise and alcohol use were associated with telehealth acceptance, whereas smoking status was not. A study by Jagielo et al [38] noted that during the COVID-19 pandemic, the stay-at-home order was more predictive of telehealth use than race or ethnicity among smokers. Among people with alcohol use disorders, researchers reported no differences in preference for telehealth or in-person treatment [39,40]. While beyond the scope of this study, it is imperative to note the rapid usage of telehealth and other digital technologies for smoking and alcohol cessation programs and exercise promotion. Given the aforementioned socioeconomic status of people who engaged in unhealthy behaviors, it is not surprising that challenges were noted in access to technology, digital literacy, and quiet session locations [30].

Recommendations

Glasgow et al [41] suggested that socioeconomic factors, cultural perceptions, and patient preferences significantly impact health care use. By considering patient health behaviors and preferences, providers and decision-makers can support "individual health and public health through enhanced care" [41] and gain a comprehensive understanding of the complex factors contributing to low telehealth access. We make several recommendations based on the results of this and other studies. First, primary care providers serve as the point of care and cost-efficient entry into the health care system [42]. Primary care supports access to other physician specialties, and its services are relatively more amenable to telehealth compared

to other physician specialties (42% and 35%, respectively); 73% of primary care services could be offered through telehealth [43]. Community-based primary care clinics can close the health equity gap [44]. Primary care clinics are uniquely positioned to engage communities in the sociocontent of their environment and culture. Using this strategy, local clinics can promote culturally relevant educational strategies and encourage positive behavior change for health outcomes. Second, as we consider supporting telehealth in primary care, it is imperative that we support a digital public health infrastructure. Primary care providers serving people vulnerable to health inequities report that the lack of digital access remains a barrier. Chang et al [14] suggested that primary care providers report that as much as 70% of their patients lack digital or internet access, and 50% are uncomfortable with the technology. Digital access provides a gateway to education and employment; thus, with effective interventions, we can mitigate this social determinant of health [45]. This suggests that discussions about including digital access in the public health infrastructure and as a social determinant of health warrant priority consideration [46].

Limitations

This study used a unique population-based data source. This allowed us to examine the Hispanic population in the Office Ally and Analytics IQ databases. We studied socioeconomic, health behavior, and demographic factors in the Hispanic community and determined the odds of using telehealth during a public health crisis. The results may not be reproducible, as the data collection was during COVID-19, and patients may not have gained proficiency with the technology or lack internet access. The study used the COVID-19 Research Database and is subject to the limitations of administrative databases. The validity of the data is dependant upon the facilities to report accurate data and code visits correctly in the Office Ally database. Analytics IQ PeopleCore consumer data rely on the accuracy of reporting by the consumer. In this study, the Hispanic population considered all persons of Hispanic or Latino identification. In addition, data were not available on geography (rural and urban), access to home internet, and the ability to read and write English proficiently. In future research, it is imperative to account for variances in usage patterns between urban and rural populations, given the potential impacts of geographic disparities, technological infrastructure, and internet accessibility on health care usage. Future studies should consider measures of English health literacy and its association with health care access. Furthermore, future studies should consider patient and provider relationships in local communities to explore additional information not captured in surveys and claims data that explicates attitudes and challenges with telehealth access and use.

Conclusions

Telehealth supports favorable health outcomes across populations. However, without equity in usage, these benefits are not realized across communities. This study highlights that telehealth use among Hispanic communities can be influenced by demographic, socioeconomic, and health behavioral factors. Telehealth use among Hispanic communities is dependent upon several important factors, such as age, gender, education, socioeconomic status, current health care engagement, and health behaviors. To overcome these barriers, we recommend interdisciplinary strategies that call for medical professionals, insurance providers, and community-based services to engage meaningfully with Hispanic communities to support telehealth use. As telehealth becomes increasingly prevalent in our society, it is imperative that we support this method for accessing the health care system.

Acknowledgments

We thank the COVID-19 Research Database Consortium for their support of the study. The research was supported by the Bill and Melinda Gates Foundation. The content is exclusively the responsibility of the authors and does not necessarily represent the official views of the Bill and Melinda Gates Foundation and the COVID-19 Research Database Consortium.

Data Availability

The data sets generated and/or analyzed during the current study are not publicly available due to the consortium restrictions and governance policies. The database can be accessed by academic, scientific, and medical researchers at COVID-19 Research Database's website. Potential users must register as approved researchers and submit a proposal, including a request to access a specific database. The submitted proposal will be reviewed by the scientific steering committee and a privacy and governance review board. If approved, access will be granted to install the database environment where researchers must conduct the analysis.

Authors' Contributions

CW contributed to the study's conception and design, interpretation of results, and drafting and revising of the work. DS contributed to the study's design, interpretation, analysis, and revising of the data. HC contributed to the writing and editing of the manuscript. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Boen CE, Hummer RA. Longer-but harder-lives?: the Hispanic health paradox and the social determinants of racial, ethnic, and immigrant-native health disparities from midlife through late life. J Health Soc Behav 2019 Dec;60(4):434-452. [doi: 10.1177/0022146519884538] [Medline: 31771347]
2. Oguz T. Update on racial disparities in access to healthcare: an application of nonlinear decomposition techniques. Soc Sci Quart 2019 Feb;100(1):60-75. [doi: 10.1111/ssqu.12551]
3. Wegermann K, Wilder JM, Parish A, et al. Racial and socioeconomic disparities in utilization of telehealth in patients with liver disease during COVID-19. Dig Dis Sci 2022 Jan;67(1):93-99. [doi: 10.1007/s10620-021-06842-5] [Medline: 33507442]
4. Stevens JP, Mechanic O, Markson L, O'Donoghue A, Kimball AB. Telehealth use by age and race at a single academic medical center during the COVID-19 pandemic: retrospective cohort study. J Med Internet Res 2021 May 20;23(5):e23905. [doi: 10.2196/23905] [Medline: 33974549]
5. Ramirez AV, Ojeaga M, Espinoza V, Hensler B, Honrubia V. Telemedicine in minority and socioeconomically disadvantaged communities amidst COVID-19 pandemic. Otolaryngol Head Neck Surg 2021 Jan;164(1):91-92. [doi: 10.1177/0194599820947667] [Medline: 32720844]
6. Ghaddar S, Vatcheva KP, Alvarado SG, Mykyta L. Understanding the intention to use telehealth services in underserved Hispanic border communities: cross-sectional study. J Med Internet Res 2020 Sep 3;22(9):e21012. [doi: 10.2196/21012] [Medline: 32880579]
7. Ryskina KL, Shultz K, Zhou Y, Lautenbach G, Brown RT. Older adults' access to primary care: gender, racial, and ethnic disparities in telemedicine. J Am Geriatr Soc 2021 Oct;69(10):2732-2740. [doi: 10.1111/jgs.17354] [Medline: 34224577]
8. Zhang D, Shi L, Han X, et al. Disparities in telehealth utilization during the COVID-19 pandemic: findings from a nationally representative survey in the United States. J Telemed Telecare 2024 Jan;30(1):90-97. [doi: 10.1177/1357633X211051677] [Medline: 34633882]
9. Pierce RP, Stevermer JJ. Disparities in the use of telehealth at the onset of the COVID-19 public health emergency. J Telemed Telecare 2023 Jan;29(1):3-9. [doi: 10.1177/1357633X20963893] [Medline: 33081595]
10. Weber E, Miller SJ, Astha V, Janevic T, Benn E. Characteristics of telehealth users in NYC for COVID-related care during the coronavirus pandemic. J Am Med Inform Assoc 2020 Dec 9;27(12):1949-1954. [doi: 10.1093/jamia/ocaa216] [Medline: 32866249]
11. Qian L, Sy LS, Hong V, et al. Disparities in outpatient and telehealth visits during the COVID-19 pandemic in a large integrated health care organization: retrospective cohort study. J Med Internet Res 2021 Sep 1;23(9):e29959. [doi: 10.2196/29959] [Medline: 34351865]
12. Increasing access to health insurance benefits everyone: health consequences of being uninsured. National Immigration Law Center. 2017 Aug. URL: https://www.nilc.org/wp-content/uploads/2017/08/insurance-access-health-consequences-2017-08.pdf [accessed 2022-07-28]
13. Qian AS, Schiaffino MK, Nalawade V, et al. Disparities in telemedicine during COVID-19. Cancer Med 2022 Feb;11(4):1192-1201. [doi: 10.1002/cam4.4518] [Medline: 34989148]
14. Chang JE, Lai AY, Gupta A, Nguyen AM, Berry CA, Shelley DR. Rapid transition to telehealth and the digital divide: implications for primary care access and equity in a post-COVID era. Milbank Q 2021 Jun;99(2):340-368. [doi: 10.1111/1468-0009.12509] [Medline: 34075622]
15. Darrat I, Tam S, Boulis M, Williams AM. Socioeconomic disparities in patient use of telehealth during the coronavirus disease 2019 surge. JAMA Otolaryngol Head Neck Surg 2021 Mar 1;147(3):287-295. [doi: 10.1001/jamaoto.2020.5161] [Medline: 33443539]
16. Saeed SA, Masters RMR. Disparities in health care and the digital divide. Curr Psychiatry Rep 2021 Jul 23;23(9):61. [doi: 10.1007/s11920-021-01274-4] [Medline: 34297202]
17. Goel V, Rosella LC, Fu L, Alberga A. The relationship between life satisfaction and healthcare utilization: a longitudinal study. Am J Prev Med 2018 Aug;55(2):142-150. [doi: 10.1016/j.amepre.2018.04.004] [Medline: 29779906]
18. Anderson A, O'Connell SS, Thomas C, Chimmanamada R. Telehealth interventions to improve diabetes management among Black and Hispanic patients: a systematic review and meta-analysis. J Racial Ethn Health Disparities 2022 Dec;9(6):2375-2386. [doi: 10.1007/s40615-021-01174-6] [Medline: 35000144]
19. How patient privacy is being protected. COVID19 Research Database. URL: https://covid19researchdatabase.org/ [accessed 2024-07-23]
20. Barry K, McCarthy M, Melikian G, Almeida-Monroe V, Leonard M, De Groot AS. Responding to COVID-19 in an uninsured Hispanic/Latino community: testing, education and telehealth at a free clinic in Providence. R I Med J (2013) 2020 Nov 2;103(9):41-46. [Medline: 33126788]
21. Clare CA. Telehealth and the digital divide as a social determinant of health during the COVID-19 pandemic. Netw Model Anal Health Inform Bioinform 2021;10(1):26. [doi: 10.1007/s13721-021-00300-y] [Medline: 33842187]
22. Albon D, Van Citters AD, Ong T, et al. Telehealth use in cystic fibrosis during COVID-19: association with race, ethnicity, and socioeconomic factors. J Cyst Fibros 2021 Dec;20 Suppl 3:49-54. [doi: 10.1016/j.jcf.2021.09.006] [Medline: 34930543]
23. Liaw WR, Jetty A, Coffman M, et al. Disconnected: a survey of users and nonusers of telehealth and their use of primary care. J Am Med Inform Assoc 2019 May 1;26(5):420-428. [doi: 10.1093/jamia/ocy182] [Medline: 30865777]

24. Edward J, Morris S, Mataoui F, Granberry P, Williams MV, Torres I. The impact of health and health insurance literacy on access to care for Hispanic/Latino communities. Public Health Nurs 2018 May;35(3):176-183. [doi: 10.1111/phn.12385] [Medline: 29372751]
25. Jain V, Al Rifai M, Lee MT, et al. Racial and geographic disparities in internet use in the U.S. among patients with hypertension or diabetes: implications for telehealth in the era of COVID-19. Diabetes Care 2021 Jan;44(1):e15-e17. [doi: 10.2337/dc20-2016] [Medline: 33139408]
26. Daw JR, Kolenic GE, Dalton VK, et al. Racial and ethnic disparities in perinatal insurance coverage. Obstet Gynecol 2020 Apr;135(4):917-924. [doi: 10.1097/AOG.0000000000003728] [Medline: 32168215]
27. Chau Q, Pathak AB, Turner D, Cheun J, Noe C. Access barriers to telehealth. SMU Data Science Review 2021;5(3) [FREE Full text]
28. Agate S. Unlocking the power of telehealth: increasing access and services in underserved, urban areas. Harv J Hisp Policy 2017;29:85-96 [FREE Full text]
29. Haun JN, Panaite V, Cotner BA, et al. Primary care virtual resource use prior and post COVID-19 pandemic onset. BMC Health Serv Res 2022 Nov 18;22(1):1370. [doi: 10.1186/s12913-022-08790-w] [Medline: 36401239]
30. Vinci C, Hemenway M, Baban SS, et al. Transition to telehealth: challenges and benefits of conducting group-based smoking and alcohol treatment virtually. Contemp Clin Trials 2022 Mar;114:106689. [doi: 10.1016/j.cct.2022.106689] [Medline: 35085833]
31. Blalock DV, Calhoun PS, Crowley MJ, Dedert EA. Telehealth treatment for alcohol misuse: reviewing telehealth approaches to increase engagement and reduce risk of alcohol-related hypertension. Curr Hypertens Rep 2019 Jun 17;21(8):59. [doi: 10.1007/s11906-019-0966-3] [Medline: 31209579]
32. Kim S, Choi S, Kim J, et al. Trends in health behaviors over 20 years: findings from the 1998-2018 Korea National Health and Nutrition Examination Survey. Epidemiol Health 2021;43:e2021026. [doi: 10.4178/epih.e2021026] [Medline: 33872483]
33. Rhee TG, Lee K, Schensul JJ. Black–White disparities in social and behavioral determinants of health index and their associations with self-rated health and functional limitations in older adults. J Gerontol A Biol Sci Med Sci 2021 Mar 31;76(4):735-740. [doi: 10.1093/gerona/glaa264] [Medline: 33049033]
34. Bares CB, Kennedy A. Alcohol use among older adults and health care utilization. Aging Ment Health 2021 Nov;25(11):2109-2115. [doi: 10.1080/13607863.2020.1793903] [Medline: 32757773]
35. Li C, Mao Z, Yu C. The effects of smoking, regular drinking, and unhealthy weight on health care utilization in China. BMC Public Health 2021 Dec 11;21(1):2268. [doi: 10.1186/s12889-021-12309-z] [Medline: 34895186]
36. Jaffe DH, Lee L, Huynh S, Haskell TP. Health inequalities in the use of telehealth in the United States in the lens of COVID-19. Popul Health Manag 2020 Oct;23(5):368-377. [doi: 10.1089/pop.2020.0186] [Medline: 32816644]
37. Kim SS. A study on the acceptance factor for telehealth service according to health status by group. Indian J Sci Technol 2015;8(S1):542. [doi: 10.17485/ijst/2015/v8iS1/63141]
38. Jagielo AD, Chieng A, Tran C, et al. Predictors of patient engagement in telehealth-delivered tobacco cessation treatment during the COVID-19 pandemic. Int J Environ Res Public Health 2024 Jan 25;21(2):131. [doi: 10.3390/ijerph21020131] [Medline: 38397622]
39. Flanagan JC, Hogan JN, Sellers S, et al. Preliminary feasibility and acceptability of alcohol behavioral couple therapy delivered via home-based telehealth [Abstract]. Alcohol Clin Exp Res 2021;45(SUPPL 1):193A. [doi: 10.1111/acer.14628]
40. Tauscher JS, DePue MK, Swank J, Salloum RG. Determinants of preference for telehealth versus in-person treatment for substance use disorders: a discrete choice experiment. J Subst Use Addict Treat 2023 Mar;146:208938. [doi: 10.1016/j.josat.2022.208938] [Medline: 36880898]
41. Glasgow RE, Kwan BM, Matlock DD. Realizing the full potential of precision health: the need to include patient-reported health behavior, mental health, social determinants, and patient preferences data. J Clin Transl Sci 2018 Jun;2(3):183-185. [doi: 10.1017/cts.2018.31] [Medline: 30370072]
42. Primary care. American Academy of Family Physicians. 2022. URL: https://www.aafp.org/about/policies/all/primary-care.html#:~:text=A%20primary%20care%20practice%20serves,physician%20and%20health%20care%20team [accessed 2022-07-28]
43. Jetty A, Jabbarpour Y, Westfall M, Kamerow DB, Petterson S, Westfall JM. Capacity of primary care to deliver telehealth in the United States. J Am Board Fam Med 2021 Feb;34(Suppl):S48-S54. [doi: 10.3122/jabfm.2021.S1.200202] [Medline: 33622818]
44. Browne AJ, Varcoe C, Ford-Gilboe M, et al. Disruption as opportunity: impacts of an organizational health equity intervention in primary care clinics. Int J Equity Health 2018 Sep 27;17(1):154. [doi: 10.1186/s12939-018-0820-2] [Medline: 30261924]
45. Bauerly BC, McCord RF, Hulkower R, Pepin D. Broadband access as a public health issue: the role of law in expanding broadband access and connecting underserved communities for better health outcomes. J Law Med Ethics 2019 Jun;47(2_suppl):39-42. [doi: 10.1177/1073110519857314] [Medline: 31298126]
46. Benda NC, Veinot TC, Sieck CJ, Ancker JS. Broadband internet is a social determinant of health. Am J Public Health 2020 Aug;110(8):1123-1125. [doi: 10.2105/AJPH.2020.305784] [Medline: 32639914]

Abbreviations

HIPAA: Health Insurance Portability and Accountability Act
OR: odds ratio

Edited by C Lovis; submitted 24.02.24; peer-reviewed by A Hassan, AG Rao, S Kale; revised version received 11.05.24; accepted 12.05.24; published 24.07.24.

Please cite as:
Shang D, Williams C, Culiqi H
Telehealth Uptake Among Hispanic People During COVID-19: Retrospective Observational Study
JMIR Med Inform 2024;12:e57717
URL: https://medinform.jmir.org/2024/1/e57717
doi:10.2196/57717

Review

Characteristics of Existing Online Patient Navigation Interventions: Scoping Review

Meghan Marsh[1], MScOT; Syeda Rafia Shah[1], MScOT; Sarah E P Munce[1,2,3], PhD; Laure Perrier[4], PhD; Tin-Suet Joan Lee[2], BSc; Tracey J F Colella[2], APN, RN, PhD; Kristina Marie Kokorelias[1,5], PhD

[1]Department of Occupational Science and Occupational Therapy, University of Toronto, Toronto, ON, Canada

[2]KITE, Toronto Rehabiliation Institute, Toronto, ON, Canada

[3]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[4]Toronto Western Hospital, University Health Network, Toronto, ON, Canada

[5]Section of Geriatrics, Sinai Health and University Health Network, Toronto, ON, Canada

Corresponding Author:
Kristina Marie Kokorelias, PhD
Section of Geriatrics
Sinai Health and University Health Network
600 University Avenue
Geriatrics Department
Toronto, ON, M5G1X5
Canada
Phone: 1 4165864800 ext 4573
Email: kristina.kokorelias@sinaihealth.ca

Abstract

Background: Patient navigation interventions (PNIs) can provide personalized support and promote appropriate coordination or continuation of health and social care services. Online PNIs have demonstrated excellent potential for improving patient knowledge, transition readiness, self-efficacy, and use of services. However, the characteristics (ie, intervention type, mode of delivery, duration, frequency, outcomes and outcome measures, underlying theories or mechanisms of change of the intervention, and impact) of existing online PNIs to support the health and social needs of individuals with illness remain unclear.

Objective: This scoping review of the existing literature aims to identify the characteristics of existing online PNIs reported in the literature.

Methods: A scoping review based on the guidelines outlined in the Joanna Briggs Institute framework was conducted. A search for peer-reviewed literature published between 1989 and 2022 on online PNIs was conducted using MEDLINE, CINAHL, Embase, PsycInfo, and Cochrane Library databases. Two independent reviewers conducted 2 levels of screening. Data abstraction was conducted to outline key study characteristics (eg, study design, population, and intervention characteristics). The data were analyzed using descriptive statistics and qualitative content analysis.

Results: A total of 100 studies met the inclusion criteria. Our findings indicate that a variety of study designs are used to describe and evaluate online PNIs, with literature being published between 2003 and 2022 in Western countries. Of these studies, 39 (39%) studies were randomized controlled trials. In addition, we noticed an increase in reported online PNIs since 2019. The majority of studies involved White females with a diagnosis of cancer and a lack of participants aged 70 years or older was observed. Most online PNIs provide support through navigation, self-management and lifestyle changes, counseling, coaching, education, or a combination of support. Variation was noted in terms of mode of delivery, duration, and frequency. Only a small number of studies described theoretical frameworks or change mechanisms to guide intervention.

Conclusions: To our knowledge, this is the first review to comprehensively synthesize the existing literature on online PNIs, by focusing on the characteristics of interventions and studies in this area. Inconsistency in reporting the country of publication, population characteristics, duration and frequency of interventions, and a lack of the use of underlying theories and working mechanisms to inform intervention development, provide guidance for the reporting of future online PNIs.

(JMIR Med Inform 2024;12:e50307) doi:10.2196/50307

KEYWORDS

online; patient navigation; peer navigation; patient navigation interventions; online patient navigation interventions; scoping review; patient portals; social care services; online medical tools; eHealth; telehealth; personal support; social care; patient navigation intervention

Introduction

Background

Individuals living with chronic illness or illnesses or disability have reported increased reliance on the health care system, as well as social supports for relevant resources or services (ie, medication, equipment, therapy, and counseling), particularly emphasizing the past decade [1-4]. This poses a problem, as they also face a number of challenges when navigating the health care system. These challenges can be attributed to various factors, such as a lack of proper care coordination and continuity of health care services [5-8]. Other concerns include patients' inadequate knowledge related to their conditions or disabilities and the lack of adherence to treatment plans [2,9]. It is also specifically challenging for patients with complex health needs to find appropriate health care services as there is a lack of training in specialized care provision [10-12]. Altogether, these challenges pose a threat to the use, coordination, and continuation of health care services for patients with chronic health conditions or disabilities.

In particular, individuals struggle with coordination difficulties [13]. Literature supports this finding, with a relationship between self-reported care coordination difficulties and the level of patient engagement and chronic illness complexity being observed [13,14]. It is critical to address this gap by providing navigation services for these patients with multiple and complex chronic conditions as a lack of proper coordination and continuation of services can lead to negative outcomes related to one's health and well-being, including one's ability to integrate and participate within the community [5-8,13,15].

Patient navigation commonly involves the use of one-to-one interactions between navigators and patients or their family members and caregivers to promote recommended health care use behaviors from patients' screening, to diagnosis, to resolution [10,16-19]. Patient navigation can be provided in the form of a professional, lay, or peer (with training) navigator [20].

Current literature has identified patient navigation interventions (PNIs) as an effective care approach for populations with chronic illness or disabilities in relation to managing their care through assistance with navigating the health care system [1,10,13,16,21-27]. In a systematic review, McBrien et al [27] assessed the impact of patient navigation on patients living with chronic diseases such as cancer, diabetes, HIV or AIDS, cardiovascular disease, chronic kidney disease, and dementia. The authors found that of the included 67 randomized controlled trials (RCTs), 44 trials indicated that patient navigation improved primary outcomes, specifically those related to the patients' care or health care navigation process [27]. A meta-analysis of RCTs involving various patient populations revealed that compared to usual care, patient navigation more than doubled the likelihood of patients' health screening rates and attendance at care events [25]. Similar findings were reported in a scoping review by Kokorelias et al [10] that summarized the literature on patient navigation for adults with chronic conditions, whereby patient navigation increased a patient's overall satisfaction with their care and improved access to care, education, and adherence to medication and treatment completion. Likewise, in the context of cancer care, reviews of patient navigation have concluded that PNI programs were found to be cost-effective approaches to care when considering factors such as life expectancy, incremental cost-effectiveness ratios, and quality-adjusted life-years [28,29], thus further supporting the benefit and need for patient navigation. While informative, these reviews focused on PNIs in general and were not specific to online PNIs.

One example of patient navigation is peer navigation, which involves trained peer navigators who have lived experiences of health conditions or disabilities that they can use to provide personalized support to patients with different needs [19,23,30-33]. Personalized support in patient or peer navigation may involve the following types of support: educational or informational (sharing of advice, personal experiences, first-hand knowledge, resources, and factual information), psychosocial (provision of emotional and social support using empathy, validation, mentorship, motivation, feedback, and reflection), and instrumental (assistance with administrative activities, accessing and navigating services or resources, advocacy) [19,23,30-33].

Consistent with the theoretical underpinnings of the Social Cognitive Theory [34], the provision of such personalized support in patient navigation can promote patients' perceived self-efficacy, appropriate health care use behaviors, and related outcomes (ie, community integration, quality of life, and well-being). For instance, Cabassa et al [35] systematic review identified peer-based navigation interventions to be among the most promising interventions for improving the health outcomes of individuals with serious mental illnesses. Peer navigators with lived experience improved health outcomes by facilitating linkages between individuals seeking care and health care services [10,36-39].

One area of development that warrants further exploration is online PNIs for a breadth of chronic conditions in the adult population. Research has shown that online-based PNIs have a great potential for improved health outcomes (eg, increased patient knowledge, transition readiness, self-efficacy, and appropriate use of health care services) in various patient populations. Casillas et al [40] conducted a three-arm RCT to test the efficacy of both a peer navigation intervention and an intervention involving the use of mobile technology (ie, SMS text messaging) in promoting cancer survivorship care in adolescents and young adults. Compared to standard care, these online interventions demonstrated the following statistically significant benefits: online peer navigation improved

participants' self-efficacy in survivorship care, SMS text messaging improved survivorship-focused knowledge, and both interventions improved participants' attitudes in seeking survivorship care [40]. Specifically, the SMS text messaging group exhibited higher levels of survivorship care knowledge compared to the control group (P<.05), while the peer navigation group showed increased survivorship care self-efficacy compared to the control group (P<.05). Both intervention groups demonstrated more positive attitudes toward seeking survivor-focused care compared to the control group (SMS text messaging group: P<.05; peer navigation group: P<.05) [40]. Considering the initial efficacy observed in both interventions, each has the potential to be used in the future to educate and empower adolescent and young adult cancer survivors in accessing necessary survivorship care [40]. Online support has also been deemed a more flexible and sustainable care model when offered to individuals with intellectual disabilities, especially during the COVID-19 pandemic [41]. Moreover, online patient navigation can better reach rural, remote, and other underserved communities.

Objective

Despite the demonstrated benefits of online patient navigation for various patient populations, the extent of the literature specifically focused on online PNIs across a range of chronic conditions or disabilities is uncharted. Therefore, the purpose of this scoping review is to comprehensively search databases and summarize data from peer-reviewed publications to address the following research question: What is known from the existing literature about the key characteristics (ie, intervention type, mode of delivery, duration, frequency, outcomes and outcome measures, underlying theories or mechanisms of change of the intervention, and impact) of online PNIs used across a range of chronic conditions or disabilities?

Methods

Research Design and Methodological Framework

A scoping review methodology was used given the broad nature of the research objective and question, and the lack of previous comprehensive reviews conducted in this area. A scoping review, also known as a scoping study, serves as a form of knowledge synthesis designed to explore research questions and map key concepts, types of evidence, and research gaps related to a defined area or field. This approach involves systematic searching, selection, and synthesis of existing knowledge [42] (page 28). Thus, a scoping review was deemed suitable to help identify key concepts and evidence related to online PNIs for adults with chronic conditions or disabilities. This scoping review was guided by the framework proposed by the Joanna Briggs Institute (JBI) Manual for Evidence [43-45]. The JBI framework was selected as it was developed based on previously reported methodological frameworks by Arksey and O'Malley [46] and Levac et al [47]. This refined framework provides additional guidance and clarity on the steps involved in the collection, analysis, and dissemination of research findings [43-45]. Specifically, the JBI framework focuses on aspects of the research process that have not been addressed as extensively in previous frameworks. The methods and the findings are reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; Multimedia Appendix 1) to further enhance the reporting of findings, as consistent with the JBI methodology [43-45,48]. A protocol was not published prior to the completion of this scoping review.

Inclusion and Exclusion Criteria

All empirical study designs (eg, experimental, quasi-experimental, observational, qualitative studies, not review methodologies) reported in peer-reviewed, full text (eg, no conference abstracts) were included to increase the scope of the literature found. To address the identified gaps in the current literature on online PNIs, only peer-reviewed studies involving trained navigators and PNIs delivered using an online format, and software or application-based PNIs both with or without allocated trained navigator support were included. Our review encompasses a broad spectrum of online PNIs including those featuring hybrid formats. These interventions incorporate both online elements and face-to-face or other non–internet-related components, ensuring a comprehensive assessment of diverse intervention modalities and their characteristics. Participants in these studies had to be adults (aged 18 years and older) with chronic conditions or disabilities as recognized by the Public Health Agency of Canada (PHAC) [49], the Canadian Chronic Disease Surveillance System, and the World Health Organization. These conditions based on their PHAC categorization could include exclusively physical or mental health–based conditions, or both. Examples of common chronic diseases and conditions, as defined by the PHAC, the Canadian Chronic Disease Surveillance System, and the World Health Organization, include cardiovascular disease (eg, heart failure, hypertension, and stroke), chronic respiratory disease (eg, asthma and chronic obstructive pulmonary disease), diabetes mellitus (types combined, but not gestational diabetes), mental illnesses (alcohol or drug-induced disorders, mood and anxiety disorders, and schizophrenia), musculoskeletal disorders (eg, arthritis and osteoporosis), and neurological conditions (eg, dementia, epilepsy, multiple sclerosis, and Parkinson) [49-51]. Additionally, studies with participants living with HIV and AIDS were also included [52]. Only studies published between 1989 and 2022 and available in full text in English were included due to feasibility considerations (ie, members of the research team could only read in English) and resource constraints. The Report to the Nation on Cancer in the Poor that began work in the area of patient navigation began in 1989 [21]. Exclusion criteria included PNIs that were delivered in alternate formats (eg, face-to-face, telephone, and mail).

Data Collection and Management

Comprehensive literature search strategies based on the inclusion and exclusion criteria of this scoping review were developed in collaboration with an experienced librarian (LP). The search strategies were further informed by the Participants/Concept/Context framework as recommended in the JBI methodology. The search strategy included medical subject headings and text words related to adults with chronic conditions or disabilities and online PNIs (Multimedia Appendix 2). The search strategy was first developed, tested, and refined

in MEDLINE (OVID interface) prior to being used in other databases. The following databases were searched using the finalized MEDLINE strategy: CINAHL (EBSCO interface), Embase (OVID interface), PsycInfo (OVID interface), and Cochrane Central Register Controlled Trials (Cochrane Library). The use of multiple health care–related databases helped broaden the scope of the comprehensive literature search. Data yielded from the comprehensive literature search strategies were stored and screened using the online Covidence software program (SaaS Enterprise) [53,54]. These data were screened at 2 levels (ie, level 1 and level 2 screening). Study titles and abstracts were screened first, followed by the screening of full-text studies. Screening at both levels was conducted by 2 independent reviewers (MM and SRS) to ensure accuracy in the included results. Discrepancies were addressed through consensus between the reviewers and the senior author (KMK). Reference lists of all included studies were reviewed to determine any studies that may have been missed from the database search. Gray literature was not included.

Data Extraction and Analysis

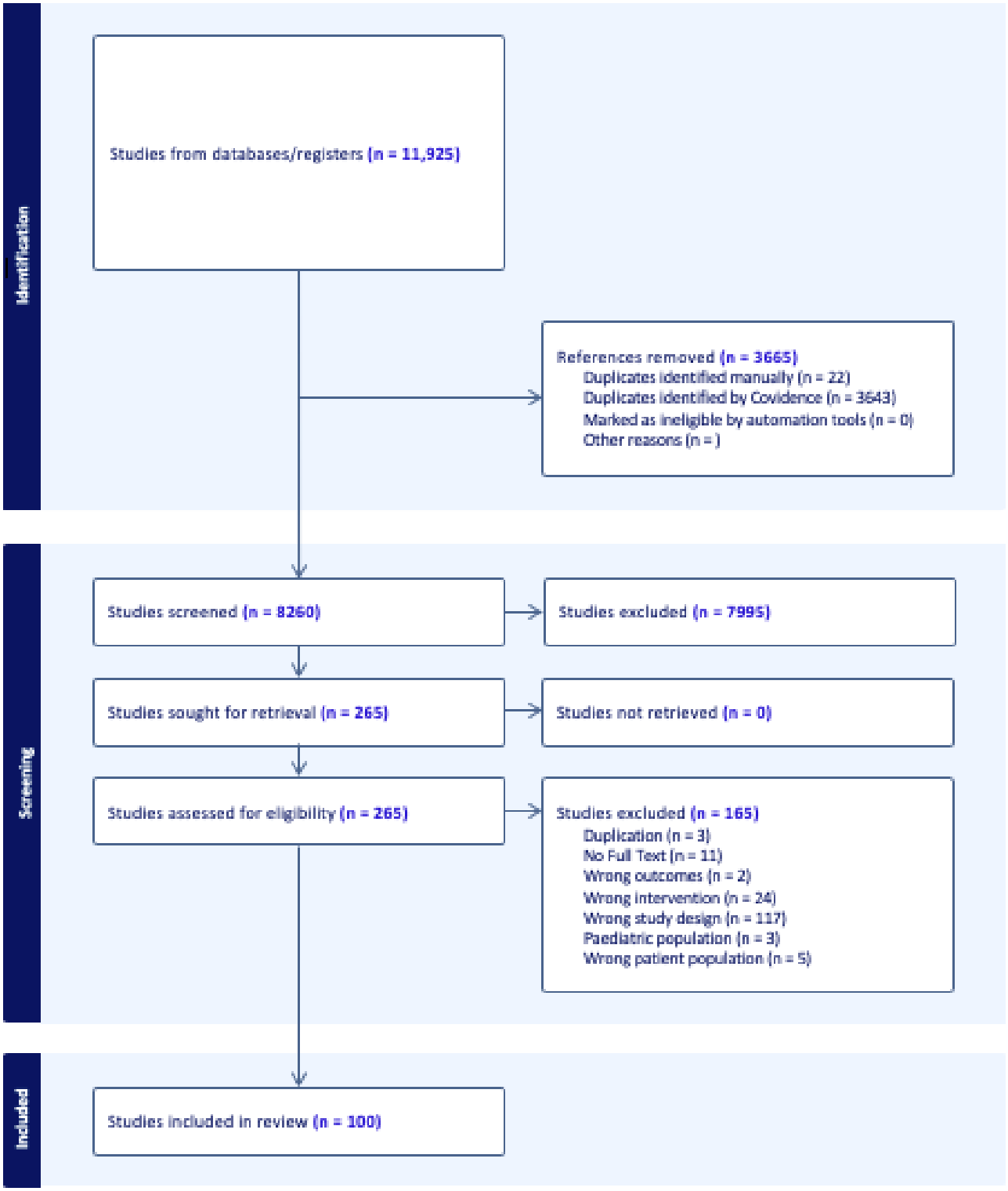
Data extraction was carried out by extracting key information or data from the included studies. A data extraction form, developed by the authors, was used to chart and record this information to ensure easy referencing and tracking of each study to ensure clarity. The form was first piloted on the first 5 included studies by all members of the research team. The extraction template was further informed by the Template for Intervention Description and Replication (TIDieR) checklist and guide, which is a framework that aims to promote replicability and implementation of interventions through the consistent reporting of key intervention characteristics [55]. The following data were extracted from the full-text studies: study characteristics (ie, title, author or authors, publication year, publication country, study purpose or objective or objectives, and study design), participant population characteristics (ie, sample size, race or ethnicity, condition or disability, age, and sex), and key characteristics of the intervention (ie, name, type, description, setting, duration, frequency, mode of delivery, underlying theories, behavior change techniques or working mechanisms, context, outcome measures used, and quantitative and qualitative outcomes). In line with scoping review methodologies, we did not evaluate the quality of included studies [56]. Data were extracted by 2 independent reviewers (MM and SRS) and any disagreements were resolved through consensus. Following data extraction, the following information was specifically summarized using descriptive statistics [57] and directed content analysis [58] to provide an accurate overview of the published literature on the key characteristics of online PNIs in adults with chronic conditions or disabilities. The research team reviewed the coded data to create a set of categories that capture the key themes, concepts, and variables relevant to the research question. This involved both inductive categories (emerging from the data) and deductive categories (informed by the TIDieR framework). The authors then began coding the selected studies according to this scheme, using Excel (Microsoft Corp) to facilitate this process. The Excel document was then reviewed by all members of the research team to identify patterns and trends. The team met over a series of meetings to determine key interpretations of the results.

Results

Overview

The PRISMA-ScR flowchart displayed in Figure 1 shows an overview of our comprehensive literature search, which yielded 11,925 studies.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flowchart.



Study Characteristics

The publication dates ranged from 2003 to 2022, with some years yielding no publications. The greatest number (n=20, 20%) were published in 2019 with 19 (19%) studies published in 2020. As for the publication country, all studies were published in Western countries. A total of 48 (48%) studies were conducted and published in the United States, followed by 23 (23%) studies from Canada, and 19 (19%) studies from the United Kingdom. There was a variation of study designs observed in the review. The most common study design included RCTs (n=39, 39%). The included studies focused on the development, implementation, and testing of online PNIs. Of the studies included, 82 (82%) studies specified their sample size. There was a significant range in the overall participant sample size (including intervention and control arm or arms), observed from 9 to 14,584 participants, with the median being 97 (IQR 342) participants. When comparing the sample sizes

of the intervention and control arms, 24 studies had an approximately equal division, while 21 studies did not. In addition, 55 studies did not specify the sample size of the intervention or control groups. Multimedia Appendix 3 outlines the study characteristics.

Population Characteristics

Among the included full-text studies, the majority (n=79, 79%) of the studies' participants had conditions or disabilities that would be classified exclusively as physical health-based conditions according to the PHAC (including cancer, heart disease, diabetes, and stroke), while only 11% (n=11) of the studies represented participants with solely mental health-based conditions (eg, mood disorders, substance use and addictions, and eating disorders). A total of 10% (n=10) of the studies represented participants who had both physical and mental health–based conditions. One study represented participants who had post–COVID-19 condition. In terms of the participants' age, 55% (n=55) of the studies provided this information. Across the studies that reported age (regardless of study design), the mean age range (with SD) of the participants included in the studies' intervention arms (ie, those who participated in online PNI) was between 19.84 (SD 1.61) and 69.6 (SD 9.1) years. When comparing the age of the participants in the intervention and control arms, 37% (n=37) of the studies had similarly aged participants (ie, 20-70 years of age). With regards to the sex of participants, 20 (20%) studies included only female participants while only 4 (4%) studies included only male participants. A total of 19 (19%) studies had an approximately equal distribution of females and males, and 14 (14%) studies did not specify the sex of the participants or inconsistently reported this information. In terms of the racial or ethnic representation of participants in studies, more than half (n=54, 54%) of the studies did not specify this information. Among those that did, in 20 (20%) studies, the majority of the participants were White, followed by the following racial groups: Black, African, or African-American (16 studies), Hispanic or Latinx (5 studies), and Asian or Asian American (4 studies; terminology as used by the original authors of the included studies). Only one study focused on the Indigenous population (in Australia).

Intervention Characteristics

Types of Intervention

Among the 100 included studies, the most common intervention type was a combination (n=61, 61%), which included a mix of peer, patient, and other navigation types including coaching, digital navigation (including mobile, eHealth, or telehealth-based interventions, or software application-based interventions), and self-management. Within the "other" category, intervention types included a mixture or combination of peer, patient, and other navigation intervention types, as well as the exclusive implementation of interventions such as coaching, counseling, health promotion, digital navigation, self-monitoring, or self-management. A total of 32 (32%) studies were exclusively patient navigation-based, and 7 (7%) studies were exclusively peer navigation-based. Other studies did not clearly specify the intervention type (ie, who the intervention was led by or what it involved).

Several studies explored diverse online PNIs for cancer-related support and care. While the interventions varied across studies, commonalities and differences emerged. One study introduced a virtual navigation tool providing comprehensive cancer information accessible 24/7, emphasizing its value in validating information and controlling exposure, while another used nurse navigators sending scripted messages, resulting in improved quality of life and symptom burden among participants [59]. A similar modality was used by other studies that used peer navigation and SMS text messaging interventions for educating cancer survivors on late effects and survivorship care, leading to improved survivorship knowledge, attitudes, and self-efficacy [40]. On the contrary, some scholars focused on self-monitoring and physical activity, noting enthusiasm and continued use among participants [60,61].

In the context of multimorbidity, the studies featured various adaptations and design elements tailored for both physical and mental health conditions and examined technology use for patients with multiple chronic conditions, focusing on communication tools, tracking medical information, and decision-making support, intending to address self-management challenges and health care navigation issues [62]. Allen et al [63] developed an internet-based health coaching intervention targeting chronic pain, depression, and mobility difficulties, emphasizing patient-clinician communication improvement and patient empowerment through goal-setting and constructive communication tools. On the other hand, some interventions focused on peer visitation, support groups, and educational materials, enhancing recovery expectations and satisfaction [64,65]. Thus, both chronic conditions, like cancer and multiple chronic conditions, used strategies such as technology integration for communication, support, and information dissemination, tailored interventions addressing specific conditions and associated challenges, peer support networks fostering engagement and optimism, and empowerment strategies encouraging collaborative patient-clinician communication, goal-setting, and self-efficacy. Multimedia Appendix 4 outlines the intervention characteristics.

Duration and Frequency

Of the included studies, the duration and frequency of interventions were varied. Of the 100 included studies, 55 (55%) studies specified a duration and only 17 (17%) studies specified a frequency of the intervention. Other studies included a variable duration or frequency that was tailored to the needs of individual patients. Of the 55 (55%) studies that specified duration, 7 (7%) studies were offered for a year. A total of 41 (41%) studies were offered for 1 month up to 11 months, 3 studies were offered weekly (eg, one time per week), and 4 (4%) studies were held on a daily basis. In addition, of the total studies included, 14 studies had a variable duration, 24 (24%) studies did not specify, and there were 7 (7%) studies where the duration was not applicable (eg, proof of concept, usability, and beta-testing). In terms of frequency, of the 17 (17%) studies that specified frequency for interventions, 3 (3%) studies were month-based, 10 (10%) studies were week-based, and 4 studies required day-to-day engagement from participants. Of the total studies included, 46 (46%) online PNIs had variable frequency, 30 (30%) did not specify frequency, and there were 7 (7%) studies

that reported interventions where the frequency was not applicable as the interventions only occurred once.

Mode of Delivery

Of the included studies, 56 (56%) of the 100 studies reported using an online mode of asynchronous or synchronous delivery for interventions, without any other components. A total of 13 (13%) studies used a format that was hybrid, with both online and offline intervention formats. In total, 22 (22%) studies included SMS text messaging as the main component of intervention delivery, of which 14 (14%) studies had SMS text messaging mixed with other intervention formats such as the use of telephone calls, educational videos, websites, and online support groups. Finally, 9 (9%) of the 100 studies used a mixed format including intervention components such as software programs and applications, telephone calls, in-person interactions, email, automated phone lines, and other online-based intervention formats.

Underlying Theories

In total, 78 (78%) studies did not specify the use of any underlying theories, models, and frameworks, that were used to guide the PNIs. Of the 22 (22%) studies that did specify an existing theory, a total of 21 different theories were identified. Four studies indicated that the intervention was based on more than one theory. Social cognitive theory followed by the self-determination theory, self-efficacy theory, behavior change theory, and community empowerment were the most common theories used. The most common working mechanism among these was a combination (n=20, 20%) of various mechanisms, which included coaching, education, peer support, navigation, self-management, and cognitive behavioral therapy among others. Self-management (n=19, 19%) and navigation (n=19, 19%) followed as other commonly identified mechanisms.

Outcome Measures

Multiple outcome measures were used in these studies; however, the most commonly used outcome measures included: the Short Form-36 survey questionnaire to measure participants' health-related quality of life, the Patient Health Questionnaire-9 to measure participants' psychological outcomes, and the Health Education Impact Questionnaire to measure participants' knowledge and self-management related outcomes. Additional standardized and nonstandardized outcome measures were used to report on common intervention outcomes such as the interventions' feasibility, acceptability, efficacy, effectiveness, uptake, use, and retention, and participants' clinical symptoms or outcomes (physical, mental, emotional or psychosocial), lifestyle or behavioral changes, quality of life, user experience or satisfaction, adherence, knowledge, attitudes, and self-efficacy.

Outcomes and Impact of Online PNIs

Out of the 76 (76%) studies that reported quantitative findings (including RCTs and non-RCTs), 46 (60.5%) studies demonstrated significant improvements. Improvements were commonly demonstrated in the following outcomes: appointment adherence, intervention retention, knowledge, self-monitoring of symptoms, and physical and mental health symptoms. Of the 23 (23%) studies that reported qualitative findings, 8 (35%) studies identified specific themes and subthemes [62,66-72]. The included qualitative studies spoke of themes that described usability or user experience, as well as participants' experience with the PNI as it related to self-management, education, knowledge, navigation, engagement, encouragement or support, and feedback.

Discussion

Principal Findings

This scoping review aimed to investigate the key characteristics of reported online PNIs to inform future intervention development and research evaluation. A total of 100 peer-reviewed studies were included. Overall, online PNIs are highly variable with various modalities for delivery, durations, frequencies, and contexts of support provided. Few studies reported participants' sex, diagnoses, age, and race or ethnicity. Moreover, most of the literature was published in Western countries, resulting in a lack of data from non-Western countries, as well as PNIs that reflect the needs of individuals from non-Western countries. Despite this, we were able to ascertain through the results of 20 RCTs (the highest level of evidence) [73] that in general, online PNIs improve outcomes of patients' self-management, knowledge, clinical symptoms (physical or mental health–based), and use and navigation of health care services.

The majority of the online PNIs were designed for physical health-based conditions (including cancer, heart disease, diabetes, and stroke), while few studies focused solely on mental health–based conditions. Our investigation revealed a notable scarcity of online PNIs specifically targeting multimorbidity of physical and mental health conditions (n=10), signifying a considerable gap in available interventions addressing the complex needs of individuals with multiple chronic conditions. This paucity carries significant implications, indicating an unmet need within the digital health landscape, that is needed to ensure comprehensive care for those navigating multifaceted health challenges. Participants in the RCTs ranged from 20 to 70 years of age. As with other reviews of digital health interventions to support the coordination of care [74,75], our review noted a lack of inclusion of particular groups of older adults (ie, 70 years and older), despite this group representing a large proportion of individuals living with chronic conditions [76] who could benefit from online health interventions [75,77]. Moreover, our review found a lack of literature exploring the impact of online PNIs on Indigenous populations and non-White populations such as Black, Asian, and Hispanic individuals, making it difficult to ascertain their unique needs to inform further online PNIs. As such, future research on online PNIs is encouraged to explore the interaction of racial and cultural factors of different groups to improve service delivery [78].

Our review highlights how future online PNIs can better support various patient populations. Only one study cited in our review noted the racial preferences of participants in which Black patients preferred the services of a Black (virtual) provider [79]. Ethnic minorities and other underserved populations often face unique barriers to accessing health services that patient navigation is able to assist with overcoming [10,80,81]. Social

and environmental factors, such as finances, health literacy, and availability of health services, influence health access [82-84]. To overcome these barriers, it is necessary to create efficient processes for referring communities affected by social and environmental factors to suitable resources, ensuring that their needs are adequately met [85,86]. Culturally appropriate patient navigation can assist with learning about the unique information needs and barriers that face particular communities and facilitate an appropriate referral and support process to services [10,87]. While online PNIs can help overcome traditional barriers to seeking support, such as transportation [88], it is important to consider that shifting to online PNIs may also increase risks to access and equity as a result of digital inequity (ie, gaps in use and participation in the use of technology) [89]. Future research efforts on online PNIs should also consider the individual needs of target populations (eg, access and geographical location, income, and digital literacy), as well as the significance of an individual-based versus group-based mode of delivery of online PNIs.

We also noted the lack of consistent reporting of intervention characteristics. For example, the duration and frequency of interventions were not reported consistently or were variable among the included studies in our review. Moreover, multiple studies did not specify the exact frequency of their intervention. Similar trends were observed in a previous review on web-based peer support interventions where the authors reported "a lack of consistency" and variation regarding the reporting of intervention characteristics such as duration and frequency [75]. The reporting of intervention doses associated with improved outcomes is important to guide other jurisdictions looking to implement or build upon existing interventions [90]. Frameworks, such as the TIDieR, have been posited as helpful for guiding researchers in reporting a full description of complex interventions [55] such as PNIs. The TIDieR can help guide the reporting of future online PNIs to ensure transparency and improve the quality of patient navigation research. Relying solely on reported intervention characteristics, however, can imply a limitation of the personalization of interventions (ie, inflexibility in the duration and frequency tailored to participant needs). While this can be a great guide for replicating the interventions, and further testing and implementation, patients with chronic illness may require individualized approaches to care [91]. Further research is needed to understand how the duration, frequency, and support provided within existing online PNIs may evolve across the illness and care trajectory of patients. Moreover, the TIDieR is only beneficial for reproducibility in the setting specified by the original individual study and therefore cannot guide researchers to implement the intervention in different contexts or settings [90]. Researchers should then reply on implementation frameworks, such as the PRACTical planning for Implementation and Scale-up guide to provide practical direction on implementing online PNIs into new sessions [92].

Despite a substantial portion (78%) of the studies not explicitly delineating underlying theories or frameworks, the 22% of studies that did highlight a diverse array of theoretical foundations (ie, Social Cognitive Theory emerged prominently, followed by self-determination theory, self-efficacy theory, behavior change theory, and community empowerment among others). This diversity underscores the need for a more comprehensive and structured integration of theoretical frameworks within the design and implementation of online PNIs. Integration of frameworks within online PNIs can help researchers understand the underlying mechanisms driving these interventions and will help to establish standardized evaluation metrics. Moving forward, comprehensive research could delve into exploring the efficacy and synergies of combining multiple theories to inform the design and implementation of online PNIs effectively. Furthermore, investigating how specific mechanisms within these theories (eg, coaching, education, and peer support) contribute to PNI outcomes can enrich our understanding and potentially optimize intervention strategies. This calls for a systematic and comparative analysis to discern the differential impact of diverse theoretical orientations on the effectiveness, sustainability, and scalability of PNIs across various health contexts and participant demographics.

Digital health interventions often incorporate elements akin to navigation programs including patient education, remote monitoring, and personalized feedback. These features aim to empower patients in self-management and facilitate communication with health care providers. In contrast, navigation programs traditionally focus on guiding patients through complex health care systems, providing support in appointment scheduling, access to resources, and continuity of care. However, as digital health evolves, distinctions between these approaches can blur. Many digital health solutions now integrate navigation functionalities such as decision support tools and care coordination platforms. This integration raises questions about the delineation between virtual care and navigation programs, particularly regarding their roles in improving health outcomes and patient experience across different chronic diseases. Moving forward, future research should explore synergies between digital health interventions and navigation programs to optimize their combined impact on chronic disease management. This includes examining the effectiveness of integrated approaches in enhancing patient adherence, reducing health care disparities, and improving the overall quality of care.

Finally, we found that online PNIs are an accessible and user-friendly option for navigational support to patients. Similar trends were observed in other studies involving peer and professional navigators [10,19,20,36,93-101]. Similarly, positive and statistically significant outcomes were also reported in another scoping review on web-based peer support, where the authors found that interventions in 4 of their 6 included RCTs improved the health navigation, emotional self-management, self-efficacy, social participation, and attitudes of adults with chronic conditions [75]. Overall, these findings demonstrate how online PNIs could play a crucial role in improving health use and navigation among adults with chronic conditions and disabilities. Additionally, a common theme that was reported by patients, specifically among the qualitative findings of studies from our review was that online PNIs provided more accessibility, engagement, and encouragement to participants navigating their health. Gaining insight on what would construe the ideal patient navigator and ideal patient navigator program

for patients with chronic health conditions is still in its infancy [102,103], and as such conducting more qualitative research with diverse patient populations would be valuable in refining and co-designing novel online PNIs and navigator roles.

Limitations

Although a systematic, comprehensive review was conducted to identify key characteristics of online PNIs, the authors acknowledge that this scoping review has some limitations. First, search results were limited to publications in English studies published after 1989, and while the broad search strategy made it unlikely that potentially eligible publications were missed as a result, we may have created a bias toward studies from English-speaking countries, which might have contributed to the majority of data coming from Western countries. We also excluded gray literature. The majority of the data extraction was not completed in duplicate, which may have affected the reliability of the extracted data. Incomplete reporting on study characteristics by original study authors also made it challenging to comment on additional participant characteristics (eg, socioeconomic status, education level, and digital literacy) that would have provided valuable information.

Conclusions

This review has mapped the existing literature on online PNIs, and in doing so, has identified several gaps that should be addressed in future research and intervention development efforts. Although many positive outcomes were reported for online PNIs, a lack of variation in included study samples, as well as a lack of consistency in reporting, was observed in the reporting of TIDieR intervention characteristics including the following: the publication country of studies, population characteristics such participants' age, sex, and racial or ethnic background, duration and frequency of interventions, and the use of underlying theories and working mechanisms to inform intervention development. Future research and development efforts should consider using theories and models, expanding inclusion criteria, and reporting key intervention characteristics more consistently.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews.
[DOCX File , 82 KB - medinform_v12i1e50307_app1.docx]

Multimedia Appendix 2
Search Strategy.
[DOC File , 31 KB - medinform_v12i1e50307_app2.doc]

Multimedia Appendix 3
Overview of study characteristics.
[DOCX File , 152 KB - medinform_v12i1e50307_app3.docx]

Multimedia Appendix 4
Intervention characteristics.
[DOCX File , 171 KB - medinform_v12i1e50307_app4.docx]

References

1. Dixit N, Rugo H, Burke NJ. Navigating a path to equity in cancer care: the role of patient navigation. Am Soc Clin Oncol Educ Book 2021;41:3-10 [FREE Full text] [doi: 10.1200/EDBK_100026] [Medline: 33830828]
2. A guideline for transition from paediatric to adult health care for youth with special needs: a national approach. CAPHC. 2016. URL: https://www.childhealthbc.ca/sites/default/files/caphc_transition_to_adult_health_care_guideline_may_2017.pdf [accessed 2024-08-03]
3. Entwistle VA, Cribb A, Owens J. Why health and social care support for people with long-term conditions should be oriented towards enabling them to live well. Health Care Anal 2018;26(1):48-65 [FREE Full text] [doi: 10.1007/s10728-016-0335-1] [Medline: 27896539]
4. Webkamigad S, Rowe R, Peltier S, Chow AF, McGilton KS, Walker JD. Identifying and understanding the health and social care needs of indigenous older adults with multiple chronic conditions and their caregivers: a scoping review. BMC Geriatr 2020;20(1):145 [FREE Full text] [doi: 10.1186/s12877-020-01552-5] [Medline: 32306912]
5. Blum RW, Garell D, Hodgman CH, Jorissen TW, Okinow NA, Orr DP, et al. Transition from child-centered to adult health-care systems for adolescents with chronic conditions. A position paper of the society for adolescent medicine. J Adolesc Health 1993 Nov;14(7):570-576. [doi: 10.1016/1054-139x(93)90143-d] [Medline: 8312295]

6. Stevens SE, Steele CA, Jutai JW, Kalnins IV, Bortolussi JA, Biggar WD. Adolescents with physical disabilities: some psychosocial aspects of health. J Adolesc Health 1996;19(2):157-164. [doi: 10.1016/1054-139X(96)00027-4] [Medline: 8863089]
7. Steinbeck KS, Brodie L, Towns SJ. Transition in chronic illness: who is going where? J Paediatr Child Health 2008;44(9):478-482. [doi: 10.1111/j.1440-1754.2008.01321.x] [Medline: 18928466]
8. Brown MM. Transitions of care. In: Chronic Illness Care. New York City: Springer; 2018:369-373.
9. Mitchell SE, Laurens V, Weigel GM, Hirschman KB, Scott AM, Nguyen HQ, et al. Care transitions from patient and caregiver perspectives. Ann Fam Med 2018;16(3):225-231. [doi: 10.1370/afm.2222] [Medline: 29760026]
10. Kokorelias KM, Shiers-Hanley JE, Rios J, Knoepfli A, Hitzig SL. Factors influencing the implementation of patient navigation programs for adults with complex needs: a scoping review of the literature. Health Serv Insights 2021;14:11786329211033267 [FREE Full text] [doi: 10.1177/11786329211033267] [Medline: 34349519]
11. Coleman EA. Falling through the cracks: challenges and opportunities for improving transitional care for persons with continuous complex care needs. J Am Geriatr Soc 2003;51(4):549-555. [doi: 10.1046/j.1532-5415.2003.51185.x] [Medline: 12657078]
12. Hudon C, Aubrey-Bassler K, Chouinard M, Doucet S, Dubois M, Karam M, et al. Better understanding care transitions of adults with complex health and social care needs: a study protocol. BMC Health Serv Res 2022;22(1):206 [FREE Full text] [doi: 10.1186/s12913-022-07588-0] [Medline: 35168628]
13. Maeng DD, Martsolf GR, Scanlon DP, Christianson JB. Care coordination for the chronically ill: understanding the patient's perspective. Health Serv Res 2012;47(5):1960-1979 [FREE Full text] [doi: 10.1111/j.1475-6773.2012.01405.x] [Medline: 22985032]
14. Berntsen G, Høyem A, Lettrem I, Ruland C, Rumpsfeld M, Gammon D. A person-centered integrated care quality framework, based on a qualitative study of patients' evaluation of care in light of chronic care ideals. BMC Health Serv Res 2018;18(1):479 [FREE Full text] [doi: 10.1186/s12913-018-3246-z] [Medline: 29925357]
15. Chen L, Xiao LD, Chamberlain D. An integrative review: challenges and opportunities for stroke survivors and caregivers in hospital to home transition care. J Adv Nurs 2020;76(9):2253-2265. [doi: 10.1111/jan.14446] [Medline: 32511778]
16. Aboumatar H, Pitts S, Sharma R, Das A, Smith BM, Day J, et al. Patient engagement strategies for adults with chronic conditions: an evidence map. Syst Rev 2022;11(1):39 [FREE Full text] [doi: 10.1186/s13643-021-01873-5] [Medline: 35248149]
17. Reid AE, Doucet S, Luke A, Azar R, Horsman A. The impact of patient navigation: a scoping review protocol. JBI Database System Rev Implement Rep 2019;17(6):1079-1085. [doi: 10.11124/JBISRIR-2017-003958] [Medline: 31021974]
18. Valaitis RK, Carter N, Lam A, Nicholl J, Feather J, Cleghorn L. Implementation and maintenance of patient navigation programs linking primary care with community-based health and social services: a scoping literature review. BMC Health Serv Res 2017;17(1):116 [FREE Full text] [doi: 10.1186/s12913-017-2046-1] [Medline: 28166776]
19. Kelly K, Doucet S, Luke A. Exploring the roles, functions, and background of patient navigators and case managers: a scoping review. Int J Nurs Stud 2019;98:27-47. [doi: 10.1016/j.ijnurstu.2019.05.016] [Medline: 31271977]
20. Reid AE, Doucet S, Luke A. Exploring the role of lay and professional patient navigators in Canada. J Health Serv Res Policy 2020;25(4):229-237. [doi: 10.1177/1355819620911679] [Medline: 32188293]
21. Freeman HP, Rodriguez RL. History and principles of patient navigation. Cancer 2011;117(S15):3539-3542 [FREE Full text] [doi: 10.1002/cncr.26262] [Medline: 21780088]
22. Freeman HP. The history, principles, and future of patient navigation: commentary. Semin Oncol Nurs 2013;29(2):72-75. [doi: 10.1016/j.soncn.2013.02.002] [Medline: 23651676]
23. Lorhan S, Cleghorn L, Fitch M, Pang K, McAndrew A, Applin-Poole J, et al. Moving the agenda forward for cancer patient navigation: understanding volunteer and peer navigation approaches. J Cancer Educ 2013;28(1):84-91. [doi: 10.1007/s13187-012-0424-2] [Medline: 23104142]
24. Kelly E, Fulginiti A, Pahwa R, Tallen L, Duan L, Brekke JS. A pilot test of a peer navigator intervention for improving the health of individuals with serious mental illness. Community Ment Health J 2014;50(4):435-446. [doi: 10.1007/s10597-013-9616-4] [Medline: 23744292]
25. Ali-Faisal SF, Colella TJ, Medina-Jaudes N, Benz Scott L. The effectiveness of patient navigation to improve healthcare utilization outcomes: a meta-analysis of randomized controlled trials. Patient Educ Couns 2017;100(3):436-448. [doi: 10.1016/j.pec.2016.10.014] [Medline: 27771161]
26. Munce SEP, Shepherd J, Perrier L, Allin S, Sweet SN, Tomasone JR, et al. Online peer support interventions for chronic conditions: a scoping review protocol. BMJ Open 2017;7(9):e017999 [FREE Full text] [doi: 10.1136/bmjopen-2017-017999] [Medline: 28947464]
27. McBrien KA, Ivers N, Barnieh L, Bailey JJ, Lorenzetti DL, Nicholas D, et al. Patient navigators for people with chronic disease: a systematic review. PLoS One 2018;13(2):e0191980 [FREE Full text] [doi: 10.1371/journal.pone.0191980] [Medline: 29462179]
28. Bernardo BM, Zhang X, Hery CMB, Meadows RJ, Paskett ED. The efficacy and cost‐effectiveness of patient navigation programs across the cancer continuum: a systematic review. Cancer 2019;125(16):2747-2761. [doi: 10.1002/cncr.32147] [Medline: 31034604]

29. Tan CHH, Wilson S, McConigley R. Experiences of cancer patients in a patient navigation program: a qualitative systematic review. JBI Database System Rev Implement Rep 2015;13(2):136-168. [doi: 10.11124/jbisrir-2015-1588] [Medline: 26447039]
30. Raut A, Thapa P, Citrin D, Schwarz R, Gauchan B, Bista D, et al. Design and implementation of a patient navigation system in rural Nepal: improving patient experience in resource-constrained settings. Healthc (Amst) 2015;3(4):251-257. [doi: 10.1016/j.hjdsi.2015.09.009] [Medline: 26699353]
31. Jean-Pierre P, Hendren S, Fiscella K, Loader S, Rousseau S, Schwartzbauer B, et al. Understanding the processes of patient navigation to reduce disparities in cancer care: perspectives of trained navigators from the field. J Cancer Educ 2011;26(1):111-120 [FREE Full text] [doi: 10.1007/s13187-010-0122-x] [Medline: 20407860]
32. Krulic T, Brown G, Bourne A. A scoping review of peer navigation programs for people living with HIV: form, function and effects. AIDS Behav 2022;26(12):4034-4054 [FREE Full text] [doi: 10.1007/s10461-022-03729-y] [Medline: 35672548]
33. Bender JL, Flora PK, Soheilipour S, Dirlea M, Maharaj N, Parvin L, et al. Web-based peer navigation for men with prostate cancer and their family caregivers: a pilot feasibility study. Curr Oncol 2022;29(6):4285-4299 [FREE Full text] [doi: 10.3390/curroncol29060343] [Medline: 35735452]
34. Bandura A. Social cognitive theory: an agentic perspective. Annu Rev Psychol 2001;52:1-26. [doi: 10.1146/annurev.psych.52.1.1] [Medline: 11148297]
35. Cabassa LJ, Camacho D, Vélez-Grau CM, Stefancic A. Peer-based health interventions for people with serious mental illness: a systematic literature review. J Psychiatr Res 2017;84:80-89. [doi: 10.1016/j.jpsychires.2016.09.021] [Medline: 27701013]
36. Reid AE. Exploring the Role of Lay and Professional Patient Navigators in Canada. New Brunswick: Graduate Academic Unit of Interdisciplinary studies, University of New Brunswick; 2019.
37. Rocque G, Dionne-Odom J, Williams C, Jackson BE, Taylor R, Pisu M, et al. Implementation and impact of lay navigator-led advance care planning for cancer patients (FR440C). J Pain Symptom Manage 2017;53(2):368. [doi: 10.1016/j.jpainsymman.2016.12.132]
38. Steinberg ML, Fremont A, Khan DC, Huang D, Knapp H, Karaman D, et al. Lay patient navigator program implementation for equal access to cancer care and clinical trials: essential steps and initial challenges. Cancer 2006;107(11):2669-2677 [FREE Full text] [doi: 10.1002/cncr.22319] [Medline: 17078056]
39. Freund KM, Haas JS, Lemon SC, White KB, Casanova N, Dominici LS, et al. Standardized activities for lay patient navigators in breast cancer care: recommendations from a citywide implementation study. Cancer 2019;125(24):4532-4540 [FREE Full text] [doi: 10.1002/cncr.32432] [Medline: 31449680]
40. Casillas JN, Schwartz LF, Crespi CM, Ganz PA, Kahn KL, Stuber ML, et al. The use of mobile technology and peer navigation to promote adolescent and young adult (AYA) cancer survivorship care: results of a randomized controlled trial. J Cancer Surviv 2019;13(4):580-592 [FREE Full text] [doi: 10.1007/s11764-019-00777-7] [Medline: 31350681]
41. Zaagsma M, Volkers K, Swart E, Schippers A, Van Hove G. The use of online support by people with intellectual disabilities living independently during COVID-19. J Intellect Disabil Res 2020;64(10):750-756 [FREE Full text] [doi: 10.1111/jir.12770] [Medline: 32830390]
42. Colquhoun HL, Levac D, O'Brien KK, Straus S, Tricco AC, Perrier L, et al. Scoping reviews: time for clarity in definition, methods, and reporting. J Clin Epidemiol 2014;67(12):1291-1294. [doi: 10.1016/j.jclinepi.2014.03.013] [Medline: 25034198]
43. Peters M, Godfrey C, Khalil H, McInerney P, Parker D, Soares C. Guidance for conducting systematic scoping reviews. Int J Evid Based Healthc 2015;13(3):141-146. [doi: 10.1097/XEB.0000000000000050] [Medline: 26134548]
44. Peters MD, Godfrey CM, McInerney P, Soares CB, Khalil H, Parker D. The Joanna Briggs Institute reviewers' manual 2015: methodology for JBI scoping reviews. TJBI 2015:1-24. [doi: 10.46658/jbimes-24-09]
45. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth 2020;18(10):2119-2126. [doi: 10.11124/JBIES-20-00167] [Medline: 33038124]
46. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Social Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
47. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
48. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
49. Diseases and conditions. Canada PHAo. URL: https://www.canada.ca/en/public-health/services/diseases.html [accessed 2024-08-03]
50. Betancourt M, Roberts K, Bennett T, Driscoll E, Jayaraman G, Pelletier L. Monitoring chronic diseases in Canada: the chronic disease indicator framework. Chronic Dis Inj Can 2014;34(1):1-30. [doi: 10.24095/hpcdp.34.s1.01f]
51. Preventing Chronic Diseases: A Vital Investment. Geneva, Switzerland: World Health Organization; 2005.
52. Liddy C, Shoemaker ES, Crowe L, Boucher LM, Rourke SB, Rosenes R, et al. How the delivery of HIV care in Canada aligns with the chronic care model: a qualitative study. PLoS One 2019;14(7):e0220516 [FREE Full text] [doi: 10.1371/journal.pone.0220516] [Medline: 31348801]

53. Macdonald M, Misener RM, Weeks L, Helwig M. Covidence vs Excel for the title and abstract review stage of a systematic review. Int J Evidence-Based Healthcare 2016;14(4):200-201. [doi: 10.1097/01.xeb.0000511346.12446.f2]
54. Babineau J. Product review: covidence (Systematic Review Software). J Can Health Libr Assoc 2014;35(2):68-71. [doi: 10.5596/c14-016]
55. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348(mar07 3):g1687. [doi: 10.1136/bmj.g1687]
56. Peters M, Marnie C, Tricco A, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Implement 2021;19(1):3-10. [doi: 10.1097/XEB.0000000000000277] [Medline: 33570328]
57. George D, Mallery P. Descriptive statistics. In: IBM SPSS Statistics 25 Step by Step. UK: Routledge; 2018:126-134.
58. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
59. Nahm ES. Advancing telemedicine outreach within a comprehensive cancer center: lessons learned from rapid transformation of care practice and operational changes associated with the COVID-19 pandemic. JONS 2021;12 [FREE Full text]
60. Ormel HL, van der Schoot GGF, Westerink NDL, Sluiter WJ, Gietema JA, Walenkamp AME. Self-monitoring physical activity with a smartphone application in cancer patients: a randomized feasibility study (SMART-trial). Support Care Cancer 2018;26(11):3915-3923 [FREE Full text] [doi: 10.1007/s00520-018-4263-5] [Medline: 29785635]
61. Monteiro-Guerra F, Signorelli GR, Tadas S, Zubiete ED, Romero OR, Fernandez-Luque L, et al. A personalized physical activity coaching app for breast cancer survivors: design process and early prototype testing. JMIR Mhealth Uhealth 2020;8(7):e17552 [FREE Full text] [doi: 10.2196/17552] [Medline: 32673271]
62. Zulman DM, Jenchura EC, Cohen DM, Lewis ET, Houston TK, Asch SM. How can eHealth technology address challenges related to multimorbidity? Perspectives from patients with multiple chronic conditions. J Gen Intern Med 2015;30(8):1063-1070 [FREE Full text] [doi: 10.1007/s11606-015-3222-9] [Medline: 25691239]
63. Allen M, Iezzoni LI, Huang A, Huang L, Leveille SG. Internet-based coaching to improve patient-clinician communication in primary care. AMIA Annu Symp Proc 2007 Oct 11:861. [Medline: 18693963]
64. Simske NM, Breslin MA, Hendrickson SB, York KP, Vallier HA. Implementing recovery resources in trauma care: impact and implications. OTA Int 2019;2(4):e045 [FREE Full text] [doi: 10.1097/OI9.0000000000000045] [Medline: 33937673]
65. Westergaard RP, Genz A, Panico K, Surkan PJ, Keruly J, Hutton HE, et al. Acceptability of a mobile health intervention to enhance HIV care coordination for patients with substance use disorders. Addict Sci Clin Pract 2017;12(1):1-9 [FREE Full text] [doi: 10.1186/s13722-017-0076-y] [Medline: 28441962]
66. Gehrke A, Lee SS, Hilton K, Ganster B, Trupp R, McCullough C, et al. Development of the cancer survivor profile-breast cancer (CSPro-BC) app: patient and nurse perspectives on a new navigation tool. J Cancer Surviv 2018;12(3):291-305. [doi: 10.1007/s11764-017-0668-2] [Medline: 29524014]
67. Nitsch M, Dimopoulos CN, Flaschberger E, Saffran K, Kruger JF, Garlock L, et al. A guided online and mobile self-help program for individuals with eating disorders: an iterative engagement and usability study. J Med Internet Res 2016;18(1):e7 [FREE Full text] [doi: 10.2196/jmir.4972] [Medline: 26753539]
68. Moradian S, Krzyzanowska MK, Maguire R, Morita PP, Kukreti V, Avery J, et al. Usability evaluation of a mobile phone-based system for remote monitoring and management of chemotherapy-related side effects in cancer patients: mixed-methods study. JMIR Cancer 2018;4(2):e10932 [FREE Full text] [doi: 10.2196/10932] [Medline: 30578238]
69. Fredriksen EH, Harris J, Moland KM. Web-based discussion forums on pregnancy complaints and maternal health literacy in norway: a qualitative study. J Med Internet Res 2016;18(5):e113 [FREE Full text] [doi: 10.2196/jmir.5270] [Medline: 27230094]
70. Loiselle CG, Peters O, Haase KR, Girouard L, Körner A, Wiljer D, et al. Virtual navigation in colorectal cancer and melanoma: an exploration of patients' views. Support Care Cancer 2013;21(8):2289-2296. [doi: 10.1007/s00520-013-1771-1] [Medline: 23519565]
71. Sánchez-Ortiz VC, House J, Munro C, Treasure J, Startup H, Williams C, et al. "A computer isn't gonna judge you": a qualitative study of users' views of an internet-based cognitive behavioural guided self-care treatment package for bulimia nervosa and related disorders. Eat Weight Disord 2011;16(2):e93-e101. [doi: 10.1007/BF03325314] [Medline: 21989103]
72. Hinchliffe A, Mummery WK. Applying usability testing techniques to improve a health promotion website. Health Promot J Austr 2008;19(1):29-35. [doi: 10.1071/he08029] [Medline: 18481929]
73. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg 2011;128(1):305-310 [FREE Full text] [doi: 10.1097/PRS.0b013e318219c171] [Medline: 21701348]
74. Kokorelias KM, Nelson ML, Tang T, Gray CS, Ellen M, Plett D, et al. Inclusion of older adults in digital health technologies to support hospital-to-home transitions: secondary analysis of a rapid review and equity-informed recommendations. JMIR Aging 2022;5(2):e35925 [FREE Full text] [doi: 10.2196/35925] [Medline: 35475971]
75. Hossain SN, Jaglal SB, Shepherd J, Perrier L, Tomasone JR, Sweet SN, et al. Web-based peer support interventions for adults living with chronic conditions: scoping review. JMIR Rehabil Assist Technol 2021;8(2):e14321 [FREE Full text] [doi: 10.2196/14321] [Medline: 34032572]

76. Fong JH. Disability incidence and functional decline among older adults with major chronic diseases. BMC Geriatr 2019;19(1):323 [FREE Full text] [doi: 10.1186/s12877-019-1348-z] [Medline: 31752701]
77. Heponiemi T, Jormanainen V, Leemann L, Manderbacka K, Aalto A, Hyppönen H. Digital divide in perceived benefits of online health care and social welfare services: national cross-sectional survey study. J Med Internet Res 2020;22(7):e17616 [FREE Full text] [doi: 10.2196/17616] [Medline: 32673218]
78. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020;22(6):e19361 [FREE Full text] [doi: 10.2196/19361] [Medline: 32452816]
79. Wilson-Howard D, Vilaro MJ, Neil JM, Cooks EJ, Griffin LN, Ashley TT, et al. Development of a credible virtual clinician promoting colorectal cancer screening via telehealth apps for and by black men: qualitative study. JMIR Form Res 2021;5(12):e28709. [doi: 10.2196/28709] [Medline: 34780346]
80. Greene GJ, Reidy E, Felt D, Marro R, Johnson AK, Phillips G, et al. Implementation and evaluation of patient navigation in Chicago: insights on addressing the social determinants of health and integrating HIV prevention and care services. Eval Program Plann 2022;90:101977. [doi: 10.1016/j.evalprogplan.2021.101977] [Medline: 34373116]
81. Freund KM. Implementation of evidence-based patient navigation programs. Acta Oncol 2017;56(2):123-127. [doi: 10.1080/0284186X.2016.1266078] [Medline: 28033027]
82. Friel S, Marmot MG. Action on the social determinants of health and health inequities goes global. Annu Rev Public Health 2011;32(1):225-236. [doi: 10.1146/annurev-publhealth-031210-101220]
83. Sousa C, Hagopian A, Stoller N. Addressing the social determinants of health through public health policy: a case study of US-based advocacy efforts for health justice in occupied Palestinian territory. The Lancet 2019;393(Supplement 1):S49. [doi: 10.1016/s0140-6736(19)30635-x]
84. Bierman AS, Dunn JR. Swimming upstream. Access, health outcomes, and the social determinants of health. J Gen Intern Med 2006;21(1):99 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.00317.x] [Medline: 16423133]
85. Artiga S, Hinton E. Beyond health care: the role of social determinants in promoting health and health equity. Health 2019;20(10):1-13 [FREE Full text]
86. Bambra C, Gibson M, Sowden A, Wright K, Whitehead M, Petticrew M. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. J Epidemiol Community Health 2010;64(4):284-291 [FREE Full text] [doi: 10.1136/jech.2008.082743] [Medline: 19692738]
87. Jandorf L, Braschi C, Ernstoff E, Wong CR, Thelemaque L, Winkel G, et al. Culturally targeted patient navigation for increasing African Americans' adherence to screening colonoscopy: a randomized clinical trial. Cancer Epidemiol Biomarkers Prev 2013;22(9):1577-1587 [FREE Full text] [doi: 10.1158/1055-9965.EPI-12-1275] [Medline: 23753039]
88. Oluyede L, Cochran AL, Wolfe M, Prunkl L, McDonald N. Addressing transportation barriers to health care during the COVID-19 pandemic: perspectives of care coordinators. Transp Res Part A Policy Pract 2022;159:157-168 [FREE Full text] [doi: 10.1016/j.tra.2022.03.010] [Medline: 35283561]
89. Hamerly D. Review of "Inequity in the technopolis: race, class, gender, and the digital divide in Austin" (University of Texas Press, 2012). FM 2012;17(11). [doi: 10.5210/fm.v17i11.4286]
90. Cotterill S, Knowles S, Martindale AM, Elvey R, Howard S, Coupe N, et al. Getting messier with TIDieR: embracing context and complexity in intervention reporting. BMC Med Res Methodol 2018;18(1):12 [FREE Full text] [doi: 10.1186/s12874-017-0461-y] [Medline: 29347910]
91. Wilcox B, Bruce SD. Patient navigation: a "win-win" for all involved. Oncol Nurs Forum 2010;37(1):21-25. [doi: 10.1188/10.ONF.21-25] [Medline: 20044337]
92. Koorts H, Eakin E, Estabrooks P, Timperio A, Salmon J, Bauman A. Implementation and scale up of population physical activity interventions for clinical and community settings: the PRACTIS guide. Int J Behav Nutr Phys Act 2018;15(1):51 [FREE Full text] [doi: 10.1186/s12966-018-0678-0] [Medline: 29884236]
93. Robinson KL, Watters S. Bridging the communication gap through implementation of a patient navigator program. Pa Nurse 2010;65(2):19-22. [Medline: 20666161]
94. Rozario MA, Walton A, Kang M, Padilla BI. Colorectal cancer screening: a quality improvement initiative using a bilingual patient navigator, mobile technology, and fecal immunochemical testing to engage hispanic adults. CJON 2021;25(4):423-429. [doi: 10.1188/21.cjon.423-429]
95. Vogel WB, Morris HL, Muller K, Huo T, Parish A, Stoner D, et al. Cost-effectiveness of the wellness incentives and navigation (WIN) program. Value Health 2021;24(3):361-368 [FREE Full text] [doi: 10.1016/j.jval.2020.06.019] [Medline: 33641770]
96. Jolly SE, Navaneethan SD, Schold JD, Arrigain S, Konig V, Burrucker YK, et al. Development of a chronic kidney disease patient navigator program. BMC Nephrol 2015;16:69 [FREE Full text] [doi: 10.1186/s12882-015-0060-2] [Medline: 26024966]
97. Willis A, Reed E, Pratt-Chapman M. Development of a framework for patient navigation: delineating roles across navigator types. J Oncol Navig Surviv 2013;4(6):20 [FREE Full text]
98. Luckett R, Pena N, Vitonis A, Bernstein MR, Feldman S. Effect of patient navigator program on no-show rates at an academic referral colposcopy clinic. J Womens Health (Larchmt) 2015;24(7):608-615. [doi: 10.1089/jwh.2014.5111] [Medline: 26173000]

99. Ranaghan C, Boyle K, Meehan M, Moustapha S, Fraser P, Concert C. Effectiveness of a patient navigator on patient satisfaction in adult patients in an ambulatory care setting: a systematic review. JBI Database System Rev Implement Rep 2016;14(8):172-218. [doi: 10.11124/JBISRIR-2016-003049] [Medline: 27635752]
100. Sullivan C, Dolata J, Barnswell K, Greenway K, Kamps C, Marbury Q, et al. Experiences of kidney transplant recipients as patient navigators. Transplant Proc 2018;50(10):3346-3350 [FREE Full text] [doi: 10.1016/j.transproceed.2018.02.090] [Medline: 30577205]
101. Lubetkin EI, Lu W, Krebs P, Yeung H, Ostroff JS. Exploring primary care providers' interest in using patient navigators to assist in the delivery of tobacco cessation treatment to low income, ethnic/racial minority patients. J Community Health 2010;35(6):618-624. [doi: 10.1007/s10900-010-9251-8] [Medline: 20336355]
102. Kokorelias KM, DasGupta T, Hitzig SL. Designing the ideal patient navigation program for older adults with complex needs: a qualitative exploration of the preferences of key informants. J Appl Gerontol 2022;41(4):1002-1010. [doi: 10.1177/07334648211059056] [Medline: 34905440]
103. Kokorelias KM, Markoulakis R, Hitzig SL. Considering a need for dementia-specific, family-centered patient navigation in Canada. J Appl Gerontol 2023;42(1):19-27. [doi: 10.1177/07334648221125781] [Medline: 36503280]

Abbreviations

JBI: Joanna Briggs Institute
PHAC: Public Health Agency of Canada
PNI: patient navigation intervention
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
TIDieR: Template for Intervention Description and Replication

Edited by C Lovis; submitted 26.06.23; peer-reviewed by W van Harten, B Crocker; comments to author 25.10.23; revised version received 19.12.23; accepted 30.06.24; published 19.08.24.

Please cite as:
Marsh M, Shah SR, Munce SEP, Perrier L, Lee TSJ, Colella TJF, Kokorelias KM
Characteristics of Existing Online Patient Navigation Interventions: Scoping Review
JMIR Med Inform 2024;12:e50307
URL: https://medinform.jmir.org/2024/1/e50307
doi:10.2196/50307
PMID:

Original Paper

How Patient-Generated Data Enhance Patient-Provider Communication in Chronic Care: Field Study in Design Science Research

Dario Staehelin[1,2], MA; Mateusz Dolata[1], PhD; Livia Stöckli[1], MSc; Gerhard Schwabe[1], PhD

[1]Department of Informatics, University of Zurich, Zurich, Switzerland

[2]Department for Information and Process Management, Eastern Switzerland University of Applied Sciences, St Gallen, Switzerland

Corresponding Author:
Dario Staehelin, MA
Department of Informatics
University of Zurich
Binzmühlestrasse 14
Zurich, 8050
Switzerland
Phone: 41 763103137
Email: dario.staehelin@ost.ch

Abstract

Background: Modern approaches such as patient-centered care ask health care providers (eg, nurses, physicians, and dietitians) to activate and include patients to participate in their health care. Mobile health (mHealth) is integral in this endeavor to be more patient centric. However, structural and regulatory barriers have hindered its adoption. Existing mHealth apps often fail to activate and engage patients sufficiently. Moreover, such systems seldom integrate well with health care providers' workflow.

Objective: This study investigated how patient-provider communication behaviors change when introducing patient-generated data into patient-provider communication.

Methods: We adopted the design science approach to design PatientHub, an integrated digital health system that engages patients and providers in patient-centered care for weight management. PatientHub was developed in 4 iterations and was evaluated in a 3-week field study with 27 patients and 6 physicians. We analyzed 54 video recordings of PatientHub-supported consultations and interviews with patients and physicians.

Results: PatientHub introduces patient-generated data into patient-provider communication. We observed 3 emerging behaviors when introducing patient-generated data into consultations. We named these behaviors *emotion labeling*, *expectation decelerating*, and *decision ping-pong*. Our findings show how these behaviors enhance patient-provider communication and facilitate patient-centered care. Introducing patient-generated data leads to behaviors that make consultations more personal, actionable, trustworthy, and equal.

Conclusions: The results of this study indicate that patient-generated data facilitate patient-centered care by activating and engaging patients and providers. We propose 3 design principles for patient-centered communication. Patient-centered communication informs the design of future mHealth systems and offers insights into the inner workings of mHealth-supported patient-provider communication in chronic care.

(JMIR Med Inform 2024;12:e57406) doi:10.2196/57406

KEYWORDS

patient-provider communication; patient-generated data; field study; chronic care; design science research; patient-centered care; integrated care; patient-provider collaboration; mobile phone

Introduction

Background

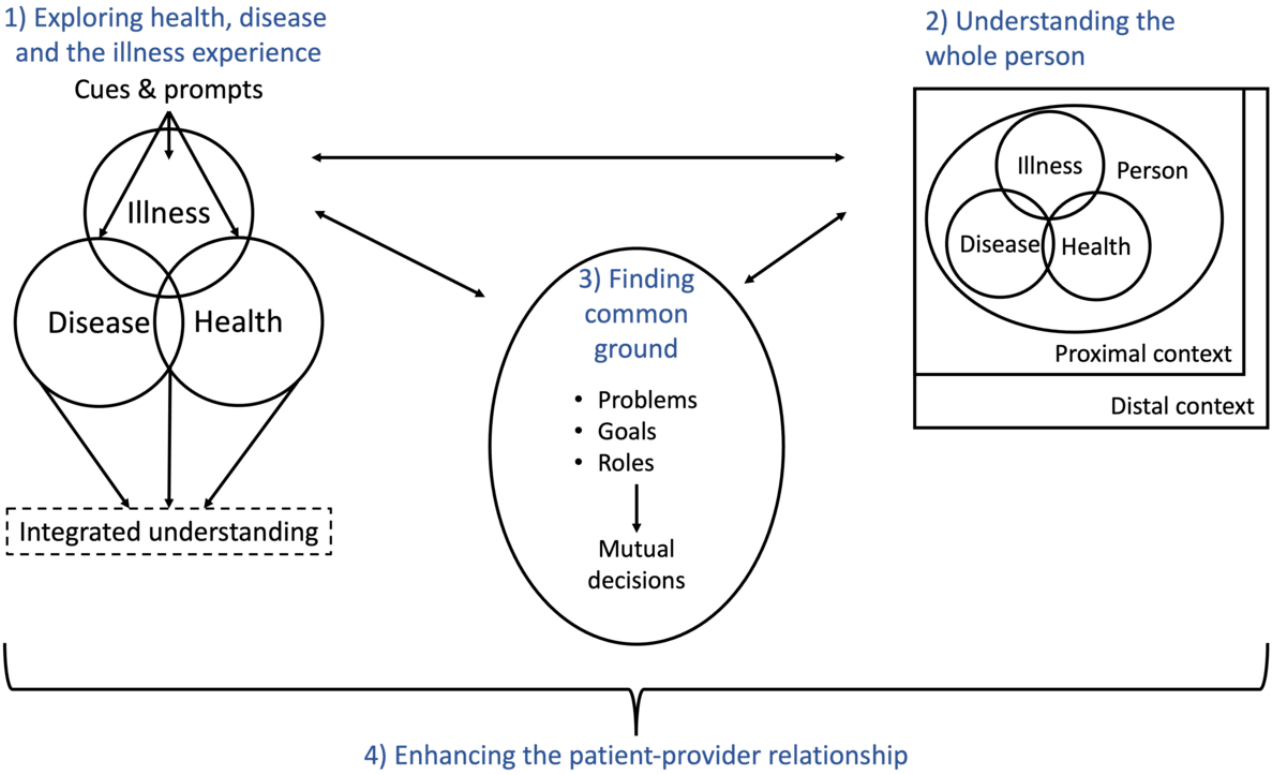
The quality of the patient-provider relationship is strongly linked to patients' increased adherence and better health outcomes [1-3]. Patient-provider communication requires exchanging accurate and relevant information to better understand patients and their preferences and context [4,5]. However, provider instructions are complex and not communicated adequately to patients [3,6] as health care providers often lack the time or communication training [7]. In turn, patients have difficulties recalling crucial information (eg, their adherence to taking medication regularly), impeding providers' ability to quickly assess their medical condition and derive actions [8,9]. Therefore, adherence and health outcomes are often subpar—especially in people with chronic conditions [6,10,11]. Due to its centrality, improving patient-provider communication is a topic of continued interest in medical research.

Patient-Centered Care

Over the last decades, the understanding of good patient-provider communication has evolved. Historically, providers have possessed most of the power in patient-provider communication [12]. These power dynamics mainly occurred due to the significant knowledge difference between providers and patients [13]. The resulting paternalistic model, in which the providers made all the decisions, led to poor adherence and increased health care costs [14,15]. Newer approaches such as shared decision-making and patient-centered care ask providers to adopt more inclusive methods focusing on the collaborative nature of patient-provider communication [4,5,16-18].

Patient-centered care proposes a holistic clinical method that centers on the patients and their preferences and contexts [4,5,19]. It is defined as "respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions" and is 1 of 6 key elements of high-quality care [16]. The clinical method of patient-centered medicine by Stewart et al [4] is among the most frequently used frameworks. It proposes a communication approach that suggests looking beyond a patient's acute problem and into their history and context (Figure 1). The framework suggests (1) *exploring health, disease, and the illness experience*; (2) *understanding the whole person*; and (3) *finding common ground* to (4) *enhance the patient-clinician relationship*. In the following paragraphs, we describe each of those dimensions.

Figure 1. Patient-centered care framework by Stewart et al [4].



Exploring health, disease, and the illness experience highlights the importance of understanding the patient's experience. People have different interpretations and unique experiences of health and illness. Someone with an asymptomatic disease may perceive themselves as healthy, whereas others feel ill without having a disease. Providers should seek to understand and support the patient's view of their situation and their experience of health and illness by listening to their concerns and feelings [4,20]

Understanding the whole person focuses on understanding the patient's proximal and distal context [4]. It enables meaningful conversations about the illness and treatment options as providers know their patients as people (eg, what is currently important to them) [20]. Younger patients might struggle more with a diagnosis than older people. A patient's relationships,

work, education, lifestyle, and culture play a significant role in their treatment.

Patients and providers work together to *find common ground* regarding the problems and priorities, the goals of the therapy, and the roles of the patient and the provider [4]. It stresses the emotional engagement with the patient and the genuinely collaborative aspect of finding common ground to arrive at mutual decisions [4].

Enhancing the patient-clinician relationship is the goal of every encounter in patient-centered medicine [4]. Health care becomes genuinely patient centered through an integrated understanding of the patient's experience, understanding them as a whole person, and mutual decisions [20]. The framework by Stewart et al [4] conceptualizes patient-provider communication to achieve patient-centered care. It proposes a mindset that places patients at the center of clinical practice.

Patient-Generated Data in Patient-Provider Communication

This shift in the mindset has attracted increasing interest in health informatics research that studies the effect of technology on patient-provider communication [21-27]. In face-to-face consultations, patient data generated on mobile health (mHealth) apps become increasingly important as they allow patients and providers to gain deeper insights into patients' routines and adherence to therapy plans. Patient-generated data are health-related data gathered or created by patients, usually through wearables and mHealth [28]. Studies in health informatics and related fields have demonstrated the potential of patient-generated data to increase patient-provider communication [22,29,30]. mHealth allows patients to generate abundant health information, such as dietary patterns, emotional conditions, or objective measures such as blood pressure [22,23,31]. These data allow for insights into the patient's health experience and journey unlike ever before [32,33].

mHealth-supported approaches have significantly improved patient-provider communication, adherence, and health outcomes in chronic care [34-36]. Studies have shown how these patient-generated data allow patients and providers to engage in collaborative sensemaking that improves decision-making [23,30,31,37]. It allows for deeper discussions about personal values [14,16] and improves patients' understanding of their condition and treatment [38]. Furthermore, introducing patient-generated data into consultations affects the role dynamics of therapeutic sessions [34]. For example, sharing clinical notes shifts power in the patient-provider relationship [36]. Other studies report how sharing patient-generated data through mHealth leads to greater disclosure and better communication in consultations, resulting in better health outcomes [30,31].

These insights have been validated for different age groups [30,39,40] and chronic conditions (eg, chronic kidney disease) [21,25,35,41]. This previous research shows the positive impact of introducing patient-generated data into consultations on adherence and health outcomes [35,36]. For example, Vitger et al [35] describe the positive impact of generating data on a smartphone app on patient activation, communication confidence, and preparedness for decision-making in patients with schizophrenia.

While existing research agrees on the vital role of patient-generated data in patient-provider communication [19,21,23], significant obstacles remain to leverage their potential. So far, structural and regulatory barriers have slowed advances [42,43]. mHealth apps seldom integrate with the provider's workflow, leading to a fragmentation of health data [44,45]. More importantly, Cozad et al [46] found that only a few mHealth apps engage and activate patients to participate in patient-centered care. Finally, most studies report on the positive effects of patient-generated data on patient-provider communication, but they often fail to investigate the communication behaviors that use patient-generated data. Accordingly, patient-provider communication remains a black box that receives patient-generated data as input and creates better communication as output (Figure 2). Little to no research focuses on the design of systems that (1) integrate patient-generated data into the provider's workflow and (2) use these data in consultations to enhance patient-provider communication. This study aimed to address this research gap by designing and evaluating an integrated digital health system that enhances the patient-provider communication.

Figure 2. The process of patient-provider communication as a black box.

Methods

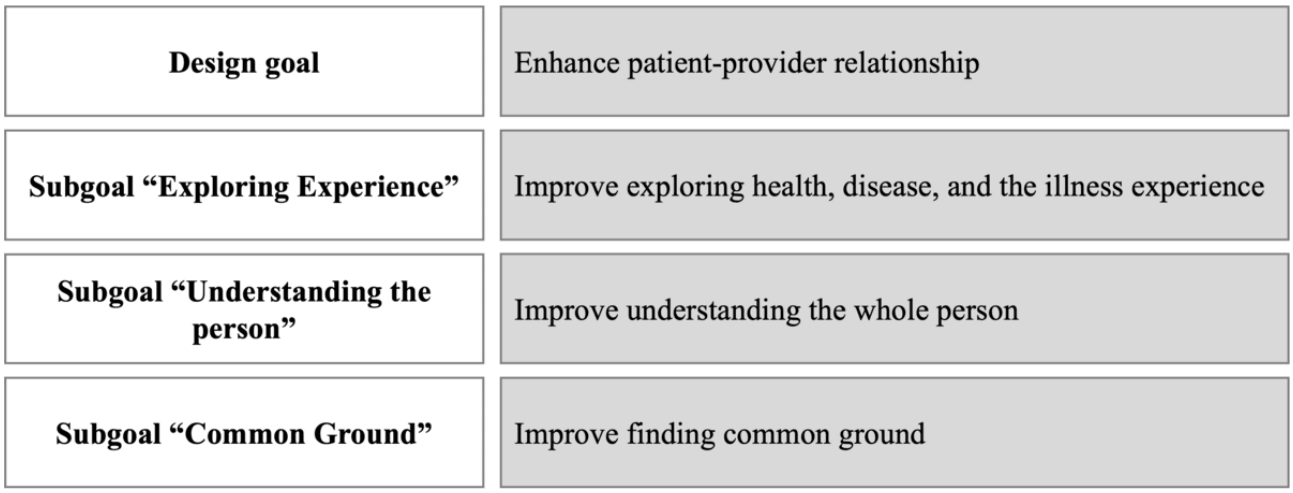
Overview

This study addressed the research gap described previously by developing PatientHub in a design science research (DSR) approach to enhance the patient-clinician relationship [47,48]. DSR is a suitable approach as it systematically solves important general problems and generates new knowledge in the form of design principles, theoretical models, approaches, and impacts of technology use [49]. DSR proposes to ground a solution's design in existing knowledge and theories, so-called kernel theories, to justify design decisions [50]. Due to these properties, design science is increasingly applied in medical informatics to study emerging technologies [48,51,52].

To address the research gap, we (1) designed PatientHub and (2) studied its impact on patient-provider communication by

adopting the patient-centered care framework proposed by Stewart et al [4] as our kernel theory. PatientHub is an integrated digital health tool that introduces patient-generated data into consultations. We build on the strong correlation between patient-generated data and improved patient-provider communication established in recent work [34-36]. While the patient-centered care framework offers a holistic foundation for improving patient-provider communication, it lacks a clear operationalization of the 3 dimensions that offer mHealth designers and health care providers guidance on implementing patient-centered care. Specifically, it is unclear how to integrate patient-generated data into the consultation process and how patients and providers use them for patient-centered care. Accordingly, we formulated a design goal and several subgoals in line with our kernel theory (Figure 3 [4]).

Figure 3. Design goal and subgoals based on the patient-centered care framework [4].



The following sections present the PatientHub design, our field study approach, and the data analysis method.

PatientHub Design

The project team designed PatientHub in 4 iterations. Our project team included a medical informatics company, a health institute specializing in chronic care, and 2 research institutions.

PatientHub's design is grounded in the patient-centered care framework [4] and leverages existing design knowledge [21-23,34,39,53]. Over the 3 preliminary iterations, we continuously evaluated and improved the design. We tested the first design iteration by applying the think-aloud method with 7 participants acting as patients [54]. After refining the design in the second iteration, a focus group of 5 domain experts from software development, medicine, and research evaluated the revised PatientHub. In the third iteration, 3 health care providers and 5 patients evaluated the design in role-plays of consultations. This paper reports on the field study evaluating the prototype with actual patients and physicians.

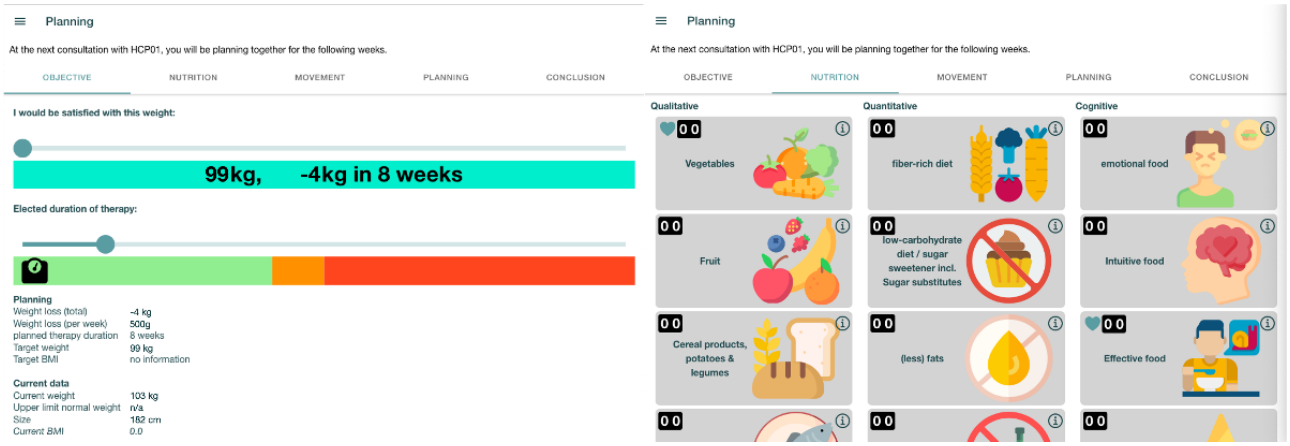
PatientHub aims to enhance patient-provider communication by integrating patient-generated data into the consultation. It consists of a patient app, where patients generate data, and a consultation app, where patient-generated data provide a foundation for discussion in the consultation. In the following section, we describe the design implementation using a scenario and screenshots of PatientHub. It represents a potential user story of PatientHub during the field study.

PatientHub Scenario

John is a patient at Laura's clinic struggling with obesity. Last week, Laura proposed that John try PatientHub to help them advance John's journey to better health. In the patient app on his smartphone, John tracked his dietary and activity habits in daily notes in a digital journal, filled out a general health questionnaire, and selected favorites for behavior change interventions as part of a 1-week preparation phase.

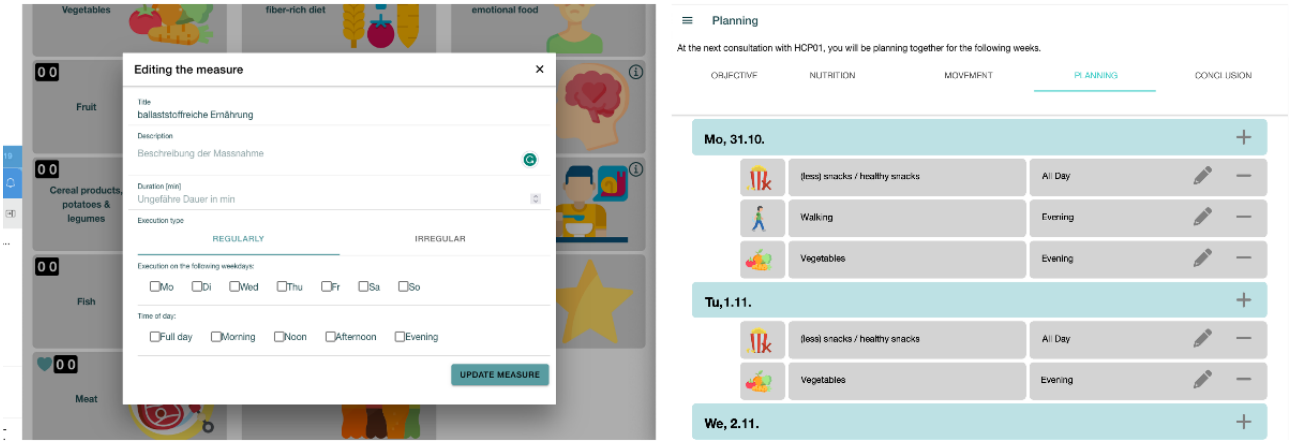
In the initial consultation, Laura and John review the journal entries, questionnaire answers, and intervention favorites in the consultation app on a tablet. The consultation app consists of 4 screens: goal setting, defining dietary and activity interventions, planning, and closing. For goal setting, John and Laura discuss target weight and therapy duration with a tablet-based visualization using sliders (Figure 4, left). As the visualization relates weight loss and duration to each other, they can discuss healthy weight loss and set realistic goals. To define dietary and activity interventions, John liked 3 interventions he would like to explore. John and Laura can discuss additional interventions from a list of obesity-friendly interventions (Figure 4, right; selected by medical professionals).

Figure 4. Goal setting (left) and dietary intervention (right) app screens.



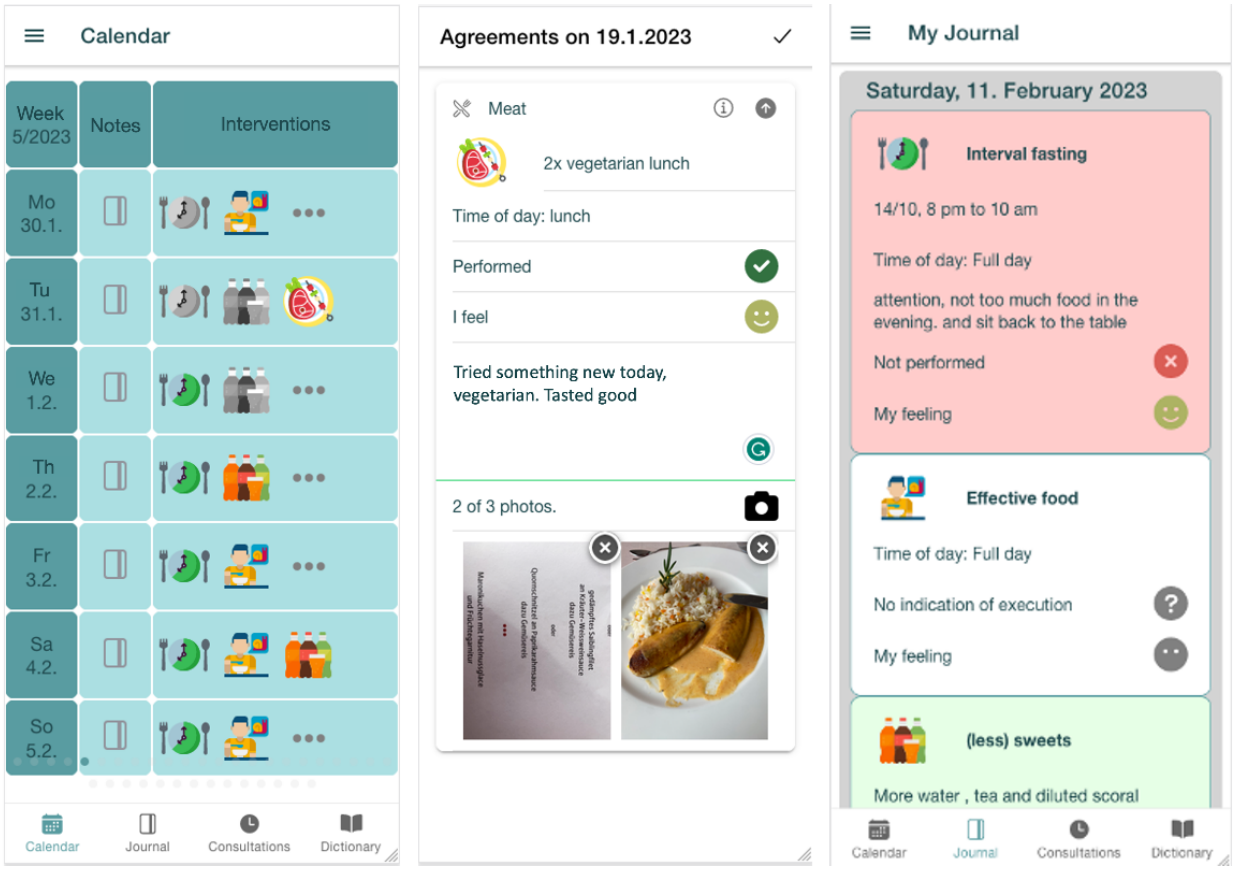
To create an intervention, they specify a name, a description (eg, 8000 steps per day), a duration (if applicable), recurrence (ie, regular or irregular), and preferred days and times (Figure 5, left). The consultation app allows them to discuss the interventions to arrive at a patient-centered therapy plan considering John's specific context. In planning, John and Laura see an overview of the interventions in calendar form (Figure 5, right). They can adjust the therapy plan if necessary (eg, move an intervention from Monday to Tuesday). Once the therapy plan is finalized, all information is automatically shared with John on the patient app.

Figure 5. Measure specification (left) and planning (right) app screens.



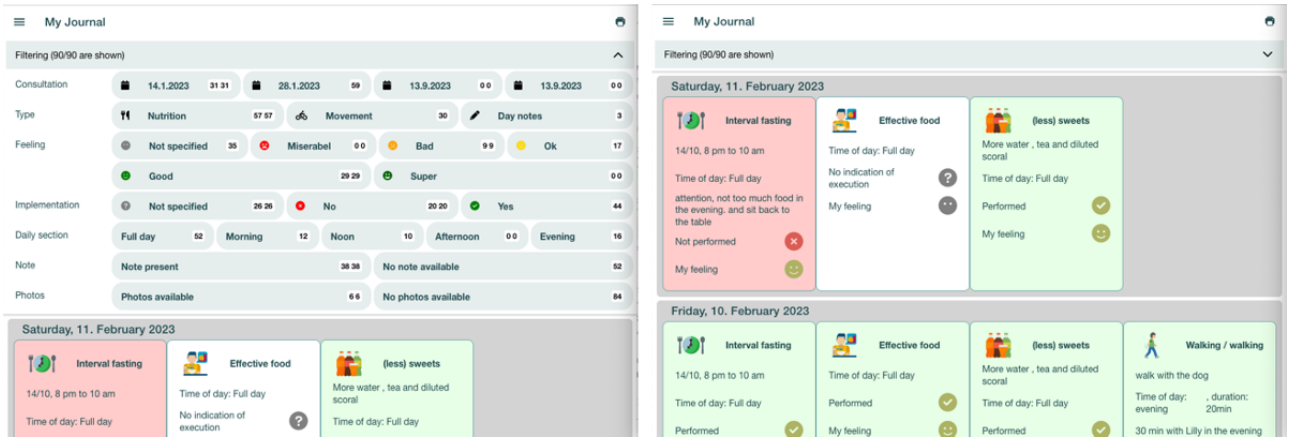
John tracks his health journey in PatientHub's journal for 2 weeks. He tracks his adherence to the therapy plan and his experience while executing it. John sees each intervention in the calendar, where open interventions are grayed out and become colored once completed (Figure 6, left). For example, John had to limit carbohydrate-dense foods today. He clicks on the gray task icon to create a task-specific entry. He marks the task as completed and sets his emotional state to medium as he missed out on dessert today. John then uploads a picture of his lunch and writes a note (Figure 6, middle). He can now review his entries in his journal (Figure 6, right). John carried out the therapy plan and kept his digital journal during the 2-week implementation phase leading up to the follow-up consultation with Laura.

Figure 6. Calendar (left), journal entry (middle), and journal overview (right) app screens.



The follow-up consultation focuses on reviewing John's journal. Laura and John apply different filters to the journal entries, such as task type (ie, diet, activity, and daily note), emotions, execution, and media type (Figure 7, left). This way, they can review interventions that were not completed or completed but not enjoyed by John (Figure 7, right). Through this discussion, they identify opportunities to improve the therapy plan and adherence. They adjust interventions as in the initial consultation by going through diet, activity, and planning before closing the consultation. Again, all data are shared across the PatientHub apps and the loop between consultations is closed. John enters a new implementation phase where he records his progress, which he will review again with Laura.

Figure 7. Filter options (left) and journal overview (right) app screens.



Data Collection

Overview

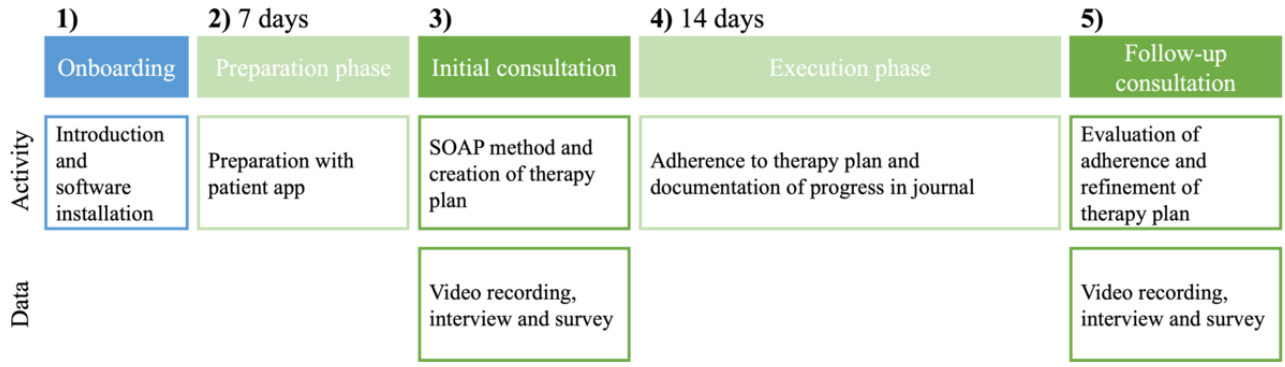
In line with the "human risk and effectiveness" strategy [55], we evaluated PatientHub in a naturalistic setting to assess its effects on patient-provider communication. For this purpose, we collected data in a 3-week field study. The data set contained video recordings of each consultation and interviews with each participant after both consultations. We analyzed 27 initial and follow-up consultations (54 recordings) to study emerging behaviors when integrating patient-generated data into the consultations. Furthermore, we analyzed interviews with all participants to study how patient-generated data enhance patient-provider communication (66 interviews in total; physicians were interviewed once per phase). In the interviews,

the participants reflected on their experiences throughout the field study. The authors and the project team developed the interview guides based on the kernel theory (ie, patient-centered care). The interview guide further included questions about the participants' experience using PatientHub and their implementation of the therapy plan. There were separate interview guides for patients and physicians and initial and follow-up consultations. English translations of the interview guides can be found in Multimedia Appendices 1-4. The interviews were held either in Swiss German or German by the lead author and experienced project members with digital health backgrounds. They were transcribed verbatim and anonymized.

The combination of video recordings and interviews provided us with a rich data set to evaluate the impact of PatientHub on patient-provider communication behaviors. Furthermore, evaluating in the field offered us valuable insights into the experience of patients and physicians when engaging with PatientHub.

The 3-week field study comprised five phases: (1) onboarding, (2) preparation phase, (3) initial consultation, (4) execution phase, and (5) follow-up consultation. Patients were asked to engage with the patient app of PatientHub during the preparation and execution phases. In the following sections, we outline the study design in detail, depicted in Figure 8.

Figure 8. Field study design. SOAP: Subjective, Objective, Assessment, Plan.



Onboarding

Patients

The study team explained the study design to the patients, including the process for the 3 weeks and the study aim, and answered any questions. All patients were interviewed and completed a survey regarding their previous experience in chronic care. The patients installed and logged into the system with help from the study team. Finally, they received instructions for the upcoming week, the so-called preparation phase.

Physicians

The physicians were introduced to the study design and goal. They received training on PatientHub before the initial consultations. While the training focused on the consultation app, we also introduced the patient app to the physicians.

Preparation Phase (Patients Only)

Patients kept a journal regarding diet and physical activity as daily notes in the patient app. They were asked to complete a general health questionnaire and select 3 favorite dietary and activity interventions.

Initial Consultation (Patients and Physicians)

The goal of the initial consultation was for the patients and physicians to create a therapy plan with dietary and activity interventions. They reviewed the patient-generated data (ie, questionnaire and journal entries). Furthermore, they discussed a desired weight loss goal and therapy duration. Finally, they created a therapy plan for the following 2 weeks. During the consultation, patients and physicians could use the consultation tool (see the aforementioned description). Physicians saw 4 to 5 patients throughout the field study, and the patient-provider matching stayed the same.

Execution Phase (Patients Only)

Over 2 weeks, patients implemented the therapy plan and generated data on their progress in the journal. Before the follow-up consultation, they filled out a shortened health questionnaire.

Follow-Up Consultation (Patients and Physicians)

The goal of the follow-up consultation was to tailor the therapy plan to the individual patients. Patients and physicians reviewed the patient's adherence to the therapy plan together. Insights from the discussion led to adjusting the therapy plan for the next execution phase. This marked the end of the field study.

Participants

The field study included 28 patients and 6 health care providers (ie, physicians in this study). Only 4% (1/28) of the patients dropped out after the initial consultation (patient 24), whereas all physicians completed the 3-week study. Before the field study, we conducted a pretest in December 2022 with 2 participants to uncover and resolve software bugs and flaws in the study design. The actual field study took place in 2 clinics in Switzerland in 2 rounds between January 2023 and April 2023 for logistic reasons. The first round was conducted in January 2023 with 18 patients and 4 physicians, and the second round started in March 2023 involving 10 patients and 2 physicians. We maintained the same study design and evaluated the same prototype.

To evaluate the effects of PatientHub on patient-provider communication in the most realistic setting, the inclusion criteria for patients were (1) age of ≥18 years, (2) a BMI of >25 kg/m^2

or specific medical indications for weight loss (eg, diabetes), (3) ability to communicate in written and verbal German, and (4) ownership of a computer and smartphone and adequate handling of both. Physicians were selected through personal contacts and had to meet the following inclusion criteria: (1) licensed physicians in Switzerland, (2) experience counseling patients with overweight, (3) ability to speak German, and (4) familiar with computers and tablets in consultations. Tables 1 and 2 present the demographic data of the participating patients and physicians, respectively.

Table 1. Demographic data of participating patients.

Participant	Sex	Age (y)	BMI (kg/m^2)	Occupation
Patient 01	Male	63	36.8	Electrician
Patient 02	Female	72	35.4	Retired
Patient 03	Female	81	34.4	Retired
Patient 04	Male	41	27.6	Lecturer
Patient 05	Female	72	29.3	Librarian
Patient 06	Male	84	30.8	Retired
Patient 07	Female	78	30.5	Retired
Patient 08	Male	62	31.9	Auditor
Patient 09	Male	81	32.5	Production manager
Patient 10	Female	37	46.3	Bank clerk
Patient 11	Male	65	25.7	Electrician
Patient 12	Female	55	41.5	Office clerk
Patient 13	Male	76	30.6	Musician
Patient 14	Female	84	35.6	Homemaker
Patient 15	Female	65	30.7	Retired
Patient 16	Male	58	36.8	Driver
Patient 17	Male	60	22.3	Teacher
Patient 18	Female	61	35.3	Chairwoman
Patient 19	Male	56	30.5	Lecturer
Patient 20	Male	57	36.6	Scientific assistant
Patient 21	Female	71	40.4	Retired
Patient 22	Female	58	37.6	Depositary
Patient 23	Male	66	31.1	Retired
Patient 25	Female	59	41.1	Office clerk
Patient 26	Male	53	27.2	Remedial teacher
Patient 27	Male	75	39.5	Office clerk
Patient 28	Female	65	26.7	Architect

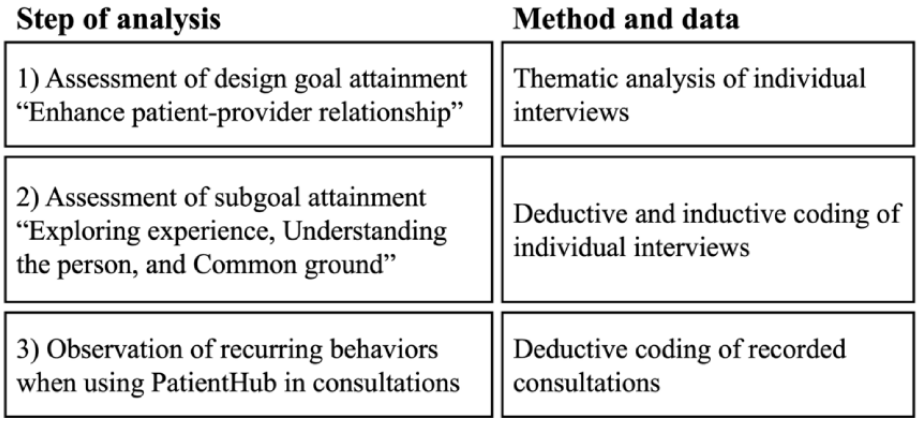
Table 2. Demographic data of physicians.

Participant	Sex	Age (y)	Workplace	Discipline
Provider 01	Female	28	Hospital	Surgery
Provider 02	Female	27	Hospital	Psychosomatics
Provider 03	Male	59	Hospital	General medicine
Provider 04	Female	58	Private clinic	General internal medicine
Provider 05	Male	52	Private clinic	General medicine
Provider 06	Female	44	Private clinic	General internal medicine

Data Analysis

The framework for evaluation in DSR proposes a continuum from formative to summative evaluation [55]. The framework offers 4 evaluation strategies, from which this study adopted the "human risk and effectiveness" strategy as the addressed problem is social and user centered [55]. Our analysis consisted of summative and formative elements as the study aimed to improve on the process under evaluation (ie, patient-provider communication). Accordingly, we applied deductive and inductive coding methods in 3 steps for the data analysis, as depicted in Figure 9.

Figure 9. Data analysis approach.



To study the effect of PatientHub on patient-provider communication, we analyzed individual interviews with patients and physicians and 54 video recordings of face-to-face initial and follow-up consultations (27 recordings each). This three-step analysis allowed us to assess (1) whether and (2) how PatientHub enhances patient-provider communication (interviews) and to (3) observe recurring behaviors when engaging with patient-generated data in consultations (video recordings). The research team consisted of a graduate student with a medical and information systems background (coding author), a PhD student in digital health (lead author), and 2 senior researchers in DSR.

First, we assessed the design goal attainment "enhance patient-provider relationship" by analyzing the interviewees' accounts regarding their perception of their relationship with the physicians (summative). After determining whether PatientHub had enhanced the patient-provider relationship, the coding author conducted a thematic analysis of the interviews [56]. This allowed us to identify the aspects that characterize high-quality patient-provider communication when using PatientHub (formative). The coding author created in vivo codes focusing on the effects on patient-provider communication attributed to patient-generated data in the consultation. The lead author conducted quality assurance by reviewing and revising the codes. In an iterative process, the lead and coding authors then grouped the codes into 4 characteristics of high-quality patient-provider communication: *personalization*, *actionability*, *trustworthiness*, and *equality*.

Second, we assessed the attainment of our subgoals "exploring experience," "understanding person," and "common ground" by applying a mixed deductive and inductive analysis [57,58]. We developed an initial coding scheme from related work on our kernel theory, patient-centered care [4,20]. We complemented the coding scheme with codes derived from the literature on patient-provider communication [22,23,38,39,59]. Again, the coding author created in vivo codes to capture emerging phenomena. She coded approximately 20% of the interviews before discussing the results with the lead author to refine the coding strategy. The coding author then finished coding all interviews. Finally, the lead author reviewed and refined the coding by discussing discrepancies with the coding author. In a workshop, the author group synthesized how PatientHub improved each dimension of patient-centered care.

Third, we deductively analyzed the video recordings regarding emerging behaviors [58]. The coding author analyzed the recordings based on the coding scheme applied and refined in the interview analysis. The coding scheme sensitized us to patient-centered behaviors enabled through PatientHub. During the coding of the videos, the coding scheme was expanded with in vivo codes to include emerging phenomena [58]. The video recordings were analyzed in 2 rounds on-site at one of the clinics (to ensure data privacy). We extracted emerging behaviors from the first coding round by drawing sequential processes. Initial drafts of these processes were discussed in a workshop including the author group and members from the project consortium. In the second coding round, we iterated on all video recordings based on the identified behaviors to refine our understanding and transcribed relevant sequences from the recordings. This step allowed us to formalize the 3 communication behaviors described in the *Results* section.

Ethical Considerations

While the ethics committee of the canton of Zurich confirmed that this study is not subject to the Swiss Human Research Act (BASEC-Nr. Req-2018-00847), we still decided to obtain written consent from all patients and physicians before the field study. The informed consent form educated participants about their rights and responsibilities, data use, and privacy measures following the World Medical Association Declaration of Helsinki [60]. The participants had the possibility to opt-out at any time during the study. The form also informed the participants' data will be gathered deidentified by assigning each participant with an identifier (ie, patient 1 to patient 28

and provider 1 to provider 6). Multimedia Appendix 5 provides the authors' positionality statement to give the readers a better understanding of the authors' backgrounds.

The participating clinics selected patients using a purposeful sampling approach [61]. The compensation for patients was CHF 50 (approximately US $60) and a raffle ticket for a restaurant voucher worth CHF 200 (US $230.15).

Results

Overview

In this section, we present our findings. First, we examine the attainment of the design goal (ie, enhance the patient-provider relationship) and discuss the 4 characteristics of high-quality patient-provider communication identified in the analysis. We then examine how patients and providers perceived patient-centered care in these consultations (ie, subgoals). Finally, we introduce 3 emerging communication behaviors when engaging with patient-generated data.

Enhancing the Patient-Provider Relationship With PatientHub

Overview

When asked about their relationship with the providers, the patients reported high satisfaction with their interaction compared than previous experiences. They often described an intimate connection that would feel significantly older than the 3 weeks of the valuation. In total, 4 central characteristics emerged from our data analysis that explain this enhanced patient-provider communication. Patients and providers commonly raised these characteristics when asked about their perception of the new approach supported by PatientHub compared with previous experiences. In the following sections, we explore each characteristic in more detail. Textbox 1 provides quotes from patients and providers for each characteristic.

Textbox 1. Characteristics of enhanced patient-provider communication with exemplary quotes.

Personalization of health care

- "She has tried to address the fact that I am not allowed to be overburdened with walking because the knees just do not work" (patient 07). Others highlighted the enriching discussions they had with the physicians "because I noticed that these are not standard answers.... He looks at you [the patient] as a human being, as an individual" (patient 25).

Actionability of interventions

- "It is simply a clean data basis. Now we're talking facts and not 'How did you perceive it?' or 'How was it for you?' but Bam! There! Suddenly, 'how many times did you go on the cross trainer?' 'how many times did you get the interval fasting done?'" (patient 04).
- "I was really happy to see the results. Because I remember the last time you said that you do so much, and you don't see any results, and now we have the result" (provider 01).

Trustworthiness of communication

- "She was prepared. So, she read my brainy entries [laughs].... So, she obviously prepared for me. She looked at the questionnaire that I filled out [during the preparation phase]. She wrote down questions about it. That really feels good" (patient 17).
- "They were unbelievably more trusting. They revealed so many, many things. So, I think it was a very different level of trust already compared to last time" (provider 05).

Equality of partners

- "I came here prepared and I already had ideas. If I had to choose favorites now [in the consultation], I would have come and I would have accepted [the physician's proposal]. Then you are externally steered" (patient 26).

Personalization of Health Care

Patient-generated data support patients and providers in personalizing health care as they facilitate in-depth discussions about the patients, their context, and their experience with their health. The data provide a solid foundation based on facts instead of gut feelings and memory. Therefore, the mutually agreed upon therapy plans consider the patient to be a person with their preferences, needs, and limitations. The physicians were understanding when proposing interventions and considered the patients' circumstances. All patients in the interviews highly appreciated this (Textbox 1).

Actionability of Interventions

Patients and providers appraised the concreteness of the discussions facilitated by the consultation tool with the intervention screens (Figure 4, right). Many patients reported previous frustrating experiences in which providers stayed abstract in their recommendations (eg, "eat less sweets"). Due to the more integrated and holistic understanding of the patients, providers and patients could concretely discuss problems and priorities, goals, and expectations regarding each other's roles. Many specifically highlighted how the patient-generated data allowed them to agree on actionable interventions. Therefore, patients perceived providers as more empathic and engaged in their health journey. For example, patient 26 "felt joy from the provider.... I think she was very motivating and also praised that I had done well." The providers proved the perception right as many were pleased with their patients' progress, such as provider 01 and patient 02 (Textbox 1).

Trustworthiness of Communication

Patients and providers believed that sharing patient-generated data requires trust in the first place and creates trustworthy communication. Patients perceived it as appreciation (Textbox 1, patient 17). The providers reciprocated this appreciation. When asked about the relationship between her and the patients after only 2 consultations, provider 05 highlighted how PatientHub created a trusting foundation that made the discussions in the consultations much more meaningful (Textbox 1, provider 05).

Approximately half (15/27, 56%) of the patients raised the topic of surveillance concerning sharing their data. However, most patients appreciated the subtle surveillance as it made the therapy plan more binding. Only 19% (5/27) of the patients felt uncomfortable sharing too much personal information. Accordingly, they only shared what they felt comfortable with in the journal.

Equality of Partners

Finally, PatientHub leads to a shift in the perceived roles of patients and providers. Many patients perceived control over the decisions made in the consultation. This perceived control leads to the feeling of cooperation between equal parties in the decision-making process. The patients felt strengthened in their position as they were the experts on their data (Textbox 1, patient 26).

The preparation allowed patient 26 to have an opinion instead of mindlessly accepting the provider's proposition. Furthermore, in the follow-up consultations, patients defended their standpoints and argued for changes to the therapy plan. For example, patient 18 demanded the reintroduction of carbohydrates into her diet due to her physically demanding job. As a result of this approach, patient 23 experienced the consultation as "an open conversation and not somehow top-down. On the same level and friendly." Many providers, too, remarked on the shift in power balance.

The Process of Patient-Centered Care

Our analysis elicited the 4 characteristics of high-quality patient-provider communication. In the following sections, we explore the process of enhancing patient-provider communication by discussing the 3 dimensions of patient-centered care (ie, subgoals)

Finding Common Ground

Overall, we observed that the consultations centered on the 3 aspects of finding common ground: problems and priorities, goals, and roles. PatientHub introduced patient-generated data into the natural consultation process through screens for goal setting, interventions, and planning. Patients and providers reported that the tool was a significant part of the consultation as it formed the starting point for exploring problems, priorities, and roles. The data also served as a reference to argue for or against a proposition, thereby shaping the individual roles. Therefore, patients and providers arrived at mutual decisions regarding all 3 aspects of finding common ground. When asked about the reasons for the positive impact of PatientHub, provider 01 answered the following:

> *You are pulling in the same direction. And are in the same reality. And that makes a much better team. And just have a more balanced, I do not want to say power balance, but a more balanced decision-making.*

However, patient-generated data did not solely support finding common ground directly. Each behavior explored the other 2 components to indirectly inform mutual decisions made in the consultations.

Exploring Health, Disease, and the Illness Experience

The general health questionnaire and journal entries allowed patients and providers to discuss the patients' unique perceptions of their health in both consultations (ie, the "exploring experience" subgoal). For example, provider 02 recognized in the journal overview (Figure 7, right) that patient 08 ate too few vegetables and drank too much alcohol, which the patient agreed with. Instead of staying abstract about the consumed amount of alcohol, they had a clear impression of the number of alcoholic beverages that the patient drank in a week. Several patients realized during the consultation that they were emotional eaters. The patient-generated data prompted the provider or patient to highlight such experiences. In addition to behaviors, they often discussed emotional aspects of the patient's experience. For example, 2 patients said that they did not like swimming because they did not want to show themselves in bathing suits. Patient 11 mentioned in the questionnaire that he feared the health problems associated with obesity. During the consultation, provider 03 could follow-up on this answer by asking why the patient was afraid. Patients and providers explored the target weight reported in the questionnaire during goal setting. Often, they discussed the origin of this specific target, such as a feeling of well-being or a historic weight they had during a significant part of their life (eg, before they became parents).

Patient-generated data had an even more profound impact during the follow-up consultation. Patients and providers could gather an integrated understanding of the patients' experience in the execution phase. The patients documented their emotions and thoughts using emojis, pictures, and text in journal entries (Figure 6). These data provided patients and providers with a rich foundation for discussions in the follow-up consultation. Instead of relying on the patients' memory and accuracy, providers had in-depth insights into the adherence and patient experience. Sometimes, patients highlighted a journal entry because they believed that it was significant for their (lack of) success in following the therapy plan. For example, patient 07 referred to the picture of an icy peer to explain her nonadherent behavior to the "walking" intervention. Most providers emphasized the benefit of recording a patient's emotional state. This way, they could inquire about negative feelings related to a specific intervention. For example, provider 01 could identify a potential correlation between patient 01's emotional state and his adherence to intermittent fasting. Together, they explored that patient 01 was under pressure at work during the execution phase. This led to him feeling tense and not sleeping well. Therefore, he was not motivated to adhere to intermittent fasting. However, they realized that the patient indeed felt better on

days when he could adhere to the intervention, as provider 01 recalled in the interview:

> *And then you could break that down nicely and say, hey, you did it. The mood was good. Look at the app. It was ALWAYS good for you to do [intermittent fasting]. And then it really came back from the patients like this: Yes, that's right.*

Understanding the Whole Person

Traditionally, consultation time is limited to a few minutes per patient. This limited time often does not allow providers to ask questions not directly associated with the presented problem. Consequently, understanding the whole person often falls victim to other, more pressing matters. However, the journal entries and the corresponding overview and filters allowed providers to understand the patient's daily life ("understanding person" subgoal). In addition, the journal entries served as the foundation to further gain a better understanding of the whole person during the consultation (Figure 7). For example, provider 03 and patient 12 discussed her consumption of vegetables, where patient 12 said the following:

> *Probably too little in proportion. Because I have to be honest, I don't like to cook..... And many times, it is so my partner works irregularly. And when we come home in the evening, something should just quickly be on the table. And I don't want to stand two hours in the kitchen when I have worked all day.*

This quote illustrates how patient-generated data prompted provider 03 to learn about patient 12's experience and her proximal context—her partner working shifts might interfere with regular habits. In general, the available information and the subsequent discussion yielded interesting insights that providers usually would not obtain, as they all said during the interviews. For example, it became evident that chocolate yogurt was a central piece of patient 18's diet. Patient 28 preferred to walk alone as he was an only child. Patients 01 and 25 had dogs, but another household member usually walked them. Patient 03 cooked for her husband and did not think he would want to eat less meat or try different grains. In addition, she drove him to therapy and, therefore, had less time for cooking.

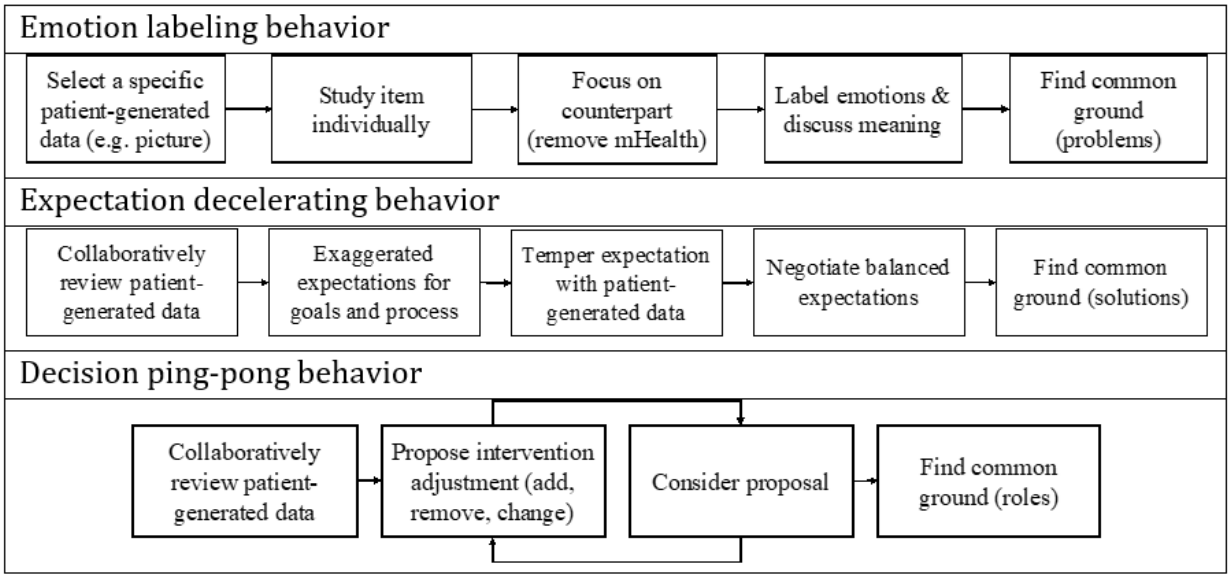
In general, patients believed that their data helped providers obtain a better understanding of them. For example, patient 16 liked that the provider had more background information before the initial consultation. Patient 10 shared this opinion as she believed that it would be impossible to obtain such a deep understanding in such a short time. The providers agreed with the patients and explicitly mentioned PatientHub's advantages to understand the whole person better. Providers 01 and 04 referred to "look beneath the patient's surface" as a significant advantage.

Emerging Patient-Provider Communication Behaviors

Overview

In the following sections, we outline 3 behaviors that we could repeatedly observe across consultations when using PatientHub. First, we describe each behavior. Then, we provide examples of the behaviors and highlight how patients and providers perceived them. Figure 10 provides visual representations of the sequential activities of the behaviors.

Figure 10. Visual representations of the 3 emerging patient-provider communication behaviors.



Emotion Labeling

This behavior is called emotion labeling as patients and providers discussed patients' experiences based on patient-generated data and attached a label to it (Figure 10). The behavior occurred at the beginning of each consultation (initial and follow-up). Patients and providers sat around the table's edge so that both had visual access to the tablet facing them. The tablet showed the journal overview with patient-generated data (Figure 7, right). The 2 variations of this behavior differed depending on the person who initiated the behavior. Providers commonly selected specific patient-generated data to start the

behavior (eg, a journal entry or a questionnaire response). Often, providers referred to pictures shared by the patients. They were picked due to the client's reported adherence or nonadherence, emotional state, or uploaded pictures. Less often, patients initiated the behavior by referring to a specific item that represented a problem or priority. The initiator pointed to the specific item and sometimes opened the journal entry. After both acknowledged the item, they moved the focus from the tablet to the other person, adjusting their seating position and turning their bodies and heads toward one another. The provider even moved the tablet out of the shared interaction space in some consultations. Then, patients and providers labeled the patients' emotions regarding the matter represented by the selected item. They discussed the item and its meaning for the patient's experience or context to arrive at a common understanding of the patient's problems and priorities.

One example of emotion labeling is patient 18, who craved chocolate yogurt. The photos and text in the journal allowed provider 04 to elicit the issue and sensibly raise the topic of the large amount of sugar in chocolate yogurt. During the discussion, they explored this craving to discover 2 reasons for it. First, chocolate yogurt was a fast and enjoyable meal after a long and stressful workday. Second, the yogurt satisfied an emotional need as it was something to look forward to. Therefore, provider 04 proposed reducing the quantity of chocolate yogurt and adding plain yogurt with fruits or jam as a low-sugar alternative. This resonated well with patient 18:

> *Ah, I'm allowed to eat chocolate yogurt (laughs). I noticed right away that she was not telling me: You're not allowed to do that anymore; from now on, there's only this and that. It's often restrictions that make things difficult. I have seen that I should lose weight over a long time, and then it doesn't have to be as radical. This way, we can mix it up.*

Patients also raised issues on their own. Patient 02 believed that one of her core problems was drinking too much alcohol, which she documented in the journal. During the interview, provider 01 explained that the journal helped her assess the quantity as unproblematic and that she would have reacted differently without the patient-generated data. They identified the guilt associated with drinking alcohol as a constant stressor.

Expectation Decelerating

Expectation decelerating (Figure 10) occurs when one party needs to reduce the momentum that the other is exhibiting. The behavior followed the emotion labeling behavior, where patients and providers mutually agreed on the patients' problems and priorities regarding weight management.

Expectation decelerating varied between the initial and follow-up consultations. In the initial consultation, the providers changed the topic to discuss target weight. They reviewed the patient's desired weight and therapy duration before the consultation. However, they first invited patients to talk about their target weight and the reason for this specific target. Upon understanding the patient's reasoning, they focused on the goal-setting screen (Figure 4, left), where the patient-generated data are prefilled (ie, target weight and therapy duration). Providers proposed an intermediate goal and explained their reasoning depending on the targeted weight loss and therapy duration. To illustrate the proposal, providers moved the weight and time slider to the proposed goal. A scale icon indicated the sustainability of the target weight and time frame on a color scale from green to red. A 500-gram weekly weight loss was considered sustainable [62]. If the chosen target was beyond this limit (ie, the scale icon was in the red area), the providers would highlight this and explore the relationship between weight and time with the patient. They simulated different scenarios to understand sustainable weight loss. In the end, they left the decision up to the patients. The exploration led to either a prolonged therapy duration or setting an intermediate goal.

In the follow-up consultation, expectation decelerating occurred after reviewing the patient's adherence using the journal entries. To conclude the emotion labeling behavior, patients or providers often proposed adding new interventions or increasing their frequency (eg, running more often). While substituting ineffective interventions usually generated mutual agreement, adding to the existing interventions was often more intensely debated. Frequently, the providers advised against a patient's request, referring to patient-generated data to highlight the continued effectiveness of the existing interventions. However, patients also had to decline similar propositions by providers. Similarly, they referred to their journal entries or insights based on these data to argue their position. The following excerpt shows a discussion regarding the therapy duration in the initial consultation between patient 06 and provider 02:

> *Provider 02: I have seen your target weight or desired weight would be 76 kg.*
>
> *Patient 06: I am also already satisfied with 77 to 78 kg.*
>
> *Provider 02: So, let's say we want to aim for 78 kg (sets target weight with slider). What is your wish when you want to reach it?*
>
> *Patient 06: Yes, that's what I am like. I want it as soon as possible. But then you gain it back again soon, right?*
>
> *Provider 02: Yes exactly. So, what does soon mean concretely? We can now play it through virtually what that means as a weekly goal. So that would be 4 kg if we say 78 kg. When do you want to have that?*
>
> *Patient 06: I don't know, I say in two months it should be possible.*
>
> *Provider 02: In two months, that would be eight weeks (Provider adjusts the slider on the tablet). That is not completely unrealistic for 4 kg, but it is relatively strict. So, if you actually don't want it to be an extra burden, an additional task, then it's too ambitious. Then I would rather put it at three months, right? (adjusts slider). Then you see, there is the scale that shows how much weight you would have to lose or how ambitious that it is. We are now well in the green zone. Concretely that would mean a little more than 300 grams decrease per week.*

Often, patients entered the consultation with overambitious goals, as they recalled themselves in the interviews. Furthermore, they were enthusiastic after the 2-week execution phase as they could follow most interventions and proposed including more interventions. Providers used patient-generated data to decelerate the patients' drive to maintain sustainable progress and prevent setbacks. While deceleration might be negatively connotated, most patients drew motivation from this behavior. For example, patient 02 mentioned that the 5 kg they decided on motivated her more than if she had tried to lose the initial 15 kg. Furthermore, patient 12 summarized the following:

> *Yes, and what really stuck with me was that he took a little bit of the pressure off. When you always have the feeling that you have to lose as much weight as possible as quickly as possible. That he then said: It's good and healthy to lose 500 grams a week. That has stayed with me very much, and has also motivated me, so that I did not think: Oh in 14 days I should be 10 kg lighter.*

Patient 16 agreed and mentioned the sliders, visualizing the relationship between weight loss and time, as essential for setting realistic goals. The providers agreed with this perception as the patient-generated data would provide a solid foundation to assess the therapy plan's effectiveness, which is usually missing in conventional consultations. Provider 06 highlighted the following:

> *Yes, I always have the feeling that I always have to make suggestions. And [with the patient-generated data] you could say, we've done a good job. We are good. We continue to do so. Done.*

Provider 02 agreed that patients usually quit because they want to achieve overly ambitious goals. With patient-generated data, they could credibly show that it is possible to lose weight with small and sustainable changes.

Decision Ping-Pong

Decision ping-pong refers to the back-and-forth process of mutual decision-making (Figure 10). It occurred when creating and adapting the therapy plan. The input for this behavior was patient-generated data created outside the consultations. During the preparation phase, patients selected favorites from a list of dietary and activity interventions and documented their daily lives in journal entries.

In the initial consultation, providers initiated the behavior by moving to the interventions screen to display the patient's favorites, marked with a "heart" icon (Figure 4, right). Providers asked the patients why they selected the specific interventions to start the decision ping-pong. Often, patients chose the interventions as they had previously engaged in an activity or started the intervention themselves only recently. Next, providers asked patients which favorites to add as an intervention. After the patients decided on an intervention, they discussed details such as frequency and duration. A back-and-forth followed to first agree on a frequency and duration (if applicable). Once agreed, patients and providers negotiated the timing of this intervention (ie, days and time of the day). While providers often proposed the frequency and duration, the patients usually initiated the discussion on timing according to their professional and private situations. This ping-pong was repeated until both were satisfied with the therapy plan. Interestingly, patients started proposing interventions themselves after some repetitions (eg, after adding 2 interventions).

In the follow-up consultation, the behavior occurred slightly differently. First, patients or providers initiated the behavior to discuss the necessity of exchanging ineffective interventions based on the journal entries crafted by the patients in the execution phase. The behavior ended if they mutually agreed not to adjust the therapy plan. If they decided to adjust, the behavior continued as in the initial consultation.

The following excerpt shows a discussion between patient 10 and provider 03 about an activity intervention:

> *Patient 10: Or aerobics would be something I would like to do.*
>
> *Provider 03: So once in the evening?*
>
> *Patient 10: Yes.*
>
> *Provider 03: Thursday is busy [with other interventions]*
>
> *Patient 10: Then, we will take Monday.*
>
> *Provider 03: Or after cleaning (both laugh).*
>
> *Patient 10: No, thank you.*
>
> *Provider 03: Monday?*
>
> *Patient 10: Yes.*
>
> *Provider 03: In the evening?*
>
> *Patient 10: Mhm.*
>
> *Provider 03: How long?*
>
> *Patient 10: Half an hour.*
>
> *Provider 03: Half an hour. I think so too. We will just put that in now.*

A total of 30% (8/27) of the patients explicitly mentioned their appreciation for the realistic goals as an outcome of this process. Patient 21 liked that the provider said that "You cannot just cancel everything overnight. Then you just stop again." After 2 weeks, patient 16 compared the results to other diet regimes he had followed in the past:

> *The whole thing is calmer and less stressful. It feels easier. It goes on for a longer time, but it is more pleasant to get through the day that way.*

The provided selection of tasks and the talk with the physician helped produce ideas that the patients usually would not have produced. Patient 04 said that he would not have chosen intermittent fasting but he did because a professional explained it. The decision ping-pong showed that it could be adapted to his situation. The process also inspired the providers as provider 05 said that he would not have all these ideas spontaneously.

Discussion

Principal Findings

Supporting patient-provider communication using patient-generated data is a growing topic of interest in health

informatics and related fields. However, the existing literature often overlooks the processes (ie, communication behaviors) when enhancing patient-provider communication using patient-generated data. More importantly, existing mHealth apps are seldom integrated into the provider's workflow and do not sufficiently engage and activate patients in patient-centered care [44-46]. Through a design science approach, we explored how to design an integrated digital health system and its impact on the communication behaviors of 27 patients and 6 providers.

Previous work has demonstrated the benefits of integrating patient-generated data into consultations [32,35,38,45,63,64]. We expand on this work by studying the effect of patient-generated data on patient-provider communication behaviors. As argued in the Introduction section, current research focuses on the input (patient-generated data) and output (enhanced patient-provider communication; Figure 2). Our analysis elicited 3 emerging behaviors that open the black box of patient-provider communication. On the basis of our theoretical foundation and empirical findings, we propose patient-centered communication in 3 design principles to operationalize the patient-centered care framework.

Facilitate Emotion Labeling to Explore Health, Disease, and the Illness Experience and Understand the Whole Person

The emotion labeling behavior highlights how the input is initially processed. Patient-generated data enable patients and providers to reflect on patients' emotions related to their health. For example, patients speak about their insecurities when engaging in physical activities in public due to being overweight. Prompted by photos, emojis, and text, patients and providers explore patients' unique experience with their health and disease ("exploring experience" subgoal). Furthermore, emotion labeling uncovers obstacles in the patients' journeys to better health. As described in the Results section, journal entries allowed patient 03 and provider 01 to uncover and discuss the reason for the patient's nonadherence. Patient-generated data allowed the patient and provider to understand the patient as a whole person and their proximal context ("understanding person" subgoal).

Existing research has shown how visualizations of health data increase health literacy in patients [38]. Our results show how patient-generated data are used to not only educate patients but also collaboratively generate new insights. We argue that patient-generated data enhance patient-provider communication by stimulating the exploration of a patient's experience and context. Providers can usually only scratch the surface of a patient's story in consultations. PatientHub introduces them to their patients' world, allowing them to dive below the surface and gain an integrated understanding. The patients themselves also profit from their data. Our results show that generating data often initiates self-reflection that fosters patients' understanding of themselves. Therefore, patient-provider communication becomes more patient centric. Patients and providers explore the patient's experience and context to find common ground regarding problems and goals.

Facilitate Expectation Decelerating to Find Common Ground on Problems and Solutions

The expectation decelerating behavior processes patient-generated data differently. Instead of beginning with a patient's experience or context, patients and providers used patient-generated data to assess problems or goals. Our results show 2 areas in which patient-generated data decelerated patients and providers: setting goals and developing therapy plans. Providers referred to these data in the former to decelerate the patients' ambitions. They suggested either more long-term planning or a more achievable intermediate goal. When developing therapy plans, patients and providers resisted the urge to add more or exchange interventions too quickly due to patient-generated data ("common ground" subgoal). As seen in the Results section, patient 01 could successfully argue his standpoint on keeping intermittent fasting in his therapy plan due to his data. The data allowed them to assess the therapy plan's effectiveness, resulting in the mutual decision not to adapt it if the effectiveness persisted. Existing research demonstrates the value of patient-generated data for problem-solving and decision-making [23,64]. While we agree with their findings, the saying "It's a marathon, not a sprint" highlights the importance of expectation decelerating. Our results show how patient-generated data are applied in consultations to resist the urge to hurry long-term behavior change. Accordingly, expectation decelerating might counteract the common pressure of a quantified self to always strive for more [65]. Quantified humans use data to continuously find inefficiencies and improve on those, often resulting in unhealthy pressure. Instead of expanding the therapy plan, patients and providers could assess the effectiveness of the current plan in the follow-up consultation. They had a reliable foundation for decision-making instead of hampering long-term success with impulsive actions. They defined achievable targets with a personalized therapy plan tailored to the patient. Therefore, patients might have more endurance in their marathon race to change their lifestyle.

Facilitate Decision Ping-Pong Between Patients and Providers to Find Common Ground on Their Roles

Decision ping-pong reflects on how patient-generated data affect the dynamics of patient-provider communication. Existing research in medicine and human-computer interaction postulates the importance of shared decision-making [4,17,21,23,53]. Our results expand on these insights by examining how patient-generated data include the patient in decision-making as an expert and how this changes the consultation dynamics ("common ground" subgoal). Many patients noticed how their position had changed in the initial consultation compared with their previous experience. We argue that introducing patient-generated data goes beyond facilitating informed decision-making. With these data, patients provide "proof" of their adherence, for example, through pictures. Moreover, they have a much deeper knowledge of PatientHub's content. It contains patient-generated data and evokes memories that might not be documented in the system. Equipped with this knowledge, patients become experts in their own domain: their health journey.

In shared decision-making, the implicit understanding is that providers assess the patient's state and offer options [17]. Existing research often sees patient-generated data as crucial information for providers to offer better health services [22,64]. Studies explore ways to design technology to make large amounts of patient-generated data consumable for providers [23]. Implicitly, this focus places the responsibility for data interpretation on the providers. However, we believe that patient-generated data could partially relieve providers from this burden. Patient-centered care emphasizes mutual decisions; therefore, patients should carry this burden with their providers.

Our results show that most patients willingly accepted such a role as they felt taken more seriously, could defend their position, and perceived having control over the decision-making process. Integrating patient-generated data into the consultation empowers patients to assume a more active role in their health care. We argue that providers must not be solely responsible as our results highlight how patients become experts in their health journey. Instead, the responsibility is shared between patients and providers according to their expertise. While providers assess the data against their medical knowledge, patients interpret the data in the context of their lives and experiences. Therefore, their communication becomes truly "respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions" [16].

Conclusions

This study evaluated communication behaviors that emerge when introducing patient-generated data into patient-provider communication. We studied how these behaviors in patient-provider communication actualize the potential of patient-generated data to increase patient-centeredness. Our analysis uncovered 3 communication behaviors in medical consultations when using PatientHub. Furthermore, we demonstrated how enhanced patient-provider communication is necessary for patient-centered care. On the basis of our findings, we believe that this study contributes to research in 2 ways. First, we emphasize the value of patient-provider communication. The identified behaviors demonstrate how data-supported patient-provider communication creates value, not the technology and data themselves. Second, the behaviors offer actionable insights into implementing patient-centered care.

However, this study does not come without limitations. First, while we evaluated PatientHub in the most realistic setting, its applicability in the real world depends on regulatory and security frameworks. Second, the generalizability could be further increased with a larger sample size and a randomized controlled trial. Future research could also investigate how patient-generated data empower patients individually to study the changing role dynamics in medical consultations.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide: physician initial consultation (English translation from German).
[DOCX File , 32 KB - medinform_v12i1e57406_app1.docx]

Multimedia Appendix 2
Interview guide: patient initial consultation (English translation from German).
[DOCX File , 33 KB - medinform_v12i1e57406_app2.docx]

Multimedia Appendix 3
Interview guide: physician follow-up consultation (English translation from German).
[DOCX File , 35 KB - medinform_v12i1e57406_app3.docx]

Multimedia Appendix 4
Interview guide: patient follow-up consultation (English translation from German).
[DOCX File , 31 KB - medinform_v12i1e57406_app4.docx]

Multimedia Appendix 5
Authors' positionality statement.
[DOCX File , 17 KB - medinform_v12i1e57406_app5.docx]

References

1. Kelley JM, Kraft-Todd G, Schapira L, Kossowsky J, Riess H. The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. PLoS One 2014 Apr 09;9(4):e94207 [FREE Full text] [doi: 10.1371/journal.pone.0094207] [Medline: 24718585]

2. Step MM, Rose JH, Albert JM, Cheruvu VK, Siminoff LA. Modeling patient-centered communication: oncologist relational communication and patient communication involvement in breast cancer adjuvant therapy decision-making. Patient Educ Couns 2009 Dec;77(3):369-378 [FREE Full text] [doi: 10.1016/j.pec.2009.09.010] [Medline: 19811883]
3. Zolnierek KB, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care 2009 Aug;47(8):826-834 [FREE Full text] [doi: 10.1097/MLR.0b013e31819a5acc] [Medline: 19584762]
4. Stewart M, Brown JB, Weston W, McWhinney IR, McWilliam CL, Freeman T. Patient-Centered Medicine: Transforming the Clinical Method. Boca Raton, FL: CRC Press; 2013.
5. Epstein RM, Street RLJ. The values and value of patient-centered care. Ann Fam Med 2011;9(2):100-103 [FREE Full text] [doi: 10.1370/afm.1239] [Medline: 21403134]
6. Vermeire E, Hearnshaw H, Van Royen P, Denekens J. Patient adherence to treatment: three decades of research. A comprehensive review. J Clin Pharm Ther 2001 Oct;26(5):331-342. [doi: 10.1046/j.1365-2710.2001.00363.x] [Medline: 11679023]
7. Costello KL. Impact of patient-provider communication on online health information behaviors in chronic illness. Proc Assoc Info Sci Tech 2016 Dec 27;53(1):1-10. [doi: 10.1002/pra2.2016.14505301060]
8. Daniels T, Goodacre L, Sutton C, Pollard K, Conway S, Peckham D. Accurate assessment of adherence: self-report and clinician report vs electronic monitoring of nebulizers. Chest 2011 Aug;140(2):425-432. [doi: 10.1378/chest.09-3074] [Medline: 21330381]
9. Stirratt MJ, Dunbar-Jacob J, Crane HM, Simoni JM, Czajkowski S, Hilliard ME, et al. Self-report measures of medication adherence behavior: recommendations on optimal use. Transl Behav Med 2015 Dec;5(4):470-482 [FREE Full text] [doi: 10.1007/s13142-015-0315-2] [Medline: 26622919]
10. Iuga AO, McGuire MJ. Adherence and health care costs. Risk Manag Healthc Policy 2014 Feb 20;7:35-44 [FREE Full text] [doi: 10.2147/RMHP.S19801] [Medline: 24591853]
11. Nieuwlaat R, Wilczynski N, Navarro T, Hobson N, Jeffery R, Keepanasseril A, et al. Interventions for enhancing medication adherence. Cochrane Database Syst Rev 2014 Nov 20;2014(11):CD000011 [FREE Full text] [doi: 10.1002/14651858.CD000011.pub4] [Medline: 25412402]
12. Pearce C, Dwan K, Arnold M, Phillips C, Trumble S. Doctor, patient and computer--a framework for the new consultation. Int J Med Inform 2009 Jan;78(1):32-38. [doi: 10.1016/j.ijmedinf.2008.07.002] [Medline: 18752989]
13. Chen Y, Cheng K, Tang C, Siek KA, Bardram JE. Is my doctor listening to me?: impact of health it systems on patient-provider interaction. In: Proceedings of the CHI '13 Extended Abstracts on Human Factors in Computing Systems. 2013 Presented at: CHI EA '13; April 27-May 2, 2013; Paris, France. [doi: 10.1145/2468356.2468791]
14. Beach MC, Duggan PS, Moore RD. Is patients' preferred involvement in health decisions related to outcomes for patients with HIV? J Gen Intern Med 2007 Aug;22(8):1119-1124 [FREE Full text] [doi: 10.1007/s11606-007-0241-1] [Medline: 17514382]
15. Sepucha K, Mulley AGJ. A perspective on the patient's role in treatment decisions. Med Care Res Rev 2009 Feb;66(1 Suppl):53S-74S. [doi: 10.1177/1077558708325511] [Medline: 19001081]
16. Institute of Medicine, Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, DC: National Academies Press; 2001.
17. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012 Oct;27(10):1361-1367 [FREE Full text] [doi: 10.1007/s11606-012-2077-6] [Medline: 22618581]
18. Bomhof-Roordink H, Gärtner FR, Stiggelbout AM, Pieterse AH. Key components of shared decision making models: a systematic review. BMJ Open 2019 Dec 17;9(12):e031763 [FREE Full text] [doi: 10.1136/bmjopen-2019-031763] [Medline: 31852700]
19. Wilcox L, Patel R, Chen Y, Shachak A. Human factors in computing systems: focus on patient-centered health communication at the ACM SIGCHI conference. Patient Educ Couns 2013 Dec;93(3):532-534. [doi: 10.1016/j.pec.2013.09.017] [Medline: 24184039]
20. Epstein RM, Fiscella K, Lesser CS, Stange KC. Why the nation needs a policy push on patient-centered health care. Health Aff (Millwood) 2010 Aug;29(8):1489-1495. [doi: 10.1377/hlthaff.2009.0888] [Medline: 20679652]
21. Berry AB, Lim CY, Hirsch T, Hartzler AL, Kiel LM, Bermet ZA, et al. Supporting communication about values between people with multiple chronic conditions and their providers. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK. [doi: 10.1145/3290605.3300700]
22. Chung CF, Dew K, Cole A, Zia J, Fogarty J, Kientz JA, et al. Boundary negotiating artifacts in personal informatics: patient-provider collaboration with patient-generated data. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. 2016 Presented at: CSCW '16; February 27-March 2, 2016; San Francisco, CA. [doi: 10.1145/2818048.2819926]
23. Raj S, Newman MW, Lee JM, Ackerman MS. Understanding individual and collaborative problem-solving with patient-generated data: challenges and opportunities. Proc ACM Hum Comput Interact 2017 Dec 06;1(CSCW):1-18. [doi: 10.1145/3134723]

24. Wilcox L, Patel R, Back A, Czerwinski M, Gorman P, Horvitz E, et al. Patient-clinician communication: the roadmap for HCI. Ext Abstr Hum Factors Computing Syst 2013 Apr 27;2013:3291-3294 [FREE Full text] [doi: 10.1145/2468356.2479669] [Medline: 28018991]
25. Zhu H, Moffa ZJ, Carroll JM. Relational aspects in patient-provider interactions: a facial paralysis case study. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376867]
26. Asan O, Choi E, Wang X. Artificial intelligence-based consumer health informatics application: scoping review. J Med Internet Res 2023 Aug 30;25:e47260 [FREE Full text] [doi: 10.2196/47260] [Medline: 37647122]
27. Dang TH, Nguyen TA, Hoang Van M, Santin O, Tran OM, Schofield P. Patient-centered care: transforming the health care system in Vietnam with support of digital health technology. J Med Internet Res 2021 Jun 04;23(6):e24601 [FREE Full text] [doi: 10.2196/24601] [Medline: 34085939]
28. Cohen DJ, Keller SR, Hayes GR, Dorr DA, Ash JS, Sittig DF. Integrating patient-generated health data into clinical care settings or clinical decision-making: lessons learned from project HealthDesign. JMIR Hum Factors 2016 Oct 19;3(2):e26 [FREE Full text] [doi: 10.2196/humanfactors.5919] [Medline: 27760726]
29. Schroeder J, Hoffswell J, Chung CF, Fogarty J, Munson S, Zia J. Supporting patient-provider collaboration to identify individual triggers using food and symptom journals. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017 Presented at: CSCW '17; February 25-March 1, 2017; Portland, OR. [doi: 10.1145/2998181.2998276]
30. Webb MJ, Wadley G, Sanci LA. Improving patient-centered care for young people in general practice with a codesigned screening app: mixed methods study. JMIR Mhealth Uhealth 2017 Aug 11;5(8):e118 [FREE Full text] [doi: 10.2196/mhealth.7816] [Medline: 28801302]
31. Wickramasinghe N, John B, George J, Vogel D. Achieving value-based care in chronic disease management: intervention study. JMIR Diabetes 2019 May 03;4(2):e10368 [FREE Full text] [doi: 10.2196/10368] [Medline: 31066699]
32. Hong MK, Lakshmi U, Olson TA, Wilcox L. Visual ODLs: co-designing patient-generated observations of daily living to support data-driven conversations in pediatric care. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC. [doi: 10.1145/3173574.3174050]
33. Tadas S, Dickson J, Coyle D. Using patient-generated data to support cardiac rehabilitation and the transition to self-care. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI '23; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3580822]
34. Stawarz K, Preist C, Tallon D, Thomas L, Turner K, Wiles N, et al. Integrating the digital and the traditional to deliver therapy for depression: lessons from a pragmatic study. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376510]
35. Vitger T, Hjorthøj C, Austin SF, Petersen L, Tønder ES, Nordentoft M, et al. A smartphone app to promote patient activation and support shared decision-making in people with a diagnosis of schizophrenia in outpatient treatment settings (momentum trial): randomized controlled assessor-blinded trial. J Med Internet Res 2022 Oct 26;24(10):e40292 [FREE Full text] [doi: 10.2196/40292] [Medline: 36287604]
36. Denneson LM, Cromer R, Williams HB, Pisciotta M, Dobscha SK. A qualitative analysis of how online access to mental health notes is changing clinician perceptions of power and the therapeutic relationship. J Med Internet Res 2017 Jun 14;19(6):e208 [FREE Full text] [doi: 10.2196/jmir.6915] [Medline: 28615152]
37. Chen Y. Health information use in chronic care cycles. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work. 2011 Presented at: CSCW '11; March 19-23, 2011; Hangzhou, China. [doi: 10.1145/1958824.1958898]
38. Schooley B, San Nicolas-Rocca T, Burkhard R. Patient-provider communications in outpatient clinic settings: a clinic-based evaluation of mobile device and multimedia mediated communications for patient education. JMIR Mhealth Uhealth 2015 Jan 12;3(1):e2 [FREE Full text] [doi: 10.2196/mhealth.3732] [Medline: 25583145]
39. Seo W, Buyuktur AG, Verma S, Kim H, Choi SW, Sedig L, et al. Learning from healthcare providers' strategies: designing technology to support effective child patient-provider communication. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445120]
40. Van den Bulck SA, Hermens R, Slegers K, Vandenberghe B, Goderis G, Vankrunkelsven P. Designing a patient portal for patient-centered care: cross-sectional survey. J Med Internet Res 2018 Oct 01;20(10):e269 [FREE Full text] [doi: 10.2196/jmir.9497] [Medline: 30287416]
41. Lee YL, Cui YY, Tu MH, Chen YC, Chang P. Mobile health to maintain continuity of patient-centered care for chronic kidney disease: content analysis of apps. JMIR Mhealth Uhealth 2018 Apr 20;6(4):e10173 [FREE Full text] [doi: 10.2196/10173] [Medline: 29678805]
42. Kim Y, Heo E, Lee H, Ji S, Choi J, Kim JW, et al. Prescribing 10,000 steps like aspirin: designing a novel interface for data-driven medical consultations. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO. [doi: 10.1145/3025453.3025570]

43. West P, Van Kleek M, Giordano R, Weal MJ, Shadbolt N. Common barriers to the use of patient-generated data across clinical settings. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC. [doi: 10.1145/3173574.3174058]
44. Staehelin D, Franke K, Huber L, Schwabe G. From persuasive applications to persuasive systems in non-communicable disease care - a systematic literature analysis. In: Proceedings of the 18th International Conference on Persuasive Technology. 2023 Presented at: PERSUASIVE 2023; April 19-21, 2023; Eindhoven, The Netherlands. [doi: 10.1007/978-3-031-30933-5_11]
45. Steele Gray C, Gill A, Khan AI, Hans PK, Kuluski K, Cott C. The electronic patient reported outcome tool: testing usability and feasibility of a mobile app and portal to support care for patients with complex chronic disease and disability in primary care settings. JMIR Mhealth Uhealth 2016 Jun 02;4(2):e58 [FREE Full text] [doi: 10.2196/mhealth.5331] [Medline: 27256035]
46. Cozad MJ, Crum M, Tyson H, Fleming PR, Stratton J, Kennedy AB, et al. Mobile health apps for patient-centered care: review of United States rheumatoid arthritis apps for engagement and activation. JMIR Mhealth Uhealth 2022 Dec 05;10(12):e39881 [FREE Full text] [doi: 10.2196/39881] [Medline: 36469397]
47. Hevner AR. A three cycle view of design science research. Scand J Inf Syst 2007 Jan;19(2):87-92.
48. Hevner AR, Wickramasinghe N. Design science research opportunities in health care. In: Wickramasinghe N, Schaffer J, editors. Theories to Inform Superior Health Informatics Research and Practice. Cham, Switzerland: Springer; Apr 21, 2018.
49. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manag Inf Syst 2014 Dec 08;24(3):45-77. [doi: 10.2753/MIS0742-1222240302]
50. Jones D, Gregor S. The anatomy of a design theory. J Assoc Inf Syst 2007;8(5):312-335. [doi: 10.17705/1jais.00129]
51. Lapão LV, Peyroteo M, Maia M, Seixas J, Gregório J, Mira da Silva M, et al. Implementation of digital monitoring services during the COVID-19 pandemic for patients with chronic diseases: design science approach. J Med Internet Res 2021 Aug 26;23(8):e24181 [FREE Full text] [doi: 10.2196/24181] [Medline: 34313591]
52. Subramanian H, Subramanian S. Improving diagnosis through digital pathology: proof-of-concept implementation using smart contracts and decentralized file storage. J Med Internet Res 2022 Mar 28;24(3):e34207 [FREE Full text] [doi: 10.2196/34207] [Medline: 35343905]
53. Chung CF, Wang Q, Schroeder J, Cole A, Zia J, Fogarty J, et al. Identifying and planning for individualized change: patient-provider collaboration using lightweight food diaries in healthy eating and irritable bowel syndrome. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Mar;3(1):7 [FREE Full text] [doi: 10.1145/3314394] [Medline: 31080941]
54. Jaspers MW, Steen T, van den Bos C, Geenen M. The think aloud method: a guide to user interface design. Int J Med Inform 2004 Nov;73(11-12):781-795. [doi: 10.1016/j.ijmedinf.2004.08.003] [Medline: 15491929]
55. Venable J, Pries-Heje J, Baskerville R. FEDS: a framework for evaluation in design science research. Eur J Inf Syst 2017 Dec 19;25(1):77-89. [doi: 10.1057/ejis.2014.36]
56. Clarke V, Braun V, Hayfield N. Thematic analysis. In: Smith JA, editor. Qualitative Psychology: A Practical Guide to Research Methods. London, UK: SAGE Publications; 2015.
57. McDonald N, Schoenebeck S, Forte A. Reliability and inter-rater reliability in qualitative research: norms and guidelines for CSCW and HCI practice. Proc ACM Hum Comput Interact 2019 Nov 07;3(CSCW):1-23. [doi: 10.1145/3359174]
58. Saldana J. The Coding Manual for Qualitative Researchers. Thousand Oaks, CA: SAGE Publications; 2009.
59. Bruce C, Harrison P, Giammattei C, Desai SN, Sol JR, Jones S, et al. Evaluating patient-centered mobile health technologies: definitions, methodologies, and outcomes. JMIR Mhealth Uhealth 2020 Nov 11;8(11):e17577 [FREE Full text] [doi: 10.2196/17577] [Medline: 33174846]
60. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
61. Palinkas LA, Horwitz SM, Green CA, Wisdom JP, Duan N, Hoagwood K. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Adm Policy Ment Health 2015 Sep;42(5):533-544 [FREE Full text] [doi: 10.1007/s10488-013-0528-y] [Medline: 24193818]
62. Purcell K, Sumithran P, Prendergast LA, Bouniu CJ, Delbridge E, Proietto J. The effect of rate of weight loss on long-term weight management: a randomised controlled trial. Lancet Diabetes Endocrinol 2014 Dec;2(12):954-962. [doi: 10.1016/S2213-8587(14)70200-1] [Medline: 25459211]
63. Luo Y, Liu P, Choe EK. Co-designing food trackers with dietitians: identifying design opportunities for food tracker customization. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK. [doi: 10.1145/3290605.3300822]
64. Mentis HM, Komlodi A, Schrader K, Phipps M, Gruber-Baldini A, Yarbrough K, et al. Crafting a view of self-tracking data in the clinical visit. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 2017 Presented at: CHI '17; May 6-11, 2017; Denver, CO. [doi: 10.1145/3025453.3025589]
65. Sharon T. Self-tracking for health and the quantified self: re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. Philos Technol 2016 Apr 18;30:93-121. [doi: 10.1007/s13347-016-0215-5]

Abbreviations

DSR: design science research
mHealth: mobile health

Edited by C Lovis; submitted 16.02.24; peer-reviewed by M Cozad, HM Lim; comments to author 02.06.24; revised version received 25.06.24; accepted 21.07.24; published 10.09.24.

Please cite as:
Staehelin D, Dolata M, Stöckli L, Schwabe G
How Patient-Generated Data Enhance Patient-Provider Communication in Chronic Care: Field Study in Design Science Research
JMIR Med Inform 2024;12:e57406
URL: https://medinform.jmir.org/2024/1/e57406
doi:10.2196/57406
PMID:

Application of Information Link Control in Surgical Specimen Near-Miss Events in a South China Hospital: Nonrandomized Controlled Study

Tingting Chen, BSc; Xiaofen Tang, BSc; Min Xu, BSc; Yue Jiang, BSc; Fengyan Zheng, BSc

Operating Room, The First Affiliated Hospital of Wenzhou Medical University, Nanbaixiang Street, Ouhai District, Wenzhou, Zhejiang Province, China

Corresponding Author:
Fengyan Zheng, BSc

Abstract

Background: Information control is a promising approach for managing surgical specimens. However, there is limited research evidence on surgical near misses. This is particularly true in the closed loop of information control for each link.

Objective: A new model of surgical specimen process management is further constructed, and a safe operating room nursing practice environment is created by intercepting specimen near-miss events through information safety barriers.

Methods: In a large hospital in China, 84,289 surgical specimens collected in the conventional information specimen management mode from January to December 2021 were selected as the control group, and 99,998 surgical specimens collected in the information safety barrier control surgical specimen management mode from January to December 2022 were selected as the improvement group. The incidence of near misses, the qualified rate of pathological specimen fixation, and the average time required for specimen fixation were compared under the 2 management modes. The causes of 2 groups of near misses were analyzed and the near misses of information safety barrier control surgical specimens were studied.

Results: Under the information-based safety barrier control surgical specimen management model, the incidence of adverse events in surgical specimens was reduced, the reporting of near-miss events in surgical specimens was improved by 100%, the quality control quality management of surgical specimens was effectively improved, the pass rate of surgical pathology specimen fixation was improved, and the meantime for surgical specimen fixation was shortened, with differences considered statistically significant at $P<.05$.

Conclusions: Our research has developed a new mode of managing the surgical specimen process. This mode can prevent errors in approaching specimens by implementing information security barriers, thereby enhancing the quality of specimen management, ensuring the safety of medical procedures, and improving the quality of hospital services.

(JMIR Med Inform 2024;12:e52722) doi:10.2196/52722

KEYWORDS

near misses; technical barriers; process barriers; surgical specimens; information

Introduction

Overview

Pathological specimens are an important basis for judging the outcome of a patient's disease [1]. Surgical specimen management involves many aspects, and errors in any part of the specimen management process may lead to serious consequences. The management of surgical pathological specimens directly affects the quality and safety of nursing management in the operating room. Therefore, how to reduce the occurrence of abnormal events has become the focus of nursing in the operating room.

A near miss is a commonly used term in clinical nursing, referring to a kind of nursing abnormal event that may damage the life, health, and safety of patients but has not yet developed to the end point [1,2]. It has similar causes and development paths with adverse events (the event caused inconvenience or harm to the patient), but the number is far greater than adverse events. At the same time, it has little harm to patients and hospitals and can give early warning before harm to patients, help the medical management system to carry out forward-looking and proactive risk assessment and prevention, which is regarded as a better learning resource and a risk management method advocated by the hospital's fine management. Studies have shown that specimen approach failure is related to a variety of factors, such as human negligence, unreasonable workflow, and communication problems [3-7]. In recent years, information management of surgical specimens has gradually become the core issue that operating room managers continue to explore.

Surgical specimen informatization refers to the digital management process of specimen information obtained during surgery. It enhances data traceability, improves management efficiency, and ensures the accuracy and security of medical information. Specimen informatization involves digitally performing specimen collection, identification, transport, and reception [8]. Surgical specimen control refers to the effective management and control of specimen collection, processing, recording, and storage during surgery. It aims to ensure the quality of surgical specimens and data integrity. Process control should be implemented at critical nodes, with risk control points established throughout the entire process. Furthermore, problems occurring at all stages should be closely monitored and supervised [9].

Study Purpose

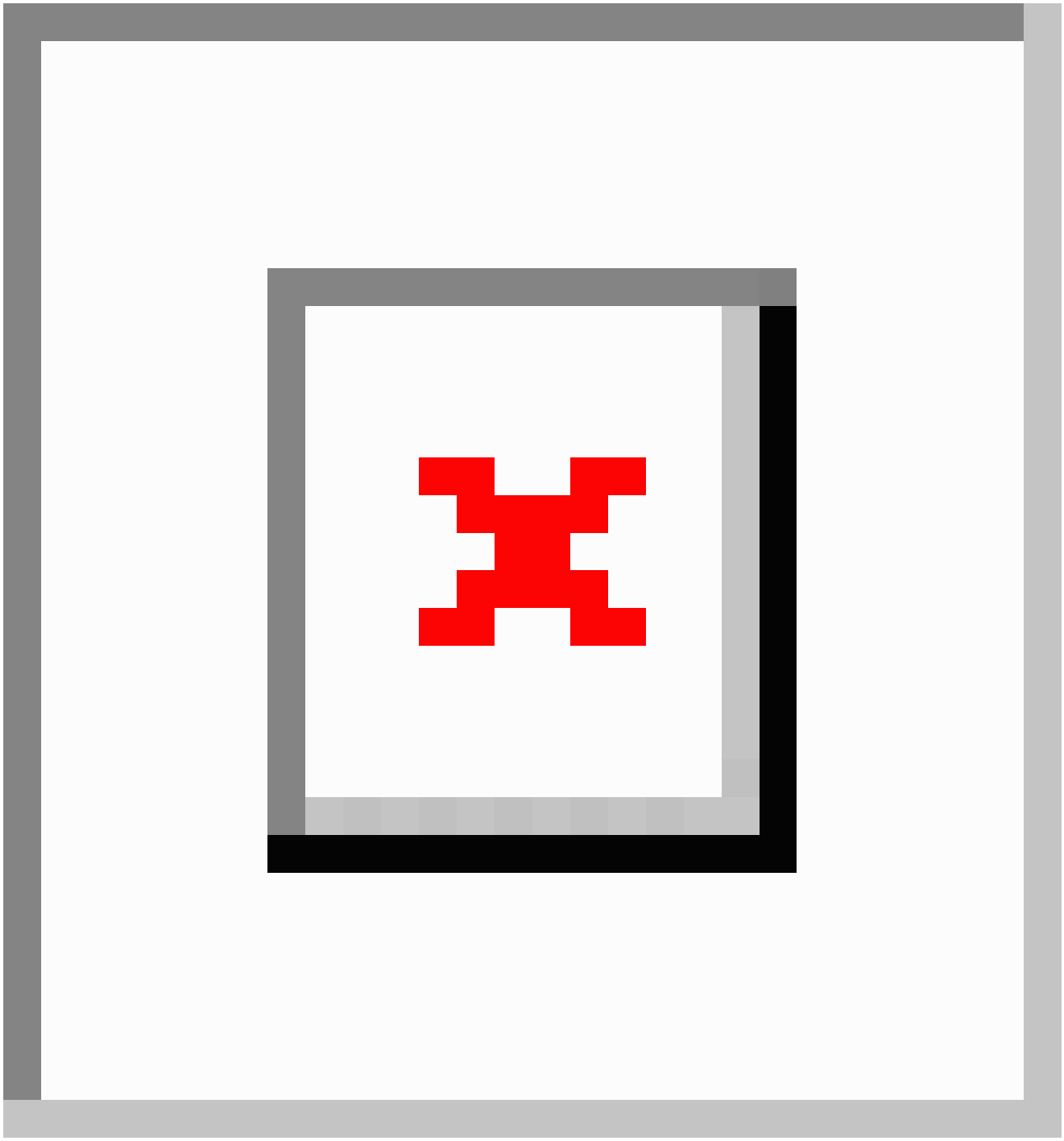
In the quality and safety of nursing care in the operating room, it has always been a hot issue to explore the application of intelligent management to the surgical specimen inspection process and safety management [10]. The purpose of this study is to further explore the effectiveness evaluation and improvement methods of information safety barriers, and fill the gap in the research field of the quality and safety impact of surgical specimens. At the same time, a new model of surgical specimen process management is further constructed and a safe operating room nursing practice environment is created by intercepting specimen approach error events through information safety barriers.

Methods

Study Design and Participants

The objects of this study were different transportation modes of surgical specimens from 2021 to 2022. The acquisition and preliminary processing of surgical specimens were carried out by standardized trained surgeons and surgical nurses. The exclusion criteria for participants in this study were (1) there was no fixed time information for labeled specimens and (2) no specimen transport time was indicated. In this research, the information transportation process management model of surgical specimens in 2021 was classified as the control group (group A) and the improved information transportation process management model of surgical specimens in 2022 was classified as the experimental group (group B). The data related to these specimens were retrospectively analyzed to explore the advantages of the new mode over the traditional mode of information transportation process management of surgical specimens (Figure 1).

Figure 1. Flowchart of pathological sample test in operating room.

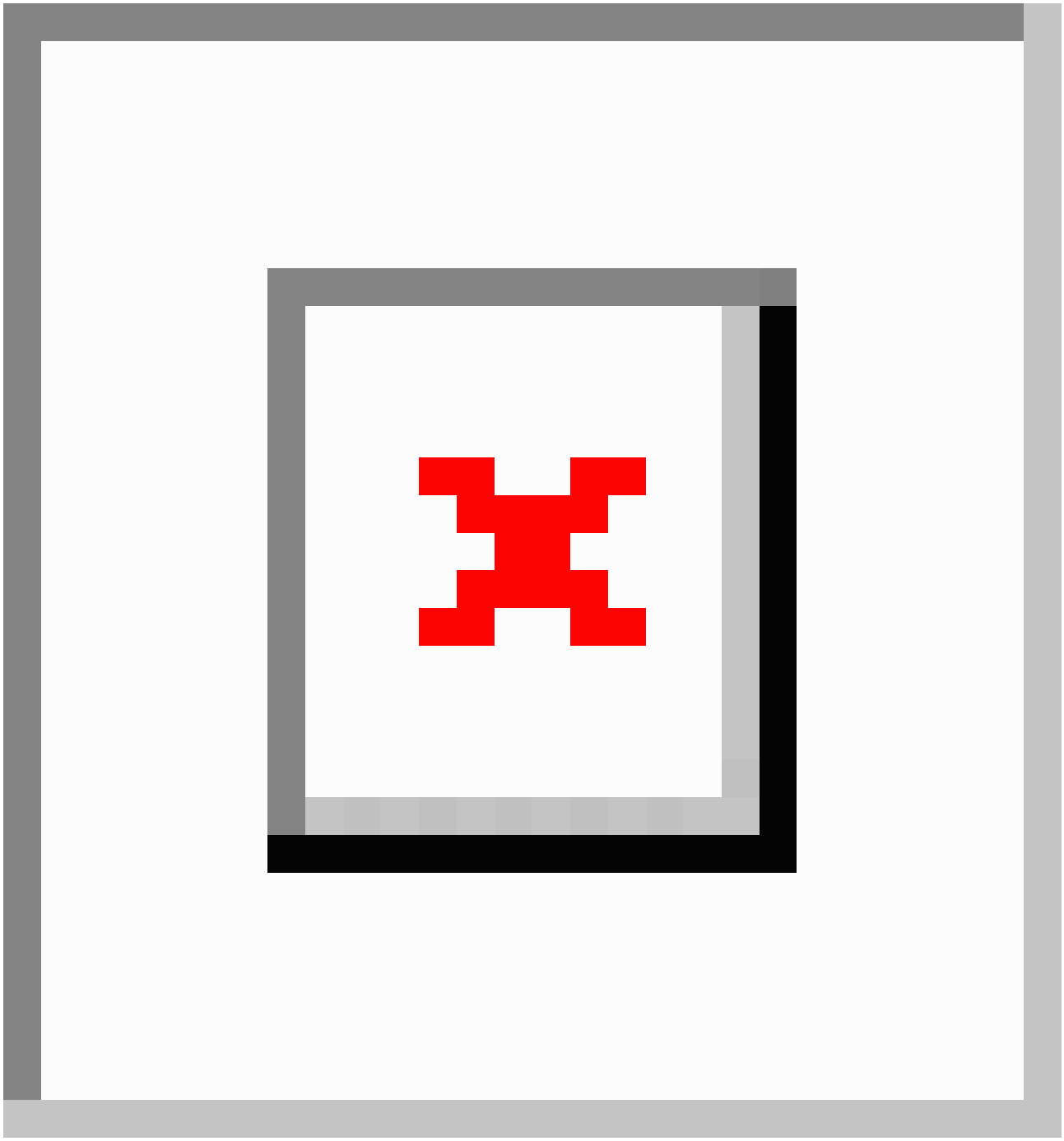


Platform Run Time

The system uses a C/S (Client/Server) structure, using PowerBuilder (version 11.5; SAP) development tools and Oracle database (Oracle Corporation) development. It follows a 3-layer technical architecture mode, with each application module functioning independently to support multiuser concurrent operations. Users access the system through the computer client. The platform consists of 4 ports, each with specific components and functions illustrated in Table S1 in Multimedia Appendix 1.

Surgical specimen process management modes are first, label—handwashing nurse, itinerate nurse, and the attending surgeon tripartite the type and name of the surgical specimen sent for examination and print the primary and secondary barcodes of specimen information. Second, transport and fixation—dedicated surgical specimen delivery personnel. Finally, reception—the pathology department staff scanned the 2D barcode of the surgical specimen, received the surgical specimen, and scanned the code for traceability in all links. All links are scanned for traceability (Figure 2).

Figure 2. Flowchart of specimen inspection in the operating room.



Procedure for Each Port Specimen

Overview

The computer operation of the nurse in the operating room is as follows. First, the itinerant nurse selects the type of surgical specimen in the operation interface of the operating room management system nurse (Figure S1A in Multimedia Appendix 2). Then the itinerant nurse prints the first-level barcode (Figure S1B in Multimedia Appendix 2) and the second-level barcode (Figure S1C in Multimedia Appendix 2) of the specimen information and pastes the second-level barcode on the specimen while checking the type and name of the specimen confirmed by the handwashing nurse, the itinerant nurse, and the surgeon. Finally, the system automatically pops up a box to select the person who submitted for testing and pushes the corresponding specimen submission method and information into the operating room (Figure S1D in Multimedia Appendix 2).

The operation of sending and storing specimens on the hospital's mobile app is as follows. First, the transport personnel used formalin for specimen fixation. Then the transport personnel scans the 2D barcode code on the surgical specimen on the hospital's mobile app (Figure S1E in Multimedia Appendix 2) and confirms it (Figure S1F in Multimedia Appendix 2). Finally, the transport personnel sent the surgical specimen to the pathology department for confirmation.

Confirmation of Receiving Surgical Specimens in the Pathology Department

The personnel of the pathology department scan the 2D code of surgical specimens at the receiving window to confirm receipt and import into the pathology reporting system (Figure S1G in Multimedia Appendix 2).

Maintenance of Basic Items of the System and Approval of Pathological Application Form Modification

The operating room management system only allows the head nurse to approve the application form modification of incorrect surgical specimens. The manager of the pathology department can modify the application form approved by the head nurse and query the history record (Figure S1H in Multimedia Appendix 2).

Testing of Platform Functions

Before launching the platform, the research and development team repeatedly simulated tests to improve platform functions, simplify operation steps, beautify the appearance of the interface, increase the output traffic of the base station, and solve problems, such as slow network speed and running lag.

Presurgical Specimen Analysis Stage

Link Control Step Diagram

The pictures below describe the control of each step in detail (Figure S2A in Multimedia Appendix 2).

Step 1

Control the error of filling in surgical specimen information. (1) Enter the operation specimen selection page and (2) select the specimen type (Figure S2B in Multimedia Appendix 2).

Step 2

Control patient identity information errors. (1) Label—surgical specimens, (2) the itinerant nurse enters the specimen's name, (3) the system bounce window 1 prompts "Please scan the wristband 2D barcode code to confirm," (4) the information is correct, and (5) the scan code is approved (Figure S2C in Multimedia Appendix 2).

Step 3

Control the error of surgical specimen information. (1) The system bounce window 2 prompts "The key information of the newly added surgical specimen site is inconsistent with the diagnostic site," (2) the system automatic bounce box to please check with the surgical doctor again "Please enter the checker's account password," (3) double-check the information, (4) itinerant nurse input the surgeon's account with the doctor's consent, and (5) the surgeon can print the pathology bar code after confirming the information is corrects (Figure S2D in Multimedia Appendix 2).

Step 4

Control the qualified rate of specimen fixation. In the fixation process of ordinary pathological specimens, the system will automatically remind the unfixed surgical specimens after 20 minutes (Figure S2E in Multimedia Appendix 2).

Step 5

Control the error of the diagnosis site of the pathology report. Selecting the automatic import and retention of surgical specimens (Figure S2F in Multimedia Appendix 2).

Step 6

Control abnormal error tracking management of surgical specimens. The status of surgical specimens in each step is distinguished by different color blocks. The full-time specimen staff and the full-time receiving staff of the pathology department check whether each surgical specimen has reached the closed-loop state. If the surgical specimen is not received more than half an hour after isolation, the system will automatically warn that the specimen has not received the prompt. In case of surgical specimen execution error, the itinerant nurse should report the cause of the failure in applying for close proximity on the pathology application sheet interface, and the specimen should be correctly executed after the examination and verification by the leader nurse. Analyze and rectify each data fetching exception problem (Figure S2G in Multimedia Appendix 2).

Primary Outcomes

The objective of this study is to compare the effectiveness of implementing information safety barriers in reducing the incidence of surgical specimen errors. These errors encompass inaccuracies in patient information registration, specimen identification and localization, as well as errors in the submission of specimens for examination. The error rate of surgical specimen' submission is equal to the number of surgical specimen submission errors divided by the total number of surgical specimen submissions divided by 100%.

Secondary Outcomes

This study aimed to compare the timeliness of fixing common pathological specimens before and after implementing the surgical specimen management mode with information security barrier control. The timeliness of fixation was assessed based on the passing rate and average fixation time of common pathological specimens. The fixed pass rate of common pathological specimens is equal to the number of fixed specimens completed within 30 minutes divided by the total number of fixed specimens multiplied by 100%. The average fixation time of common pathological specimens is equal to the total fixation time of a single common pathological specimen divided by the total fixation time of all common pathological specimens.

The reporting rate of surgical specimen near misses was compared between the 2 groups—the near miss events included wrong patient information registration, wrong specimen name and location, and wrong specimen type (such near miss events were corrected before the pathology report was issued by the pathology department, which did not cause serious consequences to patients, but there were great safety risks). The close error reporting rate of surgical specimens in the 2 groups was compared. The near error reporting rate is equal to the actual number of reported cases divided by the total number of reported close errors multiplied by 100%. The actual number of reported cases was the number of cases actually reported after the

occurrence of abnormal events by revising the information on surgical specimens through written records.

The list of all surgeons and operating room nurses was obtained through the hospital information system, and 60 surgeons and 60 operating room nurses were randomly selected by using the random number table method to investigate the satisfaction of surgeons and operating room nurses on the surgical specimen disposal process and specimen management mode, and the satisfaction evaluation was collected at the end of specimen data collection in the control group and the improved group. Satisfaction is rated on a 5-point scale from very dissatisfied to very satisfied.

Ethical Considerations

The study design and procedures conformed to the Declaration of Helsinki. This study was approved by the ethics review board of the First Affiliated Hospital of Wenzhou Medical University (KY2023-R098).

Statistical Analysis

SPSS (version 26.0; IBM Corp) statistical software was used for statistical analysis. Use cases (percentage) and (mean, SD) were used to represent counting and measuring data, respectively. The *t* test was used for comparison among measurement data groups, the chi-square test was used for counting data, and the rank sum test was used for rank data.

Results

Primary Outcomes

From January 1, 2021, to December 31, 2021, the total number of pathological specimens is 84,289, of which 55,545 are ordinary, 1096 are paraffin accelerated, and 27,648 are frozen. From January 1, 2022, to December 31, 2022, the total number of pathological specimens was 99,998, of which 62,383 were ordinary, 5744 were paraffin accelerated, and 31,871 were frozen (Table 1). Before the implementation, there were 31 errors surgical specimen in submitting surgical specimens, including 17 errors in the name and location of surgical specimens, 6 errors in the type of specimens submitted, 2 errors in the information of registered patients, and 6 errors in the left and right sides of pathology reports. After the implementation, there were 4 errors in surgical specimens, including 4 errors in the name and location of surgical specimens (Table 2). The chi-square test showed that the error rate of surgical specimens' submission was significantly lower than that before implementation and the difference was statistically significant (χ^2_1=25.9; *P*<.001).

Table . Summary table of surgical specimen types and quantities for 2021 - 2022.

Year	Paraffin specimen, n	Frozen specimen, n	Paraffin acceleration, n	Total, n
2021	55,545	27,648	1096	84,289
2022	62,383	31,871	5744	99,998

Table . Information on the occurrence of misses from 2021 - 2022.

Near miss events	Number (2021/2022), n/N	Key points of miss
Incorrect registration of patient information	2/0	Errors in electronic surgery notification checks
The specimen's name was entered in the wrong place	17/4	Communication and checking failures between medical and nursing staff
Wrong type of surgical specimen sent for examination	6/0	Roving nurse operator error
Pathology report error	6/0	Pathologist makes error in entering surgical site in pathology report

Secondary Outcomes

After the implementation of the new surgical specimen process management mode, the fixed pass rate of common pathological specimens increased to 99%, and the average fixation time was reduced to 10.221 (SD 12.552) minutes, while the differences were statistically significant (*P*<.001) with chi-square test and *t* test. At the same time, we found that only 13 of the 31 (41.9%) near-miss events in 2021 were reported, compared to all of the 4 (100%) near-miss events in 2022. We used Fisher precision probability test to measure the increase in the rate of near-miss reporting, which showed a statistically significant difference between 2021 and 2022 (*P*=.045; Table 3).

Table . Fixed qualification rate and average qualification time of common pathological specimens.

Indicators	2021	2022	Chi-square (*df*)/*t* test	*P* value
Fixed pass rate, n/N (%)	53,001/55,545 (95.5)	61,730/62,383 (98.9)	114,798.6 (1)[a]	<.001
Fixed meantime (minutes), mean (SD)	12.597 (13.032)	10.221 (12.552)	34.3 (232,657)[b]	<.001
The rate of near miss reporting, n	41.9	100	N/A[c]	.045

[a]Denotes the chi-square value.

[b]Denotes the *t* test value.

[c]N/A: not applicable.

Comparison of the satisfaction degree of doctors and nurses in relevant clinical departments with the procedure and management of surgical specimens before and after implementation. After the implementation, the satisfaction score of doctors in relevant clinical departments with the quality of pathological specimen management was 93.3% (56/60), higher than 58.3% (35/60) before the implementation, and the difference was statistically significant with rank sum test (P<.001; Table 4).

Table . Comparison of satisfaction with surgical specimen handling process and management between the 2 groups of medical staff.

Group	Great satisfaction, n	Satisfaction, n	Ordinary, n	Dissatisfaction, n	Be very dissatisfied, n	*z* score	*P* value
Doctor satisfaction						−5.133	<.001
Control group (n=60)	15	20	22	2	1		
Improvement group (n=60)	45	11	4	0	0		
Nurse satisfaction						−4.848	<.001
Control group (n=60)	14	22	20	3	1		
Improvement group (n=60)	46	10	4	0	0		

Discussion

Principal Findings and Contributions

The goal of this study is to develop a new surgical specimen operating standard and procedure that can effectively reduce the incidence of surgical specimen errors. Standardized procedures and standardized operations can simplify the work of nurses and medical teams and improve the quality of work [8]. In the traditional model mode, during the practice of the specimen identification stage, the surgeon's verbal error, nursing interruption, input error, verification and communication failure, and other links would occur [11]. Key errors in surgical specimen management will directly lead to errors in surgical specimens, resulting in incomplete or inaccurate diagnoses of patients by clinicians. This can lead to inappropriate and inaccurate treatment and care, causing temporary or permanent physical and psychological harm to patients. Therefore, we have developed a new operating standard and procedure for surgical specimens, which can effectively reduce the occurrence of surgical specimen errors, greatly improve the working efficiency of the operating room, reduce the risk of surgical specimen route errors, and ultimately greatly improve the quality of medical care and guarantee the medical safety of patients.

There were 17 errors in identifying parts of surgical specimen names in our research in 2021, mainly on the left and right sides and the upper, middle, and lower parts. And the early warning information safety barrier in our new management model can intercept the wrong specimen names. When the specimen name is inconsistent with the diagnosis site during operation, the system will automatically pop-up to remind that the interception barrier is inconsistent with the diagnosis site of the patient (left and right side and other key controls such as upper, middle, and lower). If correct, the specimen name bar code will be printed. If it is inconsistent with the diagnosis site, the system will display the red large font box "specimen name is inconsistent with the diagnosis, please check the specimen name again is correct" to play a role in checking again. In particular, errors on the right and left sides of key points can lead to a series of pathological diagnosis errors leading to very serious adverse events. The results showed that in 2022, the error rate of specimen name identification was significantly reduced to 4 cases, among which 4 cases of specimen name error were mainly caused by the failure of bilateral surgical site verification for the same incision. Meanwhile, the error rate of surgical specimens was significantly lower than that before the operation (χ^2_1=25.9, P<.001).

Theoretical and Practical Significances

In the quality and safety of nursing care in the operating room, it has always been a hot issue to explore the application of intelligent management to the surgical specimen inspection process and safety management [10]. The process of specimen inspection was improved to achieve information-based closed-loop management, and barcode technology was used to track and record the whole process of specimen inspection to form closed-loop management, so as to improve the accuracy and traceability of intraoperative pathological specimen information [9]. The traditional specimen inspection is done by scanning the 2D barcode code, but there is no information reminder function. In this paper, the link of specimen fixation is proposed. If the surgical specimen is not fixed 20 minutes after registration, the computer pop-up window will remind "Abnormal specimen submission," display the name of the corresponding operating room, itinerant nurses and patient information, and urge the itinerant nurses to submit pathological specimens for examination in time to prevent late and missed inspection of surgical specimens. Barcode technology is used to track and record the whole process of surgical specimen submission, forming closed-loop management and improving the accuracy and traceability of intraoperative pathological specimen information [12], accurately implement closed-loop management of specimens, transport personnel pays attention to the status of surgical specimens in real time, whether it has closed loop, and promptly urge itinerant and hand-washing nurses to implement the specimen inspection work in strict accordance with the surgical specimen inspection system. If the time from marking to receiving of surgical specimen exceeds half an hour, the system will show a red warning module that the specimen is not alerted to the pathology department, which can effectively reduce the incidence of surgical specimen delivery near misses. A shorter time for surgical specimen examination can reduce the difficulty of pathological diagnosis [13]. The qualified rate of fixation of common pathological specimens increased to 99%, and the differences were statistically significant (P<.001). Controlling the closed-loop management of the information of each link of surgical specimens, the number of specimens sent for surgery at each stage, and the status of specimens at each link are distinguished by different color blocks. The full-time specimen-sending personnel and the full-time receiving personnel of the pathology department checks whether each surgical specimen has reached the closed-loop state. It achieves accurate control of each link in the process of submitting surgical specimens, refines the standardized process of submitting surgical specimens, and ensures the controllability of each link, thus ensuring the correctness and simplification of the entire process of submitting surgical specimens, improving the qualification rate of handling surgical specimens, and reducing the risk of approaching errors [14].

With the high-quality development of hospitals, the informatization of operating rooms has been fully popularized and entered the era of a new paperless model [15]. It is particularly important to track and manage the surgical specimens and analyze the abnormal process. The traditional mode refers to the manual way to collect information manually, and the information filled in manually may have problems such as missing filling, wrong filling, mistransmission, and so forth. This paper uses network technology to realize information sharing and captures data through nursing managers to modify surgical specimens and near misses. In case of surgical specimen execution error, the traveling nurse should report the cause of the incident near the error on the interface of the pathology application form, and the specimen will be correctly executed after approval and verification by the head nurse. After using the new management mode, the near-miss reporting rate increased to 100% (P=.045). Meanwhile, this is also one of the ways to collect surgical specimen errors. Through information collection, we can find the causes of specimen errors and analyze and improve the key links of information. In the event of a surgical specimen near miss, the nurse on duty applied for specimen review, entered the reasons for the error, and made an application. After the review by the regulator, the correct name of the specimen can be obtained, the analysis and rectification of each data capture abnormal problem can be carried out, and scientifically use the quality tracking management tool of specimen approaching error anomaly to analyze the root cause and learn from it. It is helpful to improve the correct execution rate of surgical specimens and enhance the awareness of specimen management [2].

IT has been used to effectively improve the process of surgical specimen inspection [16], which has a sensitive and accurate perception function and effective verification, supervision, and control function, and improves work efficiency and the overall quality of specimen inspection. The satisfaction of medical staff had significantly increased to 93.3% (P<.001), and the efficiency of surgical specimen inspection was improved. The research on the prevention of surgical specimens near misses through effective information control link technology and the establishment of intelligent closed-loop management mode of pathological specimens in intelligent operating rooms is still rare, and there are no relevant reports at home and abroad. Through statistical analysis of medical satisfaction, this paper shows that the work of nurses and medical teams can be simplified and the quality of work can be improved [7,8]. The purpose of this study is to construct a new model of surgical specimen process management through the information security barrier to prevent the specimen from approaching the wrong event to create a safe operating room nursing practice environment.

Limitations

First, different personnel receive different education levels and ages, so different personnel have different ability to operate the system, so there will be deviations. Furthermore, different types of surgery may require different methods of recording and managing specimen information. There is also a Hawthorne effect in this study, which affects the results. In addition, inadequate information management systems can lead to data underreporting or inaccurate reporting, which can have serious implications for patient diagnosis and treatment. Finally, the direct relationship between the management model and patient outcomes was not discussed in our study and we will conduct this part in future studies. In conclusion, all of these issues may lead to limitations in the study.

Conclusions

We have developed a novel mode of managing the surgical specimen process. This new model effectively controls and manages the entire process, including the preanalysis stage, analysis stage, and fault analysis stage, of the information security barrier. It significantly reduces the risk of near misses associated with surgical specimens. Moreover, the new management model serves as a valuable reference for clinical decision makers and enables multiple hospitals to enhance operating room efficiency, reduce the occurrence of near misses in surgical specimens, and ultimately improve the quality of medical care while ensuring patient safety.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Module.
[DOCX File, 21 KB - medinform_v12i1e52722_app1.docx]

Multimedia Appendix 2
Procedure for each port specimen.
[DOCX File, 842 KB - medinform_v12i1e52722_app2.docx]

References

1. Zhang HSW, Jian W, Zhang M, Wu Z, Yao L. The relationship of professional practice environment and patient safety (In Chinese). Chin Nurs Manag 2018;18. [doi: 10.3969/j.issn.1672-1756.2018.10.003]
2. Feng TT, Zhang X, Tan LL, Liu HP. A literature review of nursing management for near miss (In Chinese). Chin J Nurs 2020;55:153-157. [doi: 10.3761/j.issn.0254-1769.2020.01.027]
3. Joint Commission International Accreditation Standards for Hospitals, 6th Edition: International Joint Commission; 2017. URL: https://www.jointcommissioninternational.org/jci-accreditation-standards-for-hospitals-6th-edition/2018 [accessed 2024-10-04]
4. Cheng JYL, Cao Y, Liao Z, et al. Causes and counter measures of adverse events in active medical equipment (In Chinese). Chin J Nurs 2018;53:1363-1366. [doi: 10.3761/j.issn.0254-1769.2018.11.017]
5. Claffey C. Near-miss medication errors provide a wake-up call. Nursing (Auckl) 2018 Jan;48(1):53-55. [doi: 10.1097/01.NURSE.0000527615.45031.9e] [Medline: 29280844]
6. 6th annual world patient safety, science & technology summit. World Health Organization. URL: https://www.who.int/director-general/speeches/detail/6th-annual-world-patient-safety-science-technology-summit2018 [accessed 2024-09-06]
7. Xu DGZ, Zhang X, Song Y, Wang R. Occurrence status of near-miss events and adverse events based on nursing information system:a longitudinal study (In Chinese). Chin Nurs Res 2021;35:2295-2298. [doi: 10.12102/j.issn.1009-6493.2021.13.007]
8. Wang XQH, Pan M. Practice and effect analysis of optimizing surgical specimen management by using information technology (In Chinese). Hosp Manag Forum 2021;38:86-88. [doi: 10.3969/j.issn.1671-9069]
9. Bai YHX, Tang S. The practice of whole-process closed-loop management of tumor surgical pathological specimens based on information technology (In Chinese). Clin Med Pract 2023;32:957-959. [doi: 10.16047/j.cnki.cn14-1300/r.2023.12.016]
10. Wendi N. Construction and application of smart operating room based on hospital information platform (In Chinese). Dig Tech Appl 2023;41:71-73. [doi: 10.19695/j.cnki.cn12-1369.2023.06.22]
11. Kinlaw TS, Whiteside D. Surgical specimen management in the preanalytic phase: perioperative nursing implications. AORN J 2019 Sep;110(3):237-250. [doi: 10.1002/aorn.12782] [Medline: 31465576]
12. Brennan PA, Brands MT, Caldwell L, et al. Surgical specimen handover from the operating theatre to laboratory—can we improve patient safety by learning from aviation and other high‐risk organisations? J Oral Pathol Med 2018 Feb;47(2):117-120. [doi: 10.1111/jop.12614]
13. Link T. Guidelines in practice: specimen management. AORN J 2021 Nov;114(5):443-455. [doi: 10.1002/aorn.13518] [Medline: 34706085]
14. Heher YK, Chen Y, VanderLaan PA. Pre-analytic error: a significant patient safety risk. Cancer Cytopathol 2018 Aug;126 Suppl 8:738-744. [doi: 10.1002/cncy.22019] [Medline: 30156766]
15. Wang XXJ, Lu K, Zhang M, Gu S, Lin Y, Pan S. Discussion on risk management of outpatient surgery pathology sample safety based on failure mode and effects analysis (In Chinese). Hosp Manag Forum 2020;37:48-50. [doi: 10.3969/j.issn.1671-9069.2020.05.015]
16. Kedar S, Mate AP. The antidote to fragmented health care. Harvard Business Review. 2014. URL: https://hbr.org/2014/12/the-antidote-to-fragmented-health-care [accessed 2014-12-15]

Edited by C Lovis; submitted 13.09.23; peer-reviewed by D Singh, J Drott; revised version received 16.04.24; accepted 07.07.24; published 14.10.24.

Please cite as:
Chen T, Tang X, Xu M, Jiang Y, Zheng F
Application of Information Link Control in Surgical Specimen Near-Miss Events in a South China Hospital: Nonrandomized Controlled Study
JMIR Med Inform 2024;12:e52722
URL: https://medinform.jmir.org/2024/1/e52722
doi:10.2196/52722

Reducing Firearm Access for Suicide Prevention: Implementation Evaluation of the Web-Based "Lock to Live" Decision Aid in Routine Health Care Encounters

Julie Angerhofer Richards[1,2], MPH, PhD; Elena Kuo[1], MPH, PhD; Christine Stewart[1], PhD; Lisa Shulman[1], MSW; Rebecca Parrish[3], LICSW; Ursula Whiteside[4,5], PhD; Jennifer M Boggs[6], MSW, PhD; Gregory E Simon[1,3,5], MD, MPH; Ali Rowhani-Rahbar[7,8], MD, MPH, PhD; Marian E Betz[9], MD, MPH

1

2

3

4

5

6

7

8

9

Corresponding Author:
Julie Angerhofer Richards, MPH, PhD

Abstract

Background: "Lock to Live" (L2L) is a novel web-based decision aid for helping people at risk of suicide reduce access to firearms. Researchers have demonstrated that L2L is feasible to use and acceptable to patients, but little is known about how to implement L2L during web-based mental health care and in-person contact with clinicians.

Objective: The goal of this project was to support the implementation and evaluation of L2L during routine primary care and mental health specialty web-based and in-person encounters.

Methods: The L2L implementation and evaluation took place at Kaiser Permanente Washington (KPWA)—a large, regional, nonprofit health care system. Three dimensions from the RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) model—*Reach*, *Adoption*, and *Implementation*—were selected to inform and evaluate the implementation of L2L at KPWA (January 1, 2020, to December 31, 2021). Electronic health record (EHR) data were used to purposefully recruit adult patients, including firearm owners and patients reporting suicidality, to participate in semistructured interviews. Interview themes were used to facilitate L2L implementation and inform subsequent semistructured interviews with clinicians responsible for suicide risk mitigation. Audio-recorded interviews were conducted via the web, transcribed, and coded, using a rapid qualitative inquiry approach. A descriptive analysis of EHR data was performed to summarize L2L reach and adoption among patients identified at high risk of suicide.

Results: The initial implementation consisted of updates for clinicians to add a URL and QR code referencing L2L to the safety planning EHR templates. Recommendations about introducing L2L were subsequently derived from the thematic analysis of semistructured interviews with patients (n=36), which included (1) "have an open conversation," (2) "validate their situation," (3) "share what to expect," (4) "make it accessible and memorable," and (5) "walk through the tool." Clinicians' interviews (n=30) showed a strong preference to have L2L included by default in the EHR-based safety planning template (in contrast to adding it manually). During the 2-year observation period, 2739 patients reported prior-month suicide attempt planning or intent and had a documented safety plan during the study period, including 745 (27.2%) who also received L2L. Over four 6-month subperiods of the observation period, L2L adoption rates increased substantially from 2% to 29% among primary care clinicians and from <1% to 48% among mental health clinicians.

Conclusions: Understanding the value of L2L from users' perspectives was essential for facilitating implementation and increasing patient reach and clinician adoption. Incorporating L2L into the existing system-level, EHR-based safety plan template reduced the effort to use L2L and was likely the most impactful implementation strategy. As rising suicide rates galvanize the urgency of prevention, the findings from this project, including L2L implementation tools and strategies, will support efforts to promote safety for suicide prevention in health care nationwide.

(JMIR Med Inform 2024;12:e48007) doi:10.2196/48007

KEYWORDS

suicide prevention; firearm; internet; implementation; suicide; prevention; decision aid; risk; feasible; support; evaluation; mental health; electronic health record; tool

Introduction

Firearm-related suicide accounts for approximately half of the suicide deaths in the United States annually [1]. Firearms are common in Americans' lives [2]; about one-third of Americans report owning firearms [3], and an additional 10% report living in a household with a firearm [4], with higher rates in western states [2], among veterans [5], and in rural areas [6]. Moreover, the rate of ownership of new firearms appears to have increased recently among women, Black people, and Hispanic people [7]. Suicide attempts by firearm are highly lethal; researchers estimate that 85% to 95% of individuals who attempt suicide by firearm do not survive [8,9], and people with access to firearms, particularly if firearms are kept loaded and unlocked [10,11], have increased suicide risk [12,13]. Clinicians may have opportunities to intervene with patients at risk for firearm-related suicide because about 50% of individuals who die by suicide see a clinician in the month before death, and over 80% see one in the year before death [14]. Moreover, clinician-initiated discussions about reducing access to firearms have demonstrated effectiveness for improving firearm security practices (particularly in combination with free safe storage devices) [15-17], as well as promising findings for reducing suicide attempts [18,19].

Despite its potential benefits, clinician-initiated dialogue about limiting access to firearms is an uncommon practice across many primary care and mental health specialty practices [18,20]. Common barriers include time, clinicians' unfamiliarity with firearms, and concerns about negatively impacting relationships or alienating patients [21]. "Lock to Live" (L2L) is a self-directed, anonymous, web-based decision aid that was designed to address these barriers. L2L was developed in collaboration with clinicians, firearm owners, and people who had experienced suicidal thoughts and attempts [22]. Consistent with international design standards [23,24], the L2L decision aid steps users through various considerations regarding in-home and out-of-home firearm storage options, such as types of storage, costs of storage, and background check requirements, with a goal of encouraging storage solution discussions that are consistent with the users' values and preferences [22]. Two subsequent research studies demonstrated promising results for the feasibility and acceptability of offering L2L emergency care encounters [25] and for the uptake of L2L when it was offered via secure patient portal messages after outpatient care encounters [26]. Though L2L appears to be a useful tool for supporting suicide prevention in clinical practice, little is known about how to use L2L during routine health care encounters outside of a research context.

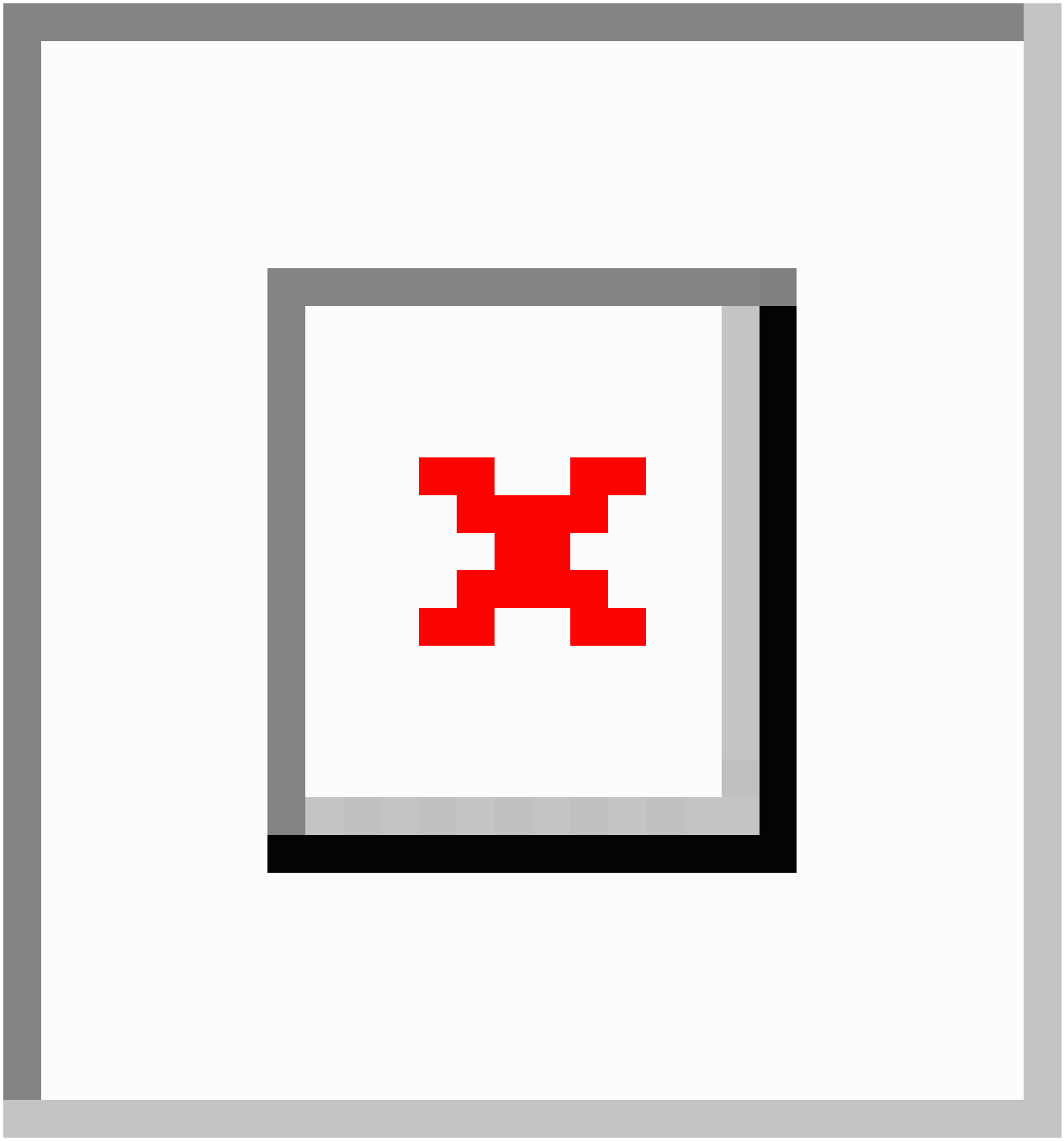
The goal of this project was to use mixed methods (qualitative and statistical evaluations) to support the implementation and evaluation of L2L during primary care and mental health specialty encounters in a large, regional health care system. Specifically, this project used semistructured interviews with clinicians and patients to support implementation, as well as statistical analyses to evaluate the reach and adoption of L2L over a 2-year period. The evaluation findings will inform considerations for implementing L2L nationwide to support suicide prevention in health care systems.

Methods

Setting

L2L implementation and evaluation took place at Kaiser Permanente Washington (KPWA)—1 of 8 regional Kaiser Permanente health care systems, which together form one of the nation's largest nonprofit health care organizations and serve 12.5 million people [27]. At the time of this evaluation, KPWA had provided comprehensive medical care to approximately 700,000 members across Washington State via employer-sponsored insurance plans, individual insurance plans, or capitated Medicaid or Medicare programs. In 2016, KPWA augmented standard clinical workflows to support the identification and engagement of patients at high risk of suicide attempts (Figure 1) [28,29]. Specifically, a system-level electronic health record (EHR) template was created to support clinician-initiated safety planning among patients who are identified as at high risk of suicide during primary care and mental health specialty encounters [28,30]. Nationally, safety planning is a widely recommended best practice [31], and KPWA had an established process for safety planning that included addressing access to lethal means but did not offer any specific resources to clinicians or patients about firearm storage options. Consistent with the goal of L2L, step 6 of this safety plan template was designed to support patients in limiting access to lethal means, such as firearms and prescription medications.

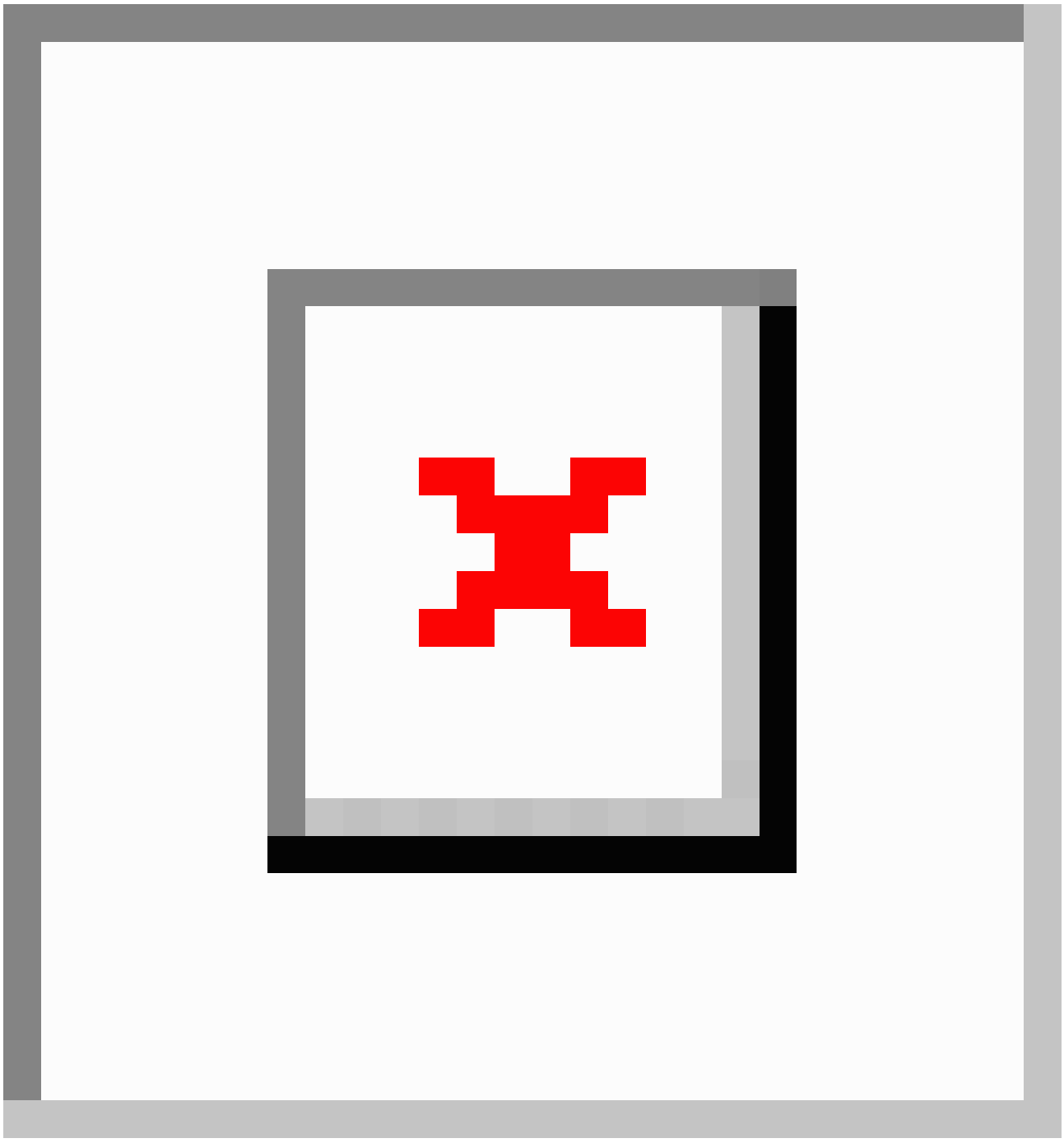
Figure 1. Clinical workflow for supporting the identification and engagement of patients at high risk of suicide during primary and mental health specialty encounters at Kaiser Permanente Washington.



Implementation and Evaluation Framework, Data Sources, and Study Design

Three dimensions from the RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) model—*Reach*, *Adoption*, and *Implementation* [32,33]—were selected to both inform and evaluate the implementation of L2L at KPWA (Multimedia Appendix 1) over a 2-year observation period (January 1, 2020, to December 31, 2021). Specifically, a qualitative, team-based, formative evaluation [34] was used, involving semistructured interviews with purposefully sampled patients and clinicians to facilitate implementation tools and strategies. Descriptive statistical analyses were used to evaluate the reach of L2L among patients identified at high risk of suicide and L2L adoption over a 2-year observation period. The findings were stratified by primary care and mental health specialty service settings due to the variation in the timing of the L2L trainings across these settings (Figure 2).

Figure 2. L2L implementation tools and strategies used over four 6-month subperiods of the 2-year observation period (January 1, 2020, to December 31, 2021). Tools are shown in blue, strategies are shown in green, and capped lines indicate semistructured interviews. EHR: electronic health record; KPWA: Kaiser Permanente Washington; L2L: Lock to Live; LICSW: licensed clinical social worker.



Semistructured Qualitative Interviews

EHR data were used to purposefully recruit adult patients to participate in semistructured interviews that included questions to elicit suggestions about introducing L2L, as part of a broader interview that focused on exploring perceptions regarding and experiences with firearm access assessment [35]. An invitation letter was mailed to sampled patients (age≥18 y) who had received a standardized question about firearm access ("Do you have access to guns? Yes/No") in the prior 2 weeks on a mental health questionnaire [30]. A stratified sampling distribution was used to recruit approximately equal numbers of patients in 3 groups, including those who (1) reported firearm access, (2) reported no firearm access, and (3) did not respond (ie, left the question blank). The three sampling groups were also designed to purposefully include patients who had reported thoughts about self-harm in the prior 2 weeks via the ninth question of the 9-item Patient Health Questionnaire (PHQ-9) [36]. Interviewers attempted to reach all invitees for 2 weeks to invite them to participate in a phone interview. Following the portion of the interview guide that focused on firearm access assessment [35], interviewers described how patients reporting suicidal thoughts would soon be receiving L2L and elicited feedback about how to introduce this tool in a way that would make patients more likely to use it (Multimedia Appendix 2). This portion of the interview transcript was extracted into Excel

(Microsoft Corporation) and analyzed by using a team-based, rapid, qualitative inquiry approach [37]. This involved an iterative data analysis wherein 2 coders independently coded the L2L portion of the transcript by using a combination of deductive and inductive content analyses, with codes developed a priori from the interview guide as well as codes that emerged from the interviews [38]. This was followed by several rounds of discussions with 2 additional team members who reviewed the coded data and reconciled themes iteratively for the purpose of using the summarized themes to facilitate implementation.

Following the completion of the qualitative analysis that focused on patient-informed L2L implementation, interviewers initiated clinician recruitment activities, which were also more broadly focused on firearm access assessment. At the recommendation of care delivery leaders, interviewers outreached to the following two groups of clinicians responsible for engaging patients identified at risk of suicide in risk mitigation (ie, safety planning [39]): (1) licensed clinical social workers (LICSWs) supporting integrated mental health in primary care [28] and (2) consulting nurses (registered nurses) responsible for connecting patients (ie, those reporting suicidality after business hours via telephone) to telephone-based follow-up care. The L2L portion of the interview guide included questions informed by patient interviews (Multimedia Appendix 2) and was analyzed by using the same rapid, qualitative inquiry approach [37] that was used for patient interviews for the purpose of further facilitating L2L implementation.

Descriptive Statistical Analyses

L2L Reach and Adoption

EHR data were used to summarize L2L "reach," which was defined as the proportion of patients identified as at high risk of suicide via routine screening and assessment clinical workflows (Figure 1) and received the web-based decision aid. Specifically, we described characteristics of patients who had a documented safety plan during the 2-year observation period and characteristics of patients who had a safety plan that included a reference to L2L. Next, we described the adoption of L2L by primary care and mental health specialty clinicians over four 6-month subperiods of the observation period by calculating the proportions of patients identified at high risk of suicide (via suicide risk assessment; described in the *Measures* section) who had a documented safety plan with a reference to L2L and those who had a documented safety plan without a reference to L2L. We selected 6-month subperiods as the most helpful way to visually describe L2L adoption over time, since implementation paused during the initial COVID-19 outbreak (described in the *Implementation Timeline, Tools, and Strategies* section). We stratified by service setting due the variation in the timing of L2L trainings for these groups of clinicians.

Measures

The Columbia Suicide Severity Rating Scale (C-SSRS) was used to measure suicide risk, as per current clinical workflows. Specifically, patients reporting some level of suicide attempt planning or intent in the past month (ie, answering "yes" to C-SSRS question 3 or higher) were considered to be at "high risk" and alerted clinicians (via EHR prompts) to initiate safety planning. Distinctive phrases from standard EHR-based templates were used to detect safety plans documented in the text of clinical notes among patients identified at high risk (Multimedia Appendix 3). Sociodemographic and clinical characteristics of interest, including those known to be associated with firearm ownership and suicide risk [40,41], were measured by using the following administrative and diagnostic EHR data: age (continuous); sex (male or female); race and ethnicity (Asian, Black, Hispanic or Latinx, White, other, or unknown); insurance type (commercial, Medicare, Medicaid, or other); rurality (urban, large suburban, small suburban, or mostly rural) [42,43]; and prior year mental health, substance use, and self-harm diagnoses derived from the *International Classification of Diseases, Tenth Revision, Clinical Modification*. Reported firearm access was measured based on a positive response to the question on the mental health questionnaire [30] that was also used for qualitative interview recruitment (described in the *Semistructured Qualitative Interviews* section).

Ethical Considerations

The project team received approval from the KPWA Region Institutional Review Board (review number: 1826198) to conduct this study. Patients who agreed to participate in the phone interview provided oral consent, including permission for the interview to be audio-recorded and professionally transcribed, and they received a US $50 cash incentive for participation. During clinician recruitment activities, clinicians received up to 3 email invitations, which included a study information sheet and instructions for opting out of participation and further contact. Participating clinicians verbally consented to participation and received a US $50 gift card for participation.

Results

Implementation Timeline, Tools, and Strategies

Over the 2-year implementation period (January 1, 2020, to December 31, 2021), a team of researchers and care delivery leaders took a pragmatic approach to iteratively creating and refining L2L implementation tools and strategies for primary care and mental health specialty clinicians (Figure 2). Initially, tools included an EHR-based macro (ie, EPIC SmartPhrase [Epic Systems Corporation]) for clinicians to easily add a URL and QR code referencing L2L to safety planning templates and a 1-page quick-start guide (ie, "Huddlecard") with information on how to use the new SmartPhrase during routine clinic meetings (ie, "huddles"). In February 2020, the LICSWs who supported mental health care delivery in primary care [28] received information about L2L during a brief, web-based staff training session. Additional trainings that were planned for mental health specialty clinicians were put on hold during the widespread service disruption that subsequently occurred in response to the initial COVID-19 outbreak in March 2020. Additional tools and strategies were used, following recommendations from the care delivery leaders responsible for primary care and mental health service recovery and from the patients and clinicians who participated in semistructured qualitative interviews (detailed in the *Findings From Semistructured Qualitative Interviews* section).

Findings From Semistructured Qualitative Interviews

Of 76 patients who were purposefully sampled during 2 waves of recruitment, 36 were interviewed from November 18, 2019, to February 10, 2020 (Table 1). Five organizing themes were derived from the portion of the interview that elicited perceptions and suggestions about L2L and were used to create a handout for clinicians, with suggestions about how to introduce L2L to their patients at risk of suicide (Multimedia Appendix 4), including recommendations to (1) "have an open conversation," (2) "validate their situation," (3) "share what to expect," (4) "make it accessible and memorable," and (5) "walk through the tool" (Table 2). In addition to these recommendations, patients expressed a preference for receiving information about L2L from "trusted" and "caring" clinicians.

Table . Characteristics of patient (n=36) and clinician (n=30) semistructured interview participants.

		Patients	Clinicians
Sex, n (%)			
	Female	17 (47)	24 (80)
	Male	19 (53)	6 (20)
Age (y), mean (SD)		47.3 (17.9)	44.3 (12.1)
Age category (y), n (%)			
	19-29	8 (22)	1 (3)
	30-49	11 (31)	20 (67)
	50-64	9 (25)	6 (20)
	≥65	8 (22)	3 (10)
Race and ethnicity, n (%)			
	American Indian or Alaska Native	0 (0)	1 (3)
	Black	3 (8)	2 (7)
	Asian or Pacific Islander	3 (8)	5 (17)
	Latinx or Hispanic	1 (3)	4 (13)
	Unknown	2 (6)	0 (0)
	White	27 (75)	18 (60)
Reported firearm access[a,b], n (%)		16 (44)	N/A[c]
Reported thoughts about self-harm (prior 2 wk)[a,d], n (%)		15 (42)	N/A

[a]Patients' responses recorded on the Kaiser Permanente Washington mental health monitoring questionnaire used for criterion sampling within the 2 wk prior to the recruitment initiation.

[b]"Do you have access to guns? Yes/No."

[c]N/A: not applicable.

[d]Ninth question on the 9-item Patient Health Questionnaire: "Thoughts that you would be better off dead, or of hurting yourself."

Table . Thematic analysis of semistructured interviews with patients (n=36) and recommendations for introducing "Lock to Live" (L2L).

Themes	Illustrative quotes[a]	Recommendation
Show caring and compassion, ask permission, and respect autonomy	• "I think it's important to just take a breath, sit down with them, hold their hand, look them in the eye - 'how can I help you? Help me help you. What's going on? Tell me. What are you thinking? How are you feeling? How can I help you?' Instead of an assembly line and 'I only have a few minutes,' so they [providers] don't take the time." (Patient B029) • "I would hope the provider is very warm and caring and explains it's a safety precaution, it's for your better health and it insures you'll be safer…basically it's another part of your little toolbox to keep yourself well." (Patient A032) • "Probably compassionately, potentially generalized to begin with, to find out if the person is resistant right upfront… Maybe you could put a question, 'would you be willing to consider options for storing or access to lethal means, whether it's firearms or medication? Is it something you would be willing to discuss and look into if you were experiencing suicidal thoughts?'" (Patient A005) • "We're offering you the means to protect yourself, this is not an us decision, this is a you decision. So here is the website, the online information and we encourage you to look at it, but it's your decision….Nobody can force you to do things. So bringing it up more as like – not we're taking it [firearm] away from you, but letting you decide what to do with it." (Patient A024) • "Reassuring people that their firearm ownership will not end because they're going through a rough patch in life; their ability to have their own authority to hold onto their possessions [will not end], firearms or not." (Patient B036)	"Have an open conversation": patients were more willing to listen and try a tool if a clinician took the time to connect, showed compassion for people's unique experiences, and showed respect for autonomy.
Frame as helpful resource and normalize experiences	• "I think overall education about the topic to start out with, just to say…'this is what we have found is helpful, in these situations.' Moving more into letting people know what their resources could be. Just education to be begin with, so it's not so threatening. I think anytime somebody's in a vulnerable place emotionally, they're already possibly feeling threatened and they may not want to trust a lot of people." (Patient A005) • "I would hope it would be pretty real, like a conversation…'based on your health concerns you're showing, we've got some important information we'd like to share with you,' especially if that person has a relationship or feels responsible with the person presenting it, would stay there together and talk about it afterwards….Also statistics to help a person realize how more common this is." (Patient A011)	"Validate their situation": normalize their experience, share how common suicidal thoughts are, and be nonjudgmental in your approach; people have a variety of gun beliefs.

Themes	Illustrative quotes[a]	Recommendation
Address privacy and security	• "[The provider] would have to explain what it does, how it's going to work and how private it is - nobody can get into your part of the website anyway, your personal page, where you go. So she has to reassure them about that….I just don't feel that being online is that secure." (Patient B004) • "privacy is probably number one and an assurance that you're not being turned in….I would be concerned in our surveillance state that disclosing things to a website about my firearm use might somehow come close to violating some kind of civil right to privacy." (Patient B008) • "As long as people don't have to put in information which can be tracked, I can see lots of people using it. I mean the minute [you have to enter] your name, address, phone number, medical ID number, whatever else, people are going to go – eh." (Patient A033)	"Share what to expect": address privacy and how information is stored if patients visit the website; assure patients that L2L is anonymous.
Accessibility is key	• "You don't want to make it hard to find on a website because it doesn't take me very long. If something's really hard to find on a website, I'm out of there." (Patient B004) • "If there are hoops to jump through before you can access it, if you have to log in, go through a bunch of pages - maybe if it was right there, ready to access at any time, I'd say that'd be better." (Patient B008) • "I'd love to see it everywhere. Have little cards that could be given out, a billboard, having my doctor [send it]." (Patient A033) • "If I knew it existed, I would probably try it. advertise it." (Patient A032) • "Highlight it in your After Visit Summary too." (Patient A002)	"Make it accessible and memorable": have multiple routes for sharing the website and sending reminders (after-visit summary, message, website, pamphlet).
Demonstrate and "show, don't just tell"	• "I think when you're in a pit of despair, to go and do it on your own, some people will do that and other people will not. They need to be taken by the hand and go, 'what do you think about this?' Read it together." (Patient B036) • "I'm more keen to follow somebody who's like 'I'm offering you the opportunity to maybe do this together,' instead of 'I'm watching out for you.'" (Patient A024) • "Being shown an example would be nice, showing it off briefly. Knowing more specifically what it does or how it could be helpful as opposed to just knowing it exists." (Patient B033) • "I think showing the patient or at least offering, would you like me to show you? Not just telling somebody, because short term memory is only like 30 s or a couple minutes and then you forget about it." (Patient A002)	"Walk through the tool": most patients said that a website walk-through, rather than simply having a conversation, would be helpful to overcome the barrier of trying something new, especially if already depressed.

[a]Identifier "A": patients in the first wave of interviews; identifier "B": patients in the second wave of interviews (grammatic edits, noted in brackets, were added to clarify intended meaning).

Of 51 purposefully sampled clinicians responsible for safety planning with patients identified at high risk of suicide, 30 were interviewed from July 7, 2020, to October 8, 2020 (Table 1), including 25 LICSWs and 5 registered nurses. During the

interviews with LICSWs, only 3 had actually used L2L with a patient—9 were unfamiliar with L2L, and 12 were familiar with but had not yet used L2L. Most clinicians saw clear benefits to L2L as an option for supporting both clinicians and patients. Several clinicians expressed concern about using the tool to replace dialogue about lethal means, and most supported the idea of a walk-through, as patients had recommended. Clinicians also expressed a strong preference to have L2L information included by default in the EHR-based safety planning template, in contrast to having clinicians remember to add it (via SmartPhrase). A clinician also suggested automatically including L2L in after-visit summaries when patients reported thoughts about self-harm on the PHQ-9. The implementation team worked with clinical partners to update the system-level, EHR-based safety plan template to include L2L information and updated the Huddlecard to communicate this change (Multimedia Appendix 5). After-visit summaries were used to provide safety plans to patients who were seen via a secure, web-based patient portal, and L2L was automatically included after the template change. Several clinicians also requested follow-up trainings or refreshers about L2L. The team therefore conducted a round of brief trainings, which were presented during routine clinic huddles with mental health specialty clinicians, and created a 3-minute training video (Multimedia Appendix 6).

Findings From Descriptive Statistical Analyses

During the study period, 2739 adult patients reported some prior-month suicide attempt planning or intent via routine suicide risk assessment workflows during primary care or mental health specialty encounters and had a documented safety plan, including 745 (27.2%) who also received L2L. Overall, there were no major differences in the demographic and clinical characteristics between patients who received L2L and the broader population that was identified as at risk of suicide and had a documented safety plan (Table 3).

Table . Characteristics of patients who received "Lock to Live" (L2L; n=745) and were among patients with a documented safety plan (n=2739) during the implementation period (January 1, 2020, to December 31, 2021).

		Patients who received L2L, n (%)	Patients with a documented safety plan[a], n (%)
Age[b] (y)			
	18-39	513 (68.9)	1817 (66.3)
	40-64	187 (25.1)	732 (26.7)
	≥65	45 (6)	190 (6.9)
Sex[c]			
	Female	445 (59.7)	1753 (64)
	Male	300 (40.3)	986 (36)
Race and ethnicity[c]			
	American Indian or Alaska Native	14 (1.9)	75 (2.7)
	Asian	54 (7.2)	199 (7.3)
	Black	52 (7)	166 (6.1)
	Hawaiian or Pacific Islander	8 (1.1)	45 (1.6)
	Hispanic or Latinx	59 (7.9)	201 (7.3)
	Unknown	93 (12.5)	294 (10.7)
	White	465 (62.4)	1759 (64.2)
Insurance[c]			
	Commercial	530 (71.1)	1831 (66.8)
	Medicare	83 (11.1)	342 (12.5)
	Medicaid	54 (7.2)	228 (8.3)
	Not enrolled	78 (10.5)	338 (12.3)
Rural or urban [c], [d]			
	Urban	301 (40.4)	1010 (36.9)
	Large suburban	205 (27.5)	799 (29.2)
	Smaller suburban	186 (25)	802 (29.3)
	Mostly rural	31 (4.2)	96 (3.5)
Mental health diagnoses[e]			
	Depression	675 (90.6)	2502 (91.3)
	Anxiety	652 (87.5)	2434 (88.9)
	Serious mental illness	144 (19.3)	586 (21.4)
	Substance use disorder	196 (26.3)	752 (27.5)
	Suicide attempt diagnosis	39 (5.2)	175 (6.4)
Reported firearm access[e]		150 (20.1)	501 (18.3)

[a]Includes safety plans with L2L.

[b]At evaluation midpoint (January 1, 2021).

[c]At first encounter.

[d]Missing information for 22 patients.

[e]During implementation period.

The adoption of L2L increased substantially over the 2-year observation period (Tables 4 and 5). During this time, rates of documented safety plans among patients identified at high risk of suicide (C-SSRS score≥3) remained fairly consistent—51.2%

to 55.2% of primary care patients and 73.4% to 78.4% of mental health specialty patients had a documented safety plan. However, over four 6-month subperiods of the observation period, L2L adoption rates increased substantially from 2% to 29% among primary care clinicians and <1% to 48% among mental health clinicians, increasing primarily after L2L was integrated into the EHR-based safety planning template.

Table . Proportions of primary care patients who were identified as at high risk of suicide and had a documented safety plan during primary care encounters over the implementation period (January 1, 2020, to December 31, 2021).

Subperiods of implementation period	Patients with a documented safety plan that did not include "Lock to Live," %	Patients with a documented safety plan that did include "Lock to Live," %
Months 1-6	53.5	1.6
Months 7-12	50	4.1
Months 13-18	41.7	11.1
Months 19-24	22.2	29.1

Table . Proportions of primary care patients who were identified as at high risk of suicide and had a documented safety plan during mental health specialty encounters over the implementation period (January 1, 2020, to December 31, 2021).

Subperiods of implementation period	Patients with a documented safety plan that did not include "Lock to Live," %	Patients with a documented safety plan that did include "Lock to Live," %
Months 1-6	78.1	0.3
Months 7-12	74.8	1.4
Months 13-18	53.8	19.6
Months 19-24	25.9	48.4

Discussion

This novel study used mixed methods to support the implementation and evaluation of a web-based decision aid that was designed to help patients at risk of suicide limit access to firearms. Specifically, findings from semistructured interviews with patients and clinicians were used to facilitate L2L implementation, while statistical analyses were used to describe rates of reach among patients identified at risk of suicide and increased adoption by clinicians who cared for them during the 2-year observation period.

L2L development centered users' values and preferences in the design process [22]. Similarly, the tools and strategies developed for this project used information from semistructured interviews with people who were the most likely to be impacted by L2L implementation, including firearm owners, patients experiencing suicidality, and the clinicians who care for them. Clinicians have reported a lack of experience with handling firearms and have expressed apprehension about discussing firearm safety due to concerns about damaging relationships with patients [44-46]. Likewise, patients have expressed apprehension about disclosing access to firearms due to concerns about privacy, autonomy, and firearm ownership rights [47,48]. For these reasons, patients and clinicians perceive firearm access assessment as challenging but also as valuable for supporting suicide prevention [35]. This implementation project showed that clinicians, that is, those responsible for engaging at-risk primary care and mental health patients in suicide risk mitigation, willingly adopted the use of L2L to support safety planning.

This study also has important implementation implications. Unsurprisingly, the rates of L2L adoption increased after L2L was incorporated into the existing system-level safety planning template as a default (primarily in the latter half of year 2). This finding underscores the importance of removing barriers to the adoption of web-based decision aids and making adoption "easy" [49]. In contrast, those seeking change often focus on amplifying benefits or "selling" their new idea or innovation; however, it may be equally as important or more important to focus on "friction," that is, "psychological forces that oppose and undermine change," such as inertia, effort, emotion, and reactance [50]. In the case of L2L, reducing the effort required for clinicians to remember to use L2L appeared to be the main driver of its adoption. However, the tools and strategies that were designed to communicate about the benefits of using L2L (eg, training, video, Huddlecard, and newsletter information) were likely necessary for leaders to understand L2L's value to patients and clinicians and approve the system-level change that was required to make L2L easier to use for clinicians.

This study has important clinical implications for supporting suicide prevention in health care. First, L2L supports clinicians who engage patients identified at risk of suicide in collaborative safety planning and lethal means counseling, which are evidence-based suicide risk mitigation practices that are recommended by the Zero Suicide Institute and follow the principles outlined in the National Strategy for Suicide Prevention [29,51-53]. Moreover, the recommendations from interview participants ("have an open conversation," "share what to expect," and "walk through the tool") support a motivational interviewing approach to lethal means counseling and align with the recommendations of the Veterans Health Administration [54]. Second, L2L was developed by patients with lived experiences of suicidality and firearm ownership; therefore, L2L supports cultural competency in health care as

a culturally aligned intervention [55]. Finally, this technology-based, EHR-embedded approach to addressing lethal means supports all 6 aims of health care quality that are outlined by the Institute of Medicine—*safe*, *effective*, *patient-centered*, *timely*, *efficient*, and *equitable* [56,57].

There are several limitations of this project that have implications for future research. First, the implementation of L2L at KPWA occurred during the initial outbreak of a global pandemic, which impacted the original implementation plans while health care systems responded to the pandemic and rapidly shifted toward providing web-based mental health care [58]. Semistructured interviews with patients took place prior to this shift. Future research should explore optimizing mental health care delivery workflows that support web-based suicide risk identification (ie, screening and assessment) [59] and incorporating L2L in web-based care encounters via secure telehealth platforms that are designed to support patient engagement. Second, L2L recognizes and addresses firearm policies related to background checks and how these policies might influence the legality of temporary firearm transfers for addressing suicide risk, but it does not address specific state laws. Additional work to understand the legality of recommendations about firearm safety practices may be helpful for health care systems that implement L2L. Third, this project was not designed to measure the specific impact of individual implementation strategies or determine whether L2L was effective in helping patients reduce access to firearms for suicide prevention purposes. Measuring the effectiveness of this tool, which was designed to support population-based suicide prevention, would require extending the implementation of L2L to other large health care systems nationwide and conducting other analyses that are designed to measure key functions of suicide prevention practices, including risk identification, engagement in evidence-based risk mitigation and treatment, and supportive care transitions [29]. Finally, L2L is meant to support adult patients at risk of suicide reduce access to firearms and other lethal means; additional tools and strategies are required to support youth at risk of suicide. Notably, there is a similar web-based decision aid that is available for this purpose; "Lock and Protect" was designed to help parents and caregivers reduce access to lethal means for youth suicide risk mitigation [60]. Similarly, the "Safety in Dementia" web-based decision aid was developed to support caregivers in addressing firearm access among individuals with Alzheimer disease and related dementias [61]. Future research should evaluate the implementation of these tools in routine care delivery.

In conclusion, incorporating L2L into the existing system-level safety plan template reduced the effort required to use L2L and was likely the most impactful implementation strategy for increasing clinician adoption and patient reach. However, understanding the value of L2L from the users' perspectives was essential for effectively amplifying the suicide risk mitigation benefits. As rising suicide rates galvanize the urgency of prevention [62], the implementation tools and strategies developed for this project will be useful for health care systems nationwide.

Acknowledgments

This implementation evaluation was funded by Kaiser Permanente's Office of Community Health and Center for Gun Violence Research and Education, as part of its Firearm Injury Prevention Program, and the Centers for Disease Control and Prevention (R01CE003460). The views and opinions expressed in this article are the responsibility of the authors and do not necessarily represent the official views of Kaiser Permanente or the Centers for Disease Control and Prevention. The authors acknowledge that this research would not be possible without the people who receive health care from Kaiser Permanente Washington and all the clinicians and staff who support the organization.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The three dimensions of RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) selected to inform and evaluate the implementation of "Lock to Live."
[DOCX File, 29 KB - medinform_v12i1e48007_app1.docx]

Multimedia Appendix 2
Patient and clinician interview questions that focused on "Lock to Live" implementation.
[DOCX File, 33 KB - medinform_v12i1e48007_app2.docx]

Multimedia Appendix 3
Elements included in the search for safety plans documented in clinical notes text.
[DOCX File, 31 KB - medinform_v12i1e48007_app3.docx]

Multimedia Appendix 4
Introducing Lock2Live.org: a guide for clinicians.

[DOCX File, 295 KB - medinform_v12i1e48007_app4.docx]

Multimedia Appendix 5
“Lock to Live” Huddlecard.
[DOCX File, 589 KB - medinform_v12i1e48007_app5.docx]

Multimedia Appendix 6
“Lock to Live” video for Kaiser Permanente Washington clinicians.
[MP4 File, 13608 KB - medinform_v12i1e48007_app6.mp4]

References

1. WISQARS — your source for U.S. injury statistics. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/injury/wisqars/facts.html [accessed 2022-04-26]
2. McCracken H, Okuley H, Floyd L. Gun ownership in America. RAND Corporation. URL: https://www.rand.org/research/gun-policy/gun-ownership.html [accessed 2022-01-10]
3. Kalesan B, Villarreal MD, Keyes KM, Galea S. Gun ownership and social gun culture. Inj Prev 2016 Jun;22(3):216-220. [doi: 10.1136/injuryprev-2015-041586] [Medline: 26124073]
4. Parker K, Horowitz JM, Igielnik R, Oliphant JB, Brown A. America’s complex relationship with guns. Pew Research Center. 2017 Jun 22. URL: https://www.pewresearch.org/social-trends/2017/06/22/americas-complex-relationship-with-guns/ [accessed 2022-03-29]
5. Cleveland EC, Azrael D, Simonetti JA, Miller M. Firearm ownership among American veterans: findings from the 2015 National Firearm Survey. Inj Epidemiol 2017 Dec 19;4(1). [doi: 10.1186/s40621-017-0130-y] [Medline: 29256160]
6. Nordstrom DL, Zwerling C, Stromquist AM, Burmeister LF, Merchant JA. Rural population survey of behavioral and demographic risk factors for loaded firearms. Inj Prev 2001 Jun;7(2):112-116. [doi: 10.1136/ip.7.2.112] [Medline: 11428557]
7. Miller M, Zhang W, Azrael D. Firearm purchasing during the COVID-19 pandemic: results from the 2021 National Firearms Survey. Ann Intern Med 2022 Feb;175(2):219-225. [doi: 10.7326/M21-3423] [Medline: 34928699]
8. Anestis MD. Guns and Suicide: An American Epidemic: Oxford University Press; 2018.
9. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Fatal Injury Reports, National, Regional and State, 1981-2017, Web-Based Injury Statistics Query and Reporting System (WISQARS). : Centers for Disease Control and Prevention; 2019.
10. Kellermann AL, Rivara FP, Somes G, et al. Suicide in the home in relation to gun ownership. N Engl J Med 1992 Aug 13;327(7):467-472. [doi: 10.1056/NEJM199208133270705] [Medline: 1308093]
11. Shenassa ED, Rogers ML, Spalding KL, Roberts MB. Safer storage of firearms at home and risk of suicide: a study of protective factors in a nationally representative sample. J Epidemiol Community Health 2004 Oct;58(10):841-848. [doi: 10.1136/jech.2003.017343] [Medline: 15365110]
12. Anglemyer A, Horvath T, Rutherford G. The accessibility of firearms and risk for suicide and homicide victimization among household members: a systematic review and meta-analysis. Ann Intern Med 2014 Jan 21;160(2):101-110. [doi: 10.7326/M13-1301] [Medline: 24592495]
13. Mann JJ, Michel CA. Prevention of firearm suicide in the United States: what works and what is possible. Am J Psychiatry 2016 Oct 1;173(10):969-979. [doi: 10.1176/appi.ajp.2016.16010069] [Medline: 27444796]
14. Ahmedani BK, Simon GE, Stewart C, et al. Health care contacts in the year before suicide death. J Gen Intern Med 2014 Jun;29(6):870-877. [doi: 10.1007/s11606-014-2767-3] [Medline: 24567199]
15. Rowhani-Rahbar A, Simonetti JA, Rivara FP. Effectiveness of interventions to promote safe firearm storage. Epidemiol Rev 2016;38(1):111-124. [doi: 10.1093/epirev/mxv006] [Medline: 26769724]
16. Simonetti JA, Rowhani-Rahbar A, King C, Bennett E, Rivara FP. Evaluation of a community-based safe firearm and ammunition storage intervention. Inj Prev 2018 Jun;24(3):218-223. [doi: 10.1136/injuryprev-2016-042292] [Medline: 28642248]
17. Runyan CW, Becker A, Brandspigel S, Barber C, Trudeau A, Novins D. Lethal means counseling for parents of youth seeking emergency care for suicidality. West J Emerg Med 2016 Jan;17(1):8-14. [doi: 10.5811/westjem.2015.11.28590] [Medline: 26823923]
18. Boggs JM, Beck A, Ritzwoller DP, Battaglia C, Anderson HD, Lindrooth RC. A quasi-experimental analysis of lethal means assessment and risk for subsequent suicide attempts and deaths. J Gen Intern Med 2020 Jun;35(6):1709-1714. [doi: 10.1007/s11606-020-05641-4] [Medline: 32040838]
19. Monuteaux MC, Azrael D, Miller M. Association of increased safe household firearm storage with firearm suicide and unintentional death among US youths. JAMA Pediatr 2019 Jul 1;173(7):657-662. [doi: 10.1001/jamapediatrics.2019.1078] [Medline: 31081861]

20. Boggs JM, Quintana LM, Powers JD, Hochberg S, Beck A. Frequency of clinicians' assessments for access to lethal means in persons at risk for suicide. Arch Suicide Res 2022;26(1):127-136. [doi: 10.1080/13811118.2020.1761917] [Medline: 32379012]
21. Walters H, Kulkarni M, Forman J, Roeder K, Travis J, Valenstein M. Feasibility and acceptability of interventions to delay gun access in VA mental health settings. Gen Hosp Psychiatry 2012;34(6):692-698. [doi: 10.1016/j.genhosppsych.2012.07.012] [Medline: 22959420]
22. Betz ME, Knoepke CE, Siry B, et al. 'Lock to Live': development of a firearm storage decision aid to enhance lethal means counselling and prevent suicide. Inj Prev 2019 Sep;25(Suppl 1):i18-i24. [doi: 10.1136/injuryprev-2018-042944] [Medline: 30317220]
23. Stacey D, Bennett CL, Barry MJ, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2011 Oct 5;10(10):CD001431. [doi: 10.1002/14651858.CD001431.pub3] [Medline: 21975733]
24. Légaré F, O'Connor AC, Graham I, et al. Supporting patients facing difficult health care decisions: use of the Ottawa Decision Support Framework. Can Fam Physician 2006 Apr;52(4):476-477. [Medline: 17327891]
25. Betz ME, Knoepke CE, Simpson S, et al. An interactive web-based lethal means safety decision aid for suicidal adults (Lock to Live): pilot randomized controlled trial. J Med Internet Res 2020 Jan 29;22(1):e16253. [doi: 10.2196/16253] [Medline: 32012056]
26. Boggs JM, Quintana LM, Beck A, et al. "Lock to Live" for firearm and medication safety: feasibility and acceptability of a suicide prevention tool in a learning healthcare system. Front Digit Health 2022 Sep 6;4:974153. [doi: 10.3389/fdgth.2022.974153] [Medline: 36148209]
27. Fast facts. Kaiser Permanente. URL: https://about.kaiserpermanente.org/who-we-are/fast-facts [accessed 2021-11-08]
28. Richards JE, Parrish R, Lee A, Bradley K, Caldeiro R. An integrated care approach to identifying and treating the suicidal person in primary care. Psychiatric Times. URL: https://www.psychiatrictimes.com/view/integrated-care-approach-identifying-and-treating-suicidal-person-primary-care [accessed 2020-01-31]
29. Richards JE, Simon GE, Boggs JM, et al. An implementation evaluation of "Zero Suicide" using normalization process theory to support high-quality care for patients at risk of suicide. Implement Res Pract 2021 Jan 1;2:26334895211011769. [doi: 10.1177/26334895211011769] [Medline: 34447940]
30. Richards JE, Kuo E, Stewart C, et al. Self-reported access to firearms among patients receiving care for mental health and substance use. JAMA Health Forum 2021 Aug 6;2(8):e211973. [doi: 10.1001/jamahealthforum.2021.1973] [Medline: 35977197]
31. National Academies of Sciences, Engineering, and Medicine. Health Systems Interventions to Prevent Firearm Injuries and Death: Proceedings of a Workshop: The National Academies Press; 2019. [doi: 10.17226/25354]
32. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
33. Holtrop JS, Rabin BA, Glasgow RE. Qualitative approaches to use of the RE-AIM framework: rationale and methods. BMC Health Serv Res 2018 Mar 13;18(1):177. [doi: 10.1186/s12913-018-2938-8] [Medline: 29534729]
34. Stetler CB, Legro MW, Wallace CM, et al. The role of formative evaluation in implementation research and the QUERI experience. J Gen Intern Med 2006 Feb;21(Suppl 2):S1-S8. [doi: 10.1111/j.1525-1497.2006.00355.x] [Medline: 16637954]
35. Richards JE, Kuo ES, Whiteside U, et al. Patient and clinician perspectives of a standardized question about firearm access to support suicide prevention: a qualitative study. JAMA Health Forum 2022 Nov 4;3(11):e224252. [doi: 10.1001/jamahealthforum.2022.4252] [Medline: 36416815]
36. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
37. Beebe J. Rapid Qualitative Inquiry: A Field Guide to Team-Based Assessment: Rowman & Littlefield; 2014.
38. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
39. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. Cogn Behav Pract 2012 May;19(2):256-264. [doi: 10.1016/j.cbpra.2011.01.001]
40. Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and suicidal behavior. Epidemiol Rev 2008;30(1):133-154. [doi: 10.1093/epirev/mxn002] [Medline: 18653727]
41. Morgan ER, Gomez A, Rowhani-Rahbar A. Firearm ownership, storage practices, and suicide risk factors in Washington State, 2013-2016. Am J Public Health 2018 Jul;108(7):882-888. [doi: 10.2105/AJPH.2018.304403] [Medline: 29771611]
42. Goldstick JE, Carter PM, Cunningham RM. Current epidemiological trends in firearm mortality in the United States. JAMA Psychiatry 2021 Mar 1;78(3):241-242. [doi: 10.1001/jamapsychiatry.2020.2986] [Medline: 32965479]
43. Ingram DD, Franco SJ. 2013 NCHS Urban-Rural Classification Scheme for Counties. Vital Health Stat 2 2014 Apr(166):1-73. [Medline: 24776070]
44. Ketterer AR, Poland S, Ray K, Abuhasira R, Aldeen AZ. Emergency providers' familiarity with firearms: a national survey. Acad Emerg Med 2020 Mar;27(3):185-194. [doi: 10.1111/acem.13849] [Medline: 31957230]
45. Farcy DA, Doria N, Moreno-Walton L, et al. Emergency physician survey on firearm injury prevention: where can we improve? West J Emerg Med 2021 Feb 8;22(2):257-265. [doi: 10.5811/westjem.2020.11.49283] [Medline: 33856309]

46. Wintemute GJ, Betz ME, Ranney ML. You can: physicians, patients, and firearms. Ann Intern Med 2016 Aug 2;165(3):205-213. [doi: 10.7326/M15-2905] [Medline: 27183181]
47. Richards JE, Hohl SD, Segal CD, et al. "What will happen if I say yes?" perspectives on a standardized firearm access question among adults with depressive symptoms. Psychiatr Serv 2021 Aug 1;72(8):898-904. [doi: 10.1176/appi.ps.202000187] [Medline: 33940947]
48. Khazanov GK, Keddem S, Hoskins K, et al. Stakeholder perceptions of lethal means safety counseling: a qualitative systematic review. Front Psychiatry 2022 Oct 20;13:993415. [doi: 10.3389/fpsyt.2022.993415] [Medline: 36339871]
49. Service O, Hallsworth M, Halpern D, et al. EAST: four simple ways to apply behavioural insights. The Behavioural Insights Team. URL: https://www.bi.team/wp-content/uploads/2015/07/BIT-Publication-EAST_FA_WEB.pdf [accessed 2023-10-12]
50. Nordgren L, Schonthal D. The Human Element: Overcoming the Resistance That Awaits New Ideas: Wiley; 2021.
51. Zero Suicide Institute. Education Development Center. URL: https://solutions.edc.org/solutions/zero-suicide-institute [accessed 2023-10-12]
52. U.S. Department of Health and Human Services (HHS) Office of the Surgeon General and National Action Alliance for Suicide Prevention. 2012 National Strategy for Suicide Prevention: Goals and Objectives for Action: HHS; 2012.
53. Layman DM, Kammer J, Leckman-Westin E, et al. The relationship between suicidal behaviors and Zero Suicide organizational best practices in outpatient mental health clinics. Psychiatr Serv 2021 Oct 1;72(10):1118-1125. [doi: 10.1176/appi.ps.202000525] [Medline: 33730886]
54. Lethal means safety counseling. US Department of Veterans Affairs. URL: https://www.mirecc.va.gov/visn19/lethalmeanssafety/counseling/ [accessed 2023-10-12]
55. TIP 59: improving cultural competence. Substance Abuse and Mental Health Services Administration. 2015. URL: https://store.samhsa.gov/product/TIP-59-Improving-Cultural-Competence/SMA15-4849 [accessed 2023-10-12]
56. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press; 2001. [doi: 10.17226/10027]
57. Six domains of healthcare quality. Agency for Healthcare Research and Quality. 2015 Feb. URL: https://www.ahrq.gov/talkingquality/measures/six-domains.html [accessed 2024-03-20]
58. Wosik J, Fudim M, Cameron B, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020 Jun 1;27(6):957-962. [doi: 10.1093/jamia/ocaa067] [Medline: 32311034]
59. Simon GE, Stewart CC, Gary MC, Richards JE. Detecting and assessing suicide ideation during the COVID-19 pandemic. Jt Comm J Qual Patient Saf 2021 Jul;47(7):452-457. [doi: 10.1016/j.jcjq.2021.04.002] [Medline: 33994334]
60. Asarnow JR, Zullo L, Ernestus SM, et al. "Lock and Protect": development of a digital decision aid to support lethal means counseling in parents of suicidal youth. Front Psychiatry 2021 Oct 6;12:736236. [doi: 10.3389/fpsyt.2021.736236] [Medline: 34690841]
61. McCarthy V, Portz J, Fischer SM, et al. A web-based decision aid for caregivers of persons with dementia with firearm access (Safe at Home study): protocol for a randomized controlled trial. JMIR Res Protoc 2023 Jan 31;12:e43702. [doi: 10.2196/43702] [Medline: 36719721]
62. Barry E. Following a two-year decline, suicide rates rose again in 2021. The New York Times. 2023 Feb 11. URL: https://www.nytimes.com/2023/02/11/health/suicide-rates-cdc.html#:~:text=Suicide%20increased%20among%20younger%20Black,reported [accessed 2023-02-13]

Abbreviations

C-SSRS: Columbia Suicide Severity Rating Scale
EHR: electronic health record
KPWA: Kaiser Permanente Washington
L2L: Lock to Live
LICSW: licensed clinical social worker
PHQ-9: 9-item Patient Health Questionnaire
RE-AIM: Reach, Effectiveness, Adoption, Implementation, Maintenance

Edited by J Hefner; submitted 07.04.23; peer-reviewed by G Khazanov, J Sung; revised version received 12.10.23; accepted 27.02.24; published 22.04.24.

Please cite as:
Richards JA, Kuo E, Stewart C, Shulman L, Parrish R, Whiteside U, Boggs JM, Simon GE, Rowhani-Rahbar A, Betz ME
Reducing Firearm Access for Suicide Prevention: Implementation Evaluation of the Web-Based "Lock to Live" Decision Aid in Routine Health Care Encounters
JMIR Med Inform 2024;12:e48007
URL: https://medinform.jmir.org/2024/1/e48007
doi:10.2196/48007

Original Paper

Effect of Performance-Based Nonfinancial Incentives on Data Quality in Individual Medical Records of Institutional Births: Quasi-Experimental Study

Biniam Kefiyalew Taye[1,2], MSc; Lemma Derseh Gezie[3], PhD; Asmamaw Atnafu[4], PhD; Shegaw Anagaw Mengiste[5], Prof Dr; Jens Kaasbøll[6], Prof Dr; Monika Knudsen Gullslett[7], Prof Dr; Binyam Tilahun[1], PhD

[1]Department of Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

[2]Ministry of Health, The Federal Democratic Republic of Ethiopia, Addis Ababa, Ethiopia

[3]Department of Epidemiology and Biostatistics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

[4]Department of Health System and Policy, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

[5]Management Information Systems, University of South-Eastern Norway, Drammen, Norway

[6]Department of Informatics, University of Oslo, Oslo, Norway

[7]Faculty of Health & Social Sciences, Science Center Health & Technology, University of South-Eastern Norway, Notodden, Norway

Corresponding Author:
Biniam Kefiyalew Taye, MSc
Ministry of Health, The Federal Democratic Republic of Ethiopia
Zambia street
Addis Ababa
Ethiopia
Phone: 251 910055867
Email: bini.bhi2013@gmail.com

Abstract

Background: Despite the potential of routine health information systems in tackling persistent maternal deaths stemming from poor service quality at health facilities during and around childbirth, research has demonstrated their suboptimal performance, evident from the incomplete and inaccurate data unfit for practical use. There is a consensus that nonfinancial incentives can enhance health care providers' commitment toward achieving the desired health care quality. However, there is limited evidence regarding the effectiveness of nonfinancial incentives in improving the data quality of institutional birth services in Ethiopia.

Objective: This study aimed to evaluate the effect of performance-based nonfinancial incentives on the completeness and consistency of data in the individual medical records of women who availed institutional birth services in northwest Ethiopia.

Methods: We used a quasi-experimental design with a comparator group in the pre-post period, using a sample of 1969 women's medical records. The study was conducted in the "Wegera" and "Tach-armacheho" districts, which served as the intervention and comparator districts, respectively. The intervention comprised a multicomponent nonfinancial incentive, including smartphones, flash disks, power banks, certificates, and scholarships. Personal records of women who gave birth within 6 months before (April to September 2020) and after (February to July 2021) the intervention were included. Three distinct women's birth records were examined: the integrated card, integrated individual folder, and delivery register. The completeness of the data was determined by examining the presence of data elements, whereas the consistency check involved evaluating the agreement of data elements among women's birth records. The average treatment effect on the treated (ATET), with 95% CIs, was computed using a difference-in-differences model.

Results: In the intervention district, data completeness in women's personal records was nearly 4 times higher (ATET 3.8, 95% CI 2.2-5.5; *P*=.02), and consistency was approximately 12 times more likely (ATET 11.6, 95% CI 4.18-19; *P*=.03) than in the comparator district.

Conclusions: This study indicates that performance-based nonfinancial incentives enhance data quality in the personal records of institutional births. Health care planners can adapt these incentives to improve the data quality of comparable medical records, particularly pregnancy-related data within health care facilities. Future research is needed to assess the effectiveness of nonfinancial incentives across diverse contexts to support successful scale-up.

(JMIR Med Inform 2024;12:e54278) doi:10.2196/54278

KEYWORDS

individual medical records; data quality; completeness; consistency; nonfinancial incentives; institutional birth; health care quality; quasi-experimental design; Ethiopia

Introduction

Background

Maternal mortality, a pressing global health concern, is particularly prevalent in low- and middle-income countries [1-5]. The existing research attributes the persistence of maternal deaths largely to inadequate health care quality during labor, delivery, and immediate postpartum care in health facilities [6,7]. Almost every low- and middle-income country implements the Routine Health Information System (RHIS) to address this challenge [8-10]. The RHIS has gained prominence for its practical roles in improving the quality of services, including (1) facilitating evidence-based action by enabling the early detection of pregnancy-related complications, (2) serving as a repository for clients' data to ensure the continuity of pregnancy-related care, and (3) functioning as the primary data source essential for health monitoring and evaluation at all levels of the public health system [11-16]. Despite its potential, the performance of RHIS remains suboptimal, primarily because of incomplete and inaccurate data, hindering its effective use by decision makers [17-20].

In Ethiopia, the introduction of the RHIS dates back to 2008 [21,22]. Ongoing efforts are in place to enhance the data quality of the RHIS in Ethiopia through interventions such as the Performance Monitoring Team (PMT), lot quality assurance sampling (LQAS), and the Capacity Building and Mentorship Program (CBMP) [23-25]. However, despite these efforts, the quality of RHIS data still lags in Ethiopia [15,26]. This challenge is pertinent to institutional birth, as shown by some previous studies in Ethiopia. For instance, a study [27] reported a completeness rate of only 18.4%. Another study from Ethiopia found that 66% of health facilities managed to produce accurate data within an acceptable range [28]. Furthermore, comparing the data from health facilities with external sources such as the Ethiopian Demographic Health Survey reveals concerns regarding data quality in Ethiopian RHIS [26].

Incentives and Its Impact on Health Care Quality

Previous studies have shown that offering incentives for personnel responsible for data collection and management can improve data quality in the RHIS. According to some studies, incentives are essential for addressing the negative attitudes and values that undermine data quality within the RHIS, which are primary challenges to achieving desired quality of RHIS data [29].

The effectiveness of incentives in health care is grounded in theoretical and empirical evidence. Theories like the theory of planned behavior emphasize the connection between motivation and improved health care quality [30]. Some studies demonstrated that incentive-based interventions can predict up to 48% of desired health care behavior [31,32].

Despite the growing interest in using incentives in health care [32-38], determining the most effective approach remains a research priority. Incentives can be financial [39,40] or nonfinancial. Financial incentives have been extensively studied globally [33,37-39,41-43], but there is limited evidence supporting their consistent impact on health care quality [40]. Some studies have even cited the counterproductive effects of financial incentives on health care [44].

Compared to financial incentives, the impact of nonfinancial incentives on health care quality has been minimally studied [45]. Nonfinancial incentive schemes offer noncash rewards or benefits to motivate recipients using approaches that involve recognition through public profiling or reporting; career advancement opportunities; providing certificates to top performers; and ensuring improved working conditions, such as vacations, grading systems, and packaging interventions with in-kind items [46-53]. Previous studies have demonstrated the effectiveness of nonfinancial incentives in enhancing the quality of health services. For instance, nonfinancial incentive schemes in the United States, India, El Salvador, and Tanzania have been reported to enhance the performance of health care providers, including enhanced root cause data analysis of medical errors, expanded community outreach services, better maternal and child care services, and higher quality health care consultations [50,51,54,55].

Objectives

This study aimed to evaluate the effect of performance-based nonfinancial incentives (PBNI) on enhancing the quality of institutional birth data, measured by the completeness and consistency of data within women's individual medical records (IMRs).

Methods

Study Design and Period

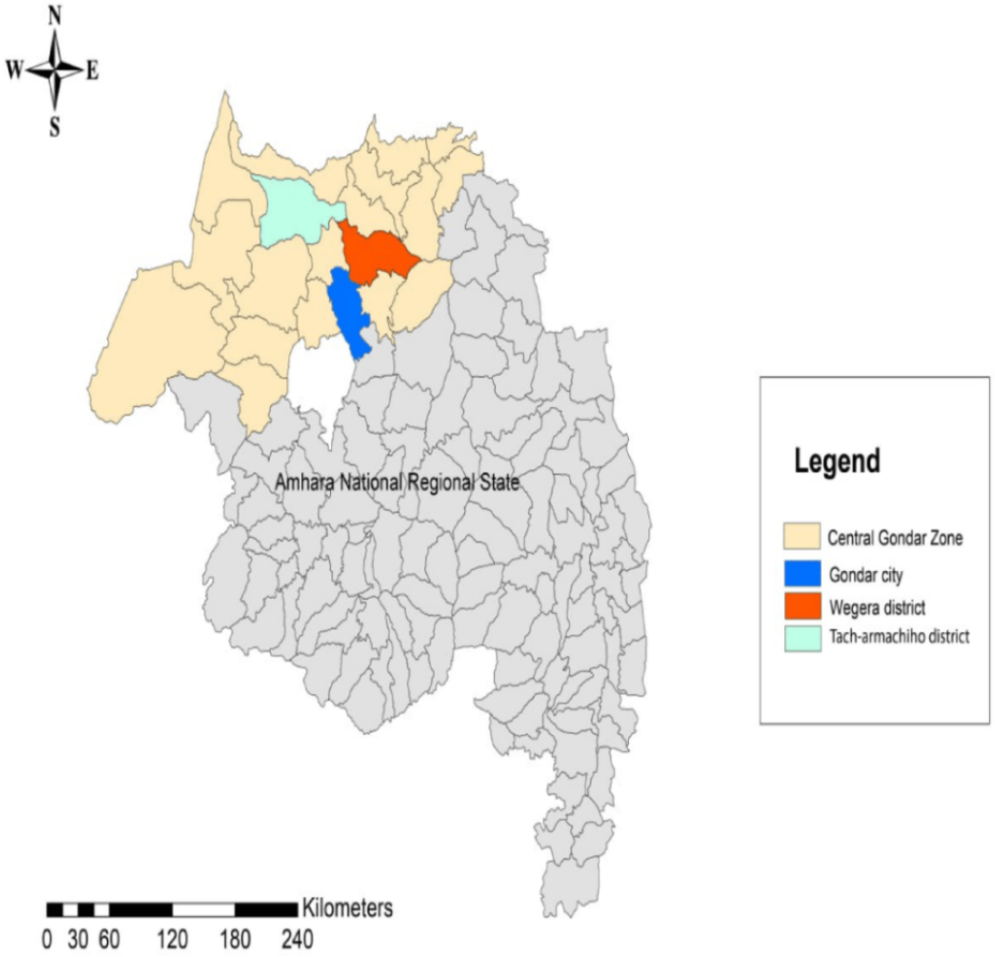
This study used a quasi-experimental design with a comparator group in the pre-post period to examine the effect of PBNI on the data quality of IMRs of institutional births. A cross-sectional survey within an institutional setting was used to review institutional birth–related medical records. The study included the IMRs of women who gave birth within 6 months before (April to September 2020) and after (February to July 2021) the intervention.

Study Setting

This study was conducted in the Amhara National Regional State, specifically in the Central Gondar Zone districts of "Wegera" and "Tach-armachiho" in northwest Ethiopia (Figure 1). According to the Ethiopian Central Statistical Agency projection, the total population in both districts was approximately 398,350 in 2021, with an estimated 93,214 women in the reproductive age group (15-49 years) [56,57].

The districts hosted 2 primary hospitals, 13 health centers, and 72 health posts. The intervention district had 678 health care providers and 215 support staff, whereas the comparator district had 202 health care providers and 141 support staff [58]. These districts were chosen for their involvement in the CBMP, a collaborative initiative between the Ethiopian Ministry of Health and the University of Gondar [24,59,60]. As part of the CBMP, the University of Gondar provides ongoing technical support, including training, supervision, and mentorship to health facilities in both districts, to enhance data quality and information use [59]. This study evaluated the effect of PBNI in the "Wegera" district, with the "Tach-armachiho" district serving as a comparator.

Figure 1. Map of the study area.



Intervention

Intervention Aim and Design

The PBNI intervention was implemented in the "Wegera" district between October 2020 and July 2021 to improve the data quality and use of the RHIS. PBNI was designed as a package with multiple nonfinancial incentive components. Incentives were offered across 3 levels: health facilities, departments, and individual health care workers. Individual health workers were offered nonfinancial incentives, including smartphones, flash disks, power banks, and scholarships. Desktop computers were offered at the department and health facility levels. Health workers, departments, and health facilities that earned nonfinancial incentives in each round were awarded certificates of recognition [61].

Target Areas for PBNI

The study included 6 (75%) out of the 8 health centers in "Wegera." The departments involved in the PBNI program include Maternal and Child Health, Outpatient Department, under-5 Outpatient Department, the Health Management Information System (HMIS), and the Medical Record Unit (MRU). Health workers who participated in the PBNI program included various experts, such as medical record personnel, health IT (HIT) personnel, health officers, midwives, nurses, and personnel involved in laboratories and pharmacies.

Awardees Selection Procedures

The selection of the best performers was conducted through 2 approaches: a subjective and an objective approach.

Subjective Approach

The subjective approach involved requesting management authorities in the intervention district to nominate the best-performing employees. The subjective approach was chosen owing to practical constraints, as the quantitative measurement of all health workers' performance was infeasible owing to limited resources. Accordingly, the number of potential

awardees was reduced to a manageable level, allowing us to concentrate on objectively evaluating the candidates.

The subjective approach was conducted in 2 phases. In the initial phase, health office department managers in the intervention district nominated 12 individuals, selecting 2 from each of the 6 participating health centers. The second phase mirrored the first phase, except that the selection process took place at the level of each health center, where the heads of each health center were tasked with nominating the best performers. With 2 nominees selected by the heads of each health center, another 12 individuals were identified. Consequently, 24 individuals were identified using a subjective approach.

Objective Approach

Previous research indicates that effective health care incentives depend on rewarding specific performance [62,63]. In this study, for the purpose of incentivizing 3 entities—health centers, departments, and individual health workers—we used a flexible approach that used objective measures to identify the best performers. For health centers, 14 quantitative performance measures, each of which was established with specific targets and points to be earned, were used (Textbox 1). The allocation of point values and performance targets took priority for the RHIS activities, as defined by the Ethiopian Ministry of Health [22].

The performance of departments and the 24 individuals selected during the subjective phases was objectively evaluated, aligning the 14 quantitative performance measures with their relevant roles and job descriptions. Further details on the performance measures used are described in prior studies [58,61,64,65].

Textbox 1. The performance indicators used to determine the awardees of nonfinancial incentives, northwest Ethiopia, 2021.

Indicators and points (total points: 90)

1. Source documents completeness rate: 10
2. Report timeliness: 5
3. Lot quality assurance sampling performance: 6
4. Data consistency among registers and reports: 12
5. Health centers established by Performance Monitoring Team: 8
6. Action plan implemented regularly: 10
7. Conducted internal supervision: 5
8. Gaps identified and prioritized by Performance Monitoring Team: 5
9. Conducted root cause analysis: 5
10. Feedback provided for case teams by health centers: 4
11. Number of feedback entries provided to health posts by health centers 5:
12. Information display status: 5
13. Report completeness: 5
14. Consistency among medical records: 5

The Awarding Processes

Initially, the team from the University of Gondar visited the health office department and health centers in the intervention district to communicate the commencement of the program. During this announcement phase, a banner illustrating the nonfinancial rewards was displayed within the compounds of the health facilities. Nonfinancial incentives were offered to the recipients through 3 ceremonial award programs that took place bimonthly. The attendees of the PBNI ceremonial award include representatives from the University of Gondar, Federal Ministry of Health, Amhara Regional Health Bureau, Central Gondar Zone, and “Wegera” District Health Office departments. These representatives comprised health experts and administrative personnel. Officials from the Federal Ministry of Health and University of Gondar rewarded the top-3 individuals, departments, and health centers. Certificates of recognition were presented to awardees during these bimonthly forums. Ceremonial events were also accompanied by presentations detailing the performance measurement and award selection procedures by the professionals from the University of Gondar.

Study Participants

Overview

The participants in this study were women who had given birth in the health centers at the study sites. The IMR sets of these individuals were examined. Thus, for each woman, there would be a set of 3 types of records: delivery register, integrated individual folder (IIF), and integrated card. These 3 sets of IMRs were combined to form a single study cohort. The 3 types of IMRs evaluated in this study, designed to record institutional birth data following the guidelines established in Ethiopia [66], are described in the following sections.

Integrated Card

The integrated card captures data on pregnant women throughout antenatal, labor, delivery, and postnatal care. It facilitates the recording of medical histories, physical examination results,

and other clinical data for both women and newborns, allowing health care providers to complete it upon birth.

Delivery Register

The delivery register is a serial-long register designed to contain a list of all women who give birth at the facility, with data abstracted from the integrated card.

Women's IIF

The IIF is designed to consolidate the entirety of a woman's personal records, including the integrated card. It ensures the convenient access to comprehensive medical data, with the front section containing personal identification data filled out during registration and the inner part featuring a summary sheet completed by service providers after each visit.

Sample Size Calculation

The required sample size for this study was calculated using StatCalc (Epi Info version 7.0; Centers for Disease Control and Prevention), incorporating assumptions to detect differences in completeness and consistency rates between the intervention and comparator groups. The assumptions included 80.3% completeness and 29.5% consistency from a prior unpublished pilot study (Taye, BK, unpublished data, September 2021), a 1:1 ratio of the intervention to the comparator group, a 5% anticipated change in the intervention group [55], 80% power, 95% CI, and a 10% nonresponse rate. Separate calculations yielded approximately 1969 participants (985 in each group) for completeness and approximately 3007 (1504 in each group) for consistency. From the 2 computed samples, we chose 1969 participants, considering the available resources for feasibility.

Inclusion and Exclusion Criteria

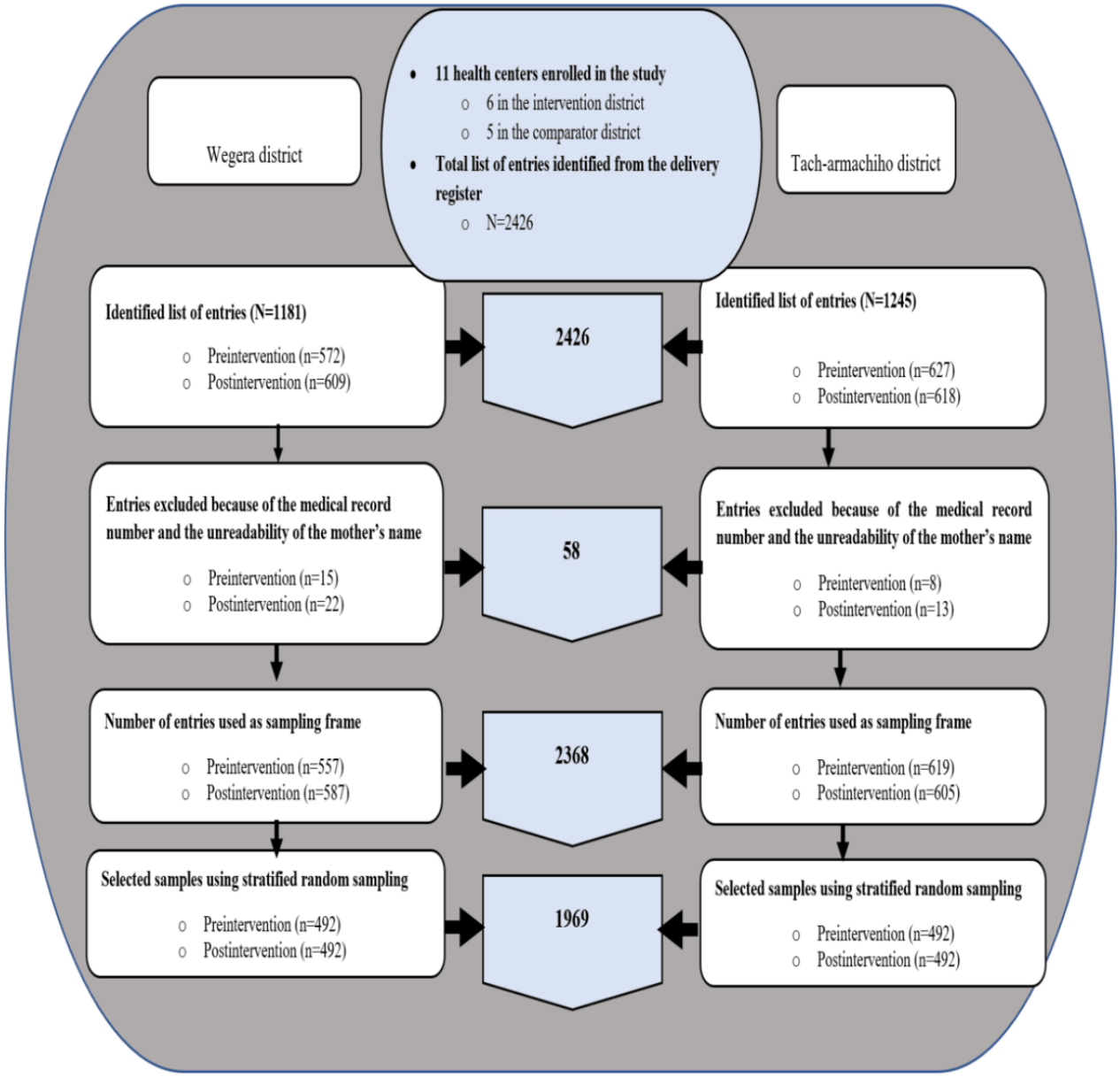
A list of women who gave birth during the study period was used as a sampling frame to select the study sample. Three distinct IMRs in combination (delivery register, IIF, and integrated card) were necessary for each participant. The inclusion criteria in this study mandated that women's "medical record numbers" (MRNs) and the mother's full name be legibly recorded on the delivery register. The readability of these 2 variables was necessary to match and retrieve women's IIF data from the MRU.

Sampling Procedure

A total of 11 health centers were included in the pre- and postintervention periods, of which 6 (55%) were in the intervention district and 5 (45%) were in the comparator district. In total, 2426 women's entries were identified from the delivery register across the baseline and end line periods.

Of the 2426 entries, we excluded 58 (2%) due to unreadability of the MRNs and the full names of the mothers, resulting in 2368 (98%) entries of women on the delivery register. The list of 2368 women in the delivery register was used as a sampling frame to select 1969 (83.1%) samples through stratified random sampling. Subsequently, the women's IIF was retrieved from the MRU (Figure 2).

Figure 2. A diagram illustrating the process of study sample selection.



Variables and Measurements

Outcome Variables

Completeness

Completeness refers to the presence of data elements in the targeted data set [67]. Within the RHIS, assessing the completeness of data elements mandates the presence of its corresponding services or medical procedures in health facilities [68]. Accordingly, this research evaluated the completeness of the 33 data elements where the guideline necessitates their recording [66]. Of the 33 data elements, 16 (48%) were from the integrated card (mother's name, gravida, para, MRN, date and time of admission, ruptured membranes, date and time of delivery, mode of delivery, placenta details, newborn's sex, newborn outcome, single or multiple births, term or preterm status, and the name and signature of providers), 7 (21%) were from the individual folder (facility name, registration date, client's sex, mother's age, date of delivery, Department-provided service, and serial number), and 10 (30%) were from the delivery register (serial number, age, address, date and time of delivery, newborn outcome, mode of delivery, maternal status, newborn's sex, and the name and signature of providers).

Consistency

According to previous research [69], consistency is a measure of data accuracy (the extent to which data elements accurately represent the true numbers), commonly assessed in RHIS through data verification (agreement of data among data sources). This research crosschecked the agreement between the data elements in the source documents to assess consistency. The delivery register served as the gold standard, allowing a comparison with 12 data elements abstracted from the IIF and integrated card, including the serial number, MRN, mother's name, delivery date and time, mode of delivery, newborn's sex, provider's name and signature, Apgar score, newborn's weight, mother's HIV test acceptance, and mother's HIV test results.

In this study, data collectors judged the completeness and consistency of data elements, and the decision adhered to the guidelines [66]. For completeness, data collectors observed the presence of data elements in personal records and declared and coded each data element as 1 for "Yes" if recorded and 0 for "No" if unrecorded. Regarding consistency, data collectors verified elements from comparable records for agreement, coding them as 1 for "Yes" if consistently recorded and 0 for "No" otherwise.

Independent Variables

The characteristics of health facilities supposed to be associated with the data quality of institutional birth were included in the analysis, drawing from pertinent literature and guidelines in the field [22,66,70-72]. These independent variables were the presence of HIT personnel, availability of data recording tools, availability of trained providers, supportive supervision from higher officials, existence of PMT, PMT per membership standard, monthly PMT meetings, monthly conducted LQAS, conducted root cause analysis (RCA) on the identified gaps, internal supervision, and availability of HMIS guidelines. Details of the measurement and data management for outcomes and independent variables have been outlined in a previous study [73].

Data Collection

In total, 11 data collectors, HIT personnel, and related health sciences graduates were recruited for the data collection. The data collectors received a 3-day training that covered the study's objectives, methodology, and ethical considerations. The principal investigator and the other 2 supervisors supervised the data collection process.

Statistical Analysis

Data were entered into the EpiData (version 3.1; Epi Data Association) and exported into Stata (version17.0; StataCorp) for analysis.

Descriptive Statistics

The frequency distribution of women's IMRs is presented based on the characteristics of the health facilities. Pearson chi-square tests were computed to assess the comparability of the data on the study participants' baseline characteristics between the intervention and comparator groups.

The completeness and consistency proportions were computed for specific data elements, with the average rates calculated for the study participants. The completeness proportion of data elements was determined by dividing the number of participants with recorded data elements by the total number of study participants. Meanwhile, the mean completeness proportion was calculated by dividing the number of data elements recorded per participant by the total expected number of data elements. Likewise, the consistency proportion of data elements was determined by dividing the number of participants with consistently recorded data elements by the total number of study participants, and the mean consistency proportion was calculated by dividing the number of data elements consistently recorded per participant by the total expected number of data elements.

Changes in completeness and consistency proportions between the study groups were compared by computing absolute differences and their corresponding 95% CIs, using a 2-sample proportion test.

Difference-in-Differences Analysis

This study used difference-in-differences (DID) to estimate the average effect of PBNI on data completeness and consistency [74,75]. The average treatment effect on the treated (ATET) was computed using the DID model. The DID model used can be described as follows:

(1)

where γ_i represents the outcome (either completeness or consistency) for each study participant. The variable 1(g = intervention) is an indicator for the study group, taking the value of "1" if the participant belongs to the intervention group and "0" otherwise. Similarly, the variable 1(t = post) is a binary indicator, taking the value of "1" for the participants sampled during the end line period and 0 otherwise. $Z_{igt}\beta$ represents the independent variables, that is, the characteristics of health facilities included in the DID model. D represents the intervention in this study (PBNI). σ represents the coefficient of average treatment effect on the intervention group (ATET), providing estimates of the average effect of the PBNI on the outcome variables, and  are the residual errors.

Ethical Considerations

Ethics clearance was obtained from the University of Gondar's Institutional Ethical Review Board (RNO: V/P/RCS/05/861-2021). As this study used medical records rather than human participants, obtaining informed consent from the participants was not feasible.

Results

Description of Women's IMRs by Health Facilities' Baseline Characteristics

In this study, 91.92% (1810/1969) of the samples met the analysis criteria for the coexistence of all 3 distinct IMRs. Of the analysis sample, 49.78% (901/1810) were sampled from the intervention district and 50.22% (909/1810) from the comparator district, whereas 49.56% (897/1810) were enrolled at baseline and 50.44% (913/1810) were enrolled at end line.

Among the baseline samples involved in the analysis, 51% (458/897) were enrolled from the intervention district and 48.9% (439/897) were from the comparator district. When comparing study participants according to baseline health facility characteristics, the distribution was identical (897/897, 100%) across the following variables: the presence of HIT personnel, the existence of PMT, the PMT established per membership standard, monthly PMT meetings in the last 6 months, and monthly LQAS in the previous 6 months.

Most of the IMRs (425/458, 92.8%) in the intervention district were from health centers where the data recording tools were fully available, compared with 12.3% (54/439) in the comparator

district (*P*<.001). In the intervention district, 69% (316/458) of the IMRs were from health centers where most providers had received training on data quality, compared with 65.6% (288/439; *P*=.28) in the comparator district. In the intervention district, 39.7% (182/458) of the IMRs were from health centers with at least 3 supportive supervision visits by higher officials, compared with 65.6% (289/439; *P*<.001) in the comparator district. Nearly one-third of the IMRs (145/458, 31.6%) in the intervention district were from health centers that conducted at least 2 internal supervisions, compared with 4.8% (21/439; *P*<.001) in the comparator district. Less than three-fourth of the IMRs (303/458, 66.2%) in the intervention district were from health facilities that conducted RCA at least 3 times, compared with 25.2% (111/439; *P*<.001) in the comparator district. In the intervention district, 92.8% (425/458) of the IMRs were from health centers with fully available HMIS guidelines, compared with 91.8% (403/439; *P*=.58) in the comparator district (Table 1).

Table 1. Baseline comparison between the study groups by the characteristics of health facilities, northwest Ethiopia, 2021.

Variables	Intervention (N=458), n (%)	Comparator (N=439), n (%)	*P* value
Presence of health information technology personnel	458 (100)	439 (100)	<.001
Existence of PMT[a]	458 (100)	439 (100)	<.001
The PMT per membership standard	458 (100)	439 (100)	<.001
Monthly PMT meeting	458 (100)	439 (100)	<.001
Monthly conducted lot quality assurance sampling	458 (100)	439 (100)	<.001
Availability of data recording tools			
Fully	425 (92.8)	54 (12.3)	<.001
Partially	33 (7.2)	385 (87.7)	<.001
Availability of trained providers			
Mostly	316 (69)	288 (65.6)	.28
Partially	142 (31)	151 (34.4)	.28
Supportive supervisions from higher officials			
<3 times	276 (60.2)	151 (34.4)	<.001
At least 3 times	182 (39.7)	289 (65.6)	<.001
Conducted root cause analysis on the identified gap			
Yes	303 (66.2)	111 (25.2)	<.001
No	155 (33.8)	328 (74.7)	<.001
Internal supervision			
At least 2 times	145 (31.6)	21 (4.8)	<.001
<2 times	313 (68.3)	418 (95.2)	<.001
Availability of the Health Management Information System guidelines			
Fully available	425 (92.8)	403 (91.8)	.58
Partially available	33 (7.2)	36 (8.2)	.58

[a]PMT: Performance Monitoring Team.

Specific Data Elements Completeness Across Study Groups

Table 2 compares the intervention and comparator districts regarding the completeness of specific data elements across the 3 IMRs. Concerning the data elements from the integrated card, the "Name of the mother" showed 95.6% (861/901) completeness in the intervention district compared with 92.6% (842/909; *P*=.004) in the comparator district. The completeness proportion of "Gravida" was 91.7% (826/901) in the intervention district, compared with 90.4% (822/909; *P*=.18) in the comparator district. The completeness proportion of "MRN" was 92.7% (835/901) in the intervention district, compared with 84.9% (771/909; *P*<.001) in the comparator district. The completeness proportion of "Time of Admission" was 92.5% (833/901) in the intervention district, compared with 88.7% (806/909; *P*=.003) in the comparator district. The completeness proportion of "time of delivery" was 94.9% (855/901) in the intervention district, compared with 89.6% (814/909; *P*<.001) in the comparator district. The completeness proportion of "Name and signature of providers" was 90.1% (812/901) in the intervention district, compared with 86.8% (789/909; *P*=.01) in the comparator district.

In the IIF, the "date of delivery" completeness proportion was 39.3% (354/901) in the intervention district, compared with 29.4% (268/909; *P*<.001) in the comparator district. The completeness proportion of "Date of registration" was 99.8%

(900/901) in the intervention district, compared with 93.4% (849/909; *P*<.001) in the comparator district. The completeness proportion of "Department-provided service" was 37.1% (334/901) in the intervention district, compared with 29.4% (267/909; *P*<.001) in the comparator district. The completeness proportion of "Serial Number" was 35.2% (318/901) in the intervention district, compared with 31% (282/909; *P*=.03) in the comparator district.

Regarding the data elements in the delivery register, the "Serial Number" and "Age" were found to be recorded for all study participants across the intervention and comparator districts. In the intervention district, the "time of delivery" showed 92.7% (835/901) completeness, compared with the comparator district, which showed 62.6% (571/909; *P*<.001) completeness. In the intervention district, the "date of delivery" showed 99.7% (899/901) completeness, compared with the comparator district, which showed 95.3% (867/909; *P*<.001) completeness. The completeness proportion of "sex of newborn" was 99.4% (896/901) in the intervention district, compared with 99% (900/909; *P*=.14) in the comparator district. The "Name and signature of providers" completeness proportion was 78.3% (706/901) in the intervention district, compared with 80.9% (736/909; *P*=.92) in the comparator district.

Table 2. Specific data elements' completeness in individual medical records of institutional births across intervention and comparator districts, northwest Ethiopia, 2021.

Completeness	Intervention (N=901), n (%)	Comparator (N=909), n (%)	Difference[a] (95% CI)	*P* value[b]
Integrated card				
Name of the mother	861 (95.6)	842 (92.6)	2.9 (0.76 to 5)	.004
Gravida	826 (91.7)	822 (90.4)	1.2 (−1.48 to 3.8)	.18
Para[c]	858 (95.2)	820 (90.2)	5 (2.6 to 7.49)	<.001
Medical record number	835 (92.7)	772 (84.9)	7.7 (4.8 to 10.6)	<.001
Date of admission	841 (93.3)	822 (90.4)	2.9 (0.4 to 5.4)	.01
Time of admission	833 (92.5)	806 (88.7)	3.8 (1 to 6.4)	.003
Ruptured membranes	664 (73.7)	760 (83.6)	9.9 (−13.6 to −6.26)	>.99
Date of delivery	861 (95.6)	843 (92.7)	2.8 (0.6 to 4.9)	.005
Time of delivery	855 (94.9)	814 (89.6)	5.3 (2.8 to 7.89)	<.001
Mode of delivery	834 (92.6)	806 (88.7)	3.8 (1.2 to 6.6)	.002
Placenta	842 (93.5)	804 (88.5)	5 (2.4 to 7.6)	<.001
Sex of the newborn	851 (94.5)	813 (89.4)	5 (2.5 to 7.5)	<.001
Newborn outcome	850 (94.3)	822 (90.4)	3.9 (1.5 to 6.3)	<.001
Single or multiple	756 (83.9)	802 (88.2)	4.3 (−7.5 to −1.1)	>.99
Term or preterm	786 (87.2)	801 (88.1)	0.8 (−3.9 to 2.1)	.72
Name and signature	812 (90.1)	789 (86.8)	3.3 (0.38 to 6.26)	.01
Integrated individual folder				
Name of the facility	887 (98.4)	807 (88.7)	9.6 (7.4 to 11.8)	<.001
Date of registration	900 (99.8)	849 (93.3)	6.4 (4.8 to 8.1)	<.001
Sex of the client	899 (99.7)	858 (94.3)	5.4 (3.8 to 6.9)	<.001
Age of the mother	843 (93.5)	860 (94.6)	1 (−3.2 to 1.1)	.83
Date of delivery	354 (39.2)	268 (29.4)	9.8 (5.4 to 14.2)	<.001
Department-provided service	334 (37)	267 (29.3)	7.6 (3.4 to 12)	<.001
Serial number	318 (35.2)	282 (31)	4.3 (−0.01 to 8.6)	.03
Delivery register				
Serial number	901 (100)	909 (100)	—[d]	<.001
Age	901 (100)	909 (100)	—	<.001
Address	900 (99.8)	900 (99)	0.8 (0.2 to 1.6)	.005
Date of delivery	899 (99.7)	867 (95.3)	4.4 (2.9 to 5.8)	<.001
Time of delivery	835 (92.6)	571 (62.8)	29.8 (26.2 to 33.4)	<.001
Newborn outcome	897 (99.5)	905 (99.5)	—	<.001
Mode of delivery	898 (99.6)	899 (98.8)	0.8 (−0.01 to 1.5)	.05
Maternal status	897 (99.6)	909 (100)	0.4 (−0.8 to –0.01)	.98
Sex of the newborn	896 (99.4)	900 (99)	0.4 (−0.37 to 1.24)	.14
Name and signature	706 (78.3)	736 (80.9)	2.6 (−6.31 to 1.09)	.92

[a]The absolute difference is calculated by subtracting the completeness proportion of the comparator group from that of the intervention group.

[b]*P* value based on 2 independent sample proportion tests.

[c]A number of times a woman has given birth to a viable child.

[d]No difference among intervention and comparator group.

Average Data Completeness Across the Pre- and Postintervention Periods

For the integrated card, the average completeness increased from 86.2% (95% CI 83.9%-88.57%) at the baseline to 96.6% (95% CI 96%-97.1%) at the end line in the intervention district; however, in the comparator district, it showed a decrease from 91.1% (95% CI 89.4%-92.7%) at the baseline to 87% (95% CI 84.2%-89.7%) at the end line.

The average completeness of the IIF increased from 58.9% (95% CI 57.6%-60.2%) at the baseline to 85.3% (95% CI 83.6%-87%) at the end line in the intervention district, whereas the comparator district showed a change from 63.5% (95% CI 61.5%-65.4%) to 68.1% (95% CI 65.6%-70.6%).

In the intervention district, the mean completeness proportion of the delivery register increased from 94.6% (95% CI 93.9%-95.2%) at the baseline to 99.3% (95% CI 99%-99.5%) at the end line. In comparison, the comparator district showed a change from 93.5% (95% CI 92.9%-94%) to 93.6% (95% CI 92.9%-94.3%).

In the intervention district, the average data completeness proportion across the 3 individual IMRs was 82.9% (95% CI 81.88%-4.1%) at the baseline, and it increased to 95% (95% CI 94.6%-95.5%) at the end line. In the comparator district, the average data completeness proportion across the 3 IMRs was 86% (95% CI 84.96%-86.97%) at the baseline but decreased to 84.97% (95% CI 83.28%-86.66%) at the end line (Multimedia Appendix 1).

Effect of PBNI on Data Completeness

In the intervention district, the "integrated card" resulted in an average 2.6 percentage-point increase in completeness compared with the comparator district (ATET 2.67, 95% CI 0.7-4.4; *P*=.04). On average, the intervention district showed a 3.8 percentage-point increase in the completeness of the delivery register compared with the comparator district (ATET 3.8, 95% CI 2.9-4.8; *P*=.01). The intervention district showed a 6.8 percentage-point increase in the average completeness of the IIFs compared with the comparator district (ATET 6.8, 95% CI 4.55-9; *P*=.02). Overall, on average, the intervention district showed a 3.8 times higher chance of complete recording of IMRs compared with the comparator district (ATET 3.8, 95% CI 2.2-5.5; *P*=.02; Table 3).

Table 3. Effect of performance-based nonfinancial incentives on the data completeness in individual medical records of institutional births, northwest Ethiopia, 2021.

Completeness	Intervention, mean (SD)	Comparator, mean (SD)	Intervention effect, ATET[a] (95% CI)[b]	*P* value
Integrated card	91.3 (18.8)	88.9 (24.9)	2.6 (0.7-4.4)	.04
Integrated individual folder	71.9 (21.1)	65.8 (24.8)	6.8 (4.6-9)	.02
Delivery register	96.8 (5.8)	93.6 (6.9)	3.8 (2.9-4.8)	.01
Overall	88.8 (11.2)	85.4 (15.3)	3.8 (2.2-5.5)	.02

[a]ATET: average treatment effect on the treated.

[b]ATET estimates adjusted for covariates (presence of health information technology personnel, availability of data recording tools, availability of trained providers, supportive supervision from higher officials, existence of the Performance Monitoring Team [PMT], PMT per membership standard, monthly PMT meeting, monthly conducted lot quality assurance sampling, conducted root cause analysis, internal supervision, and availability of Health Management Information System guidelines).

Consistency of Specific Data Elements Across Study Groups

Regarding the delivery register and IIF, the "date of delivery" showed a consistency proportion of 82.3% (742/901) in the intervention district, compared with 58.7% (534/909; *P*<.001) in the comparator district. The "Serial Number" showed a consistency proportion of 42.1% (380/901) in the intervention district, compared with 45.5% (414/909; *P*=.92) in the comparator district.

Comparing the delivery register and integrated card, the "MRN" exhibited a consistency proportion of 87% (784/901) in the intervention district, compared with 74.9% (681/909; *P*<.001) in the comparator district. The "time of delivery" showed a consistency proportion of 88.2% (795/901) in the intervention district, compared with 56.7% (516/909; *P*<.001) in the comparator district. The "Name and signature of providers" showed a consistency proportion of 89.5% (807/901) in the intervention district, compared with 77.7% (707/909; *P*<.001) in the comparator district. The "newborn weight" had a consistency proportion of 85.7% (773/901) in the intervention district, compared with 82.1% (746/909; *P*=.01) in the comparator district. The "Women's HIV test accepted" showed a consistency proportion of 39.9% (360/901) in the intervention district and 40.5% (368/909; *P*=.59) in the comparator district (Table 4).

Table 4. Consistency of specific data elements across the intervention and comparator districts, northwest Ethiopia, 2021.

Consistency	Intervention (N=901), n (%)	Comparator (N=909), n (%)	Difference[a] (95% CI)	*P* value[b]
Delivery register vs integrated individual folder				
Date of delivery	742 (82.3)	534 (58.7)	23.6 (19.55 to 27.66)	<.001
Serial number	380 (42.1)	414 (45.5)	3.36 (−7.93 to 1.2)	.92
Delivery register vs integrated card				
Medical record number	784 (87)	681 (74.9)	12.09 (8.52 to 15.66)	<.001
Name of the mother	828 (91.8)	709 (77.9)	13.90 (10.67 to 17.12)	<.001
Date of delivery	859 (95.3)	803 (88.3)	6.99 (4.50 to 9.49)	<.001
Time of delivery	795 (88.2)	516 (56.7)	31.46 (27.62 to 35.31)	<.001
Mode of delivery	831 (92.2)	796 (87.5)	4.66 (1.89 to 7.42)	<.001
Sex of the newborn	846 (93.9)	806 (88.6)	5.22 (2.64 to 7.81)	<.001
Name and signature	807 (89.5)	707 (77.7)	11.78 (8.42 to 15.14)	<.001
Apgar score	781 (86.6)	746 (82.1)	4.61 (1.27 to 7.95)	.003
Newborn weight	773 (85.7)	746 (82.1)	3.72 (0.34 to 7.1)	.01
Women’s HIV test accepted	360 (39.9)	368 (40.4)	0.52 (−5.04 to 3.98)	.59
Women’s HIV test result	615 (68.2)	381 (41.9)	26.34 (21.92 to 30.76)	<.001

[a]The absolute difference is calculated by subtracting the consistency proportion of the comparator group from that of the intervention group.

[b]*P* value based on 2 independent sample proportion tests.

Pre- and Postintervention Changes in Average Data Consistency

In the intervention district, the average consistency proportion increased from 71.6% (95% CI 69.6%-73.6%) to 89.2% (95% CI 88.2%-90.2%) after the intervention. In the comparator district, it increased from 68% (95% CI 66.2%-69.8%) to 70.8% (95% CI 67.9%-73.6%) post intervention. Overall, the average consistency proportion increased from 69.8% (95% CI 68.5%-71.2%) to 79.6% (95% CI 78%-81.3%) after the intervention.

Effect of PBNI on Data Consistency

On average, the intervention district showed an 11.2 percentage-point increase in the consistency of data among the delivery register and IIF compared with the comparator district (ATET 11.2; 95% CI 9.6- 12.87; *P*=.007). The intervention district showed an 11.6 percentage-point increase in the average consistency of data among the delivery register and the integrated card compared with the comparator district (ATET 11.6; 95% CI 3.1-20.1; *P*=.04). Overall, the average consistency of data among IMRs in the intervention district was 11.6 times higher than that of the comparator district (ATET 11.6; 95% CI 4.2- 19; *P*=.03; Table 5).

Table 5. Effect of performance-based nonfinancial incentives on data consistency in individual medical records of institutional births, northwest Ethiopia, 2021.

Consistency	Intervention, mean (SD)	Comparator, mean (SD)	Intervention effect, ATET[a] (95% CI)[b]	*P* value
Delivery register vs integrated individual folder	62.2 (36.4)	51.1 (45.5)	11.2 (9.6-12.8)	.007
Delivery register vs integrated card	83.5 (19.9)	72.6 (26.0)	11.6 (3.1-20.1)	.04
Overall	80.2 (19.1)	69.4 (26.1)	11.6 (4.2-19)	.03

[a]ATET: average treatment effect on the treated.

[b]ATET estimates adjusted for covariates (presence of health information technology personnel, availability of data recording tools, availability of trained providers, supportive supervision from higher officials, existence of the Performance Monitoring Team [PMT], PMT per membership standard, monthly PMT meeting, monthly conducted lot quality assurance sampling, conducted root cause analysis, internal supervision, and availability of Health Management Information System guidelines).

Discussion

Principal Findings

This study evaluated the effect of PBNI on the quality of institutional birth data in northwest Ethiopia. PBNI improved both data completeness and consistency. The intervention district showed a 12% increase in data completeness compared with the comparator district, which showed a 1% decrease. Regarding data consistency, the intervention district improved by 18%, whereas the comparator district saw a 3% improvement. Controlling for other variables in the DID analysis, women’s

IMRs from the intervention district exhibited nearly 4 times higher data completeness and approximately 12 times greater data consistency than the comparator district.

Comparison With Prior Work

This study revealed a positive effect of PBNI on the data completeness and consistency of women's IMRs for institutional births. This finding aligns with that of previous studies that demonstrate the effectiveness of nonfinancial incentives in improving different aspects of health care quality. For instance, nonfinancial incentives have been proven to enhance the quality of medical error RCA in a US study [50]. Furthermore, studies from India and El Salvador [51,54,76] have shown an increase in the equitable and quality provision of maternal and child services. Nonfinancial incentives have also been reported to enhance quality consultations, according to studies from Tanzania [55,77]. The demonstrated effectiveness across contexts suggests the adaptability of nonfinancial incentives to improve the data quality and the quality of pregnancy-related services at health care facilities. These findings are particularly relevant for resource-limited settings where poor health care quality is associated with persistent mortality rates among mothers and children [3,5,78].

In this study, PBNI induced a greater extent of change compared with that in previous studies [51,54]. This difference may be due to differences in the incentive structures among the studies. In contrast to prior research that used team-based incentives, this study provided incentives at 3 levels: health facilities, departments, and individual health workers. Notably, the similarity between previous and current studies is apparent in the use of team-based incentives, reflected in this study's provision of incentives at the departmental level. Some earlier studies support the efficacy of team-based incentives in health care, emphasizing their role in fostering collective engagement [79-84]. Despite variations in the magnitude of the effect, the findings of this study do not contradict earlier research on the effectiveness of team-based incentives. Instead, the findings assert the potential for increased effectiveness by combining team-based, individual, and facility-level incentives.

According to this study, the effect of PBNI on data quality varies across the 3 women's records (integrated card, IIF, and delivery register). For example, although the integrated card saw a 3% increase in data completeness, women's folders increased by approximately 7%. These variations may suggest that the effectiveness of PBNI varies across health workers' professions. In Ethiopia, for instance, nonprofessional health workers are largely responsible for women's folder data, whereas midwives and other professionals are responsible for recording integrated card data [66]. Previous research in India has also demonstrated that the effectiveness of nonfinancial incentives varies depending on the health workers' professions, with frontline health workers experiencing a greater performance than supervisors [54]. Another study in northwest Ethiopia found a strong correlation between health worker motivation and their professional category [85]. These findings indicate the importance of recognizing the differences in the effectiveness of incentives and tailoring interventions to specific groups of health workers. Hence, policy makers and health care managers need to consider these variations when designing incentive programs, adopting a flexible approach that accounts for diverse roles and responsibilities.

This study reinforces the existing evidence that favors nonfinancial incentives in health care over financial incentives [86-90]. Concerns about financial incentives contradicting health care providers' intrinsic motivation to deliver quality care are widespread [86,90-100]. Therefore, this study suggests the practical use of nonfinancial incentives to enhance health care quality, especially in countries such as Ethiopia with limited financial capability. Prior studies in African countries have also indicated the importance of nonfinancial incentives in improving health care quality [44,101,102].

Policy and Research Implications

This study introduces PBNI as an effective intervention to improve the quality of institutional birth data. These findings underscore the potential of PBNI to complement established interventions to enhance RHIS performance, such as supportive supervision, mentorship, training, and feedback.

As this study evaluated the effect of PBNI on institutional birth data—a core indicator of maternal health care quality—the implications extend to broader RHIS data related to maternal and child services. These findings indicate the potential of PBNI to improve the quality of health services, which can contribute to maternal and child morbidity and mortality reduction. Hence, health care planners can consider adapting PBNI to improve the quality of maternal and child health services.

This study examined the effectiveness of PBNI in the context of health workers in health centers. Future studies are essential to understand the impact of PBNI on health staff across diverse settings, including health posts and hospitals.

Strengths and Limitations of the Study

One of the strengths of this study lies in its evaluation of the effect of PBNI on data quality within women's IMR in institutional births. Unlike most previous studies on RHIS data quality, which studied health facilities as the unit of analysis, this study delved into the individual level of data quality, which is essential to understanding how PBNI influences client-level service quality. In addition, we attempted to detect the minimum effect of the PBNI, using a sufficient study sample, and compared the intervention with comparator sites, increasing the robustness of the findings. Furthermore, to establish a causal effect of PBNI, the study used DID analysis, a recognized causal analysis technique in nonrandomized studies. Nevertheless, it is essential to recognize the limitations of this study. First, the retrospective design prevented randomization in the selection of the study participants. Second, interviewer bias is possible during the completeness and consistency assessments, as data collectors judged these aspects despite training to reduce bias. Although we attempted to disentangle the effect of PBNI from other potential factors, unmeasured confounders may still exist. Moreover, the security issues in Northern Ethiopia might have disrupted the effectiveness of the PBNI, as the intervention coincided with security problems in the adjacent regions.

Conclusions

This study shows that PBNI improves institutional birth data quality, as demonstrated by enhanced completeness and consistency. The effectiveness of PBNI can be extended to enhancing comparable RHIS data in maternal and child care and improving service quality at health care facilities. Health care planners can consider PBNI to enhance the quality of maternal and child health services in health care facilities. Future studies are essential to understand the impact of PBNI in diverse health care settings.

Acknowledgments

This work was financially supported by the Doris Duke Charitable Foundation (grant 2017187). The mission of the Doris Duke Charitable Foundation is to improve the quality of people's lives through grants supporting the performing arts, environmental conservation, medical research, and child well-being and through the preservation of the cultural and environmental legacy of Doris Duke's properties. The authors would like to express their gratitude to the supervisors, data collectors, and health office departments of the Central Gondar Zone, Wegera district, and Tach-armachiho district.

Authors' Contributions

BKT conceived and designed the study, performed the data collection, analyzed and interpreted the data, and drafted the manuscript. LDG, AA, SAM, JK, MKG, and BT analyzed and interpreted the data, and contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Average data completeness across individual medical records of institutional births during pre- and postintervention periods, northwest Ethiopia, 2021.
[PNG File , 35 KB - medinform_v12i1e54278_app1.png]

References

1. New global targets to prevent maternal deaths. World Health Organization. 2021 Oct 05. URL: https://www.who.int/news/item/05-10-2021-new-global-targets-to-prevent-maternal-deaths [accessed 2023-05-05]
2. Say L, Chou D, Gemmill A, Tunçalp Ö, Moller AB, Daniels J, et al. Global causes of maternal death: a WHO systematic analysis. Lancet Glob Health 2014 Jun;2(6):e323-e333 [FREE Full text] [doi: 10.1016/S2214-109X(14)70227-X] [Medline: 25103301]
3. Maternal mortality. World Health Organization. 2023 Feb 22. URL: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality [accessed 2022-11-17]
4. Sk MI, Paswan B, Anand A, Mondal NA. Praying until death: revisiting three delays model to contextualize the socio-cultural factors associated with maternal deaths in a region with high prevalence of eclampsia in India. BMC Pregnancy Childbirth 2019 Aug 28;19(1):314 [FREE Full text] [doi: 10.1186/s12884-019-2458-5] [Medline: 31455258]
5. Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/population division. World Health Organization. 2023 Feb 23. URL: https://www.who.int/publications/i/item/9789240068759 [accessed 2024-03-15]
6. Tunçalp , Were WM, MacLennan C, Oladapo OT, Gülmezoglu AM, Bahl R, et al. Quality of care for pregnant women and newborns-the WHO vision. BJOG 2015 Jul;122(8):1045-1049 [FREE Full text] [doi: 10.1111/1471-0528.13451] [Medline: 25929823]
7. van den Broek NR, Graham WJ. Quality of care for maternal and newborn health: the neglected agenda. BJOG 2009 Oct;116 Suppl 1:18-21. [doi: 10.1111/j.1471-0528.2009.02333.x] [Medline: 19740165]
8. SDG Target 3.1 Reduce the global maternal mortality ratio to less than 70 per 100 000 live births. World Health Organization. URL: https://www.who.int/data/gho/data/themes/topics/sdg-target-3-1-maternal-mortality [accessed 2023-02-09]
9. Wagenaar BH, Sherr K, Fernandes Q, Wagenaar AC. Using routine health information systems for well-designed health evaluations in low- and middle-income countries. Health Policy Plan 2016 Feb;31(1):129-135 [FREE Full text] [doi: 10.1093/heapol/czv029] [Medline: 25887561]
10. Shamba D, Day LT, Zaman SB, Sunny AK, Tarimo MN, Peven K, et al. Barriers and enablers to routine register data collection for newborns and mothers: EN-BIRTH multi-country validation study. BMC Pregnancy Childbirth 2021 Mar 26;21(Suppl 1):233 [FREE Full text] [doi: 10.1186/s12884-020-03517-3] [Medline: 33765963]
11. Framework and standards for country health information systems, 2nd edition. World Health Organization. 2023 Apr 24. URL: https://www.who.int/publications/i/item/9789241595940 [accessed 2024-03-15]

12. Standards for improving quality of maternal and newborn care in health facilities. World Health Organization. 2016. URL: https://pesquisa.bvsalud.org/portal/resource/pt/per-3087?lang=en [accessed 2022-11-08]
13. Dehnavieh R, Haghdoost A, Khosravi A, Hoseinabadi F, Rahimi H, Poursheikhali A, et al. The District Health Information System (DHIS2): a literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries. Health Inf Manag 2019 May;48(2):62-75. [doi: 10.1177/1833358318777713] [Medline: 29898604]
14. Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Serv Res 2020 Aug 25;20(1):790 [FREE Full text] [doi: 10.1186/s12913-020-05660-1] [Medline: 32843033]
15. Shama AT, Roba HS, Abaerei AA, Gebremeskel TG, Baraki N. Assessment of quality of routine health information system data and associated factors among departments in public health facilities of Harari region, Ethiopia. BMC Med Inform Decis Mak 2021 Oct 19;21(1):287 [FREE Full text] [doi: 10.1186/s12911-021-01651-2] [Medline: 34666753]
16. Lippeveld T. Routine health facility and community information systems: creating an information use culture. Glob Health Sci Pract 2017 Sep 28;5(3):338-340 [FREE Full text] [doi: 10.9745/GHSP-D-17-00319] [Medline: 28963169]
17. O'Hagan R, Marx MA, Finnegan KE, Naphini P, Ng'ambi K, Laija K, et al. National assessment of data quality and associated systems-level factors in Malawi. Glob Health Sci Pract 2017 Sep 28;5(3):367-381 [FREE Full text] [doi: 10.9745/GHSP-D-17-00177] [Medline: 28963173]
18. Mwinnyaa G, Hazel E, Maïga A, Amouzou A. Estimating population-based coverage of reproductive, maternal, newborn, and child health (RMNCH) interventions from health management information systems: a comprehensive review. BMC Health Serv Res 2021 Oct 25;21(Suppl 2):1083 [FREE Full text] [doi: 10.1186/s12913-021-06995-z] [Medline: 34689787]
19. Begum T, Khan SM, Adamou B, Ferdous J, Parvez MM, Islam MS, et al. Perceptions and experiences with district health information system software to collect and utilize health data in Bangladesh: a qualitative exploratory study. BMC Health Serv Res 2020 May 26;20(1):465 [FREE Full text] [doi: 10.1186/s12913-020-05322-2] [Medline: 32456706]
20. Day LT, Ruysen H, Gordeev VS, Gore-Langton GR, Boggs D, Cousens S, et al. "Every Newborn-BIRTH" protocol: observational study validating indicators for coverage and quality of maternal and newborn health care in Bangladesh, Nepal and Tanzania. J Glob Health 2019 Jun;9(1):010902 [FREE Full text] [doi: 10.7189/jogh.09.010902] [Medline: 30863542]
21. Alaro T, Sisay S, Samuel S. Implementation level of health management information system program in governmental hospitals of Ethiopia. Int J Intell Inf Syst 2019 Apr;8(2):52-57. [doi: 10.11648/j.ijiis.20190802.13]
22. HMIS indicator reference guide. International Institute for Primary Health Care Ethiopia. 2017. URL: http://repository.iifphc.org/handle/123456789/392 [accessed 2022-11-17]
23. Lemma S, Janson A, Persson L, Wickremasinghe D, Källestål C. Improving quality and use of routine health information system data in low- and middle-income countries: a scoping review. PLoS One 2020 Oct 8;15(10):e0239683 [FREE Full text] [doi: 10.1371/journal.pone.0239683] [Medline: 33031406]
24. Alemu MB, Atnafu A, Gebremedhin T, Endehabtu BF, Asressie M, Tilahun B. Outcome evaluation of capacity building and mentorship partnership (CBMP) program on data quality in the public health facilities of Amhara National Regional State, Ethiopia: a quasi-experimental evaluation. BMC Health Serv Res 2021 Oct 05;21(1):1054 [FREE Full text] [doi: 10.1186/s12913-021-07063-2] [Medline: 34610844]
25. Kanfe SG, Endehabtu BF, Ahmed MH, Mengestie ND, Tilahun B. Commitment levels of health care providers in using the district health information system and the associated factors for decision making in resource-limited settings: cross-sectional survey study. JMIR Med Inform 2021 Mar 04;9(3):e23951 [FREE Full text] [doi: 10.2196/23951] [Medline: 33661133]
26. Adane A, Adege TM, Ahmed MM, Anteneh HA, Ayalew ES, Berhanu D, et al. Routine health management information system data in Ethiopia: consistency, trends, and challenges. Glob Health Action 2021 Jan 01;14(1):1868961 [FREE Full text] [doi: 10.1080/16549716.2020.1868961] [Medline: 33446081]
27. Endriyas M, Kawza A, Alano A, Lemango F. Quality of medical records in public health facilities: a case of Southern Ethiopia, resource limited setting. Health Informatics J 2022;28(3):14604582221112853 [FREE Full text] [doi: 10.1177/14604582221112853] [Medline: 35793497]
28. Arsenault C, Yakob B, Kassa M, Dinsa G, Verguet S. Using health management information system data: case study and verification of institutional deliveries in Ethiopia. BMJ Glob Health 2021 Aug;6(8):e006216 [FREE Full text] [doi: 10.1136/bmjgh-2021-006216] [Medline: 34426404]
29. Hoxha K, Hung YW, Irwin BR, Grépin KA. Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: a systematic review. Health Inf Manag 2022 Sep;51(3):135-148. [doi: 10.1177/1833358320928729] [Medline: 32602368]
30. Straus S, Tetroe J, Graham ID. Knowledge Translation in Health Care: Moving from Evidence to Practice. Hoboken, NJ: Wiley; 2009.
31. Godin G, Bélanger-Gravel A, Eccles M, Grimshaw J. Healthcare professionals' intentions and behaviours: a systematic review of studies based on social cognitive theories. Implement Sci 2008 Jul 16;3:36 [FREE Full text] [doi: 10.1186/1748-5908-3-36] [Medline: 18631386]

32. Eccles MP, Johnston M, Hrisos S, Francis J, Grimshaw J, Steen N, et al. Translating clinicians' beliefs into implementation interventions (TRACII): a protocol for an intervention modeling experiment to change clinicians' intentions to implement evidence-based practice. Implement Sci 2007 Aug 16;2:27 [FREE Full text] [doi: 10.1186/1748-5908-2-27] [Medline: 17705824]
33. Flodgren G, Eccles MP, Shepperd S, Scott A, Parmelli E, Beyer FR. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. Cochrane Database Syst Rev 2011 Jul 06;2011(7):CD009255 [FREE Full text] [doi: 10.1002/14651858.CD009255] [Medline: 21735443]
34. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344 [FREE Full text] [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
35. Eccles M, Grimshaw J, Walker A, Johnston M, Pitts N. Changing the behavior of healthcare professionals: the use of theory in promoting the uptake of research findings. J Clin Epidemiol 2005 Feb;58(2):107-112. [doi: 10.1016/j.jclinepi.2004.09.002] [Medline: 15680740]
36. Saint-Lary O, Plu I, Naiditch M. Ethical issues raised by the introduction of payment for performance in France. J Med Ethics 2012 Aug;38(8):485-491 [FREE Full text] [doi: 10.1136/medethics-2011-100159] [Medline: 22493186]
37. Basinga P, Gertler PJ, Binagwaho A, Soucat AL, Sturdy J, Vermeersch CM. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. Lancet 2011 Apr 23;377(9775):1421-1428 [FREE Full text] [doi: 10.1016/S0140-6736(11)60177-3] [Medline: 21515164]
38. Eichler R, Agarwal K, Askew I, Iriarte E, Morgan L, Watson J. Performance-based incentives to improve health status of mothers and newborns: what does the evidence show? J Health Popul Nutr 2013 Dec;31(4 Suppl 2):36-47. [Medline: 24992802]
39. Mendelson A, Kondo K, Damberg C, Low A, Motúapuaka M, Freeman M, et al. The effects of pay-for-performance programs on health, health care use, and processes of care: a systematic review. Ann Intern Med 2017 Mar 07;166(5):341-353 [FREE Full text] [doi: 10.7326/M16-1881] [Medline: 28114600]
40. Diaconu K, Falconer J, Verbel A, Fretheim A, Witter S. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. Cochrane Database Syst Rev 2021 May 05;5(5):CD007899 [FREE Full text] [doi: 10.1002/14651858.CD007899.pub3] [Medline: 33951190]
41. Bufalino V, Peterson ED, Burke GL, LaBresh KA, Jones DW, Faxon DP, et al. Payment for quality: guiding principles and recommendations: principles and recommendations from the American Heart Association's Reimbursement, Coverage, and Access Policy Development Workgroup. Circulation 2006 Feb 28;113(8):1151-1154. [doi: 10.1161/CIRCULATIONAHA.105.171760] [Medline: 16401766]
42. Soeters R, Peerenboom PB, Mushagalusa P, Kimanuka C. Performance-based financing experiment improved health care in the Democratic Republic of Congo. Health Aff (Millwood) 2011 Aug;30(8):1518-1527. [doi: 10.1377/hlthaff.2009.0019] [Medline: 21821568]
43. Huntington D, Zaky HH, Shawky S, Fattah FA, El-Hadary E. Impact of a service provider incentive payment scheme on quality of reproductive and child-health services in Egypt. J Health Popul Nutr 2010 Jun;28(3):273-280 [FREE Full text] [doi: 10.3329/jhpn.v28i3.5556] [Medline: 20635638]
44. Lagarde M, Huicho L, Papanicolas I. Motivating provision of high quality care: it is not all about the money. BMJ 2019 Sep 23;366:l5210 [FREE Full text] [doi: 10.1136/bmj.l5210] [Medline: 31548200]
45. Leonard KL, Masatu MC. Professionalism and the know-do gap: exploring intrinsic motivation among health workers in Tanzania. Health Econ 2010 Dec;19(12):1461-1477. [doi: 10.1002/hec.1564] [Medline: 19960481]
46. Ochenge NC, Susan W. Role of reward systems in employee motivation in Kenyan deposit taking micro finance institutions: a case study of Faulu Kenya. Int J Soc Sci Manag Entrep 2014;1(2):203-220.
47. Burgess S, Metcalfe R, Sadoff S. Understanding the response to financial and non-financial incentives in education: field experimental evidence using high-stakes assessments. Econ Educ Rev 2021 Dec;85:102195 [FREE Full text] [doi: 10.1016/j.econedurev.2021.102195]
48. Ashraf N, Bandiera O, Jack BK. No margin, no mission? A field experiment on incentives for public service delivery. J Public Econ 2014 Dec;120:1-17 [FREE Full text] [doi: 10.1016/j.jpubeco.2014.06.014]
49. Ashraf N, Bandiera O, Lee SS. Do-gooders and go-getters: career incentives, selection, and performance in public service delivery. Harvard Business School. 2015 Mar. URL: https://www.hbs.edu/faculty/Pages/item.aspx?num=46043 [accessed 2024-03-25]
50. Bagian JP, King BJ, Mills PD, McKnight SD. Improving RCA performance: the Cornerstone Award and the power of positive reinforcement. BMJ Qual Saf 2011 Nov;20(11):974-982. [doi: 10.1136/bmjqs.2010.049585] [Medline: 21775506]
51. Bernal P, Martinez S. In-kind incentives and health worker performance: experimental evidence from El Salvador. J Health Econ 2020 Mar;70:102267 [FREE Full text] [doi: 10.1016/j.jhealeco.2019.102267] [Medline: 32028090]
52. Bufalino V, Peterson ED, Krumholz HM, Burke GL, LaBresh KA, Jones DW, et al. Nonfinancial incentives for quality: a policy statement from the American Heart Association. Circulation 2007 Jan 23;115(3):398-401. [doi: 10.1161/CIRCULATIONAHA.106.180202] [Medline: 17179024]

53. Cacace M, Geraedts M, Berger E. Public reporting as a quality strategy. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies. Copenhagen, Denmark: European Observatory on Health Systems and Policies; 2019.
54. Grant C, Nawal D, Guntur SM, Kumar M, Chaudhuri I, Galavotti C, et al. 'We pledge to improve the health of our entire community': improving health worker motivation and performance in Bihar, India through teamwork, recognition, and non-financial incentives. PLoS One 2018 Aug 30;13(8):e0203265 [FREE Full text] [doi: 10.1371/journal.pone.0203265] [Medline: 30161213]
55. Brock JM, Lange A, Leonard KL. Giving and promising gifts: experimental evidence on reciprocity from the field. J Health Econ 2018 Mar;58:188-201 [FREE Full text] [doi: 10.1016/j.jhealeco.2018.02.007] [Medline: 29524793]
56. Population size of towns by sex, region, zone and Weredas as of July 2021. Ethiopian Statistical Service. URL: https://www.statsethiopia.gov.et/wp-content/uploads/2020/08/Population-of-Towns-as-of-July-2021.pdf [accessed 2024-03-25]
57. National guideline for family planning services in Ethiopia. Ministry of Health, Federal Democratic Republic of Ethiopia. 2011 Feb. URL: https://pdf4pro.com/view/national-family-planning-guideline-phe-ethiopia-1876ed.html [accessed 2023-10-20]
58. Asmamaw A, Tesfahun H, Berhanu FE, Lemma DG, Adane M, Teklehaymanot G, et al. Implementation outcomes of performance based non- financial incentive: using RE-AIM framework. Ethiop J Health Dev 2023;37(1):1-10.
59. Capacity building and mentorship program (CBMP). eHealthlab Ethiopia. URL: https://ehealthlab.org/cbmp/ [accessed 2024-03-25]
60. Chanyalew MA, Yitayal M, Atnafu A, Mengiste SA, Tilahun B. The effectiveness of the capacity building and mentorship program in improving evidence-based decision-making in the Amhara Region, Northwest Ethiopia: difference-in-differences study. JMIR Med Inform 2022 Apr 22;10(4):e30518 [FREE Full text] [doi: 10.2196/30518] [Medline: 35451990]
61. Tilahun B, Endehabtu BF, Hailemariam T, Derseh Gezie L, Mamuye A, Gebrehiwot T, et al. Effectiveness of performance-based non-financial incentive for improved health data quality and information use at primary health care units, northwest Ethiopia. Ethiop J Health Dev 2023 Nov 16;37(1):1-9. [doi: 10.20372/ejhd.v37i1.5839]
62. Witter S, Bertone MP, Diaconu K, Bornemisza O. Performance-based financing versus "unconditional" direct facility financing - false dichotomy? Health Syst Reform 2021 Jan 01;7(1):e2006121. [doi: 10.1080/23288604.2021.2006121] [Medline: 34874806]
63. Quentin W, Partanen VM, Brownwood I, Klazinga N. Measuring healthcare quality. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies. Copenhagen, Denmark: European Observatory on Health Systems and Policies; 2019.
64. Gezie LD, Endehabtu BF, Hailemariam T, Atnafu A, Mamuye A, Mohamed M, et al. Barriers and facilitators of implementing performance- based non- financial incentives to improve data quality and use: using a consolidated framework for implementation research. Ethiop J Health Dev 2023;37(1):1-10. [doi: 10.20372/ejhd.v37i1.5838]
65. Amare G, Minyihun A, Atnafu A, Endehabtu BF, Derseh L, Hailemariam T, et al. Cost-effectiveness of performance-based non-financial incentive (PBNI) intervention to improve health information system performance at Wogera district in northwest Ethiopia. Ethiop J Health Dev 2023 Nov 16;37(1):1-11. [doi: 10.20372/ejhd.v37i1.5837]
66. Ethiopian health management information system: data recording and reporting procedures manual. Federal Ministry of Health Ethiopia. 2017 Jul. URL: http://dataverse.nipn.ephi.gov.et/bitstream/handle/123456789/293/HMIS%20Recording%20and%20Reporting%20Procedures.pdf?sequence=1 [accessed 2022-11-17]
67. Zozus MN, Hammond WE, Green BB, Kahn MG, Richesson RL, Rusincovitch SA, et al. Assessing data quality for healthcare systems data used in clinical research. NIH Pragmatic Trials Collaboratory. 2014. URL: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf [accessed 2023-10-20]
68. Data quality assurance: module 3: site assessment of data quality: data verification and system assessment. World Health Organization. 2023 Jan 30. URL: https://www.who.int/publications/i/item/9789240049123 [accessed 2023-10-20]
69. Maïga A, Jiwani SS, Mutua MK, Porth TA, Taylor CM, Asiki G, et al. Generating statistics from health facility data: the state of routine health information systems in Eastern and Southern Africa. BMJ Glob Health 2019 Sep 29;4(5):e001849 [FREE Full text] [doi: 10.1136/bmjgh-2019-001849] [Medline: 31637032]
70. PRISM: performance of routine information system management series. Measure Evaluation. URL: https://tinyurl.com/bdhap69j [accessed 2024-03-17]
71. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy Plan 2009 May;24(3):217-228 [FREE Full text] [doi: 10.1093/heapol/czp010] [Medline: 19304786]
72. Health data quality training module participant manual. Federal Democratic Republic Of Ethiopia, Ministry of Health. 2018. URL: https://tinyurl.com/3rc5uts3 [accessed 2024-02-11]
73. Taye BK, Gezie LD, Atnafu A, Mengiste SA, Tilahun B. Data completeness and consistency in individual medical records of institutional births: retrospective crossectional study from Northwest Ethiopia, 2022. BMC Health Serv Res 2023 Oct 31;23(1):1189 [FREE Full text] [doi: 10.1186/s12913-023-10127-0] [Medline: 37907881]
74. Luedicke J. Difference-in-differences estimation using Stata. Stata Users Group. 2022. URL: https://ideas.repec.org/p/boc/dsug22/06.html [accessed 2024-02-11]

75. Introduction to difference-in-differences estimation. In: Causal Inference and Treatment-Effects Estimation Reference Manual. College Station, TX: Stata Press; 2023.
76. Carmichael SL, Mehta K, Raheel H, Srikantiah S, Chaudhuri I, Trehan S, et al. Effects of team-based goals and non-monetary incentives on front-line health worker performance and maternal health behaviours: a cluster randomised controlled trial in Bihar, India. BMJ Glob Health 2019 Aug 26;4(4):e001146 [FREE Full text] [doi: 10.1136/bmjgh-2018-001146] [Medline: 31543982]
77. Brock MJ, Lange A, Leonard KL. Generosity norms and intrinsic motivation in health care provision: evidence from the laboratory and the field. European Bank for Reconstruction and Development, Office of the Chief Economist. 2012. URL: https://ideas.repec.org/p/ebd/wpaper/147.html [accessed 2024-02-11]
78. Newborn mortality. World Health Organization. 2024 Mar 14. URL: https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021 [accessed 2024-03-25]
79. Mbindyo P, Gilson L, Blaauw D, English M. Contextual influences on health worker motivation in district hospitals in Kenya. Implement Sci 2009 Jul 23;4:43 [FREE Full text] [doi: 10.1186/1748-5908-4-43] [Medline: 19627590]
80. Benjamin L, Cotté FE, Philippe C, Mercier F, Bachelot T, Vidal-Trécan G. Physicians' preferences for prescribing oral and intravenous anticancer drugs: a discrete choice experiment. Eur J Cancer 2012 Apr;48(6):912-920. [doi: 10.1016/j.ejca.2011.09.019] [Medline: 22033327]
81. Sharma R, Webster P, Bhattacharyya S. Factors affecting the performance of community health workers in India: a multi-stakeholder perspective. Glob Health Action 2014 Oct 13;7:25352 [FREE Full text] [doi: 10.3402/gha.v7.25352] [Medline: 25319596]
82. Kivimäki M, Vanhala A, Pentti J, Länsisalmi H, Virtanen M, Elovainio M, et al. Team climate, intention to leave and turnover among hospital employees: prospective cohort study. BMC Health Serv Res 2007 Oct 23;7:170 [FREE Full text] [doi: 10.1186/1472-6963-7-170] [Medline: 17956609]
83. Borkum E, Rangarajan A, Rotz D, Sridharan S, Sethi S, Manorajini M. Evaluation of the team-based goals and performance-based incentives (TBGI) innovation in Bihar. Mathematica Policy Research Reports. URL: https://ideas.repec.org/p/mpr/mprres/d8e1097122ff47a6bf42580c82677834.html [accessed 2024-02-11]
84. Lee TH, Bothe A, Steele GD. How Geisinger structures its physicians' compensation to support improvements in quality, efficiency, and volume. Health Aff (Millwood) 2012 Sep;31(9):2068-2073. [doi: 10.1377/hlthaff.2011.0940] [Medline: 22949457]
85. Weldegebriel Z, Ejigu Y, Weldegebreal F, Woldie M. Motivation of health workers and associated factors in public hospitals of West Amhara, Northwest Ethiopia. Patient Prefer Adherence 2016 Feb 15;10:159-169 [FREE Full text] [doi: 10.2147/PPA.S90323] [Medline: 26929608]
86. Lee TH. Financial versus non-financial incentives for improving patient experience. J Patient Exp 2015 May;2(1):4-6 [FREE Full text] [doi: 10.1177/237437431500200102] [Medline: 28725809]
87. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ 2016 Jan;50(1):36-60. [doi: 10.1111/medu.12817] [Medline: 26695465]
88. Lagarde M, Blaauw D. Pro-social preferences and self-selection into jobs: evidence from South African nurses. J Econ Behav Organ 2014 Nov;107(A):136-152 [FREE Full text] [doi: 10.1016/j.jebo.2014.09.004]
89. Ashraf N, Bandiera O. Altruistic capital. Am Econ Rev 2017 May;107(5):70-75. [doi: 10.1257/aer.p20171097]
90. Attema AE, Galizzi MM, Groß M, Hennig-Schmidt H, Karay Y, L'Haridon O, et al. The formation of physician altruism. J Health Econ 2023 Jan;87:102716 [FREE Full text] [doi: 10.1016/j.jhealeco.2022.102716] [Medline: 36603361]
91. Khullar D, Wolfson D, Casalino LP. Professionalism, performance, and the future of physician incentives. JAMA 2018 Dec 18;320(23):2419-2420. [doi: 10.1001/jama.2018.17719] [Medline: 30476944]
92. Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, Furler J, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database Syst Rev 2011 Sep 07(9):CD008451. [doi: 10.1002/14651858.CD008451.pub2] [Medline: 21901722]
93. Witter S, Fretheim A, Kessy FL, Lindahl AK. Paying for performance to improve the delivery of health interventions in low- and middle-income countries . Cochrane Database Syst Rev 2012 Feb 15(2):CD007899. [doi: 10.1002/14651858.CD007899.pub2] [Medline: 22336833]
94. Berwick DM. Era 3 for medicine and health care. JAMA 2016 Apr 05;315(13):1329-1330. [doi: 10.1001/jama.2016.1509] [Medline: 26940610]
95. Delfgaauw J. Dedicated doctors: public and private provision of health care with altruistic physicians. SSRN Preprint posted online February 5, 2007. [doi: 10.2139/ssrn.958693]
96. Liu T, Ma CT. Health insurance, treatment plan, and delegation to altruistic physician. J Econ Behav Org 2013 Jan;85:79-96 [FREE Full text] [doi: 10.1016/j.jebo.2012.11.002]
97. Mannion R, Davies HT. Payment for performance in health care. BMJ 2008 Feb 09;336(7639):306-308 [FREE Full text] [doi: 10.1136/bmj.39463.454815.94] [Medline: 18258966]
98. Chen LM, Epstein AM, Orav EJ, Filice CE, Samson LW, Joynt Maddox KE. Association of practice-level social and medical risk with performance in the medicare physician value-based payment modifier program. JAMA 2017 Aug 01;318(5):453-461 [FREE Full text] [doi: 10.1001/jama.2017.9643] [Medline: 28763549]

99. Heath I, Hippisley-Cox J, Smeeth L. Measuring performance and missing the point? BMJ 2007 Nov 24;335(7629):1075-1076 [FREE Full text] [doi: 10.1136/bmj.39377.387373.AD] [Medline: 18033930]
100. McDonald R, Harrison S, Checkland K, Campbell SM, Roland M. Impact of financial incentives on clinical autonomy and internal motivation in primary care: ethnographic study. BMJ 2007 Jun 30;334(7608):1357 [FREE Full text] [doi: 10.1136/bmj.39238.890810.BE] [Medline: 17580318]
101. Borghi J, Little R, Binyaruka P, Patouillard E, Kuwawenaruwa A. In Tanzania, the many costs of pay-for-performance leave open to debate whether the strategy is cost-effective. Health Aff (Millwood) 2015 Mar;34(3):406-414. [doi: 10.1377/hlthaff.2014.0608] [Medline: 25732490]
102. Mathauer I, Imhoff I. Health worker motivation in Africa: the role of non-financial incentives and human resource management tools. Hum Resour Health 2006 Aug 29;4(1):24 [FREE Full text] [doi: 10.1186/1478-4491-4-24] [Medline: 16939644]

Abbreviations

ATET: average treatment effect on the treated
CBMP: Capacity Building and Mentorship Program
DID: difference-in-differences
HIT: health IT
HMIS: Health Management Information System
IIF: integrated individual folder
IMR: individual medical record
LQAS: lot quality assurance sampling
MRN: medical record number
MRU: Medical Record Unit
PBNI: performance-based nonfinancial incentives
PMT: Performance Monitoring Team
RCA: root cause analysis
RHIS: Routine Health Information System

Edited by C Perrin; submitted 03.11.23; peer-reviewed by T Wonde, T Tefera; comments to author 11.01.24; revised version received 20.01.24; accepted 05.02.24; published 05.04.24.

Please cite as:
Taye BK, Gezie LD, Atnafu A, Mengiste SA, Kaasbøll J, Gullslett MK, Tilahun B
Effect of Performance-Based Nonfinancial Incentives on Data Quality in Individual Medical Records of Institutional Births: Quasi-Experimental Study
JMIR Med Inform 2024;12:e54278
URL: https://medinform.jmir.org/2024/1/e54278
doi:10.2196/54278
PMID:38578684

Addressing Information Biases Within Electronic Health Record Data to Improve the Examination of Epidemiologic Associations With Diabetes Prevalence Among Young Adults: Cross-Sectional Study

Sarah Conderino[1], DrPH; Rebecca Anthopolos[1], DrPH; Sandra S Albrecht[2], PhD; Shannon M Farley[3], DrPH; Jasmin Divers[1,4], PhD; Andrea R Titus[1], PhD; Lorna E Thorpe[1], PhD

1

2

3

4

Corresponding Author:
Sarah Conderino, DrPH

Abstract

Background: Electronic health records (EHRs) are increasingly used for epidemiologic research to advance public health practice. However, key variables are susceptible to missing data or misclassification within EHRs, including demographic information or disease status, which could affect the estimation of disease prevalence or risk factor associations.

Objective: In this paper, we applied methods from the literature on missing data and causal inference to assess whether we could mitigate information biases when estimating measures of association between potential risk factors and diabetes among a patient population of New York City young adults.

Methods: We estimated the odds ratio (OR) for diabetes by race or ethnicity and asthma status using EHR data from NYU Langone Health. Methods from the missing data and causal inference literature were then applied to assess the ability to control for misclassification of health outcomes in the EHR data. We compared EHR-based associations with associations observed from 2 national health surveys, the Behavioral Risk Factor Surveillance System (BRFSS) and the National Health and Nutrition Examination Survey, representing traditional public health surveillance systems.

Results: Observed EHR-based associations between race or ethnicity and diabetes were comparable to health survey-based estimates, but the association between asthma and diabetes was significantly overestimated (OR_{EHR} 3.01, 95% CI 2.86-3.18 vs OR_{BRFSS} 1.23, 95% CI 1.09-1.40). Missing data and causal inference methods reduced information biases in these estimates, yielding relative differences from traditional estimates below 50% ($OR_{MissingData}$ 1.79, 95% CI 1.67-1.92 and OR_{Causal} 1.42, 95% CI 1.34-1.51).

Conclusions: Findings suggest that without bias adjustment, EHR analyses may yield biased measures of association, driven in part by subgroup differences in health care use. However, applying missing data or causal inference frameworks can help control for and, importantly, characterize residual information biases in these estimates.

(JMIR Med Inform 2024;12:e58085) doi:10.2196/58085

KEYWORDS

information bias; electronic health record; EHR; epidemiologic method; confounding factor; diabetes; epidemiology; young adult; cross-sectional study; risk factor; asthma; race; ethnicity; diabetic; diabetic adult

Introduction

Understanding patterns and risk factors of chronic disease burden is a key function of public health practice. Electronic health record (EHR) data have numerous strengths that can be leveraged for this purpose, including near real-time information on large patient populations, allowing for improved precision and timeliness when estimating patterns or trends in disease burden [1]. EHRs also contain clinically based diagnoses, laboratory results, and physical measurements [2]. Often, researchers use these clinical data to classify patients' disease status using rule-based computable phenotypes or prespecified logic-based criteria [3]. Using EHR-defined measures of disease offers promise for improving the estimation of patterns or associations compared to traditional surveillance systems (eg, health surveys), several of which have poor validity from

self-reported disease status and limited reliability for small subgroups or geographies [4].

Despite these strengths, the classification of disease status using EHR data is susceptible to information biases. Misclassification or measurement errors in key variables can bias estimates of epidemiologic associations. Within EHR data, a diagnostic suspicion bias could occur if certain patients are disproportionately screened for health outcomes compared to others (eg, those who are obese may be more likely to be screened for diabetes with A_{1c} testing) [5], and an informed presence bias could occur if, for example, patients who visit the health system more frequently are sicker or have more opportunities to receive a diagnosis than infrequent visitors [6]. These different biases can distort our understanding of patterns or risk factors for disease burden, such as by underestimating the relative burden among individuals with limited or more fragmented access to care. Notably, selection biases whereby the EHR sample is not generalizable to the target population can also affect estimates of epidemiologic associations, and misclassification itself can be framed as a selection bias issue [7]. Improving our capacity to address misclassification, framed as an information bias issue herein, is a critical step to providing valid inferences.

A variety of statistical approaches and frameworks have been developed to help address misclassification, including treating misclassification as a missing data problem, a causal inference problem, or both [8,9]. With EHR-based computable phenotypes, researchers often rely on variables that are not identifiably missing (eg, lack of diagnosis codes are assumed to mean the absence of disease). For example, pertinent evidence of the disease may be absent from a single EHR due to patients receiving care across multiple distinct health care systems, which could lead to individuals being falsely classified as disease-free [5]. Internal or external data sets can allow for the validation and correction of the computable phenotype's performance, but these data sets are often costly or resource-intensive to obtain [8,10]. Under a missing data framework, the observed health outcome can be assumed to have some level of misclassification and the true health outcome for some or all of the patients can be treated as missing [8,11]. Traditional missing data methods, such as regression calibration, multiple imputation, and inverse probability weighting (IPW), can then be used to address this misclassification [8,10-12].

Misclassification in disease status can also be conceptualized using directed acyclic graphs (DAGs), illustrating factors that affect the causal relationship between the true and observed exposure and outcome [8,13]. Researchers can use traditional epidemiologic methods to control for variables that act as confounders of the observed exposure and outcome. For example, researchers often hypothesize that the number of health care encounters affects the misclassification of EHR-defined disease status through an informed presence bias, with EHR samples further restricted to those with at least 1 encounter [6]. Prior studies have demonstrated that conditioning on the number of encounters can sometimes reduce confounding from differential misclassification but also has the potential to induce a smaller Berkson's or M-bias if the number of encounters is a common effect of the exposure and outcome, particularly when computable phenotypes are highly sensitive [14]. Nevertheless, DAGs offer a promising approach to address misclassification.

In this paper, we applied missing data and causal inference frameworks to evaluate the impact of information biases in EHR data on longstanding epidemiologic research questions about diabetes among young adults. Diabetes is a serious, chronic condition that is still relatively rare within this age group, affecting an estimated 3% of those aged 18 - 44 years [15] but is increasing in both prevalence and incidence [16-18]. We first aimed to estimate the age and sex-adjusted odds of diabetes by race or ethnicity, given established differences in diabetes prevalence across racial or ethnic groups [15,19,20]. Second, we characterized the association between asthma and diabetes, presuming a causal relationship between these chronic conditions [21]. For the target population of in-care US adults aged 18 - 44 years, we compared estimated associations using these frameworks to estimates based on probability samples from national health surveys. This work informs the broader discussion on how to address differential misclassification of disease outcomes within EHR data.

Methods

EHR Sample

The EHR sample comprises patients from NYU Langone Health, a large academic medical center with primary service areas in the New York City boroughs of Manhattan, Brooklyn, and Queens. EHR data were pulled from the Epic Clarity database for all New York City–resident patients aged 18 - 44 years who had an inpatient or outpatient encounter from 2017 to 2019.

Demographic variables of age, sex (male or female), race or ethnicity (White, Black, Latino, Asian, and other), most recent insurance status (Medicaid vs non-Medicaid), and neighborhood poverty level (<10%, 10%-<20%, 20%-<30%, and ≥30% living in poverty within resident census tract) were defined for each patient. Race or ethnicity was imputed for those with unknown race or ethnicity (19.4%, n=88,102) using the Bayesian Improved Surname Geocoding (BISG) methods through the "wru" R package (R Core Team) [22]. Neighborhood poverty level was assigned using zip code tabulation area poverty group when census tract level data were unavailable (1%). Those with an unknown or other age or sex (<1%) were excluded from all analyses.

Health care use variables of total encounters, duration within the NYU Langone Health system, presence of at least 1 routine health exam (2024 ICD-10-CM: Z00.*), presence of at least 1 diabetes-related laboratory (fasting glucose, random glucose, or A_{1c}), and presence of at least 1 encounter with an endocrinology review of systems, or inventory of signs or symptoms of diseases related to the endocrine system, were also defined. Endocrinology review of systems were identified using keyword text searches of history and physical examination notes and progress notes (Multimedia Appendix 1).

Patients were classified as having prevalent obesity, asthma, and diabetes if they had evidence supporting these chronic conditions, using all historical EHR data through 2019, and

were classified as not having each respective health outcome without such evidence. In alignment with diabetes definitions from national health surveys, EHR-defined diabetes included all diabetes types. Evidence of diabetes included at least 2 encounter diagnoses for diabetes (ICD-10-CM: E08.*, E09.*, E10.*, E11.*, and E13.*); or 1 encounter diagnosis and at least 2 elevated A_{1c} laboratory results ≥6.5%; or at least 1 antidiabetes prescription medication (not including metformin or acarbose) [23]. Evidence of asthma was defined as at least 2 encounter diagnoses for asthma (ICD-10-CM: J45*–J46*) or at least 2 prescriptions for asthma-related medications [24]. To maintain consistency across chronic disease classification methods, evidence of obesity was defined as a most recent BMI ≥30 kg/m^2, with no naïve corrections for those who were missing BMI, height, or weight measurements (19.1%, n=86,709).

EHR-Based Estimation

We estimated odds ratios (OR) for diabetes by race or ethnicity and asthma status under 4 EHR-based estimation methods that we describe herein. First, "naïve" models were estimated by fitting a logistic regression model for observed diabetes status (DM*) on the total sample (n=454,612). ORs for race or ethnicity were adjusted for age and sex and ORs for asthma were adjusted for the potential confounders of age, sex, race or ethnicity, Medicaid insurance status, obesity, and neighborhood poverty level, as informed by existing literature [25-27].

Second, "sufficiency" models were estimated among the subset of patients who we hypothesized to have sufficient data, defined as those with at least 1 encounter with an endocrinology review of systems or those who were classified as diabetic through the above definition (n=181,036). Since diabetes is a rare disease within the young adult population, we assumed that the specificity of the classification was near-perfect and all patients who were classified as diabetic had sufficient data [28]. Sensitivity analyses tested this assumption and varied the definition of sufficient data to incorporate information related to the other health outcomes (eg, having a diabetes-related laboratory, BMI measurement, and respiratory review of systems; Multimedia Appendix 1). Sufficiency models were estimated by fitting a logistic regression model for DM* using the same covariates as the naïve models.

Third, using "IPW" models, we hypothesized that missing health outcomes would be predicted by demographics (eg, differential screening by race or ethnicity), health care use (eg, informed presence bias), and neighborhood (eg, degree of continuity of care within the health system by catchment area). We estimated the probability of having sufficient data using a multilevel logistic regression model including all demographic and health care use variables and a random intercept for neighborhood defined by public use microdata areas. Stabilized IPW weights were then calculated as the inverse of the predicted probability of having sufficient data multiplied by the overall probability of having sufficient data [10]. The final models were estimated by fitting a logistic regression model for DM* on the subset of patients defined as having sufficient data (n=181,036), weighted for the stabilized IPW weights and using the same covariates as the naïve models.

Fourth, using DAG models (Multimedia Appendix 1), we hypothesized that total encounters would both be associated with differential misclassification of health outcomes through an informed presence bias and be a common effect of the health outcomes, consistent with prior research [6,14]. DAG models were estimated by fitting a logistic regression model for DM* using the total sample (n=454,612), controlling for total encounters and the covariates included in the naïve model. Further details on these models can be found in Multimedia Appendix 1.

Comparison to Survey-Based Estimation

For comparison to traditional surveillance systems, samples were obtained from 2 publicly available national health surveys, the 2019 Behavioral Risk Factor Surveillance System (BRFSS) and the pooled 2013-March 2020 National Health and Nutrition Examination Survey (NHANES). BRFSS is a cross-sectional telephone survey conducted by the Centers for Disease Control and Prevention annually on a sample of over 400,000 US adults [29]. Within BRFSS data, diabetes and asthma were defined by self-reported prior diagnosis from a medical provider, and obesity was defined by a BMI ≥30 kg/m^2 based on self-reported height and weight. Demographic variables of 5-year age group, sex (male or female), race or ethnicity (White, Black, Latino, Asian, and other), insurance status (uninsured vs insured), and income level (<US $50,000, US $50,000-<US $75,000, ≥US $75,000) were defined for each respondent. To reduce the effects of undiagnosed diabetes on misclassification of self-reported diabetes status, the BRFSS survey data were subset to those aged 18 - 44 years who were in care, as defined as those who reported having a personal health care provider.

NHANES is a cross-sectional survey involving interviews and physical examinations that is conducted by the Centers for Disease Control and Prevention annually on a sample of approximately 5000 US children and adults [30]. Within NHANES data, diabetes was defined by a self-reported prior diagnosis or elevated laboratory results (A_{1c}≥6.5% or fasting glucose≥126 mm Hg) [15,31]. Sensitivity analyses varied this definition to be based solely on self-reported prior diagnosis. Asthma was defined by self-reported prior diagnosis from a medical provider and obesity was defined by a measured BMI≥30 kg/m^2. Demographic variables of age, sex (male or female), race or ethnicity (White, Black, Latino, Asian, and other), insurance status (Medicaid vs non-Medicaid), and income level (≤130%, 130% - 350%, >350% of the federal poverty level) were defined for each respondent. NHANES data were subset to those aged 18 - 44 years.

Survey-based estimates were obtained by fitting logistic regression models for diabetes accounting for the unequal probability sample, stratification, and clustering in the complex sample designs [32,33]. Survey-based ORs for race or ethnicity were adjusted for age and sex and ORs for asthma were adjusted for age, sex, race or ethnicity, obesity, insurance status, and income level. Relative differences between EHR-based and survey-based ORs were calculated as percent differences.

Ethical Considerations

This study was approved by the NYU Langone Health Institutional Review Board (i20-01338) and Columbia University Institutional Review Board (AAAU5390) and informed consent was waived. Participants were not compensated.

Results

EHR Sample

The EHR sample comprised 454,612 patients seen within the NYU Langone Health system from 2017 to 2019 (Table 1). A total of 37.8% (n=171,968) of patients were male and 22.2% (n=100,979) had Medicaid insurance. The largest racial or ethnic group within the sample was White, with 41.8% (n=190,225) having a White race or ethnicity recorded within the EHR, and 52.1% (n=237,057) classified as White through BISG imputation. Approximately one-quarter (n=115,249) of patients had a routine medical exam and one-half (n=205,408) had a DM-related laboratory. Within the full sample, 3.1% (n=14,044) of patients were classified as having diabetes, 17.5% (n=79,580) were classified as being obese, and 4.2% (n=19,240) were classified as having asthma.

Table . Descriptive summary of the New York University patient population by data sufficiency status[a,b].

Variables	Total sample (N=454,612)	Insufficient case (n=273,576)	Sufficient case[a] (n=181,036)
Age (years), mean (SD)	32.13 (7.11)	32.02 (7.13)	32.31 (7.07)
Sex (male), n (%)	171,968 (37.8)	100,964 (36.9)	71,004 (39.2)
Medicaid insurance, n (%)	100,979 (22.2)	59,001 (21.6)	41,978 (23.2)
Raw race or ethnicity, n (%)			
White	190,225 (41.8)	111,123 (40.6)	79,102 (43.7)
Black	45,509 (10.0)	24,442 (8.9)	21,067 (11.6)
Latino	62,989 (13.9)	32,157 (11.8)	30,832 (17.0)
Asian or Pacific Islander	35,262 (7.8)	20,947 (7.7)	14,315 (7.9)
Other	32,525 (7.2)	19,669 (7.2)	12,856 (7.1)
Missing	88,102 (19.4)	65,238 (23.8)	22,864 (12.6)
Imputed race or ethnicity, n (%)			
White	237,057 (52.1)	144,783 (52.9)	92,274 (51.0)
Black	57,709 (12.7)	33,439 (12.2)	24,270 (13.4)
Latino	86,679 (19.1)	49,131 (18.0)	37,548 (20.7)
Asian or Pacific Islander	49,170 (10.8)	31,288 (11.4)	17,882 (9.9)
Other	23,997 (5.3)	14,935 (5.5)	9062 (5.0)
Recorded BMI, n (%)	367,903 (80.9)	190,916 (69.8)	176,987 (97.8)
Encounters[c], mean (SD)	15.23 (23.51)	9.41 (13.15)	24.03 (31.60)
Duration[d], mean (SD)	1.84 (1.93)	1.53 (1.83)	2.31 (1.98)
Routine medical exam, n (%)	115,249 (25.4)	32,786 (12.0)	82,463 (45.6)
Diabetes-related laboratory[b], n (%)	205,408 (45.2)	80,728 (29.5)	124,680 (68.9)
Neighborhood coverage [e] , n (%)			
<10%	79,563 (17.5)	48,320 (17.7)	31,243 (17.3)
10%-<20%	143,907 (31.7)	82,148 (30.0)	61,759 (34.1)
20%-<30%	163,076 (35.9)	101,293 (37.0)	61,783 (34.1)
30%-<40%	68,066 (15.0)	41,815 (15.3)	26,251 (14.5)
Asthma, n (%)	19,240 (4.2)	6167 (2.3)	13,073 (7.2)
Obese, n (%)	79,580 (17.5)	38,819 (14.2)	40,761 (22.5)

[a]Sufficient cases defined as those with at least 1 encounter with an endocrinology review of systems or those who were classified as diabetic through the computable phenotype of having at least 2 encounter diagnoses for diabetes, 1 encounter diagnosis and at least 2 elevated A_{1c} laboratory results ≥6.5%, or at least 1 antidiabetes prescription medication.

[b]Including all A_{1c}, random blood glucose, and fasting blood glucose laboratory results.

[c]Number of encounters.

[d]Number of years in the health system.

[e]Proportion of individuals residing in the Public Use Microdata Area (PUMA) neighborhood who are present within the electronic health record system.

A total of 39.8% (n=181,036) of the patient population were classified as having sufficient data (Table 1). Patients who were classified as sufficient had greater health care use, measured by a higher average number of total encounters, greater duration within the NYU system, and greater proportion having at least 1 BMI, routine health exam, or diabetes-related laboratory. Compared to insufficient cases, a greater proportion also had a known race or ethnicity (87.4%, n=158,172) or were classified as diabetic (7.8%, n=14,044), obese (22.5%, n=40,761), or asthmatic (7.2%, n=13,073).

Within the total naïve EHR sample, those who were classified as nondiabetic consistently had lower health care use than those who were classified as diabetic (Table 2). This pattern was disrupted in the sufficiency sample, where a greater proportion of patients who are nondiabetic had at least 1 routine medical exam (46.7%, n=77,927 vs 32.3%, n=4536 of patients who are diabetic) and almost all patients had a recorded BMI regardless

of diabetes status. Within both the naïve and sufficiency samples, a lower proportion of those classified as nondiabetic were of Black or Latino race or ethnicity and were classified as having asthma or obesity compared to those classified as diabetic.

Table . Descriptive summary of the New York University patient population by diabetes status[a,b] .

Variables	Naïve—diabetes status		Sufficient—diabetes status[a]	
	Nondiabetic (n=440,568)	Diabetic (n=14,044)	Nondiabetic (n=166,992)	Diabetic (n=14,044)
Age (30 - 44 years), n (%)	273,216 (62.0)	10,843 (77.2)	104,005 (62.3)	10,843 (77.2)
Sex (male), n (%)	166,204 (37.7)	5764 (41.0)	65,240 (39.1)	5764 (41.0)
Medicaid insurance, n (%)	96,760 (22.0)	4219 (30.0)	37,759 (22.6)	4219 (30.0)
Raw race or ethnicity, n (%)				
White	185,119 (42.0)	5106 (36.4)	73,996 (44.3)	5106 (36.4)
Black	43,289 (9.8)	2220 (15.8)	18,847 (11.3)	2220 (15.8)
Latino	59,636 (13.5)	3353 (23.9)	27,479 (16.5)	3353 (23.9)
Asian or Pacific Islander	34,174 (7.8)	1088 (7.7)	13,227 (7.9)	1088 (7.7)
Other	31,378 (7.1)	1147 (8.2)	11,709 (7.0)	1147 (8.2)
Missing	86,972 (19.7)	1130 (8.0)	21,734 (13.0)	1130 (8.0)
Imputed race or ethnicity, n (%)				
White	231,412 (52.5)	5645 (40.2)	86,629 (51.9)	5645 (40.2)
Black	55,268 (12.5)	2441 (17.4)	21,829 (13.1)	2441 (17.4)
Latino	82,827 (18.8)	3852 (27.4)	33,696 (20.2)	3852 (27.4)
Asian or Pacific Islander	47,887 (10.9)	1283 (9.1)	16,599 (9.9)	1283 (9.1)
Other	23,174 (5.3)	823 (5.9)	8239 (4.9)	823 (5.9)
Recorded BMI, n (%)	354,043 (80.4)	13,860 (98.7)	163,127 (97.7)	13,860 (98.7)
Obese, n (%)	73,412 (16.7)	6168 (43.9)	34,593 (20.7)	6168 (43.9)
Encounters[c], mean (SD)	14.29 (21.16)	44.90 (54.27)	22.27 (28.19)	44.90 (54.27)
Duration[d], mean (SD)	14.29 (21.16)	2.87 (2.06)	2.26 (1.97)	2.87 (2.06)
Routine medical exam, n (%)	110,713 (25.1)	4536 (32.3)	77,927 (46.7)	4536 (32.3)
Diabetes-related laboratory[b], n (%)	193,480 (43.9)	11,928 (84.9)	112,752 (67.5)	11,928 (84.9)
Neighborhood coverage[e], n (%)				
<10%	76,463 (17.4)	3100 (22.1)	28,143 (16.9)	3100 (22.1)
10%-<20%	139,333 (31.6)	4574 (32.6)	57,185 (34.2)	4574 (32.6)
20%-<30%	158,890 (36.1)	4186 (29.8)	57,597 (34.5)	4186 (29.8)
30%-<40%	65,882 (15.0)	2184 (15.6)	24,067 (14.4)	2184 (15.6)
Asthma, n (%)	17,339 (3.9)	1901 (13.5)	11,172 (6.7)	1901 (13.5)

[a]Sufficient cases defined as those with at least 1 encounter with an endocrinology review of systems or those who were classified as diabetic through the computable phenotype of having at least 2 encounter diagnoses for diabetes, 1 encounter diagnosis and at least 2 elevated A_{1c} laboratory results ≥6.5%, or at least 1 antidiabetes prescription medication.

[b] Including all A_{1c}, random blood glucose, and fasting blood glucose laboratory results.

[c]Number of encounters.

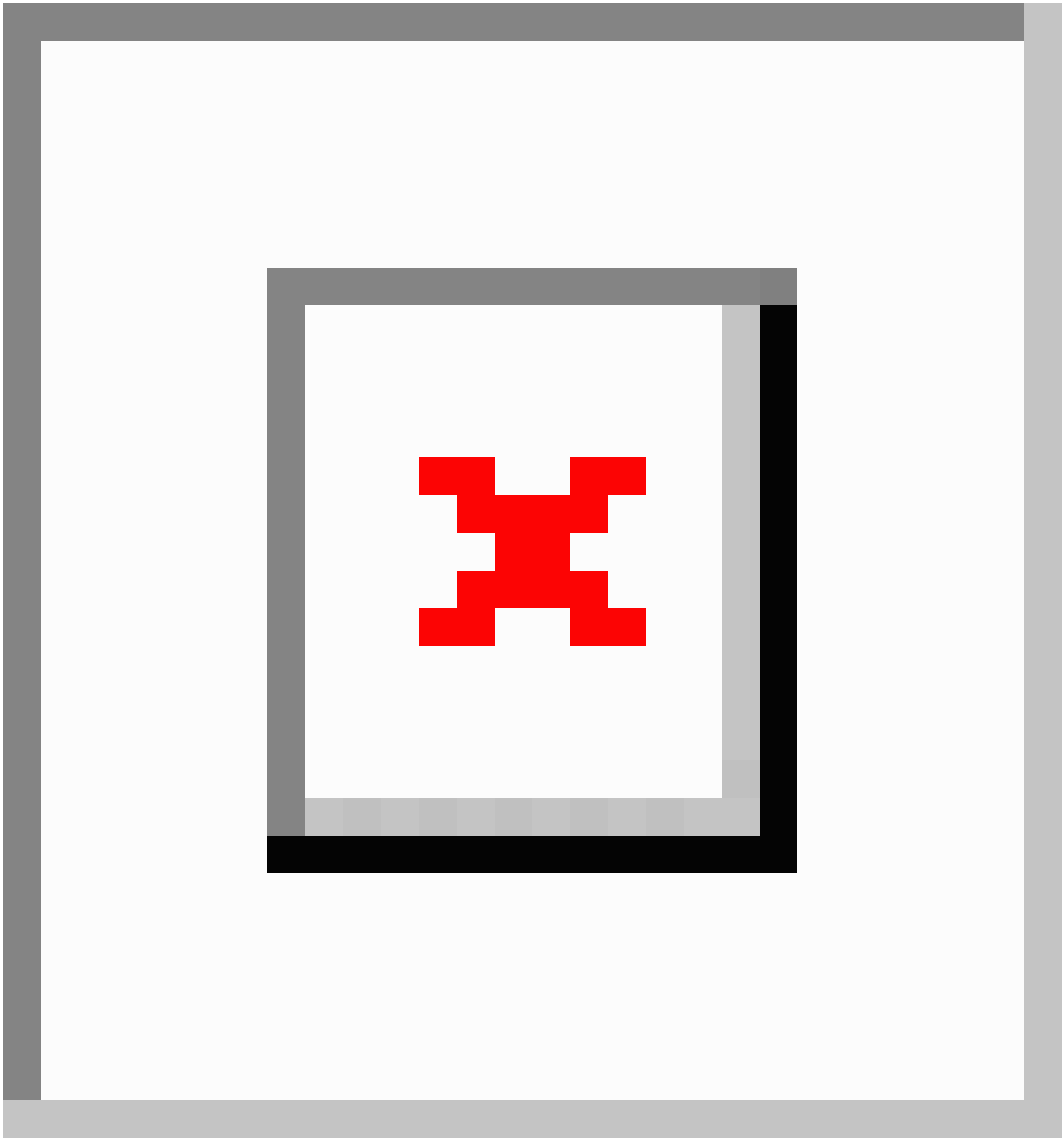
[d]Number of years in the health system.

[e]Proportion of individuals residing in the Public Use Microdata Area (PUMA) neighborhood who are present within the electronic health record system.

Estimated Associations With Race or Ethnicity

Estimated associations between race or ethnicity and diabetes were comparable across BRFSS and NHANES surveys (Figure 1). In both the survey and EHR-based analyses, respondents who were Black or Latino had significantly higher odds of diabetes compared to White respondents, controlling for age and sex. The naive EHR-based OR estimate for Latino patients was 26% higher than the BRFSS point estimate (OR_{Naive} 1.93, 95% CI 1.85 - 2.01 vs OR_{BRFSS} 1.53, 95% CI 1.32 - 1.77). Sufficiency, IPW, and causal methods reduced this association, with point estimates falling within the 95% CIs and relative differences below 15% compared to both health survey estimates. BRFSS and NHANES respondents who were Asian did not have significantly higher odds of diabetes compared to White respondents, and CIs were wide due to the small sample sizes of this subgroup. Within the EHR analyses, patients who were Asian had a significant 11% - 26% increased odds of diabetes.

Figure 1. Odds ratios for diabetes by race or ethnicity and asthma, EHR-based estimates versus health survey estimates. BRFSS: Behavioral Risk Factor Surveillance System; DAG: directed acyclic graph; EHR: electronic health record; IPW: inverse probability weighting; NHANES: National Health and Nutrition Examination Survey.



Estimated Associations With Asthma

In the BRFSS and NHANES analyses, having asthma was associated with an approximate 20% - 40% increased odds of diabetes after controlling for demographics and obesity (OR_{BRFSS} 1.23, 95% CI 1.09 - 1.40 and OR_{NHANES} 1.38, 95% CI 1.01 - 1.91). In the naive EHR analysis, asthma was strongly associated with diabetes, with those with asthma estimated to have 3 times the odds of diabetes as those without asthma after controlling for demographics and obesity (95% CI 2.86 - 3.18). This association was reduced in the sufficiency sample and the IPW-weighted sample, with corrected OR estimates falling within the 95% CI of the NHANES estimate and having an approximate 30% - 50% relative difference from the health survey point estimates. The association between asthma and diabetes was further reduced in the DAG model (OR 1.42, 95% CI 1.34 - 1.51), falling within the 95% CIs of both survey-based estimates and having a 3% relative difference from the NHANES point estimate and a 15% relative difference from the BRFSS point estimate (Figure 1). Sensitivity analyses varying the definition for sufficiency (Table S3 in Multimedia Appendix 1), varying the BRFSS inclusion criteria, or varying the NHANES diabetes definition produced similar patterns in these results (Table S1 in Multimedia Appendix 1).

Discussion

Principal Findings

Our analysis explored the potential impact of information biases on observed associations of diabetes risk factors within an EHR sample of young adults. We estimated naïve associations on the full patient sample and then attempted to address information biases using missing data and causal inference frameworks. Biases were apparent in the naïve association between asthma and diabetes, which was significantly higher than the health survey-based estimates used as a benchmark for the expected association. All EHR-based methods resulted in estimated associations between race or ethnicity and diabetes that were largely comparable to health survey-based estimates. Those who were observed to have diabetes had greater health care use than those who were classified as nondiabetic, which could reflect an information bias where those with a greater number of health care encounters may have been more likely to have documentation of an underlying diabetes or asthma diagnosis. Attempting to address this bias through missing data or causal frameworks reduced the estimated associations between asthma and diabetes, with the causal framework having the best performance in producing an estimate comparable to the benchmarks.

Within naïve analyses, the observed age and sex-adjusted ORs for diabetes among Latino patients appeared slightly inflated compared to health survey estimates. Using IPW, or controlling the number of health care encounters produced ORs that were closer to health survey estimates. Despite the potential to introduce collider bias, subsetting to those with sufficient data also led to estimated associations that were more similar to health survey estimates than the naïve method. Prior research has demonstrated that Latino and Black individuals may have increased screening for diabetes while Asian individuals may have decreased screening compared to White individuals [34]. These disparities in screening practices may partially explain the observed patterns within the EHR estimates. Increased likelihood of screening would produce a positive bias while decreased likelihood of screening would produce a negative bias in naïve EHR associations. The tested methods may have helped correct for this bias by controlling or restricting based on factors associated with the likelihood of diabetes screening, resulting in decreases in the Latino ORs and increases in the Asian OR relative to naïve estimates.

Consistent with prior research, the naïve association between 2 EHR-observed conditions, asthma and diabetes, was substantially positively biased relative to health survey estimates and prior studies from the literature [5,6,14,21]. Imposing data sufficiency criteria to subset to those for whom we had greater confidence in an accurate diabetes classification helped to reduce disparities in health care use by observed diabetes status. In addition, while all correction methods greatly reduced the estimated association between these 2 chronic conditions, the DAG method had the largest impact on this estimate, suggesting use was a strong confounder of the association between these observed health outcomes. Since the presence of either chronic condition may cause increased health care use, it is also possible that controlling for this variable induced a small collider bias, producing an estimate that was lower than the sufficiency or IPW estimates. However, prior research suggests that the magnitude of this collider bias would be small relative to the confounding bias imparted when not controlling for this variable [14]. All corrected EHR estimates were still higher than the health survey point estimates, suggesting that the NYU patient population may not be generalizable, that there are residual biases in these estimates, or that there are other inherent differences between EHR and survey-based estimates. For example, individuals interacting with the NYU hospital system may be sicker and more likely to have multiple chronic conditions than those who receive care at independent primary care practices and conditioning on sicker patients could potentially introduce a collider bias if presence within the hospital system was a common effect of these conditions. This selection bias was not addressed in this work and is an important avenue for future research.

Limitations

This study applied 2 bias correction frameworks to a large, diverse patient population and these findings can inform broader discussions on addressing misclassification of disease outcomes within epidemiologic studies using EHR data. However, there are limitations to these analyses. Importantly, internal or external validation samples were not available to inform computable phenotype sensitivity or specificity for these methods [8], and the sufficiency and IPW models relied on the strong assumption that sufficient cases had no misclassification in diabetes status. Sensitivity analyses were used to test this assumption, but it is possible that imposing sufficiency determinations generated a collider bias by selecting sicker individuals or those with diabetes [5], which could explain why these methods found higher odds of diabetes among those with asthma compared to the DAG method. That said, internal and external validation samples are often costly or time-intensive to obtain, so these

methods offer an imperfect, yet feasible, solution within resource-constrained environments. In addition, methods focused on addressing differential misclassification of health outcomes, but there is potential for misclassification within other covariates. The hypothesized DAG likely represents a simplified depiction of information biases within these data. In particular, a large proportion of patients had an unknown race or ethnicity, and the BISG imputation methods used may have differential performance by race or ethnicity or marital status [35]. Overall, determining the accuracy of the EHR-based estimates was challenging due to the wide CIs for survey-based estimates, but EHR-based associations had greatly improved precision due to the diversity and larger sample size of these data.

Additionally, comparisons were made to estimates from 2 health surveys, which have distinct biases that were not addressed in this analysis. The BRFSS is limited to self-reported health outcomes, which can be prone to misclassification. However, evidence suggests that self-reported diabetes status may have good validity relative to other chronic conditions [4,36,37]. Physical measurements from NHANES may improve the classification of diabetes status among those who are unaware or undiagnosed [15]. However, the smaller sample size of this survey requires multiyear pooled analyses, which can be biased by changes in screening or diagnostic criteria over time. For example, in 2015, the American Diabetes Association lowered the recommended BMI screening threshold for Asian American individuals to better account for the differential risk of diabetes at equivalent BMI levels, which could change the burden of undetected diabetes within this subgroup across time [38]. The complementary strengths of these 2 data sources may help to remedy these unaddressed biases. However, differences in the targets for inference could have contributed to the observed differences between the EHR and survey estimates. Survey estimates reflect national data, which may not be transportable to this New York City patient population. Although local versions of these health surveys are available, sample sizes were too small to produce reliable associations. Covariate definitions also varied across data sources. Importantly, individual-level income was available within the survey data but was unavailable in the EHR data. The use of neighborhood-level poverty likely resulted in residual confounding in all EHR-based ORs for asthma, which may have contributed to the positive relative differences compared to survey-based estimates.

Conclusions

EHRs offer a compelling data source for public health research; however, differential misclassification of disease status has the potential to bias the results of these studies. Methods to control for factors that affect misclassification using a causal framework, particularly when an informed presence bias can be hypothesized to strongly confound the exposure-outcome relationship, should be considered to help produce valid estimates of risk factor associations. The next steps include applying these methods to additional exposure-outcome relationships and incorporating the longitudinal nature of EHR data to assess causal relationships between chronic conditions.

Authors' Contributions

SC, RA, and LET developed the study concept and design. SC analyzed the data and wrote the paper. All authors reviewed and edited the paper for intellectual content. All authors reviewed and approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional details on methods and supplementary tables and figures.
[DOCX File, 92 KB - medinform_v12i1e58085_app1.docx]

References

1. Perlman SE. Use and visualization of electronic health record data to advance public health. Am J Public Health 2021 Feb;111(2):180-182. [doi: 10.2105/AJPH.2020.306073] [Medline: 33439707]
2. Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. J Med Syst 2018 Sep 29;42(11):214. [doi: 10.1007/s10916-018-1075-6] [Medline: 30269237]
3. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 2018 Jul;1:53-68. [doi: 10.1146/annurev-biodatasci-080917-013315] [Medline: 31218278]
4. Merrill RM, Richardson JS. Validity of self-reported height, weight, and body mass index: findings from the National Health and Nutrition Examination Survey, 2001-2006. Prev Chron Dis 2009 Oct;6(4):A121. [Medline: 19754997]
5. Bower JK, Patel S, Rudy JE, Felix AS. Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. Curr Epidemiol Rep 2017 Dec;4(4):346-352. [doi: 10.1007/s40471-017-0130-z] [Medline: 31223556]
6. Phelan M, Bhavsar NA, Goldstein BA. Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. EGEMS (Wash DC) 2017 Dec 6;5(1):22. [doi: 10.5334/egems.243] [Medline: 29930963]

7. Peskoe SB, Arterburn D, Coleman KJ, Herrinton LJ, Daniels MJ, Haneuse S. Adjusting for selection bias due to missing data in electronic health records-based research. Stat Methods Med Res 2021 Oct;30(10):2221-2238. [doi: 10.1177/09622802211027601] [Medline: 34445911]
8. Grace YY, Delaigle A, Gustafson P. Handbook of Measurement Error Models: CRC Press; 2021. [Medline: 1351588591]
9. Padilla MA, Divers J, Vaughan LK, Allison DB, Tiwari HK. Multiple imputation to correct for measurement error in admixture estimates in genetic structured association testing. Hum Hered 2009;68(1):65-72. [doi: 10.1159/000210450] [Medline: 19339787]
10. Sayon-Orea C, Moreno-Iribas C, Delfrade J, et al. Inverse-probability weighting and multiple imputation for evaluating selection bias in the estimation of childhood obesity prevalence using data from electronic health records. BMC Med Inform Decis Mak 2020 Jan 20;20(1):9. [doi: 10.1186/s12911-020-1020-8] [Medline: 31959164]
11. Young JC, Conover MM, Funk MJ. Measurement error and misclassification in electronic medical records: methods to mitigate bias. Curr Epidemiol Rep 2018 Dec;5(4):343-356. [doi: 10.1007/s40471-018-0164-x] [Medline: 35633879]
12. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. Epidemiology 2011 Jul;22(4):589-597. [doi: 10.1097/EDE.0b013e3182117c85] [Medline: 21487295]
13. Hernán MA, Cole SR. Invited commentary: causal diagrams and measurement bias. Am J Epidemiol 2009 Oct 15;170(8):959-962. [doi: 10.1093/aje/kwp293] [Medline: 19755635]
14. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for informed presence bias due to the number of health encounters in an electronic health record. Am J Epidemiol 2016 Dec 1;184(11):847-855. [doi: 10.1093/aje/kww112] [Medline: 27852603]
15. National diabetes statistics report, 2020. : Centers for Disease Control and Prevention, US Department of Health and Human Services; 2020. URL: https://stacks.cdc.gov/view/cdc/85309 [accessed 2024-09-24]
16. Bullard KM, Cowie CC, Lessem SE, et al. Prevalence of diagnosed diabetes in adults by diabetes type - United States, 2016. MMWR Morb Mortal Wkly Rep 2018 Mar 30;67(12):359-361. [doi: 10.15585/mmwr.mm6712a2] [Medline: 29596402]
17. Geiss LS, Wang J, Cheng YJ, et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. JAMA 2014 Sep 24;312(12):1218-1226. [doi: 10.1001/jama.2014.11494] [Medline: 25247518]
18. Gregg EW, Li Y, Wang J, et al. Changes in diabetes-related complications in the United States, 1990-2010. N Engl J Med 2014 Apr 17;370(16):1514-1523. [doi: 10.1056/NEJMoa1310799] [Medline: 24738668]
19. Link CL, McKinlay JB. Disparities in the prevalence of diabetes: is it race/ethnicity or socioeconomic status? Results from the Boston area community health (BACH) survey. Ethn Dis 2009;19(3):288-292. [Medline: 19769011]
20. Cheng YJ, Kanaya AM, Araneta MRG, et al. Prevalence of diabetes by race and ethnicity in the United States, 2011-2016. JAMA 2019 Dec 24;322(24):2389-2398. [doi: 10.1001/jama.2019.19365] [Medline: 31860047]
21. Torres RM, Souza MDS, Coelho ACC, de Mello LM, Souza-Machado C. Association between asthma and type 2 diabetes mellitus: mechanisms and impact on asthma control-a literature review. Can Respir J 2021;2021:8830439. [doi: 10.1155/2021/8830439] [Medline: 33520042]
22. Imai K, Khanna K. Improving ecological inference by predicting individual ethnicity from voter registration records. Polit anal 2016;24(2):263-272. [doi: 10.1093/pan/mpw001]
23. Avramovic S, Alemi F, Kanchi R, et al. US veterans administration diabetes risk (VADR) national cohort: cohort profile. BMJ Open 2020 Dec 4;10(12):e039489. [doi: 10.1136/bmjopen-2020-039489] [Medline: 33277282]
24. Chen T, Li W, Zambarano B, Klompas M. Small-area estimation for public health surveillance using electronic health record data: reducing the impact of underrepresentation. BMC Public Health 2022 Aug 9;22(1):1515. [doi: 10.1186/s12889-022-13809-2] [Medline: 35945537]
25. Mueller NT, Koh WP, Odegaard AO, Gross MD, Yuan JM, Pereira MA. Asthma and the risk of type 2 diabetes in the singapore Chinese health study. Diabetes Res Clin Pract 2013 Feb;99(2):192-199. [doi: 10.1016/j.diabres.2012.11.019] [Medline: 23260853]
26. Kuruvilla ME, Vanijcharoenkarn K, Shih JA, Lee FEH. Epidemiology and risk factors for asthma. Respir Med 2019 Mar;149:16-22. [doi: 10.1016/j.rmed.2019.01.014] [Medline: 30885424]
27. Winer N, Sowers JR. Epidemiology of diabetes. J Clin Pharmacol 2004 Apr;44(4):397-405. [doi: 10.1177/0091270004263017] [Medline: 15051748]
28. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res 2008 Aug;43(4):1424-1441. [doi: 10.1111/j.1475-6773.2007.00822.x] [Medline: 18756617]
29. Behavioral risk factor surveillance system survey data. : Centers for Disease Control and Prevention, US Department of Health and Human Services; 2019. URL: https://www.cdc.gov/brfss/annual_data/annual_2019.html
30. National health and nutrition examination survey data. : US Department of Health and Human Services; 2013. URL: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx
31. Antonio-Villa NE, Fernández-Chirino L, Vargas-Vázquez A, Fermín-Martínez CA, Aguilar-Salinas CA, Bello-Chavolla OY. Prevalence trends of diabetes subgroups in the United States: a data-driven analysis spanning three decades from

NHANES (1988-2018). J Clin Endocrinol Metab 2022 Feb 17;107(3):735-742. [doi: 10.1210/clinem/dgab762] [Medline: 34687306]
32. National Health and Nutrition Examination Survey, 2017–March 2020 prepandemic file: sample design, estimation, and analytic guidelines. Data evaluation and methods research. : US Department of Health and Human Services; 2022. URL: https://stacks.cdc.gov/view/cdc/115434 [accessed 2024-09-24]
33. Behavioral risk factor surveillance system (BRFSS): complex sampling weights and preparing 2019 BRFSS module data for analysis. Centers for Disease Control and Prevention, US Department of Health and Human Services. 2020 URL: https://www.cdc.gov/brfss/annual_data/2019/pdf/Complex-Smple-Weights-Prep-Module-Data-Analysis-2019-508.pdf
34. Tran L, Tran P, Tran L. A cross-sectional analysis of racial disparities in US diabetes screening at the national, regional, and state level. J Diabetes Complications 2020 Jan;34(1):107478. [doi: 10.1016/j.jdiacomp.2019.107478] [Medline: 31706806]
35. Elliott MN, Morrison PA, Fremont A, McCaffrey DF, Pantoja P, Lurie N. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Serv Outcomes Res Method 2009 Jun;9(2):69-83. [doi: 10.1007/s10742-009-0047-1]
36. Bowlin SJ, Morrill BD, Nafziger AN, Jenkins PL, Lewis C, Pearson TA. Validity of cardiovascular disease risk factors assessed by telephone survey: the behavioral risk factor survey. J Clin Epidemiol 1993 Jun;46(6):561-571. [doi: 10.1016/0895-4356(93)90129-o] [Medline: 8501483]
37. Bowlin SJ, Morrill BD, Nafziger AN, Lewis C, Pearson TA. Reliability and changes in validity of self-reported cardiovascular disease risk factors using dual response: the behavioral risk factor survey. J Clin Epidemiol 1996 May;49(5):511-517. [doi: 10.1016/0895-4356(96)00010-8] [Medline: 8636724]
38. Hsu WC, Araneta MRG, Kanaya AM, Chiang JL, Fujimoto W. BMI cut points to identify at-risk Asian Americans for type 2 diabetes screening. Diabetes Care 2015 Jan 1;38(1):150-158. [doi: 10.2337/dc14-2391]

Abbreviations

BISG: Bayesian Improved Surname Geocoding
BRFSS: Behavioral Risk Factor Surveillance System
DAG: directed acyclic graph
DM: diabetes
EHR: electronic health record
IPW: inverse probability weighting
NHANES: National Health and Nutrition Examination Survey
NYU: New York University
OR: odds ratio

Edited by J Hefner; submitted 05.03.24; peer-reviewed by C Drake, PY Chan, T Mitsuhashi; revised version received 10.07.24; accepted 10.07.24; published 01.10.24.

Please cite as:
Conderino S, Anthopolos R, Albrecht SS, Farley SM, Divers J, Titus AR, Thorpe LE
Addressing Information Biases Within Electronic Health Record Data to Improve the Examination of Epidemiologic Associations With Diabetes Prevalence Among Young Adults: Cross-Sectional Study
JMIR Med Inform 2024;12:e58085
URL: https://medinform.jmir.org/2024/1/e58085
doi:10.2196/58085

Transforming Primary Care Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study

Mathilde Fruchart[1], MSc; Paul Quindroit[1], PhD, RN; Chloé Jacquemont[2], MSc; Jean-Baptiste Beuscart[1], MD, PhD; Matthieu Calafiore[1,2], MD, PhD; Antoine Lamer[1,3], PhD

1

2

3

Corresponding Author:
Mathilde Fruchart, MSc

Abstract

Background: Patient-monitoring software generates a large amount of data that can be reused for clinical audits and scientific research. The Observational Health Data Sciences and Informatics (OHDSI) consortium developed the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to standardize electronic health record data and promote large-scale observational and longitudinal research.

Objective: This study aimed to transform primary care data into the OMOP CDM format.

Methods: We extracted primary care data from electronic health records at a multidisciplinary health center in Wattrelos, France. We performed structural mapping between the design of our local primary care database and the OMOP CDM tables and fields. Local French vocabularies concepts were mapped to OHDSI standard vocabularies. To validate the implementation of primary care data into the OMOP CDM format, we applied a set of queries. A practical application was achieved through the development of a dashboard.

Results: Data from 18,395 patients were implemented into the OMOP CDM, corresponding to 592,226 consultations over a period of 20 years. A total of 18 OMOP CDM tables were implemented. A total of 17 local vocabularies were identified as being related to primary care and corresponded to patient characteristics (sex, location, year of birth, and race), units of measurement, biometric measures, laboratory test results, medical histories, and drug prescriptions. During semantic mapping, 10,221 primary care concepts were mapped to standard OHDSI concepts. Five queries were used to validate the OMOP CDM by comparing the results obtained after the completion of the transformations with the results obtained in the source software. Lastly, a prototype dashboard was developed to visualize the activity of the health center, the laboratory test results, and the drug prescription data.

Conclusions: Primary care data from a French health care facility have been implemented into the OMOP CDM format. Data concerning demographics, units, measurements, and primary care consultation steps were already available in OHDSI vocabularies. Laboratory test results and drug prescription data were mapped to available vocabularies and structured in the final model. A dashboard application provided health care professionals with feedback on their practice.

(JMIR Med Inform 2024;12:e49542) doi:10.2196/49542

KEYWORDS

data reuse; Observational Medical Outcomes Partnership; common data model; data warehouse; reproducible research; primary care; dashboard; electronic health record; patient tracking system; patient monitoring; EHR; primary care data

Introduction

The digitalization of health care organizations has made it possible to automatically collect and reuse data from electronic health records (EHRs) for care, administrative, and research purposes [1]. Data reuse generally relies on extracting data from source databases, formatting and normalizing it in a data warehouse [2-5]. Over the last few years, hospital-based data warehouses have started to provide comprehensive overviews of patient management during a hospital stay or on a hospital ward. However, these data warehouses do not contain data on primary care or other data not related to the hospital stay. These data cover first-line services—outpatient care provided in local practices, including general practice, community pharmacy, dental care, and optometry [6,7]. Additionally, data from all individuals covered by the French national health insurance scheme are anonymously and prospectively included in the

national claims database [8]. These data are used for reimbursement purposes and are not clinical.

The reuse of EHR data is now a major topic of interest for hospital care [9,10] and primary care [4,11]. Several research groups have retrospectively reused primary care data on patients with neuromuscular diseases [12], diabetes [4,13], dermatological diseases [14,15], lung diseases [16], cancer [17], or urinary tract infections [18] or on older adult patients [19]. Several national projects aim to collect primary care data on a routine basis. The main primary care projects are the Clinical Practice Research Datalink and The Health Improvement Network in the United Kingdom [20,21], the Veterans Administration data warehouses in the United States [22], and the Canadian Primary Care Sentinel Surveillance Network in Canada [23]. Nevertheless, the reuse of health care data (and especially primary care data) faces many challenges [9,10,24-27]. The abovementioned projects collect data from millions of patients by implementing local data models. A UK project uses a national database (BioBank) to standardize vaccination data into a common data model (CDM) format [28].

The heterogeneous structure of the data and the use of country- and facility-specific vocabularies create barriers to the implementation of multicenter studies and the sharing of data, methods, and results. Initiatives such as those proposed by the Observational Health Data Sciences and Informatics (OHDSI) consortium seek to (1) standardize data structure and vocabularies; (2) promote reproducible research and collaboration; and (3) share methods, tools, and results [29,30]. The OHDSI has notably developed the Observational Medical Outcomes Partnership (OMOP) CDM [31] and provides standard data structures and vocabularies that are independent of individual software developers and countries [32].

Hospital and claims databases have already undergone the mapping process to adopt the OMOP CDM format [33-35]. Nevertheless, primary care data, specifically general practice data, which serve as a valuable addition to hospital and claims databases, are still rarely being integrated into the OMOP CDM format.

Data standardization might facilitate the development of common tools for health care professionals, such as activity dashboards and software for managing multicenter research projects. Hence, the primary objective of this study (part of the Primary Care Data Warehouse [PriCaDa] project) was to transform primary care data into the OMOP CDM format.

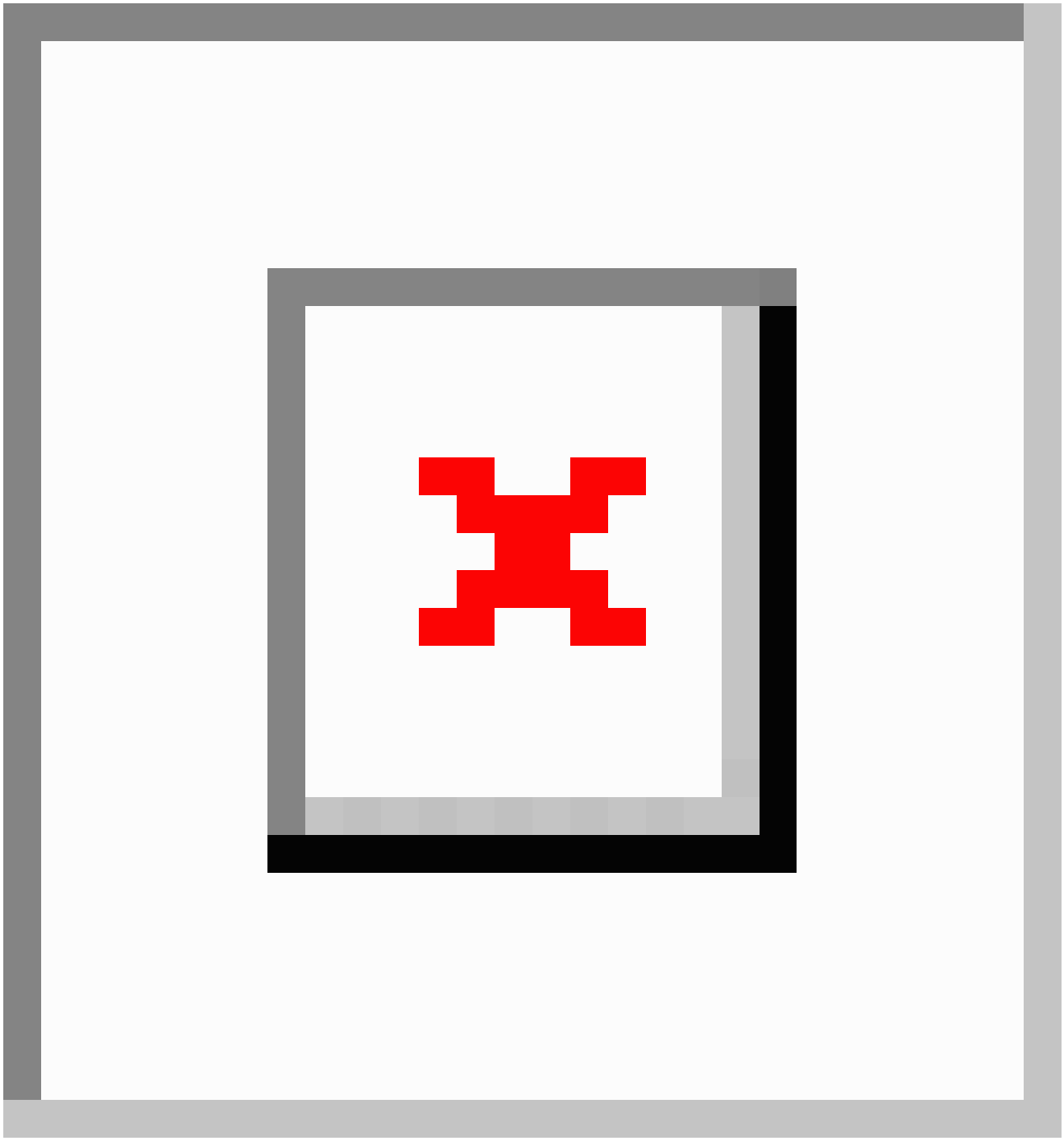
Methods

Overview

The goal of the PriCaDa project is to reuse primary care data and to provide an overview of the practices of health care professionals working in primary care (family physicians, nurses, and pharmacists).

In this study, we used primary care data from a multidisciplinary health center (MHC) located in the town of Wattrelos, France, in collaboration with the software company that produced the MHC's EHR software (Weda). Weda ranks third in terms of sales volume, boasting over 20,000 health care professionals in France as users [36]. We used an extract-transform-load (ETL) process to create a data warehouse (Figure 1), using Python and SQL scripts. The data warehouse was stored in a PostgreSQL database, using version 5.4 of the OMOP CDM [31]. The first step in the ETL process was to extract, classify, and normalize entities from the XML files. The database was then scanned and assessed with the OHDSI tool White Rabbit [37] to create a report on each raw table, attribute, and data type. The OHDSI tool Rabbit In A Hat was then used to implement the specifications of the structural mapping, which aims to associate each local table and local column to the corresponding OMOP nomenclature [38]. Then, transformation steps normalize the structure and the semantic of the data source to the OMOP CDM format. The final step is to load the data into the OMOP CDM. The development of a dashboard was proposed for a practical application.

Figure 1. Pipeline of the primary care data transformation into the OMOP CDM. CDM: Common Data Model; ETL: extract-transform-load; OMOP: Observational Medical Outcomes Partnership.



Ethical Considerations

The study was approved by the staff at the MHC in Wattrelos and the management of the Weda software house. Furthermore, the study was registered with the French National Data Protection Commission (*Commission nationale de l'informatique et des libertés*; reference 2022 - 203). In line with the French legislation on retrospective studies of anonymized data from routine clinical practice, approval by an institutional review board was not required [39]. All the data collection and analysis methods complied with relevant guidelines and regulations. Patient confidentiality was maintained at all times, and all the data were anonymized before extraction. The Weda software exports data with an identifier for each patient and consultation. This identifier is replaced during the transformations with a unique artificial ID. For this unique ID, sequences provided by the database management system (ie, PostgreSQL) were generally used to create an autoincrement primary key.

Data Extraction

Following approval by the MHC's staff, Weda exported an XML-based hierarchical structure file for each patient. A primary care consultation occurs in 4 steps. At each step, the family physician collects data on the patient's health status. In step one (arrival), the reason for consultation is recorded as free text. In step two (the interview), the physician records clinical signs and symptoms as free text. In step three (the examination),

the physician makes biometric measurements and enters the data in a structured way in the appropriate fields of the software (variables, values, and units). Step four (the outcome) includes the diagnosis (if made, as free text), any drug prescriptions (documented with the *Code Identifiant de Spécialité* [CIP; the French national drug code]), any referrals (as free text in a PDF letter), and information on vaccinations. The parts of the Weda software dedicated to these steps depend on the physician's habits; some physicians put all observations into a single field.

Furthermore, the family physician receives laboratory test results from medical laboratories and clinical reports from other specialist physicians. The test variables differ from one medical laboratory to another. For example, one laboratory might use the term "creatinine" and a specific unit, whereas another might use "create" and a different unit for the same variable. These reports were documented as free text.

Transformation Steps

Semantic Mapping

The goal of semantic mapping was to link local vocabularies to standard vocabularies that have been defined by the OHDSI community and are available in an web-based tool [32,40]. The difficulty of the mapping depends on the vocabularies available in the source data. We defined four difficulty levels: level 1 for local vocabulary items that already belonged to the standard OHDSI vocabulary; level 2 for local, nonstandard vocabulary items for which the mappings to the standard OHDSI vocabulary already existed; level 3 for local vocabulary items that did not correspond to the OHDSI vocabulary but were in a structured format (manual mapping was necessary); and level 4 for local vocabulary items that did not correspond to the OHDSI vocabulary and were in an unstructured format (free text). Natural language processing (NLP) techniques, such as fuzzy matching algorithms, SpaCy tools, and regular expressions, can be used for concept identification and information extraction. The free text (eg, reasons for consultation, symptoms, diagnosis, and clinical reports) was sometimes heterogeneous and unstructured. Indeed, a single free-text record can consolidate various types of information, including the reason for a consultation, the patient's clinical signs, and their diagnosis, all of which may vary based on the practices of the individual family physician.

For difficulty levels 2 and 3, the data were mapped manually and independently by two experts (MF and CJ). A κ score was computed as a statistical metric for assessing the degree of consensus among the annotators tasked with assigning labels to data. It gauges the degree to which the annotators' assessments align, while also considering the potential for chance agreement. The score ranges from 0 to 1 (1 for perfect agreement among annotators). A κ score below 0.4 indicates weak agreement, a κ score above 0.6 (60%) suggests moderate agreement, and a κ score exceeding 0.8 (80%) signifies strong agreement [41]. Any disagreements were settled by a third expert (PQ). This mapping provided correspondences between the local vocabulary items and standard vocabulary items. When necessary, new primary care concepts not available in the OHDSI vocabularies were loaded into the CONCEPT table. The Logical Observation Identifiers Names and Codes vocabulary was used to map laboratory test variables. The source laboratory test concepts contained the name of the laboratory test variable used by the laboratory, together with the unit. The text referencing the laboratory test variables was cleaned up by grouping equivalent source concepts. Punctuation and special characters were removed, and abbreviations or spelling differences were replaced and grouped under the full name (eg "CRP" was replaced with "C-reactive protein"). Chapter numbers or line numbers at the beginning of a line were removed. Multiple spaces were replaced by a single space, and stop words (eg, "of," "to," or "an") were removed. The Systematized Nomenclature of Medicine was used to map biometric variables, with the Unified Code for Units of Measure for units of measurement and the *International Classification of Diseases, 10th Revision* (*ICD-10*) for the patient's medical history. Drug prescriptions were recorded using CIP drug name codes. The CIP code was extracted from the Weda EHR software and mapped to the Anatomical Therapeutic Chemical (ATC) code and then the RxNorm code (ATC to RxNorm mapping is already available in the OMOP CDM).

The new concepts were integrated into the CONCEPT table with a concept identifier greater than 2,000,000,000. When a local concept is mapped, the correspondence with a standard OMOP concept is loaded into the CONCEPT_RELATIONSHIP table (resulting in a link between a local concept identifier and a OMOP standard concept identifier). As a result, the identifiers of standard concepts can be loaded into the x_concept_id column of the corresponding table for the local concept (eg, new local concepts in the MEASUREMENT table that are mapped are associated with the standard concepts loaded into the measurement_concept_id column).

Structural Mapping

The goal of structural mapping was to transform the source structure into the OMOP CDM structure. The mapping comprised two steps. In the first step, the relevant variables required for the OMOP model were selected and normalized in a table format. In the second step, the variables and table structures were transformed to match the OMOP CDM nomenclature. We observed that most of the medical histories were coded as free text, rather than nomenclature items. The DIAGNOSIS source table contained the information from the patient interview (clinical signs), the clinical examination (measurements), and the "outcome" parts of the consultation. However, this field contained data related to several types of medical information in text format. The MEASUREMENT table contained information from the "laboratory data" and "biometrics" source tables. Outliers were removed, and primary and foreign keys were identified. New artificial identifiers were created for the primary key in each table.

Data Loading and Quality Assessment

After the semantic and structural transformations, the data were loaded into the OMOP model. We used Achilles, a data quality assessment and visualization tool developed by the OHDSI community. Achilles reports the compliance of the mapping with the constraints of the OMOP CDM (the primary and foreign keys), the vocabulary (ie, the correct choice of concepts corresponding to each table), and the business rules (ie, rules

that ensure data consistency, for example, data chronology respects real life). Based on the Achilles analysis tables, the Atlas server summarizes the results of the data quality assessment in a dashboard [42]. Kahn et al [43] have developed a data quality framework for the secondary use of EHR data integrated into the Atlas quality assessment dashboard. *Conformance* describes "the compliance of the representation of data against internal or external formatting, relational, or computational definitions." *Completeness* computes "features that describe the frequencies of data attributes present in a data set without reference to data values." *Plausibility* describes "the believability or truthfulness of data values." The data quality assessment context *verification* is a strategy "for the source of expectations or comparisons of EHR data based on internal characteristics" [43].

To test the relevance and usability of primary care data in the OMOP CDM, we compared the results of several queries of the OMOP CDM with the results obtained directly from the Weda software. The queries were run by a physician (CJ) who used the software in his clinical practice. To ensure that the patient records could be checked manually, the queries were chosen to keep the resulting number of patients low. Two queries corresponded to the MHC's general activity (eg, the number of patients per family physician), two corresponded to prescription data, and a fifth query corresponded to laboratory test data.

Practical Application

Using RShiny and the *shiny*, *shinyBS*, *shinycssloaders*, *shinydashboard*, *shinyjs*, and *shinyWidgets* libraries, we implemented a dashboard to report the MHC's general activity, prescriptions, and the distribution of the laboratory test results [44].

Results

Data Extraction

The data were extracted in July 2021. The available patient profiles dated back to 1997, and data on clinical measurements, drug prescriptions, and laboratory tests were available from 2013 onward. The extracted data contained 18,395 patient files. Each patient's file contained anonymized demographic information (ie, sex, year of birth, the town or city of residence, and the country of residence), the date of the first consultation, and the name of the family physician with whom the patient was registered. It also contained the patient's medical history (documented with *ICD-10* codes or as free text), information related to the consultation, laboratory test results, and clinical reports from other physicians (also as free text).

Transformation

Semantic Mapping

In all, 17 vocabularies were mapped (Table 1).

All the new concepts were added to the CONCEPT table (n=10,221). These primary care concepts were added alongside the existing concepts in the OMOP model developed by OHDSI. The mapping between local concepts and standard concepts was integrated into the CONCEPT_RELATIONSHIP table (n=9432).

Table . The concept mapping. When a local vocabulary is not given, it means that the concept had to be created.

Feature	Local vocabulary	OMOP[a] vocabulary	Level of mapping difficulty	Concepts, n	Mapped concepts, n/N (%)	Associated records, n/N (%)
Care site	—[b]	Care site	Level 1	1	1/1 (100)	1/1 (100)
Medical histories	*ICD-10* [c]	*ICD-10*	Level 1	83	80/83 (96.4)	2252/2315 (97.3)
Visit	—	Visit	Level 1	2	2/2 (100)	592,226/592,226 (100)
Drug	ATC[d]	RxNorm	Level 2	9070	9100/9100 (100)	684,805/684,805 (100)
Drug	CIP[e] code	ATC	Level 3	9946	9100/9946 (91.5)	684,805/814,772 (84)
Biometric variables	Free text in structured fields	SNOMED[f]	Level 3	243	12/243 (4.9)	172,549/179,337 (96.2)
Laboratory test variables	Free text in structured fields	LOINC[g]	Level 3	2312	170/2312 (7.4)	829,498/941,522 (88.1)
Measurement units	Free text in structured fields	UCUM[h]	Level 3	217	65/217 (30)	863,259/1,120,859 (77)[i]
Patient characteristics	Free text in structured fields	Sex	Level 3	2	2/2 (100)	18,395/18,395 (100)
Outcome (diagnosis)	Free text	—	Level 4	—	—	—
Consultation (reason)	Free text	—	Level 4	—	—	—
Examination (measure taken)	Free text	—	Level 4	—	—	—
Clinical report by another physician	Free text	—	Level 4	—	—	—
Medical history	Free text	—	Level 4	—	—	—
Referrals	Free text	—	Level 4	—	—	—
Supplementary information	Free text	—	Level 4	—	—	—
Vaccine prescription	Free text	—	Level 4	—	—	—
Vaccination	Free text	—	Level 4	—	—	—

[a]OMOP: Observational Medical Outcomes Partnership.

[b]Not applicable.

[c] *ICD-10: International Classification of Diseases, 10th Revision.*

[d]ATC: Anatomical Therapeutic Chemical.

[e]CIP: *Code Identifiant de Spécialité* (the French national drug code).

[f]SNOMED: Systematized Nomenclature of Medicine.

[g]LOINC: Logical Observation Identifiers Names and Codes.

[h]UCUM: Unified Code for Units of Measure.

[i]23% (257,801/1,120,859) not available.

More than 80% (9100/9946, 91.5%) of the drug records and more than 90% (80/83, 96.4%) of the *ICD-10*–coded medical history records were mapped. Less than 4.9% (12/243) of the biometric measurement concepts were mapped; the latter accounted for 96.2% (172,549/179,337) of the records because a small number of concepts were used (eg, weight, height, and heart rate; Table 1). The remaining records were free-text and family physician–dependent variables. The writing style can vary and the same variable can be asked in several ways. For example, to find out whether the patient is a smoker, the family physician can input the information in the family physician–dependent variable "does my patient smoke?" "is he a smoker?" or "do you smoke?"

For level 3 mapping difficulties, drug-related concepts (coded as CIP codes) were mapped in two steps. RxNorm is the standard classification chosen by OHDSI for the OMOP model, and therefore, we had to map our local terminology (ie, CIP)

to RxNorm. First, the CIP codes were mapped to the ATC codes. Second, the ATC codes were mapped to the RxNorm codes, using the correspondences already implemented in the CONCEPT_RELATIONSHIP table.

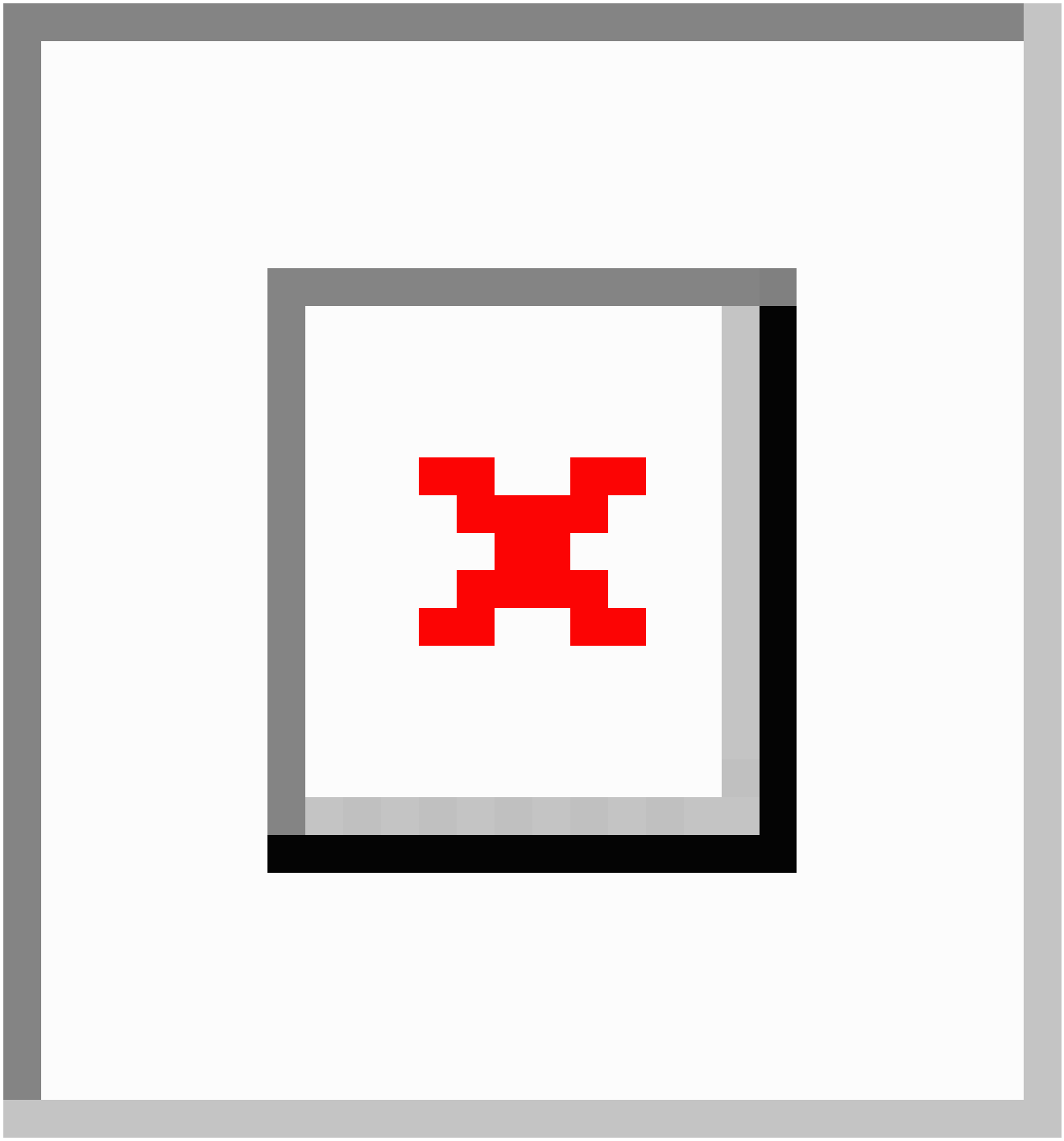
With regard to the laboratory test variables, the cleaning step reduced the number of concepts from 3003 to 2312. We restricted the mapping to the most frequently cited laboratory test concepts in the MEASUREMENT table, with the aim of covering more than 80% of the records. Disagreements over laboratory test concept mapping by experts were resolved by consensus and the involvement of a third annotator. The experts disagreed about 24.3% (37/152) of the mapped concepts, which corresponds to a κ score of 75%.

Structural Mapping

Several XML tables in the source model corresponded to tables in the OMOP CDM format (Figure 2). Information about the patient, the family physician, the profile creation date, and the health care center was stored in the PERSON, PROVIDER, OBSERVATION_PERIOD, and CARE_SITE tables, respectively. The medical history (documented with *ICD-10* codes) was stored in the OBSERVATION table. Each consultation at the health care center corresponded to a record in the VISIT_OCCURRENCE table, identified by a visit_concept_id (A*mbulatory Primary Care Clinic/Center*; concept_id=38004247). Biometric measurements were stored in the MEASUREMENT table, with a concept type corresponding to *EHR physical examination*. Laboratory test results were stored in the MEASUREMENT table, with a concept type *Lab* (concept_id=32,856). The feature measurement_source_type_id distinguishes each source table (laboratory data or biometrics measurements). Each drug prescription was stored in the DRUG_EXPOSURE table. Free-text information collected during the consultation was stored in the NOTE table with an appropriate source vocabulary concept identifier (eg, a note of clinical signs; the reason for the consultation; medical histories; the outcome of the consultation; and, in some cases, the associated diagnosis). Medical reports from a specialist physician not based in the MHC were recorded in the NOTE table (Figure 2).

All the standard concept types used to identify information in the various source tables are detailed in Multimedia Appendix 1; for example, the NOTE table contains information about the reason for the consultation, the patient's medical history, and other medical reports.

Figure 2. Structural mapping of each source variable to the OMOP CDM format. Common Data Model; OMOP: Observational Medical Outcomes Partnership.



Data Loading and Quality Assessments

Records spanning 20 years (592,226 consultations by 18,395 patients) were integrated into the OMOP CDM. The numbers of records per OMOP table before and after the ETL process and the related computing time are reported (Table 2). The CARE_SITE and DRUG_ERA tables were implemented using data transformed into the OMOP format.

The Atlas Server dashboard produced the data distributions for each table. An example of the distribution of blood potassium concentrations is shown in Multimedia Appendix 2.

The Atlas dashboard's overview tab provided a top-level summary of the total number of passes and failures by Kahn category (Figure 3). We found 28 *Conformance* failures that did not respect the specifications of the OMOP CDM. There were 21 *Completeness* failures related to potentially missing data and 23 failures related to *Plausibility* for implausible dates or measurement values. For each of these failures, the results tab (Multimedia Appendix 3) displayed one line per *verification*.

Table . The volume of each table before and after the ETL[a] process and the associated computing time.

OMOP[b] tables	Records before the ETL process, n	Records after ETL process, n	Computing time (s)	Local tables
CARE_SITE	—[c]	1	<0.001	—
DEATH	419	419	<0.001	Death
DRUG_ERA	—	1,084,012	3.75	—
DRUG_EXPOSURE	924,216	814,772	4.54	Drug
LOCATION	19,662	11,433	0.01	Address
MEASUREMENT	1,120,859	1,120,859	225.76	Biometrics and laboratory test
NOTE	2,772,809	2,091,705	4.57	Referrals, consultations, diagnoses, clinical reports from outside the MHC[d], medical history, additional information, and vaccination status
OBSERVATION	64,669	2315	0.16	Medical history
OBSERVATION_PERIOD	—	18,256	0.01	—
PERSON	18,395	18,395	0.04	Patient
PROVIDER	8	8	—	Patient
VISIT_OCCURRENCE	592,227	592,226	1.54	Consultation

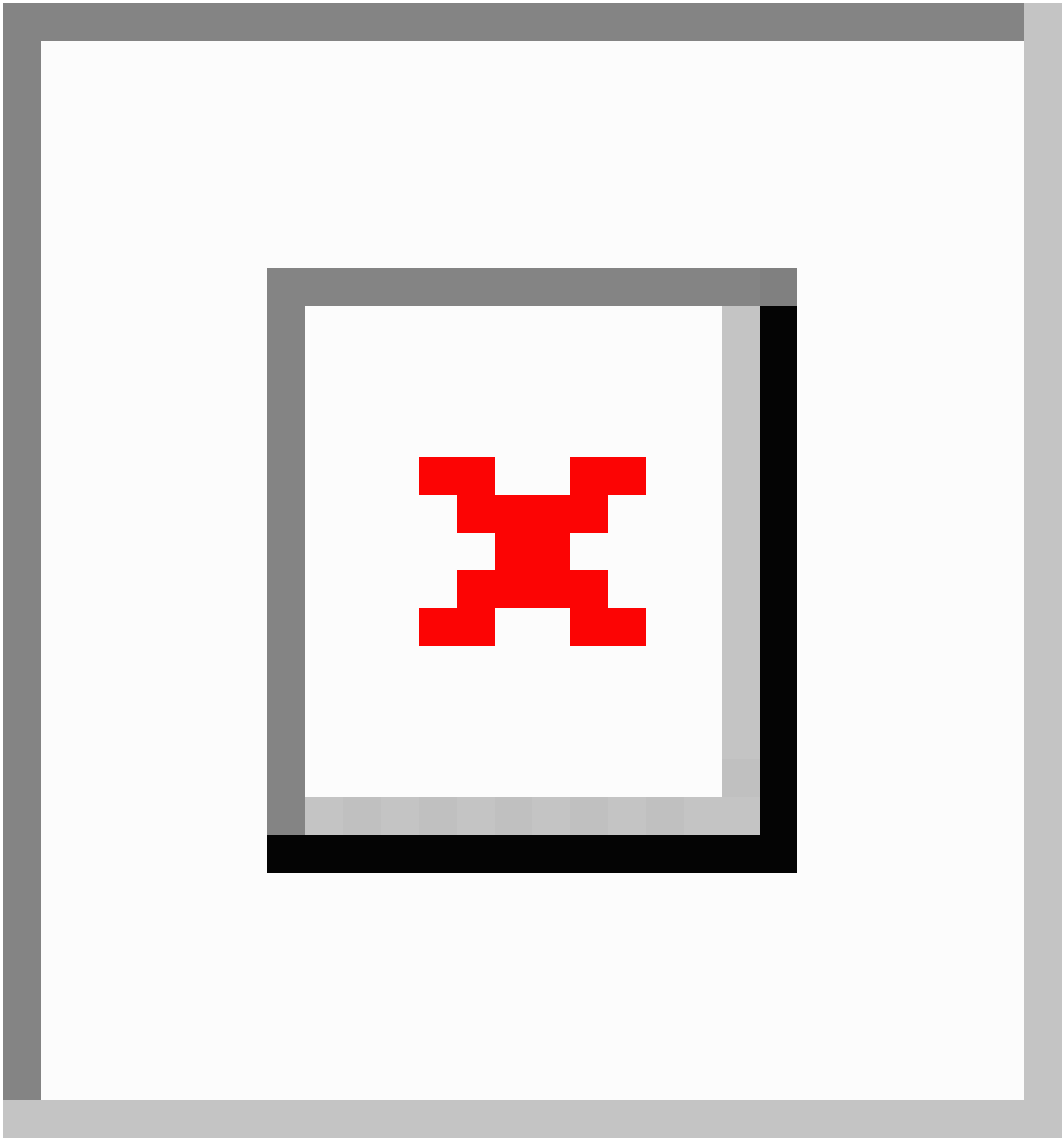
[a]ETL: extract-transform-load.

[b]OMOP: Observational Medical Outcomes Partnership.

[c]Not applicable.

[d]MHC: multidisciplinary health center.

Figure 3. The data quality assessment dashboard. NA: not available.

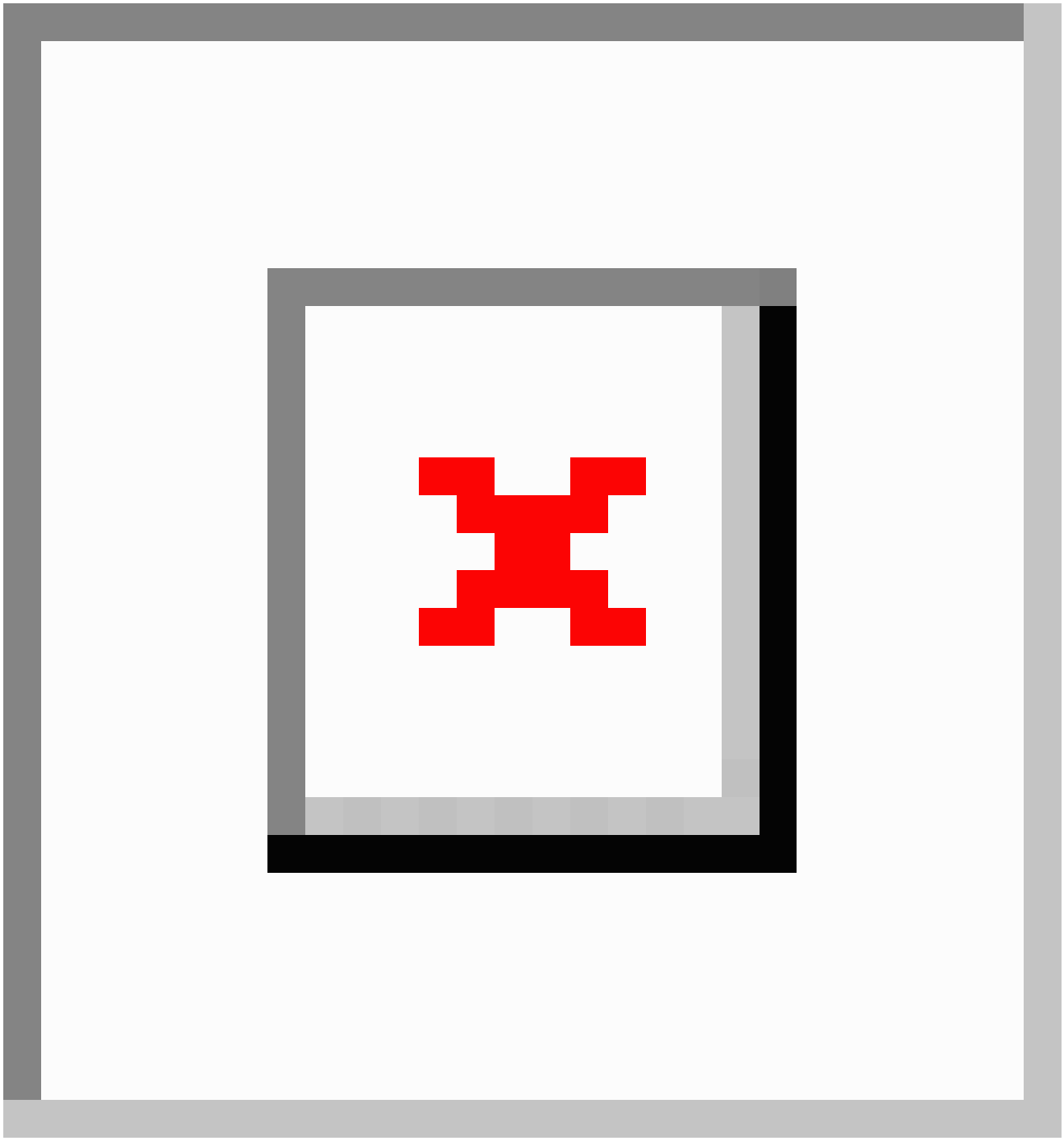


The concordance of results obtained from data quality evaluation queries between the source EHRs and the data warehouse's records proved that the process was reliable (Multimedia Appendix 4). In the event of a difference between the two (results obtained from the data warehouse in OMOP format and results from the Weda software), we had to check the patient records in the software to identify what the correct figure was. Patients whose profile had been created after 2021 had to be removed from the results produced by the Weda software because the extraction data (stored in the data warehouse) goes up to June 2021 at the latest, and queries on the software interface allow filtering for the entire year. The family physician must declare when he or she became the patient's referring family physician. This family physician declaration is dated in the Weda software. This date is not found in the data warehouse. The date of the patient's registration with the family physician was also taken into account in the queries; registrations after the extraction date were removed. Queries concerning the laboratory test results were checked manually by one of the MHC's family physicians (CJ). Some laboratory test results were saved in the "reports" part and others were saved in the "consultations" part on the software.

Practical Application

A 3-tab prototype dashboard was implemented (Figure 4). The first tab concerned the MHC's activity, the second concerned drug prescriptions, and the third concerned laboratory test results.

Figure 4. The prototype primary care dashboard. PriCaDa: Primary Care Data Warehouse.



Discussion

Principal Findings

In this study, we implemented primary care data into the OMOP CDM format from a French health care facility. Five concepts related to the consultation (reason for the consultation, diagnosis, and comments), the clinical report by another physician, some medical histories, referrals, and vaccination data in free text were stored in the NOTE table, and three structured free-text concepts were mapped to standard concepts (laboratory test results, biometrics, and measurement units). Overall, we included 592,226 consultations and 10,221 concepts over 20 years, including 9432 mappings between local concepts and standard concepts. The concept mapping was validated by three experts, including a family physician. We then used queries to validated the design of the OMOP CDM by comparing the results obtained in our data warehouse with those obtained in the source software.

The OMOP model was originally developed to answer pharmaco-epidemiologic questions by reference to hospital databases or claims databases. New types of data have since been integrated, such as those in the fields of cancer, microbiology, and anesthesia [32,33,45]. Integrating these new data types might present difficulties not foreseen in the original CDM. Integrating primary care data required the addition of 10,221 new concepts. Through the sharing of tools and methods, the OMOP model enables reproducible analyses of decentralized data [28,46].

Comparison With Previous Work

This work stands out for its integration of out-of-hospital clinical data over a long time period. In contrast to the French national health care database (*Système National des Données de Santé*), we included clinical data. Similarly, the data produced by French hospital included test results from the hospital's laboratory only. We were able to load data from outpatient treatments and out-of-hospital medical laboratories into the OMOP model; in France, these data are not documented in claims and hospital databases. We also have data on drug prescriptions and associated dosage (duration of treatment, number of refills, number of drugs per dose, and dosing period), whereas the *Système National des Données de Santé* only contains data on prescription fulfillment.

By using a CDM, we will be able to share our work with other primary care data reuse initiatives [28].

This study had a number of strengths. First, it was based on collaboration between data scientists and family physicians. Each data transformation step was approved by the MHC's family physicians. Second, working with the EHR software's developer enabled us to understand the software's structure and export format. The involvement of the software developer in the study expedited the data extraction process. This collaboration with the software developer allowed us to retrieve the data in XML format, comprehend each of the XML tags, and discern the origin of the XML information within the software. Third, our development of a dashboard gave health care professionals an overview of their practice in terms of the number of patients followed and the number of consultations carried out over a defined period.

Limitations

The first limitation is that a large proportion of the extracted primary care data had been entered as free text and required NLP methods or manual examination to be used secondarily. Moreover, the primary care EHR software provides information on drug prescriptions but not on filling or patient compliance.

The free text was difficult to map. Although the most frequent codes were mapped, further mapping and information extraction are needed. NLP methods might be able to recover the free-text information on the diagnosis established during the consultation and the symptoms mentioned.

Perspective

The next step in the project will involve sharing the ETL process with other health care facilities equipped with Weda software or other software. A qualitative study during the presentation of the dashboard to the health care professionals might improve the prototype dashboard and identify unmet needs for further development. This extension will provide us with an opportunity to conduct multicenter studies and to integrate data from other professions working in primary care (such as nurses, midwives, physiotherapists, and pharmacists).

Conclusion

We implemented primary care data from a French health care facility into the OMOP CDM format. Data concerning demographics, units, measurements, and primary care consultation steps were already available in the OHDSI vocabularies. Laboratory test results and drug prescription data were mapped with the available vocabulary and structured in the final model. However, the free text in the primary care EHR software complicates the reuse of additional clinical information such as diagnoses, symptoms, clinical reports, and reasons for consultation. A dashboard application provided health care professionals with feedback on their practice.

Acknowledgments

We thank Dr David Fraser (Biotech Communication SARL, Ploudalmézeau, France) for editorial assistance and helpful advice. The research was funded by the French government through the GIRCI Nord-Ouest program (GIRCI is a French abbreviation for "interregional groups for clinical research and innovation"; project: "REsP-IR"). The funding bodies did not participate in the design of this study; its execution, analyses, interpretation of the data; or the decision to submit the results.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standard concept types for each source table.
[DOC File, 39 KB - medinform_v12i1e49542_app1.doc]

Multimedia Appendix 2
The distribution of the blood potassium concentration values on the Achilles dashboard on the Atlas server.
[PNG File, 139 KB - medinform_v12i1e49542_app2.png]

Multimedia Appendix 3
Results of the data quality assessment dashboard for each feature.
[PNG File, 149 KB - medinform_v12i1e49542_app3.png]

Multimedia Appendix 4
List of quality requests and the associated SQL codes.
[DOC File, 27 KB - medinform_v12i1e49542_app4.doc]

References

1. Schoen C, Osborn R, Squires D, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff (Millwood) 2012 Dec;31(12):2805-2816. [doi: 10.1377/hlthaff.2012.0884] [Medline: 23154997]
2. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017 Aug;26(1):38-52. [doi: 10.15265/IY-2017-007] [Medline: 28480475]
3. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital clinical data warehouse: a 8-years follow-up experience. Int J Med Inform 2017 Jun;102:21-28. [doi: 10.1016/j.ijmedinf.2017.02.006] [Medline: 28495345]
4. Menéndez Torre EL, Ares Blanco J, Conde Barreiro S, Rojo Martínez G, Delgado Alvarez E, en representación del Grupo de Epidemiología de la Sociedad Española de Diabetes (SED). Prevalence of diabetes mellitus in Spain in 2016 according to the Primary Care Clinical Database (BDCAP). Endocrinol Diabetes Nutr (Engl Ed) 2021 Feb;68(2):109-115. [doi: 10.1016/j.endinu.2019.12.004] [Medline: 32988801]
5. Lamer A, Moussa MD, Marcilly R, Logier R, Vallet B, Tavernier B. Development and usage of an anesthesia data warehouse: lessons learnt from a 10-year project. J Clin Monit Comput 2023 Apr;37(2):461-472. [doi: 10.1007/s10877-022-00898-y] [Medline: 35933465]
6. Primary care. NHS Digital. URL: https://digital.nhs.uk/data-and-information/areas-of-interest/primary-care [accessed 2023-07-31]
7. Institute of Medicine (US) Committee on the Future of Primary Care. Defining primary care. In: Donaldson MS, Yordy KD, Lohr KN, et al, editors. Primary Care: America's Health in a New Era: National Academies Press (US); 1996. URL: https://www.ncbi.nlm.nih.gov/books/NBK232631/ [accessed 2023-07-31]
8. Gentil ML, Cuggia M, Fiquet L, et al. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. BMC Med Inform Decis Mak 2017 Sep 25;17(1):139. [doi: 10.1186/s12911-017-0538-x] [Medline: 28946908]
9. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014 Dec;52:28-35. [doi: 10.1016/j.jbi.2014.02.003] [Medline: 24534443]
10. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013 Dec;274(6):547-560. [doi: 10.1111/joim.12119] [Medline: 23952476]
11. Muller S. Electronic medical records: the way forward for primary care research? Fam Pract 2014 Apr;31(2):127-129. [doi: 10.1093/fampra/cmu009] [Medline: 24627543]
12. Carey IM, Banchoff E, Nirmalananthan N, et al. Prevalence and incidence of neuromuscular conditions in the UK between 2000 and 2019: a retrospective study using primary care data. PLoS One 2021 Dec 31;16(12):e0261983. [doi: 10.1371/journal.pone.0261983] [Medline: 34972157]
13. Boullenger L, Quindroit P, Legrand B, et al. Type 2 diabetics followed up by family physicians: treatment sequences and changes over time in weight and glycated hemoglobin, prim. Prim Care Diabetes 2022 Oct;16(5):670-676. [doi: 10.1016/j.pcd.2022.07.002]
14. Loadsman MEN, Verheij TJM, van der Velden AW. Impetigo incidence and treatment: a retrospective study of Dutch routine primary care data. Fam Pract 2019 Jul 31;36(4):410-416. [doi: 10.1093/fampra/cmy104] [Medline: 30346521]
15. Marwaha S, Dusendang JR, Alexeeff SE, et al. Comanagement of rashes by primary care providers and dermatologists: a retrospective study. Perm J 2021 Dec 13;25:20.320. [doi: 10.7812/TPP/20.320] [Medline: 35348083]
16. Milea D, Yeo SH, Nam Y, et al. Long-acting bronchodilator use in chronic obstructive pulmonary disease in primary care in New Zealand: a retrospective study of treatment patterns and evolution using the HealthStat database. Int J Chron Obstruct Pulmon Dis 2021 Apr 20;16:1075-1091. [doi: 10.2147/COPD.S290887] [Medline: 33907394]
17. Sollie A, Sijmons RH, Helsper C, Numans ME. Reusability of coded data in the primary care electronic medical record: a dynamic cohort study concerning cancer diagnoses. Int J Med Inform 2017 Mar;99:45-52. [doi: 10.1016/j.ijmedinf.2016.08.004] [Medline: 28118921]
18. Kornfält Isberg H, Hedin K, Melander E, Mölstad S, Beckman A. Increased adherence to treatment guidelines in patients with urinary tract infection in primary care: a retrospective study. PLoS One 2019 Mar 28;14(3):e0214572. [doi: 10.1371/journal.pone.0214572] [Medline: 30921411]
19. Fruchart M, Quindroit P, Patel H, Beuscart JB, Calafiore M, Lamer A. Implementation of a data warehouse in primary care: first analyses with elderly patients. Stud Health Technol Inform 2022 May 25;294:505-509. [doi: 10.3233/SHTI220510] [Medline: 35612131]
20. Clinical Practice Research Datalink (CPRD). URL: https://www.cprd.com/node/120 [accessed 2023-05-04]
21. The Health Improvement Network (THIN). URL: https://www.the-health-improvement-network.com [accessed 2023-05-04]

22. Corporate Data Warehouse (CDW). US Department of Veterans Affairs. URL: https://www.hsrd.research.va.gov/for_researchers/cdw.cfm [accessed 2023-05-04]
23. Canadian Primary Care Sentinel Surveillance Network (CPCSSN). URL: https://cpcssn.ca/ [accessed 2023-05-04]
24. Safran C. Reuse of clinical data. Yearb Med Inform 2014 Aug 15;9(1):52-54. [doi: 10.15265/IY-2014-0013] [Medline: 25123722]
25. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
26. Pascoe SW, Neal RD, Heywood PL, Allgar VL, Miles JN, Stefoski-Mikeljevic J. Identifying patients with a cancer diagnosis using general practice medical records and cancer registry data. Fam Pract 2008 Aug;25(4):215-220. [doi: 10.1093/fampra/cmn023] [Medline: 18550895]
27. Terry AL, Stewart M, Cejic S, et al. A basic model for assessing primary health care electronic medical record data quality. BMC Med Inform Decis Mak 2019 Feb 12;19(1):30. [doi: 10.1186/s12911-019-0740-0] [Medline: 30755205]
28. Papez V, Moinat M, Voss EA, et al. Transforming and evaluating the UK Biobank to the OMOP common data model for COVID-19 research and beyond. J Am Med Inform Assoc 2022 Dec 13;30(1):103-111. [doi: 10.1093/jamia/ocac203] [Medline: 36227072]
29. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578. [Medline: 26262116]
30. Observational Health Data Sciences and Informatics (OHDSI). URL: https://www.ohdsi.org/ [accessed 2023-02-13]
31. OMOP CDM v5.4. OHDSI GitHub. URL: https://ohdsi.github.io/CommonDataModel/cdm54.html [accessed 2023-02-13]
32. OMOP Standardized Vocabulary V5.0. Observational Health Data Sciences and Informatics (OHDSI). URL: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary [accessed 2022-07-13]
33. Lamer A, Abou-Arab O, Bourgeois A, et al. Transforming anesthesia data into the Observational Medical Outcomes Partnership common data model: development and usability study. J Med Internet Res 2021 Oct 29;23(10):e29259. [doi: 10.2196/29259] [Medline: 34714250]
34. Paris N, Lamer A, Parrot A. Transformation and evaluation of the MIMIC database in the OMOP common data model: development and usability study. JMIR Med Inform 2021 Dec 14;9(12):e30970. [doi: 10.2196/30970] [Medline: 34904958]
35. Haberson A, Rinner C, Schöberl A, Gall W. Feasibility of mapping Austrian health claims data to the OMOP common data model. J Med Syst 2019 Sep 7;43(10):314. [doi: 10.1007/s10916-019-1436-9] [Medline: 31494719]
36. Weda. URL: https://weda.fr/ [accessed 2023-09-14]
37. White Rabbit. OHDSI GitHub. URL: http://ohdsi.github.io/WhiteRabbit/WhiteRabbit.html [accessed 2023-02-13]
38. Rabbit in a Hat. OHDSI GitHub. URL: http://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html [accessed 2023-02-13]
39. Lemaire F. The Jardé law: what does change [Article in French]. Presse Med 2019 Mar;48(3 Pt 1):238-242. [doi: 10.1016/j.lpm.2019.01.006] [Medline: 30853280]
40. Athena. Observational Health Data Sciences and Informatics (OHDSI). URL: https://athena.ohdsi.org/search-terms/start [accessed 2023-02-13]
41. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
42. ATLAS. Observational Health Data Sciences and Informatics (OHDSI). URL: https://atlas-demo.ohdsi.org/ [accessed 2023-02-13]
43. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016 Sep 11;4(1):1244. [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
44. Shinydashboard. Rstudio GitHub. URL: https://rstudio.github.io/shinydashboard/ [accessed 2023-02-13]
45. Yoo S, Yoon E, Boo D, et al. Transforming thyroid cancer diagnosis and staging information from unstructured reports to the Observational Medical Outcome Partnership Common Data Model. Appl Clin Inform 2022 May;13(3):521-531. [Medline: 35705182]
46. Delanerolle G, Williams R, Stipancic A, et al. Methodological issues in using a common data model of COVID-19 vaccine uptake and important adverse events of interest: feasibility study of data and connectivity COVID-19 vaccines pharmacovigilance in the United Kingdom. JMIR Form Res 2022 Aug 22;6(8):e37821. [doi: 10.2196/37821] [Medline: 35786634]

Abbreviations

ATC: Anatomical Therapeutic Chemical
CDM: common data model
CIP: *Code Identifiant de Spécialité* (the French national drug code)
EHR: electronic health record
ETL: extract-transform-load

ICD-10**:** *International Classification of Diseases, 10th Revision*
MHC: multidisciplinary health center
NLP: natural language processing
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
PriCaDa: Primary Care Data Warehouse

Edited by C Lovis; submitted 01.06.23; peer-reviewed by E Sylvestre, N Ahmadi; revised version received 11.04.24; accepted 11.04.24; published 13.08.24.

Please cite as:
Fruchart M, Quindroit P, Jacquemont C, Beuscart JB, Calafiore M, Lamer A
Transforming Primary Care Data Into the Observational Medical Outcomes Partnership Common Data Model: Development and Usability Study
JMIR Med Inform 2024;12:e49542
URL: https://medinform.jmir.org/2024/1/e49542
doi:10.2196/49542

Data Flow Construction and Quality Evaluation of Electronic Source Data in Clinical Trials: Pilot Study Based on Hospital Electronic Medical Records in China

Yannan Yuan[1], MS; Yun Mei[2], MS; Shuhua Zhao[1], MS; Shenglong Dai[3], MS; Xiaohong Liu[1], MS; Xiaojing Sun[3], MA; Zhiying Fu[1], MS; Liheng Zhou[3], MS; Jie Ai[2], MS; Liheng Ma[3], MD; Min Jiang[4], MS

1

2

3

4

Corresponding Author:
Min Jiang, MS

Abstract

Background: The traditional clinical trial data collection process requires a clinical research coordinator who is authorized by the investigators to read from the hospital's electronic medical record. Using electronic source data opens a new path to extract patients' data from electronic health records (EHRs) and transfer them directly to an electronic data capture (EDC) system; this method is often referred to as eSource. eSource technology in a clinical trial data flow can improve data quality without compromising timeliness. At the same time, improved data collection efficiency reduces clinical trial costs.

Objective: This study aims to explore how to extract clinical trial–related data from hospital EHR systems, transform the data into a format required by the EDC system, and transfer it into sponsors' environments, and to evaluate the transferred data sets to validate the availability, completeness, and accuracy of building an eSource dataflow.

Methods: A prospective clinical trial study registered on the Drug Clinical Trial Registration and Information Disclosure Platform was selected, and the following data modules were extracted from the structured data of 4 case report forms: demographics, vital signs, local laboratory data, and concomitant medications. The extracted data was mapped and transformed, deidentified, and transferred to the sponsor's environment. Data validation was performed based on availability, completeness, and accuracy.

Results: In a secure and controlled data environment, clinical trial data was successfully transferred from a hospital EHR to the sponsor's environment with 100% transcriptional accuracy, but the availability and completeness of the data could be improved.

Conclusions: Data availability was low due to some required fields in the EDC system not being available directly in the EHR. Some data is also still in an unstructured or paper-based format. The top-level design of the eSource technology and the construction of hospital electronic data standards should help lay a foundation for a full electronic data flow from EHRs to EDC systems in the future.

(JMIR Med Inform 2024;12:e52934) doi:10.2196/52934

KEYWORDS

clinical trials; electronic source data; EHRs; electronic data capture systems; data quality; electronic health records

Introduction

Source data are the original records from clinical trials or all information recorded on certified copies, including clinical findings, observations, and records of other relevant activities necessary for the reconstruction and evaluation of the trial [1]. Electronic source data are data initially recorded in an electronic format (electronic source data or eSource) [2,3].

The traditional clinical trial data collection process requires a clinical research coordinator (CRC) who is authorized by the investigators to read from the hospital's electronic medical record and other clinical trial–related data from the hospital information system and then manually enter the patient's data into the electronic data capture (EDC) system. After data entry, the clinical research associate visits the site to perform source data verification and source data review. The drawbacks of collecting data by manual transcription are that data quality and timeliness cannot be guaranteed and that it is a waste of human and material resources. Using electronic source data opens a new path to extract patients' data from electronic health records (EHRs) and transfer it directly to EDC systems (often the method is referred to as eSource) [4]. eSource technology in a clinical trial data flow can improve data quality without

compromising timeliness [5]. At the same time, improved data collection efficiency reduces clinical trial costs [6].

eSource can be divided into two levels. The first level is to enable the hospital information system to obtain complete data sets; the second level is to allow direct data transfer to EDC systems based on the clinical trial patients' electronic data in hospitals to avoid the electronic data being transcribed manually again, which is the core purpose of eSource [7]. This project will explore the use of eSource technology to extract clinical trial data from EHRs, send it to the sponsor data environment, and discuss the issues and challenges occurring in its application process.

Methods

Ethics Approval

This study was approved by the Ethics Committee and Human Genetic Resource Administration of China (2020YW135). During the ethical review process, the most significant challenges were patients' informed consent, privacy protection, and data security. The B7461024 Informed Consent Form (Version 4) states that "interested parties may use subjects' personal information to improve the quality, design, and safety of this and other studies," and "Is my personal information likely to be used in other studies? Your coded information may be used to advance scientific research and public health in other projects conducted in future." This project is an exploration of using electronic source data technology instead of traditional manual transcription in the process of transferring data from hospital EHRs to EDC systems, which will improve the data quality of clinical trials and will improve the data flow in the future. Therefore, this project is within the scope of the informed consent form for study B7461024, which was approved by the ethics committee after clarification.

Project Information

This project was conducted from December 15, 2020, to November 19, 2021, which was before China's personal information protection law and data security law were introduced. The data for this project were obtained from an ongoing phase 2, multicenter, open-label, dual-cohort study to evaluate the efficacy and safety of Lorlatinib (pf-06463922) monotherapy in anaplastic lymphoma kinase (ALK) inhibitor–treated locally advanced or metastatic ALK-positive non–small cell lung cancer patients in China (B7461024), registered by the sponsor on the Drug Clinical Trials Registration and Disclosure Platform (CTR20181867). The data extraction involved 4 case report form (CRF) data modules: demographics, concomitant medication, local lab, and vital signs, which were collected in the following ways:

- Demographics: Originally entered directly into the hospital EHR then manually transcribed by the CRC to the sponsor's EDC system
- Local lab: Laboratory data collected by the hospital laboratory information management system (LIMS) and then manually transcribed by the CRC into the EDC system
- Vital signs: Hospital uses paper-based tracking form provided by the sponsor to record patients' vital signs and investigators transcribe the vital signs data into the hospital medical record
- Concomitant medication: Similar to vital signs, hospital uses the paper tracking form provided by the sponsor to record the adverse reactions and concomitant medication; investigator might also transfer the concomitant medication data into the hospital EHR, but there was no mandatory requirement to transfer these data into patients' medical records

All information was collected from 6 patients in a total of 29 fields (Textbox 1).

Textbox 1. Data collection fields.

Demographics

- Subject ID
- Date of birth
- Sex
- Ethnicity
- Race
- Age

Concomitant medication

- Combined drug name
- Whether for the treatment of adverse reactions
- Adverse event number
- Combined drug start date
- Combined drug end date
- Currently still in use

Vital signs

- Date of vital signs collection
- Weight
- Weight unit
- Body temperature
- Height
- Height unit
- Location of temperature measurement
- Systolic blood pressure
- Diastolic blood pressure
- Pulse

Local lab

- Laboratory inspection name
- Laboratory name and address
- Sponsor number
- Laboratory number
- Incomplete laboratory inspection
- Sample collection data
- Inspection results

Data Process Workflow

Overview

The study chosen in our project used the traditional manual data entry method to transcribe patients' CRF data into the EDC system. This project proposes testing the acquisition of data directly from the hospital EHR, deidentification of the patients' electronic data on the hospital medical data intelligence platform, mapping and transforming the data based on the sponsor's EDC data standard, and transferring the data into the sponsor's environment. The data was transferred from the hospital to the sponsor's data environment and compared to data that was captured by traditional manual entry methods to verify the availability, completeness, and accuracy of the eSource technology.

In the network environment of this project, the technology provider accessed the hospital network through a virtual private network (VPN) and a bastion host, and processed the data of this project as a private cloud, thus ensuring the security of the hospital data.

Data Integration

The hospital information system involved in this project has reached the national standards of "Level 3 Equivalence," "Electronic Medical Record Level 5," and "Interoperability Level 4." The medical data intelligence platform in this project is deployed in a hospital intranet, isolated from external networks. Integrated data from different information systems, including the hospital information system, LIMS, picture archiving and communication system, etc, were deidentified from the platform and transferred to a third-party private cloud platform for translation and data format conversion after authorization by the hospital through a VPN.

The scope of data collection in this project was limited to patients who signed Informed Consent Form (Version 4) for study B7461024. The structured data of four CRF data modules (demographic, concomitant medications, local lab, and vital signs) were extracted from the source data in hospital systems, and data processing was completed.

Three-Layer Deidentification of Data

In this project, three layers of deidentification were performed on the electronic source data to ensure data security. The first layer of deidentification was performed before the certified copy of data was loaded to the hospital's medical data intelligence platform. The second layer of deidentification follows the Health Insurance Portability and Accountability Act (HIPAA) by deidentifying 18 data fields at the system level. A third layer of deidentification was performed when mapping and transforming third-party databases for the clinical trial data (demographics, concomitant medications, laboratory tests, and vital signs) collected for this study, as required by the project design.

Collected data did not contain any sensitive information with personal identifiers of the patients, and all deidentification processes were conducted in the internal environment of the hospital. In addition to complying with the relevant laws and regulations, we followed the requirements of Good Clinical Practice regarding patient privacy and confidentiality, and further complied with the requirements of HIPAA to deidentify the 18 basic data fields. Data fields outside the scope of HIPAA will be deidentified and processed in accordance with the TransCelerate guidelines published in April 2015 to ensure the security of patients' personal information and to eliminate the possibility of patient information leakage [8].

The general rules for the third layer of deidentification were as follows:

- Time field: A specific time point is used as the base time, and the encrypted time value is the difference between the word time and the base time
- ID field: Categorized according to the value and only shows the category
- Age field: Categorized according to the value and only shows the category
- Low-frequency field: set to null

In addition, all data flows keep audit trails throughout and are available for audit.

Data Normalization and Information Extraction

After three layers of deidentification, the data was transferred from a hospital to a third-party private cloud platform through a VPN, where translation from Chinese to English and data format conversion were implemented. The whole transfer process was performed for the data that was collected for the clinical trial of this study. Standardization of data is a crucial task during the data preparation phase. This process involves consolidating data from different systems and structures into a consistent, comprehensible, and operable format. First, a thorough examination of data from various systems is necessary. Understanding the data structure, format, and meaning of each system is essential. The second step involves establishing a data dictionary that clearly outlines the meaning, format, and possible values of each data element. Next, selecting a data standard is necessary to ensure consistency and comparability. In this study, we adopted the Health Level 7 (HL7) standard. Additionally, data cleansing and transformation are needed to meet standard requirements, including handling missing data, resolving mismatched data formats, or performing data type conversions. Extract, transform, and load tools were used to integrate data from different systems. Data security must be ensured throughout the data integration process. This includes encrypting sensitive information and strictly managing data permissions. Data verification and validation steps were then performed by professional staff on the translated data. The data from the hospital's medical data intelligence platform were then converted from JSON format to XML and Excel formats. The processed data was transferred back to the hospital via a VPN to a designated location for final adjudication before loading to the sponsor's environment.

One-Time Data Push and Quality Assessment

After the hospital received the processed data, it was then pushed by the hospital to the sponsor's secure and controlled environment (Figure 1). All data deidentification processes were conducted in the hospital's environment, and none of the data obtained by the sponsor can be traced back to patients' personal information to ensure their privacy and information security.

The data quality of this project was assessed using industry data quality assessment rules [9], which are shown in Table 1.

Figure 1. Project operation flow. EHR: electronic health record.

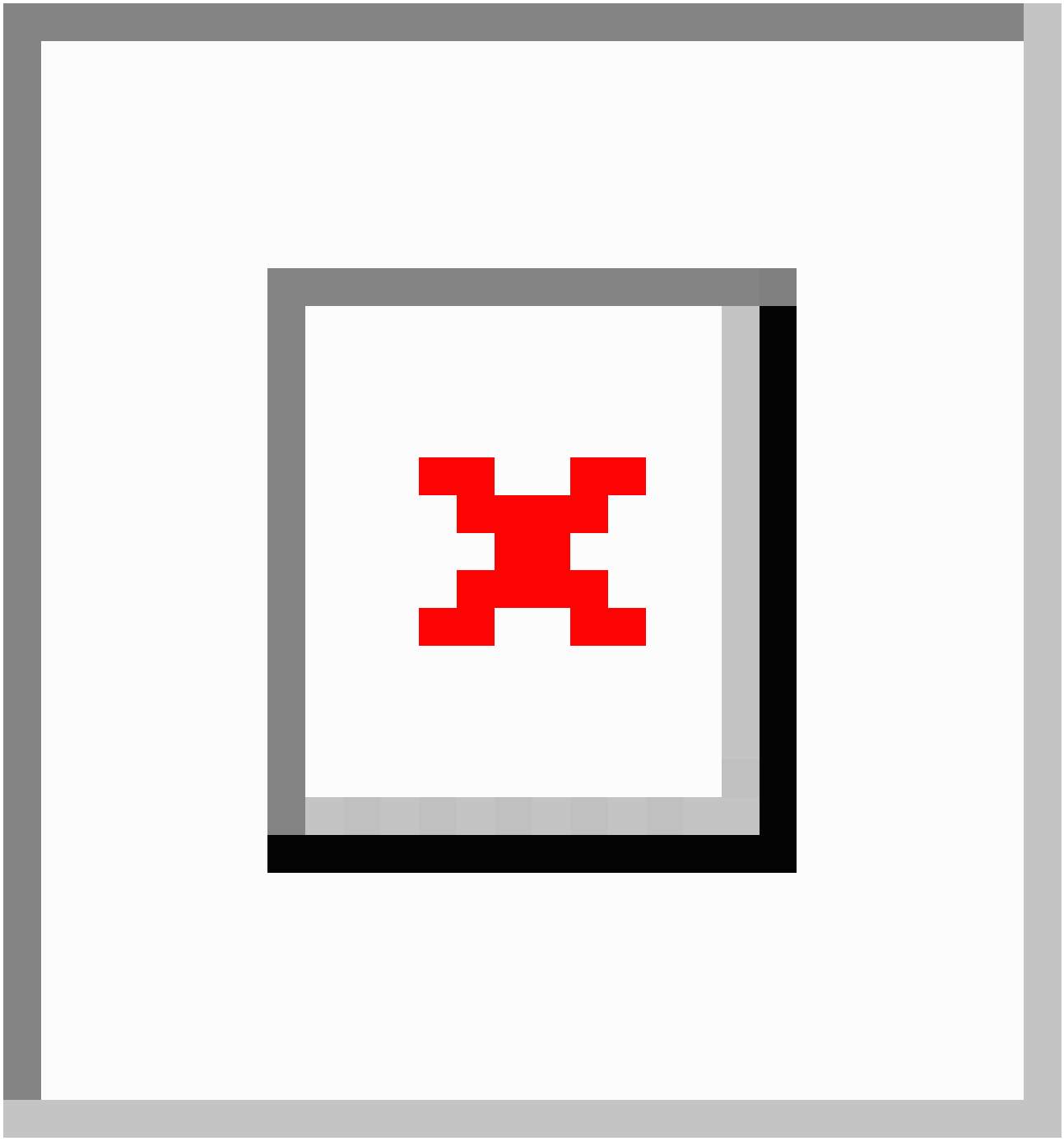


Table . Introduction of data quality assessment rules.

Data validation methods	Dimension	Method description	Cases
Data availability verification	Field dimension	The ratio of the total number of data fields in the clinical trial CRF[a] available in the hospital EHR[b] to the total number of data fields required in the electronic CRF: EHR[c]/CRF[d] × 100%	Based on the electronic CRF, 6 data fields in the demography need to be captured, and 3 of them have records in the EHR. Data availability: 3/6 × 100% = 50%
Data availability verification	Field dimension	The ratio of the total number of data fields in the clinical trial CRF (eSource) that can be transmitted electronically in the hospital's EHR to the total number of data fields required in the electronic CRF: eSource[e]/CRF[d] × 100%	Based on the electronic CRF, 6 data fields in the demography need to be captured, and 2 data fields can be captured by the eSource method. Data availability: 2/6 × 100% = 33.33%
Data completeness verification	Numerical dimension	The ratio of the total number of nonnull data (eSourceV) captured (processed and sent to the sponsor) via the eSource method to the total number of data fields requested on the electronic CRF: eSourceV[f]/CRF[d] × 100%	Based on the clinical trial design, 38 concomitant medication pages need to be collected: 7 pages were collected via eSource and 2 fields was entered per page. Data completeness: 7 × 2/(2 × 38) × 100% = 18.42%
Data accuracy verification	Numerical dimension	Matching of data field values in the hospital's EHR with data field values that can be captured by eSource (data fields that are processed and sent to the sponsor)	4 fields of demography were successfully transmitted through eSource, with 4 data points in each. After comparing with the data in the electronic data capture system, there were no mismatches for one data point. Data accuracy: 8/(2 × 4) × 100% = 100%

[a]CRF: case report form.

[b]EHR: electronic health record.

[c]Total number of data fields in the hospital's EHR.

[d]Total number of data fields requested in the electronic CRF.

[e]Total number of data fields captured (processed and sent to the sponsor) through the eSource method.

[f]Total number of nonempty data fields captured (processed and sent to the sponsor) through the eSource method.

Results

In this project, we collected patients' demographics, vital signs information, local laboratory data, and concomitant medication data from EHRs, successfully pushed the data directly to the designated sponsor environment, and evaluated the data quality from three perspectives including availability, completeness, and accuracy (Table 2).

- The eSource-CRF availability score, which is used to evaluate the ratio of fields in EHR that can be collected by eSource and used for CRF, was low for demographics, blood tests, and urine sample tests but higher for vital signs and concomitant medications.
- Data completeness, defined as the ratio of the total number of nonnull data captured by eSource to the total number of data fields required in the electronic CRF, was used to evaluate the ratio of nonnull data fields in the CRF that can be captured by eSource. In this study, the completeness score of the vital signs module was only 1.32%, and the concomitant medications and laboratory test modules also had poor performance in the data completeness evaluation.
- Data accuracy, defined as the compatibility between the data field values in the hospital EHR and the data field values that can be collected using eSource, was 100% for all modules.
- EHR-CRF availability, which is used to evaluate the ratio of fields in the EHR that can be used for the CRF, was 50%, 60%, and 66.67% for demographics, blood tests, and urine sample tests, respectively, in this study, and the rest of the data were 100% available.

Table . Metrics measured.

CRF[a] domain		CRF-EHR[b] data availability, n/N (%)	CRF-eSource data availability, n/N (%)	Data completeness (preliminary findings), n/N (%)	Data accuracy (preliminary findings), n/N (%)
Definition		Study CRF data elements available in hospital EHR	Study CRF data elements available in hospital EHR and able to be electronically transferred through eSource technology	Study CRF data elements available and entered into hospital EHR and transferred through eSource technology	Study CRF data elements available and entered into hospital EHR and transferred through eSource technology with expected result (eg, matches what was entered directly in form)
Demographics		3/6 (50.00)	2/6 (33.33)	12/12 (100.00)	12/12 (100.00)
Vital signs		10/10 (100.00)	9/10 (90.00)	24/1812 (1.32)[c]	20/20 (100.00)
Local lab					
	Blood biochemical tests	6/10 (60.00)	5/10 (50.00)	12,968/13,540 (95.78)[d]	7767/7767 (100.00)
	Urine sample tests	6/9 (66.67)	5/9 (55.56)	15/40 (37.56)	15/15 (100.00)
Concomitant medication		10/10 (100.00)	9/10 (90.00)	14/76 (18.42)[e]	6/6 (100.00)

[a]CRF: case report form.

[b]EHR: electronic health record.

[c]Checks were made with the relevant clinical research associates (CRAs) regarding the original data collection and CRF completion methods for the following reasons: vital signs were obtained using paper tracking forms provided by the sponsor as the original data source, and the data may not be transcribed into the hospital information system (HIS) by the researcher. Therefore, data from many visits are not available in the HIS.

[d]A total of 2708 blood biochemistry tests were involved.

[e]Concomitant medication uses tracking forms to record adverse event and ConMed (a paper source), and data may not be transcribed into the HIS. As confirmed by the CRA, the percentage of paper ConMed sources was approximately 80%.

Discussion

Although EHRs have been widely used, the degree of structure of EHR data varies substantially among different data modules. In EHRs, demographics, vital signs, local lab data, and concomitant medications are more structured than patient history or progress notes and often contain unstructured text [10]. Therefore, we selected these 4 well-structured data modules for exploration in this project.

For demographics data, among the 6 required fields (subject ID, date of birth, sex, ethnicity, race, and age), subject ID (subject code number/identifier in the trial, not the patient code number/identifier in the EHR system), ethnicity, and race were not available in the EHR, so the EHR-CRF availability score was 50%. Since this was an exploratory project, the date of birth field was also deidentified and thus could not be collected based on our deidentification rule, so the eSource-CRF availability score was 33%. In the future, the availability score can reach close to 100% by bidirectional design of the EHR and CRF under the premise of obtaining compliance for industrial-level applications.

The low availability score of local laboratory data on EHR-CRFs is due to the lack of required fields in the hospital system; “Lab ID” and “Not Done” do not exist in the LIMS, and for the “Clinically Significant” field, the meaning of laboratory test results needs to be manually interpreted by an investigator, so they cannot be transcribed directly. The availability score of eSource-CRFs was further decreased because the field “Laboratory Name and Address” is not an independent structured field in the EHR. The completeness score of urine sample test data was only 37.56% because during the actual clinical trial, especially amid the COVID-19 pandemic period, patients completed study-related laboratory tests at other sites, and those test results were collected via paper-based reports, so the complete data sets cannot be extracted from the site’s system.

To improve data availability in future applications, clinical trial–specific fields need to be added to EHR designs for those data that require an investigator’s interpretation such as “Clinically Significant,” and data transfer and mapping processes for the determination of the scope of data collection also needs to be optimized. Based on these two conditions, the completeness score can be improved to over 90%.

The availability and accuracy of vital signs data are ideal. However, since not all vital signs data collection was recorded by the electronic system during the actual study visit, many vital signs data were collected in “patient diary” and other types of paper-based documents during the study, resulting in a serious limitation in data completeness. With the development of more clinical trial–related electronic hardware and enhancements in products intelligence, more vital signs data will be directly collected by electronic systems, and the completeness of vital signs data transferred from EHR to EDC will be greatly improved in the future.

In the concomitant medication module, there was a good score for availability and accuracy because the standardization and structuring of prescriptions are well done in this hospital system. However, the patient's medication use period during hospitalization is recorded in unstructured text, so the data could not be captured for this study, resulting in a low completeness score of 18.42% for concomitant medication.

In summary, the accuracy score of eSource data in this study was high (100% for all fields). A study by Memorial Sloan Kettering Cancer Center and Yale University confirmed that the error rate of automatic transcription reduced from 6.7% to 0% compared to manual transcription [10]. However, data availability and completeness have not reached a good level. Data availability varies widely across studies, ranging from 13.4% in the Retrieving EHR Useful Data for Secondary Exploitation (REUSE) project [11] to 75% in The STARBRITE Proof-of-Concept Study [12], mainly related to the coverage and structure of the EHR.

National drug regulatory agencies (eg, US Food and Drug Administration [FDA], European Medicines Agency, Medicines and Healthcare products Regulatory Agency, and Pharmaceuticals and Medical Devices Agency) have developed guidelines to support the application of eSource to clinical trials [3,13-15]. The new Good Clinical Practice issued by the Center for Drug Evaluation in 2020 encourages investigators to use clinical trials' electronic medical records for source data documentation [1]. Despite this, we still encountered challenges, including ethical review and data security, during this study's implementation process. Without knowing the precedents, the project team decided to follow the requirements for clinical trials to control the quality of the study. There were no existing regulatory policies or national guidance on eSource in China at the time of this study. The project team provided explanations for inapplicable documents and communicated several times to ensure the approval of relevant institutional departments before finally becoming the first eSource technology study to be approved by the Ethics Committee and Human Genetic Resource Administration of China.

In the absence of regulatory guidelines, our eSource study, the first in China's International Multi-center Clinical Trial, navigated challenges in data deidentification. We adopted HIPAA and TransCelerate's guidelines [8]. Securing approval under "China International Cooperative Scientific Research Approval for Human Genetic Resources," we answered queries and achieved unprecedented recognition. For transferring data from the hospital to the sponsor's environment, we prioritized security and obtained necessary approvals. Iterative revisions ensured a robust data flow design. Challenges in mapping hospital EHR to EDC standards highlighted the need for a scalable mechanism. This study pioneers eSource tech integration in China, emphasizing the importance of seamless data mapping. In the process of executing data standardization, several challenges may arise, including inconsistent data definitions. Data from different systems may use different definitions due to the independent development of these systems, leading to varied interpretations of even identical concepts. To address this issue, establishing a unified data dictionary is crucial to ensure consensus on the definition of each data element. Different systems might also use distinct data formats such as text encodings. Preintegration format conversion is required, and extract, transform, and load tools or scripts can assist in standardizing these formats. During the integration of data from multiple systems, it is possible to discover data in one system that is not present in another. In the data standardization process, considerations must be made on how to handle missing data, which may involve interpolation, setting default values, etc. Quality issues like errors, duplicates, or inaccuracies may exist in data from different systems. Data cleansing, involving deduplication, error correction, logical validation, etc, is necessary to address these quality issues. Different systems may generate data based on diverse business rules and hospital use scenarios. In data standardization, unifying these rules requires collaboration with domain experts to ensure consistency.

Internationally, multiple research studies and publications have been released on regulations, guidelines, and validation of eSource. The FDA provided guidance on the use of electronic source data in clinical trials in 2013 that aims to address barriers to capturing electronic source data for clinical trials, including the lack of interoperability between EHRs and EDC systems. The European-wide Electronic Health Records for Clinical Research (EHR4CR) project was launched in 2011 to explore technical options for the direct capture of EHR data within 35 institutions, and the project was completed in 2016 [16]. The second phase of the project connected the EHRs to EDC systems [17] and aimed to realize the interoperability of EHRs and EDC systems. The US experience focuses more on improving and standardizing the existing EHRs to make them more uniform.

In Europe, the experience focuses on breaking down the technical barrier of interoperability between EHRs and EDC systems. In China, the current industry trends focus on the governance of existing EHR data in the hospital and the building of clinical data repository platforms [7]. Clinical data repository platforms focus on data integration and cleaning between EHRs and other systems in hospital environments and on unstructured data normalization and standardization by natural language processing and other AI technology [18]. At the national level, China is also actively promoting the digitization of medical big data and is committed to the formation of regional health care databases [19], which lays the foundation for the future implementation of eSource in China [20].

This study evaluates the practical application value of eSource in terms of availability, completeness, and accuracy. To improve availability, the structure of the CRF needs to be designed according to the information of the EHR data at the design stage of clinical trials. Even so, since EHRs are designed for the physicians to conduct daily health care activities, certain fields in clinical trials (eg, judgment of normal or abnormal values of laboratory tests and judgment of correlations of adverse events and combined medications) are still not available, and clinical trial–specific fields need to be added to EHR designs for those data that require investigators' interpretation to improve data availability. Completeness could be improved by the development of hospital digitalization that ensures patients' data is collected electronically rather than on paper. Additionally, 2708 blood test records were successfully collected from only 6 patients via eSource in this study, which indicates

that laboratory tests often contain large amounts of highly structured data that are suitable for eSource. EHR-EDC end-to-end automatic data extraction by eSource is suitable for laboratory examinations and can improve the efficiency and accuracy of data extraction significantly as well as reduce redundant manual transcriptions and labor costs. Processing unstructured or even paper-based data in eSource is still a big challenge. Using machine learning tools (eg, natural language processing tools) for autostructuring can be explored in the future. The goal is to have common data standards and better top-level design to facilitate data integrity, interoperability, data security, and patient privacy protection in eSource applications. During deidentification, we processed certain data with a specific logic to protect privacy. The accuracy assessment was performed during the deidentification step to ensure that the data was still sufficiently accurate while meeting privacy requirements. Reversible methods need to be used when performing deidentification as well as providing controlled access mechanisms to the data so that the raw data can be accessed when needed. It is worth noting that different regions and industries may have different privacy regulations and compliance requirements. When deidentifying, you need to ensure that you are compliant with the relevant regulations and understand the limitations of data use. This may require working closely with a legal team.

In the future, we can consider adding performance analysis, including an assessment of data import performance. This involves evaluating the speed and efficiency of data import to ensure it is completed within a reasonable timeframe. Additionally, analyzing data query performance is crucial in practical applications to ensure that the imported data meets the expected query performance in the application. For long-term applications involving a larger size of patients, it is advisable to consider adding analyses related to maintainability and cost-effectiveness. This includes implementing detailed logging and monitoring mechanisms to promptly identify and address potential issues. Furthermore, for the imported data, establishing a version control mechanism is essential for tracing and tracking changes in the data. Simultaneously, for overall resource use, evaluating the resources required during the data import process ensures completion within a cost-effective framework. It is also important to consider the value of imported data for clinical trial operations and related decision-making, providing a comparative analysis between cost and value.

Acknowledgments

This research was supported by the Capital's Funds for Health Improvement and Research (grant No. CFH2022-2Z-2153), and the Beijing Municipal Science & Technology Commission (grant No. Z211100003521008).

Conflicts of Interest

None declared.

References

1. Good clinical practice for drug clinical trial (GCP) 2020[EB/OL]. National Medical Products Administration. 2020 Apr 26. URL: https://www.nmpa.gov.cn/xxgk/fgwj/xzhgfxwj/20200426162401243.html [accessed 2024-06-07]
2. Sheng Q, Wang B, Chen J, et al. Classification and application of electronic source data in clinical trials [Article in Chinese]. Chin Food Drug Admin Magazine 2021;3:36-43 [FREE Full text]
3. Guidance for industry: electronic source data in clinical investigations. Food and Drug Administration. 2013 Sep. URL: https://www.fda.gov/media/85183/download [accessed 2024-06-07]
4. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14(1):1-9. [doi: 10.1197/jamia.M2273] [Medline: 17077452]
5. Garza M, Myneni S, Fenton SH, Zozus MN. eSource for standardized health information exchange in clinical research: a systematic review of progress in the last year. J Soc Clin Data Manag 2021;1(2). [doi: 10.47912/jscdm.66]
6. Beresniak A, Schmidt A, Proeve J, et al. Cost-benefit assessment of using electronic health records data for clinical research versus current practices: contribution of the electronic health records for clinical research (EHR4CR) European project. Contemp Clin Trials 2016 Jan;46:85-91. [doi: 10.1016/j.cct.2015.11.011] [Medline: 26600286]
7. Dong C, Yao C, Gao S, et al. Strengthening clinical research source data management in hospitals to promote data quality of clinical research in China. Chin J Evid Based Med 2019;19:11-1261 [FREE Full text]
8. Data de-identification and anonymization of individual patient data in clinical studies: a model approach. TransCelerate BioPharma. 2015. URL: http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-Data-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-1.pdf [accessed 2024-06-13]
9. Nordo A, Eisenstein EL, Garza M, Hammond WE, Zozus MN. Evaluative outcomes in direct extraction and use of EHR data in clinical trials. Stud Health Technol Inform 2019;257:333-340. [Medline: 30741219]
10. Vattikola A, Dai H, Buckley M, Maniar R. Direct data extraction and exchange of local LABS for clinical research protocols: a partnership with sites, biopharmaceutical firms, and clinical research organizations. J Soc Clin Data Manag 2021 Mar;1(1). [doi: 10.47912/jscdm.21]

11. El Fadly A, Rance B, Lucas N, et al. Integrating clinical research with the healthcare enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform 2011 Dec;44 Suppl 1:S94-S102. [doi: 10.1016/j.jbi.2011.07.007] [Medline: 21888989]
12. Kush R, Alschuler L, Ruggeri R, et al. Implementing single source: the STARBRITE proof-of-concept study. J Am Med Inform Assoc 2007;14(5):662-673. [doi: 10.1197/jamia.M2157] [Medline: 17600107]
13. Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. European Medicines Agency. 2010 Jun 9. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/reflection-paper-expectations-electronic-source-data-data-transcribed-electronic-data-collection_en.pdf [accessed 2024-06-07]
14. Technical conformance guide on electronic study data submissions. Pharmaceuticals and Medical Devices Agency. 2015 Apr 27. URL: https://www.pmda.go.jp/files/000206449.pdf [accessed 2024-06-07]
15. MHRA position statement and guidance: electronic health records. MHRA. 2015 Sep 16. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/470228/Electronic_Health_Records_MHRA_Position_Statement.pdf [accessed 2024-06-07]
16. McCowan C, Thomson E, Szmigielski CA, et al. Using electronic health records to support clinical trials: a report on stakeholder engagement for EHR4CR. Biomed Res Int 2015 Oct;2015:707891. [doi: 10.1155/2015/707891] [Medline: 26539523]
17. Sundgren M, Ammour N, Hydes D, Kalra D, Yeatman R. Innovations in data capture transforming trial delivery. Appl Clin Trials 2021 Aug 12;30(7/8) [FREE Full text]
18. Wang Q, Yingping Y. Governance and application of big data in clinical healthcare. J Med Inform 2018;39:2-6. [doi: 10.3969/j.issn.1673-6036.2018.08.001]
19. Guidance from the general office of the state council on promoting and regulating the development of health care big data applications 2016[EB/OL]. Gov.CN. 2016. URL: https://www.gov.cn/gongbao/content/2016/content_5088769.htm [accessed 2024-06-07]
20. Wang B, Lai J, Liao X, Jin F, Yao C. Challenges and solutions in implementing eSource technology for real-world studies in China: qualitative study among different stakeholders. JMIR Form Res 2023 Aug 10;7:e48363. [doi: 10.2196/48363] [Medline: 37561551]

Abbreviations

ALK: anaplastic lymphoma kinase
CRC: clinical research coordinator
CRF: case report form
EDC: electronic data capture
EHR: electronic health record
EHR4CR: Electronic Health Records for Clinical Research
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level 7
LIMS: laboratory information management system
REUSE: Retrieving EHR Useful Data for Secondary Exploitation
VPN: virtual private network

Edited by C Lovis; submitted 19.09.23; peer-reviewed by H Veldandi, Y Su; revised version received 20.12.23; accepted 18.04.24; published 27.06.24.

Please cite as:
Yuan Y, Mei Y, Zhao S, Dai S, Liu X, Sun X, Fu Z, Zhou L, Ai J, Ma L, Jiang M
Data Flow Construction and Quality Evaluation of Electronic Source Data in Clinical Trials: Pilot Study Based on Hospital Electronic Medical Records in China
JMIR Med Inform 2024;12:e52934
URL: https://medinform.jmir.org/2024/1/e52934
doi:10.2196/52934

Examining Linguistic Differences in Electronic Health Records for Diverse Patients With Diabetes: Natural Language Processing Analysis

Isabel Bilotta[1], PhD; Scott Tonidandel[2], PhD; Winston R Liaw[3], MPH, MD; Eden King[4], PhD; Diana N Carvajal[5], MPH, MD; Ayana Taylor[6], MD; Julie Thamby[7], BA; Yang Xiang[8], PhD; Cui Tao[9], PhD; Michael Hansen[10], MPH, MS, MD

1
2
3
4
5
6
7
8
9
10

Corresponding Author:
Winston R Liaw, MPH, MD

Abstract

Background: Individuals from minoritized racial and ethnic backgrounds experience pernicious and pervasive health disparities that have emerged, in part, from clinician bias.

Objective: We used a natural language processing approach to examine whether linguistic markers in electronic health record (EHR) notes differ based on the race and ethnicity of the patient. To validate this methodological approach, we also assessed the extent to which clinicians perceive linguistic markers to be indicative of bias.

Methods: In this cross-sectional study, we extracted EHR notes for patients who were aged 18 years or older; had more than 5 years of diabetes diagnosis codes; and received care between 2006 and 2014 from family physicians, general internists, or endocrinologists practicing in an urban, academic network of clinics. The race and ethnicity of patients were defined as *White non-Hispanic*, *Black non-Hispanic*, or *Hispanic or Latino*. We hypothesized that Sentiment Analysis and Social Cognition Engine (SEANCE) components (ie, negative adjectives, positive adjectives, joy words, fear and disgust words, politics words, respect words, trust verbs, and well-being words) and mean word count would be indicators of bias if racial differences emerged. We performed linear mixed effects analyses to examine the relationship between the outcomes of interest (the SEANCE components and word count) and patient race and ethnicity, controlling for patient age. To validate this approach, we asked clinicians to indicate the extent to which they thought variation in the use of SEANCE language domains for different racial and ethnic groups was reflective of bias in EHR notes.

Results: We examined EHR notes (n=12,905) of Black non-Hispanic, White non-Hispanic, and Hispanic or Latino patients (n=1562), who were seen by 281 physicians. A total of 27 clinicians participated in the validation study. In terms of bias, participants rated negative adjectives as 8.63 (SD 2.06), fear and disgust words as 8.11 (SD 2.15), and positive adjectives as 7.93 (SD 2.46) on a scale of 1 to 10, with 10 being extremely indicative of bias. Notes for Black non-Hispanic patients contained significantly more negative adjectives (coefficient 0.07, SE 0.02) and significantly more fear and disgust words (coefficient 0.007, SE 0.002) than those for White non-Hispanic patients. The notes for Hispanic or Latino patients included significantly fewer positive adjectives (coefficient −0.02, SE 0.007), trust verbs (coefficient −0.009, SE 0.004), and joy words (coefficient −0.03, SE 0.01) than those for White non-Hispanic patients.

Conclusions: This approach may enable physicians and researchers to identify and mitigate bias in medical interactions, with the goal of reducing health disparities stemming from bias.

(JMIR Med Inform 2024;12:e50428) doi:10.2196/50428

KEYWORDS

bias; sociodemographic factors; health care disparities; natural language processing; sentiment analysis; diabetes; electronic health record; racial; ethnic; diversity; Hispanic; medical interaction

Introduction

Background

Language and communication play a significant, if not primary, role in social relations across different cultures [1]. Language has increasingly been recognized as a relevant form of data that describe relations and behavior [2]. One of the most intimate forms of communication between individuals occurs between clinicians and patients during clinical visits. However, these encounters may be undermined by different forms of bias directed toward patients from certain racial and ethnic minority groups [3]. Generally, *bias* refers to an evaluation, decision, perception, or action in favor of or against a person or group compared to another. Bias can be blatant, wherein it is characterized by deliberate actions (eg, racist comments) that are intentionally and overtly discriminatory [4]. Bias can also be subtle, including "actions that are ambiguous in intent to harm, difficult to detect, low in intensity, and often unintentional but are nevertheless deleterious" to targets [4]. Subtle bias by health care clinicians is linked to negative outcomes for racial and ethnic minority patients, particularly Black non-Hispanic and Hispanic or Latino patients [5].

Race and Racial Bias in Medical Interactions

Health disparities between racial and ethnic groups have historically been attributed to varying levels of socioeconomic status, as well as genetic and biological factors that were thought to predispose groups to different medical conditions. Research has emerged over the past few decades demonstrating that in fact, there is no biological basis for racial and ethnic differences. Humans share 99.9% of their genome, and the 0.1% variation cannot be explained or elucidated by race [6]. Race describes physical traits considered socially significant, and ethnicity denotes a shared cultural heritage, such as language, practices, and beliefs [7]. As such, race and ethnicity are social constructs, and since the landmark report *Unequal Treatment* in 2002 detailed the impact of racial and ethnic discrimination in patient-clinician interactions, research interest in this area has burgeoned [8]. Relative to White non-Hispanic patients, Black non-Hispanic and Hispanic or Latino patients are less likely to ''engender empathic responses from clinicians, establish rapport with clinicians, receive sufficient information, and be encouraged to participate in medical decision making" [9]. A lack of relationship building [10], reduced positive patient and clinician affect [11], decreased patient trust [12], and fewer patient questions [13] are all more likely outcomes for Black non-Hispanic and Hispanic or Latino patients compared to White non-Hispanic patients during medical interactions. Indeed, the 2018 *National Healthcare Disparities Report* revealed that, compared to White non-Hispanic patients, Black non-Hispanic patients receive inferior care on 40% of quality measures, and Hispanic or Latino patients receive worse care on 35% of quality measures, many of which indicate biased and discriminatory behaviors by clinicians [14]. For example, indicators were worse for Black non-Hispanic and Hispanic or Latino patients than White non-Hispanic patients for measures such as "physicians sometimes or never showed respect for what they had to say" and "physicians sometimes or never spent enough time with them" [14]. Black non-Hispanic and Hispanic or Latino patients are more likely to report racial and ethnic bias and discrimination during medical encounters compared to White non-Hispanic patients [15]. Yet, less is known about the manifestations and details of such experiences during the clinician-patient interaction [16] and whether racial and ethnic discrepancies in care can be observed in the content of electronic health records (EHRs). Similar to the thesis described in *Unequal Treatment*, we hypothesized that the mitigation of bias at the clinician level is needed to improve patient outcomes for diverse racial and ethnic populations and narrow the disparities gap. To address bias, researchers need to understand how to measure its existence, and clinicians need to be informed of its manifestations.

Research Contributions

Bias can have many forms—blatant, subtle, malevolent, or benevolent—all of which can be indicated by language. With increasing access to EHR documentation and advances in natural language processing, we may be better equipped to identify differences in clinician encounters with patients of diverse racial and ethnic backgrounds. This study searched for linguistic discrepancies in EHRs using a natural language processing approach followed by linear mixed effect model analyses. EHRs are digital summaries of the clinician-patient encounter and include the clinician's assessment of the interaction, as well as the patient's health history. Since the clinician is responsible for inputting information, as well as reviewing the information inputted by other care clinicians in the EHR for each patient encounter, the contents of the EHR may be particularly useful in illuminating biases that clinicians hold toward patients of different racial and ethnic backgrounds. Although several studies have indicated that clinician bias occurs, particularly in racially and ethnically discordant interactions (ie, when the patient and clinician are of different racial and ethnic backgrounds), relatively little research has examined the ways in which the clinician may be thinking about the patient and how the clinician's sentiment and cognitions are reflected in the language of the EHR [8,17]. EHRs can include many years of patient-clinician interactions, with multiple clinicians having access to them, allowing for biases to be passed on and potentially impact future medical decisions.

Our data set contained EHR notes for a large sample of White non-Hispanic, Black non-Hispanic, and Hispanic or Latino patients with diabetes in the Southern United States. The natural language processing tool, Sentiment Analysis and Social Cognition Engine (SEANCE), was applied to assess multiple linguistic markers in the EHR text [18,19]. We then explored whether 8 of the 20 SEANCE components (see Table 1) differed for patients of different races and ethnicities [20,21].

Table . Description of SEANCE[a] components.

Component label	Indices, n[b]	Key indices[c]	Language examples
Negative adjectives	18	NRC[d] negative adjectives, NRC disgust adjectives, NRC anger adjectives, GI[e] negative adjectives, and Hu-Liu[f] negative adjectives	Unkind, bad, cruel, hurtful, and intolerant
Positive adjectives	9	Hu-Liu positive adjectives, VADER[g] positive adjectives, GI positive adjectives, and Lasswell[h] positive affect adjectives	Supportive, kind, great, and nice
Joy words	8	NRC joy adjectives, NRC anticipation adjectives, and NRC surprise adjectives	Admiration, advocacy, elated, glad, liking, and pleased
Fear and disgust words	8	NRC disgust nouns, NRC negative nouns, NRC fear nouns, and NRC anger nouns	Abnormal, adverse, attack, cringe, criticize, distress, intimidate, unequal, and stigma
Politics words	7	GI politics nouns and Lasswell power nouns	Alliance, ally, authorize, civil, concession, consent, and oppose
Respect words	4	Lasswell respect nouns	Status, honor, recognition, and prestige
Trust verbs	5	NRC trust verbs, NRC joy verbs, and NRC positive verbs	Affirm, advise, confide, and cooperating
Well-being words	4	Lasswell well-being physical nouns and Lasswell well-being total nouns	Alive, ambulance, adjust, afraid, blood, clinic, and nutrition

[a]SEANCE: Sentiment Analysis and Social Cognition Engine.

[b]Indices refer to the number of dictionary lists from which the component was developed.

[c]The key indices came from the following dictionary lists: NRC Emotion Lexicon [18,22], the Harvard-IV dictionary list used by the General Inquirer [23], the Hu-Liu polarity word lists [22,23], the Valence Aware Dictionary and Sentiment Reasoner [24], the Lasswell dictionary lists [25,26], and the Geneva Affect Label Coder database [27]. For a thorough review of the SEANCE indices and corresponding dictionaries, see Crossley et al [18].

[d]NRC: NRC Emotion Lexicon.

[e]GI: General Inquirer.

[f]Hu-Liu: Hu-Liu polarity word lists.

[g]VADER: Valence Aware Dictionary and Sentiment Reasoner.

[h]Lasswell: Lasswell dictionary lists.

We hypothesized that the SEANCE components for negative adjectives, positive adjectives, joy words, fear and disgust words, politics words, respect words, trust verbs, and well-being words and the mean word count in the notes would be indicators of bias, as these concepts have been linked to bias in nonmedical contexts. Ng's [28] review of linguistic racial bias in verbiage offers the rationale for our choice of fear and disgust words, politics words, respect words, and trust verbs as indicators of bias, whereas the work of Li et al [29] examining gender differences in standardized writing assessment provides further support for our use of SEANCE as a tool for examining biases in language. We selected positive and negative adjectives, well-being words, politics words, and word count indicators as prior research demonstrates that clinicians may be less likely to establish rapport and provide appropriate medications and are more inclined to show negative attitudes and be dismissive toward Black non-Hispanic and Hispanic or Latino patients as a result of their unconscious racial and ethnic biases [30-33].

Specifically, we investigated which aspects of communication differ and whether differences are indicative of biased interactions. Any systematic variation in language can convey differential perceptions, attitudes, and expectations. For example, words such as "resistant" or "non-compliant" could reflect bias if (all else being equal) they tend to be used more to reflect people from some racial or ethnic backgrounds than others. This work aimed to elucidate for clinicians and researchers where discrepancies in communication emerge in the EHR and whether these differences are indicative of racial and ethnic bias. We also assessed the extent to which clinicians perceive linguistic markers to be indicative of bias.

Methods

Sample

This was a cross-sectional study using EHR-derived physician notation of outpatient clinical encounters. We extracted EHR encounters (n=15,460) for patients (n=1647) who were aged 18 years or older; had more than 5 years of diabetes diagnosis codes; and received care between 2006 and 2014 from family physicians, general internists, or endocrinologists practicing in an urban, academic network of clinics. We chose this disease

because of its high prevalence (11.3% in the United States) and chose to examine outpatient visits because of the relative scope of annual outpatient visits (1 billion) relative to hospital admissions (32 million) [34-36]. The demographic variables collected were patient race and ethnicity, sex, and age. The race and ethnicity of patients were defined as *White non-Hispanic*, *Black non-Hispanic*, or *Hispanic or Latino* (see Table 2 for a summary of patient demographics).

Table . Patient demographics of the final sample.

Variable		Value (n=1562)
Age (years)		
	Mean (SD)	68.74 (13.76)
	Range	20-102
	Median (IQR)	69 (61-78)
Sex, n (%)		
	Female	871 (55.74)
	Male	691 (44.26)
Race and ethnicity, n (%)		
	White non-Hispanic	682 (43.66)
	Black non-Hispanic	755 (48.34)
	Hispanic or Latino	125 (8)

SEANCE Algorithm

SEANCE is a lexical scoring algorithm that includes over 200 word vectors (also referred to as indices or features) designed to assess sentiment, cognition, and social order, which were developed from preexisting and widely used databases such as EmoLex and SenticNet [22,37]. In addition to the core indices, SEANCE allows for several customized indices, including filtering for particular parts of speech and controlling for instances of negation [18]. Since SEANCE computes such a large quantity of indices, Crossley et al [18] developed 20 components from all the indices using principal component analysis (PCA) [18]. These components are essentially clusters of related indices in SEANCE and allow users to interpret the SEANCE output at a more macro level. This process enabled them to summarize the SEANCE indices into a smaller and more interpretable set of variables. In the PCA by Crossley et al [18], they retained even the smallest components, setting a conservative cutoff point for inclusion (ie, 1% for variance explained by each component). The analyses for this research were run on a subset of 8 of the 20 *components* that Crossley et al [18] developed. We selected these 8 components a priori (see Table 1 for a description of the selected components).

We chose SEANCE instead of other natural language processing tools, such as Linguistic Inquiry and Word Count (LIWC), because it contains a larger number of core indices taken from multiple lexicons, as well as 20 components, and is based on the most recent improvements in sentiment analysis [18]. In their validation of SEANCE, Crossley et al [18] found that SEANCE components demonstrated significantly greater accuracy than LIWC indices (P<.001) for 3 of the 4 review types examined. In addition to the core indices, SEANCE allows for several customized indices, including filtering for parts of speech (also known as "parts-of-speech tagging") and controlling for instances of negation, which LIWC does not offer. We analyzed all words in the EHR (ie, *not* single parts of speech), but we controlled for negation. For example, this means that "not good" would be recognized as *not being positive* by SEANCE, as opposed to LIWC, which would see the word "good" and count it as positive.

Validation of the Sentiment Analysis Approach

To provide validation of the sentiment analysis approach used in this study, we surveyed subject-matter experts in EHR note writing (ie, physicians, physician assistants, and nurse practitioners) to garner their perspectives on the appropriateness of the linguistic components identified in our pilot study as indicators of subtle racial and ethnic bias in EHR notes. The team of researchers for this study included industrial-organizational psychologists who have expertise in bias and discrimination; however, it was also valuable to garner opinions from clinicians who are experts in EHR note writing and who understand the differences in the types of language used. To recruit participants, we used a combination of opportunistic and snowball sampling, starting with individuals within our personal networks. Through a web-based program, we asked participants to indicate the extent to which they thought the language domains (eg, negative adjectives, fear and disgust words, etc) were reflective of bias in EHR notes. Participants were told the following:

> *One type of language that could represent bias reflects the amount of NEGATIVE ADJECTIVES contained in the electronic health record. Examples of negative adjectives include "unkind," "bad," "harmful," "intolerant," and "stupid." If these kinds of words were used to describe Black or LatinX patients more than White patients, to what extent do you think this would be indicative of racial bias? Please indicate the extent of your agreement on the 1 to 10 scale below.*

The same formatting was used for each of the linguistic components, with component-specific language examples offered so participants understood the types of sentiment that each component was designed to assess.

Cross-Classified Linear Mixed Effects Models

We used the *lme4* package in R (R Foundation for Statistical Computing) to perform linear mixed effects analyses of the relationships between the outcomes of interest (SEANCE components and word count) and patient race and ethnicity, controlling for patient age. We ran an identical analysis, treating 8 different SEANCE components and the mean word count in the EHR as the dependent variables, while leaving all other variables consistent across the models. The same steps of entering fixed and random effects were applied across all cross-classified linear mixed effects models with different dependent variables (ie, negative adjectives, positive adjectives, well-being words, trust verbs, fear and disgust words, joy words, politics words, respect words, and mean word count).

We first ran a null model with only the random intercepts. We then added random effects and applied a crossed design (vs a traditional nested structure), leading us to have intercepts for physicians and patients. Then, we ran a model with the random intercepts as well as the fixed effects. As fixed effects, we entered *race and ethnicity* and *age* (without an interaction term) into the model. For all models examined, the intercept variation can be attributed primarily to different physicians rather than patients. We used a 95% CI to determine statistical significance. To be more conservative, given that we ran multiple tests, we also computed an additional set of CIs at the 99th percentile.

Ethical Considerations

We obtained ethics approval from the University of Texas Health Science Center's Committee for the Protection of Human Subjects (HSC-MS-18-0431) and the Rice University Institutional Review Board (IRB-FY2021-325). Participants consented and received a US $25 gift card after completing the survey. EHR data were deidentified prior to the analysis.

Results

Description and Justification for Cross-Classified Analyses

An initial inspection of the data revealed that 2 physicians were extreme outliers, accounting for 16.53% (2555/15,460) of the notes in our sample. To ensure that the overrepresentation of these physicians would not bias the results, we removed those notes from the data set (taking us from our initial sample of 15,460 visits with 283 physicians and 1647 patients to 12,905 visits with 281 physicians and 1562 patients; Table 2). The distribution of visits by patients indicates an average of 8.27 visits per patient with a minimum of 1, a median of 5, and a maximum of 97. Physicians see 11.72 patients on average, with a median of 2 and a maximum of 143, suggesting a skewed distribution. Despite the relatively large number of patients seen by some physicians, these physicians accounted for substantially fewer patient notes than the 2 physicians that were previously removed. Patients see 2.11 physicians on average, with a minimum of 1 and a maximum of 12; however, the distribution suggests that 6.6% (109/1647) of patients saw 5 or more physicians. Moreover, 742 (45.1%) of the 1647 patients saw 1 physician, whereas 119 (7.2%) saw 4 physicians. In our data set, patients can have multiple visits to a variety of physicians, indicating that patient visits are not nested within physicians. Further, physicians may see different patients with no consistent overlap of patients between physicians, indicating that physicians are not nested within patients. Thus, there is no clear hierarchical nesting of patients within physicians (or vice versa), which suggests that a cross-classified design is more appropriate than a traditional, hierarchical, multilevel model structure.

Cross-Classified Linear Mixed Effects Model Results

In the negative adjective component model (Table 3), the random effects of patient (σ^2=0.02) and physician (σ^2=0.12) indicated that intercept variation in use of negative adjectives is mainly a function of the physician rather than the patient. The physician random effect was over 5 times as large as the random effect for the patient; the intraclass correlation (ICC) for physicians was 0.41 and the ICC for patients was 0.07 (ICC_{total}=0.481). This pattern of results in random effects and ICC values for patients and physicians was consistent across the other 8 models. Overall, 2 of the 5 relationships (ie, the significant difference in positive adjectives for Hispanic or Latino and White non-Hispanic patient notes, and the significant difference in trust verbs for Hispanic or Latino and White non-Hispanic patient notes) that were previously significant at the 95th percentile had CIs that included zero at the 99th percentile. For 3 of the SEANCE components—well-being, politics, and respect words—and for the overall word count, there was not a statistically significant difference between the 3 races and ethnicities. In contrast, for all the other remaining SEANCE components, there was a statistically significant race and ethnicity effect for either Black non-Hispanic or Hispanic or Latino patients relative to White non-Hispanic patients. Specifically, notes for Black non-Hispanic patients contained significantly more negative adjectives and fear and disgust words than those for White non-Hispanic patients. Notes for Hispanic or Latino patients included significantly fewer positive adjectives, trust verbs, and joy words than those for White non-Hispanic patients. As such, across most of the SEANCE components, we observed favoritism of White non-Hispanic patients in terms of note content.

Table . Fixed effects model results for negative adjectives, positive adjectives, well-being words, trust verbs, joy words, politics words, respect words, fear and disgust words, and word count.

Variables[a]	Negative adjectives	Positive adjectives	Well-being words	Trust verbs	Joy words	Politics words	Respect words	Fear and disgust words	Word count
Fixed effect estimates									
Age (years)									
β (SE)	−0.00 (0.00)	0.00 (0.00)	.0002 (0.00009)	−0.00007 (0.00008)	0.000002 (0.0002)	−0.00009 (0.00004)	−0.00004 (0.00005)	0.000005 (0.00)	−0.43 (0.68)
95% CI	−0.002 to 0.0003	−0.002 to 0.00	0.0006 to 0.0004[b]	−0.002 to 0.0008	−0.0004 to 0.0004	−0.0002 to −0.000007[b]	−0.0004 to 0.0004	−0.0001 to 0.0002	−1.76 to 0.90
Race and ethnicity									
White non-Hispanic (reference)									
β (SE)	0.42 (0.05)	−0.24 (0.017)	0.18 (0.007)	0.16 (0.007)	0.32 (0.02)	0.07 (0.003)	0.05 (0.004)	0.17 (0.007)	868.50 (54.45)
95% CI	0.32 to 0.53	−0.26 to −0.21	0.17 to 0.20	0.14 to 0.17	0.28 to 0.35	0.06 to 0.07	0.04 to 0.05	0.16 to 0.19	761.84 to 975.17
Black non-Hispanic									
β (SE)	0.07 (0.02)	0.02 (0.004)	0.004 (0.002)	−0.003 (0.002)	−0.01 (0.006)	0.001 (0.001)	−0.001 (0.001)	0.007 (0.002)	20.61 (19.01)
95% CI	0.04 to 0.11[b]	−0.006 to 0.01	−0.0007 to 0.009	−0.007 to 0.001	−0.02 to 0.0004	−0.001 to 0.004	−0.004 to 0.002	0.003 to 0.01[b]	−16.71 to 57.84
Hispanic or Latino									
β (SE)	0.02 (0.03)	−0.02 (0.007)	0.002 (0.004)	−0.009 (0.004)	−0.03 (0.01)	−0.0009 (0.003)	0.0006 (0.002)	−0.002 (0.004)	15.73 (32.30)
95% CI	−0.03 to 0.08	−0.03 to −0.004[b]	−0.007 to 0.01	−0.02 to −0.001[b]	−0.05 to −0.01[b]	−0.005 to 0.003	−0.004 to 0.005	−0.01 to 0.006	−47.61 to 78.98
Random effects, estimate (SE)									
U0 patient	0.02 (0.14)	0.0008 (0.03)	0.0004 (0.02)	0.0002 (0.02)	0.0006 (0.02)	0.00001 (0.004)	0.00002 (0.005)	0.0004 (0.02)	27,878 (167.0)
U0 physician	0.12 (0.34)	0.006 (0.08)	0.003 (0.05)	0.003 (0.05)	0.02 (0.15)	0.0002 (0.016)	0.0005 (0.02)	0.003 (0.05)	119,489 (345.7)

[a]Random effects are presented as estimate and SE. For the fixed effect estimates, cell entries are parameter (β) estimates, SE, and 95% CIs. White non-Hispanic was the reference group for race and ethnicity.

[b]Significant effects based on the 95% CIs.

Sentiment Analysis Validation

In all, 27 participants completed the surveys (see Multimedia Appendix 1 for the demographics of the participants). On a scale of 1 to 10, with 10 being extremely indicative of bias, participants rated negative adjectives as 8.63 (SD 2.06), fear and disgust words as 8.11 (SD 2.15), positive adjectives as 7.93 (SD 2.46), trust verbs as 7.56 (SD 2.64), and joy words as 6.81 (SD 2.47). The means and SDs for each of the components are reported in Table 4. The results of this preliminary analysis provide support for the validity of the linguistic components as indicators of bias in EHRs, as our sample of clinicians regard them as highly suggestive of bias if used differently for patients of diverse racial and ethnic backgrounds.

Table . Subject-matter expert assessment of bias based on specific linguistic markers.

Component	Score, mean (SD)[a]
Negative adjectives	8.63 (2.06)
Fear and disgust words	8.11 (2.15)
Positive adjectives	7.93 (2.46)
Joy words	6.81 (2.47)
Trust verbs	7.56 (2.64)
Politics words	7.07 (2.32)
Respect nouns	7.56 (2.55)
Well-being words	5.56 (2.55)
Mean word count	6.11 (2.19)

[a]Scale ranges from 1 (*Not at all indicative of bias*) to 10 (*Extremely indicative of bias*).

Discussion

Principal Findings

We found that the words that physicians use in EHR notes differ based on the racial and ethnic backgrounds of patients. Specifically, for Black non-Hispanic patients, notes consisted of words that convey negativity, fear, and disgust. When seeing Hispanic or Latino patients, physicians used fewer positive words and were less likely to use words that communicate trust and joy. Our findings are consistent with others who have documented that physicians communicate in the EHR differently (ie, more negatively) when caring for patients from some minority groups [9,17], which may ultimately result in adverse and inequitable health outcomes for patients. Our results also align with other papers that found that stigmatizing language is more commonly used in EHRs for minority populations [38-42]. Those papers used language guidelines [38] and experts [39] to identify stigmatizing language. We came to a similar conclusion by using established language dictionaries and contend that our approach allows for a more comprehensive assessment of language. For example, a prior paper used 15 descriptors [42]. In contrast, our approach encompasses tens of thousands of words, including multiple word lists, positive and negative sentiments, and emotions. Thus, this method does not merely capture the presence or absence of stigmatizing language, but rather offers a broader glimpse of the clinician-patient relationship. Furthermore, the validation survey confirmed that subject-matter experts perceive the types of words included in this study to be indicative of bias when used differentially for patients of diverse racial and ethnic backgrounds. Taken together, these findings indicate that the language used differs for patients based on racial and ethnic backgrounds and that those differences are suggestive of bias. As a result, our paper is the first to use this particular method to examine outpatient, diabetes notes. Since diabetes quality measures already exist, our analysis allows researchers to link bias to differences in quality in future studies [43].

EHR notes are important, although imperfect, assessments of physician attitudes toward their patients. With more and more time now being devoted to EHR documentation, physicians are increasingly burned out, which has led to the adoption of more efficient data entry strategies such as using templates, copy-pasting previous text, and inserting preset language [44,45]. Consequently, notes can be standardized, limiting our ability to assess physician attitudes and subconscious biases toward patients. Despite these caveats, notes remain the definitive and often sole account of what happened in the examination room, and based on these data, Black non-Hispanic and Hispanic or Latino patients are written about differently than White non-Hispanic patients.

The method described in this paper offers a scalable blueprint that provides clinicians with data about their interactions with patients and overcomes limitations of other traditional measures of bias. Existing measures require primary data collection through surveys, videotaped encounters, and confederate observations. Surveys assess perceptions of interactions and are prone to retrospective bias and socially desirable responding, whereas the time-consuming nature of encounters and observations lack scalability and limit the number of clinicians that can receive feedback at any given time. The relevance of alternative measures has also been questioned. For example, critics of the implicit association test have asked whether performance on the test is applicable to real-world contexts [46], which may explain why some change their behavior when confronted with their own biases, whereas others do not [5,47]. In contrast, our method uses data that are automatically and universally collected through the course of delivering care and generated by physicians in actual encounters.

Limitations

When interpreting our results, several limitations should be considered. First, due to limitations in our data, we are unable to determine which additional team members, including scribes, medical assistants, and residents, contributed to the notes. However, attending physicians are ultimately responsible for the content and have the authority and responsibility to modify language that is inconsistent with their values. Second, we lack information about physicians in this sample and do not have access to physician demographic characteristics (eg, their racial and ethnic backgrounds), although this would be an important next step. We attempted to account for this limitation by comparing language within rather than across physicians. Third, we included all language within notes, including physical exams,

medications, and past medical histories. These sections can be guided by templates or not actively entered by physicians. We retained these parts in case the language within these sections contributed to variation. An alternative approach could assess only the history of present illness, assessment, and plan sections of the note and could yield different results. Additional work is needed to determine whether differential word choices reflect attitudes and behaviors toward patients. EHR notes serve a wide range of purposes. They convey medical information to others, remind physicians of their impressions, communicate plans to patients, provide justification for billing codes, and serve as legal evidence [44]. Thus, specific phrases (eg, worsening, uncontrolled, or adherence) may be required for billing, compliance, and legal purposes and may not reflect bias toward patients. Finally, these results may not be generalizable to other conditions. Our findings may be unique to the language used for diabetes care and by clinicians who manage diabetes. Determining whether these results persist for different diseases (eg, cancer, heart disease, and acute injuries) is an important next step.

Directions for Future Research

Additional research is needed to interpret and provide context for this exploratory work. To determine whether these measures are associated with bias, subject-matter experts could label notes using known patterns of bias (eg, the ratio of collective to personal pronouns, the amount and level of abstraction of speech, and passive vs active voice) [48]. Further research is needed to understand whether biased language in notes reflects biased behaviors during encounters as well as inequitable health outcomes for some racial and ethnic minority populations. Conducting further experiments (eg, with research actors as patients in a mock medical visit) could help determine whether biased language in notes reflects manifestations of bias during encounters (eg, less eye contact, hostile language, or less time spent on education and counseling). If bias is confirmed, we need to determine whether clinicians who use differential language provide worse care and quality for minority patients. Ultimately, this tool may be used to identify and mitigate bias. Future studies should assess whether receiving feedback using this method leads to behavior change and whether changing the language used in EHR notes leads to changes in patient interactions. Although many strategies for reducing bias exist—such as affirming egalitarian goals, seeking common-group identities, perspective taking, and individuation—it is unclear which approach best complements our proposed method [5].

Conclusion

In this novel, exploratory work, we used natural language processing and found that compared to encounters with White non-Hispanic patients, physicians use language conveying more negativity, fear, and disgust in their encounters with some racial and ethnic minority patients. If confirmed in future studies, these features could be used to make clinicians aware of their biases with the goal of reducing racial and ethnic discrimination and the resulting health inequities.

Acknowledgments

This work was supported by a Rice Anti-Racism Research Grant through Rice University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographics of the validation study participants.
[DOCX File, 14 KB - medinform_v12i1e50428_app1.docx]

References

1. Maass A, Karasawa M, Politi F, Suga S. Do verbs and adjectives play different roles in different cultures? a cross-linguistic analysis of person representation. J Pers Soc Psychol 2006 May;90(5):734-750. [doi: 10.1037/0022-3514.90.5.734] [Medline: 16737371]
2. Boroditsky L, Schmidt LA, Phillips W. Sex, syntax, and semantics. In: Gentner D, Goldin-Meadow S, editors. Language in Mind: Advances in the Study of Language and Thought: The MIT Press; 2003. [doi: 10.7551/mitpress/4117.001.0001]
3. Hall WJ, Chapman MV, Lee KM, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am J Public Health 2015 Dec;105(12):e60-e76. [doi: 10.2105/AJPH.2015.302903] [Medline: 26469668]
4. Jones KP, Peddie CI, Gilrane VL, King EB, Gray AL. Not so subtle: a meta-analytic investigation of the correlates of subtle and overt discrimination. J Manag 2016 Jul 10;42(6):1588-1613. [doi: 10.1177/0149206313506466]
5. Zestcott CA, Blair IV, Stone J. Examining the presence, consequences, and reduction of implicit bias in health care: a narrative review. Group Process Intergroup Relat 2016 Jul;19(4):528-542. [doi: 10.1177/1368430216642029] [Medline: 27547105]
6. Ahn SM, Kim TH, Lee S, et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 2009 Sep;19(9):1622-1629. [doi: 10.1101/gr.092197.109] [Medline: 19470904]

7. Chadha N, Lim B, Kane M, Rowland B. Toward the abolition of biological race in medicine. Othering & Belonging Institute. 2020 May 13. URL: https://belonging.berkeley.edu/toward-abolition-biological-race-medicine-8 [accessed 2023-06-27]
8. Institute of Medicine. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care: National Academies Press; 2003. [doi: 10.17226/10260]
9. Ferguson WJ, Candib LM. Culture, language, and the doctor-patient relationship. Fam Med 2002 May;34(5):353-361. [Medline: 12038717]
10. Siminoff LA, Graham GC, Gordon NH. Cancer communication patterns and the influence of patient characteristics: disparities in information-giving and affective behaviors. Patient Educ Couns 2006 Sep;62(3):355-360. [doi: 10.1016/j.pec.2006.06.011] [Medline: 16860520]
11. Johnson RL, Roter D, Powe NR, Cooper LA. Patient race/ethnicity and quality of patient–physician communication during medical visits. Am J Public Health 2004 Dec;94(12):2084-2090. [doi: 10.2105/ajph.94.12.2084] [Medline: 15569958]
12. Jacobs EA, Rolle I, Ferrans CE, Whitaker EE, Warnecke RB. Understanding African Americans' views of the trustworthiness of physicians. J Gen Intern Med 2006 Jun;21(6):642-647. [doi: 10.1111/j.1525-1497.2006.00485.x] [Medline: 16808750]
13. Eggly S, Hamel LM, Foster TS, et al. Randomized trial of a question prompt list to increase patient active participation during interactions with Black patients and their oncologists. Patient Educ Couns 2017 May;100(5):818-826. [doi: 10.1016/j.pec.2016.12.026] [Medline: 28073615]
14. National Healthcare Quality & Disparities Reports. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/research/findings/nhqrdr/index.html [accessed 2023-06-27]
15. Shavers VL, Fagan P, Jones D, et al. The state of research on racial/ethnic discrimination in the receipt of health care. Am J Public Health 2012 May;102(5):953-966. [doi: 10.2105/AJPH.2012.300773] [Medline: 22494002]
16. Penner LA, Dovidio JF, West TV, et al. Aversive racism and medical interactions with Black patients: a field study. J Exp Soc Psychol 2010 Mar 1;46(2):436-440. [doi: 10.1016/j.jesp.2009.11.004] [Medline: 20228874]
17. Hagiwara N, Slatcher RB, Eggly S, Penner LA. Physician racial bias and word use during racially discordant medical interactions. Health Commun 2017 Apr;32(4):401-408. [doi: 10.1080/10410236.2016.1138389] [Medline: 27309596]
18. Crossley SA, Kyle K, McNamara DS. Sentiment Analysis and Social Cognition Engine (SEANCE): an automatic tool for sentiment, social cognition, and social-order analysis. Behav Res Methods 2017 Jun;49(3):803-821. [doi: 10.3758/s13428-016-0743-z] [Medline: 27193159]
19. Crossley SA, Skalicky S, Dascalu M. Moving beyond classic readability formulas: new methods and new models. J Res Read 2019 Nov;42(3-4):541-561. [doi: 10.1111/1467-9817.12283]
20. Hu M, Liu B. Mining and summarizing customer reviews. In: KDD '04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Association for Computing Machinery; 2004:168-177. [doi: 10.1145/1014052.1014073]
21. Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the web. In: WWW '05: Proceedings of the 14th International Conference on World Wide Web: Association for Computing Machinery; 2005:342-351. [doi: 10.1145/1060745.1060797]
22. Mohammad SM, Turney PD. Emotions evoked by common words and phrases: using Mechanical Turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text: Association for Computational Linguistics; 2010:26-34 URL: https://aclanthology.org/W10-0204/ [accessed 2024-05-10]
23. Stone PJ, Dunphy DC, Smith MS. The General Inquirer: A Computer System for Content Analysis: MIT Press; 1966.
24. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media 2014 May 16;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
25. Lasswell HD, Namenwirth J. The Lasswell Value Dictionary: Yale University Press; 1969.
26. Namenwirth J, Weber R. Dynamics of Culture: Allen & Unwin; 1987.
27. Scherer KR. What are emotions? and how can they be measured? Social Science Information 2005 Dec;44(4):695-729. [doi: 10.1177/0539018405058216]
28. Ng SH. Language-based discrimination: blatant and subtle forms. J Lang Soc Psychol 2007 Jun;26(2):106-122. [doi: 10.1177/0261927X07300074]
29. Li Z, Chen MY, Banerjee J. Using corpus analyses to help address the DIF interpretation: gender differences in standardized writing assessment. Front Psychol 2020 Jun 3;11:1088. [doi: 10.3389/fpsyg.2020.01088] [Medline: 32581944]
30. Blair IV, Steiner JF, Fairclough DL, et al. Clinicians' implicit ethnic/racial bias and perceptions of care among Black and Latino patients. Ann Fam Med 2013;11(1):43-52. [doi: 10.1370/afm.1442] [Medline: 23319505]
31. Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. J Gen Intern Med 2013 Nov;28(11):1504-1510. [doi: 10.1007/s11606-013-2441-1] [Medline: 23576243]
32. Sabin JA, Greenwald AG. The influence of implicit bias on treatment recommendations for 4 common pediatric conditions: pain, urinary tract infection, attention deficit hyperactivity disorder, and asthma. Am J Public Health 2012 May;102(5):988-995. [doi: 10.2105/AJPH.2011.300621] [Medline: 22420817]
33. Sue DW, Capodilupo CM, Torino GC, et al. Racial microaggressions in everyday life: implications for clinical practice. Am Psychol 2007;62(4):271-286. [doi: 10.1037/0003-066X.62.4.271] [Medline: 17516773]

34. Statistics about diabetes. American Diabetes Association. URL: https://diabetes.org/about-us/statistics/about-diabetes [accessed 2023-06-28]
35. Ambulatory care use and physician office visits. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/fastats/physician-visits.htm [accessed 2023-06-28]
36. Fast facts on U.S. hospitals. American Hospital Association. URL: https://www.aha.org/statistics/fast-facts-us-hospitals [accessed 2023-06-28]
37. Cambria E, Havasi C, Hussain A. SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis. In: Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2012): AAAI Press; 2012:202-207 URL: https://cdn.aaai.org/ocs/4411/4411-21497-1-PB.pdf [accessed 2024-05-10]
38. Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. JAMA Netw Open 2022 Jan 4;5(1):e2144967. [doi: 10.1001/jamanetworkopen.2021.44967] [Medline: 35084481]
39. Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. Health Aff (Millwood) 2022 Feb;41(2):203-211. [doi: 10.1377/hlthaff.2021.01423] [Medline: 35044842]
40. Barcelona V, Scharp D, Idnay BR, et al. A qualitative analysis of stigmatizing language in birth admission clinical notes. Nurs Inq 2023 Jul;30(3):e12557. [doi: 10.1111/nin.12557] [Medline: 37073504]
41. Goddu PA, O'Conor KJ, Lanzkron S, et al. Do words matter? stigmatizing language and the transmission of bias in the medical record. J Gen Intern Med 2018 May;33(5):685-691. [doi: 10.1007/s11606-017-4289-2] [Medline: 29374357]
42. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. JAMA Netw Open 2021 Jul 1;4(7):e2117052. [doi: 10.1001/jamanetworkopen.2021.17052] [Medline: 34259849]
43. Comprehensive diabetes care (CDC). National Committee for Quality Assurance. URL: http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2016-table-of-contents/diabetes-care [accessed 2017-11-29]
44. Ho YX, Gadd CS, Kohorst KL, Rosenbloom ST. A qualitative analysis evaluating the purposes and practices of clinical documentation. Appl Clin Inform 2014 Feb 26;5(1):153-168. [doi: 10.4338/ACI-2013-10-RA-0081] [Medline: 24734130]
45. Weis JM, Levy PC. Copy, paste, and cloned notes in electronic health records. Chest 2014 Mar;145(3):632-638. [doi: 10.1378/chest.13-0886] [Medline: 27845637]
46. Sukhera J, Wodzinski M, Rehman M, Gonzalez CM. The implicit association test in health professions education: a meta-narrative review. Perspect Med Educ 2019 Oct;8(5):267-275. [doi: 10.1007/s40037-019-00533-8] [Medline: 31535290]
47. van Ryn M, Hardeman R, Phelan SM, et al. Medical school experiences associated with change in implicit racial bias among 3547 students: a medical student CHANGES study report. J Gen Intern Med 2015 Dec;30(12):1748-1756. [doi: 10.1007/s11606-015-3447-7] [Medline: 26129779]
48. von Hippel W, Sekaquaptewa D, Vargas P. The linguistic intergroup bias as an implicit indicator of prejudice. J Exp Soc Psychol 1997 Sep;33(5):490-509. [doi: 10.1006/jesp.1997.1332]

Abbreviations

EHR: electronic health record
ICC: intraclass correlation
LIWC: Linguistic Inquiry and Word Count
PCA: principal component analysis
SEANCE: Sentiment Analysis and Social Cognition Engine

Edited by C Lovis; submitted 30.06.23; peer-reviewed by B Senst, M Chatzimina, X Jing; revised version received 26.09.23; accepted 23.04.24; published 23.05.24.

Please cite as:
Bilotta I, Tonidandel S, Liaw WR, King E, Carvajal DN, Taylor A, Thamby J, Xiang Y, Tao C, Hansen M
Examining Linguistic Differences in Electronic Health Records for Diverse Patients With Diabetes: Natural Language Processing Analysis
JMIR Med Inform 2024;12:e50428
URL: https://medinform.jmir.org/2024/1/e50428
doi:10.2196/50428

A Scalable Pseudonymization Tool for Rapid Deployment in Large Biomedical Research Networks: Development and Evaluation Study

Hammam Abu Attieh[1], MSc; Diogo Telmo Neves[1], BSc; Mariana Guedes[2,3,4], MSc, MD; Massimo Mirandola[5], PhD; Chiara Dellacasa[6], MSc; Elisa Rossi[6], MSc; Fabian Prasser[1], Prof Dr

1

2

3

4

5

6

Corresponding Author:
Hammam Abu Attieh, MSc

Abstract

Background: The SARS-CoV-2 pandemic has demonstrated once again that rapid collaborative research is essential for the future of biomedicine. Large research networks are needed to collect, share, and reuse data and biosamples to generate collaborative evidence. However, setting up such networks is often complex and time-consuming, as common tools and policies are needed to ensure interoperability and the required flows of data and samples, especially for handling personal data and the associated data protection issues. In biomedical research, pseudonymization detaches directly identifying details from biomedical data and biosamples and connects them using secure identifiers, the so-called pseudonyms. This protects privacy by design but allows the necessary linkage and reidentification.

Objective: Although pseudonymization is used in almost every biomedical study, there are currently no pseudonymization tools that can be rapidly deployed across many institutions. Moreover, using centralized services is often not possible, for example, when data are reused and consent for this type of data processing is lacking. We present the ORCHESTRA Pseudonymization Tool (OPT), developed under the umbrella of the ORCHESTRA consortium, which faced exactly these challenges when it came to rapidly establishing a large-scale research network in the context of the rapid pandemic response in Europe.

Methods: To overcome challenges caused by the heterogeneity of IT infrastructures across institutions, the OPT was developed based on programmable runtime environments available at practically every institution: office suites. The software is highly configurable and provides many features, from subject and biosample registration to record linkage and the printing of machine-readable codes for labeling biosample tubes. Special care has been taken to ensure that the algorithms implemented are efficient so that the OPT can be used to pseudonymize large data sets, which we demonstrate through a comprehensive evaluation.

Results: The OPT is available for Microsoft Office and LibreOffice, so it can be deployed on Windows, Linux, and MacOS. It provides multiuser support and is configurable to meet the needs of different types of research projects. Within the ORCHESTRA research network, the OPT has been successfully deployed at 13 institutions in 11 countries in Europe and beyond. As of June 2023, the software manages data about more than 30,000 subjects and 15,000 biosamples. Over 10,000 labels have been printed. The results of our experimental evaluation show that the OPT offers practical response times for all major functionalities, pseudonymizing 100,000 subjects in 10 seconds using Microsoft Excel and in 54 seconds using LibreOffice.

Conclusions: Innovative solutions are needed to make the process of establishing large research networks more efficient. The OPT, which leverages the runtime environment of common office suites, can be used to rapidly deploy pseudonymization and biosample management capabilities across research networks. The tool is highly configurable and available as open-source software.

(JMIR Med Inform 2024;12:e49646) doi:10.2196/49646

KEYWORDS

biomedical research; research network; data sharing; data protection; privacy; pseudonymization

Introduction

Background

As a response to the SARS-CoV-2 pandemic, many research projects have been rapidly set up to study the virus, its impact, and possible interventions [1,2]. This accelerated the general trend toward large collaborative networks in biomedical research [3,4]. These are motivated by the need to generate sufficiently large data sets and collections of biosamples, which are essential for developing new methods of personalized medicine and generating real-world evidence [5]. However, setting up such networks usually takes quite some time, as common tools and policies are needed to achieve interoperability and enable the required flows of data and biosamples [6,7]. One area in which this challenge is frequently encountered is the handling of personal data and the related data protection issues, which can arise in all processing steps, from collection [8] to sharing [9] and even analysis and visualization [10].

Laws and regulations, such as the European Union General Data Protection Regulation (GDPR) [11] or the US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [12], advocate for various strategies for the protection of personal data. In general terms, the GDPR prohibits the processing of sensitive categories of personal data, including medical data, unless consent is given. However, under certain conditions, processing is also possible without consent if technical and organizational safeguards are implemented [13]. Although there is no consensus on which protection methods are best suited for use in biomedical research [14], pseudonymization (also called coding or pseudo-anonymization) [15] is a common strategy, which can also be used to deidentify data under the HIPAA Privacy Rule. Pseudonymization is an essential aspect of the GDPR, as it is mentioned in multiple articles, in particular as a data minimization measure [16]. In this privacy-by-design approach, directly identifying data about study subjects are stored separately from biomedical data and biosamples, which are needed for scientific analyses [17]. The link between the different types of data and assets is established through secure identifiers, the so-called pseudonyms [18], which enable data linkage and allow the reidentification of subjects only if strictly necessary, for example, for follow-up data collection.

Objective

Although pseudonymization is done in almost any biomedical study, there are currently no pseudonymization tools that can rapidly be rolled out across many institutions. Existing tools, such as the Generic Pseudonym Administration Service (gPAS) [19] and Mainzelliste [20], are client-server applications, requiring server components to be deployed to and integrated into the institutions' IT infrastructures. Although this can have some important advantages (see the *Limitations and Future Work* section), it is usually time-consuming, for example, due to a lack of resources or efforts required to ensure compliance with local security policies. Moreover, using central services, such as the European Unified Patient Identity Management (EUPID) [21], is often not an option, for example, when data should be reused and consent is missing for this type of processing [22].

In this paper, we present the ORCHESTRA Pseudonymization Tool (OPT) that has been developed under the umbrella of the ORCHESTRA consortium. This project faced the challenges described in the previous paragraph when quickly establishing a large-scale research network as part of Europe's rapid pandemic response [23]. Hence, the OPT has been developed with the aim of supporting (1) the registration, pseudonymization, and management of study subject identities as well as biosamples; (2) rapid rollout across research network partners; and (3) scalability and simple configurability. The objective of this paper is to describe the design and implementation of the OPT and to offer insights into its usability and scalability, as evidenced by its deployment in the ORCHESTRA research network.

Methods

Ethical Considerations

The work described in this article covers the design and implementation of a generic research tool, which did not involve research on humans or human specimens and no epidemiological research with personal data. Therefore, no approval was required according to the statutes of the Ethics Committee of the Faculty of Medicine at Charité - Universitätsmedizin Berlin. However, the individual studies which use the tool usually have to apply for ethics approval. For example, the COVID HOME study within the ORCHESTRA project was approved by the Medical Ethical Review Committee of the University Medical Center Groningen (UMCG) under vote number METc 2020/158.

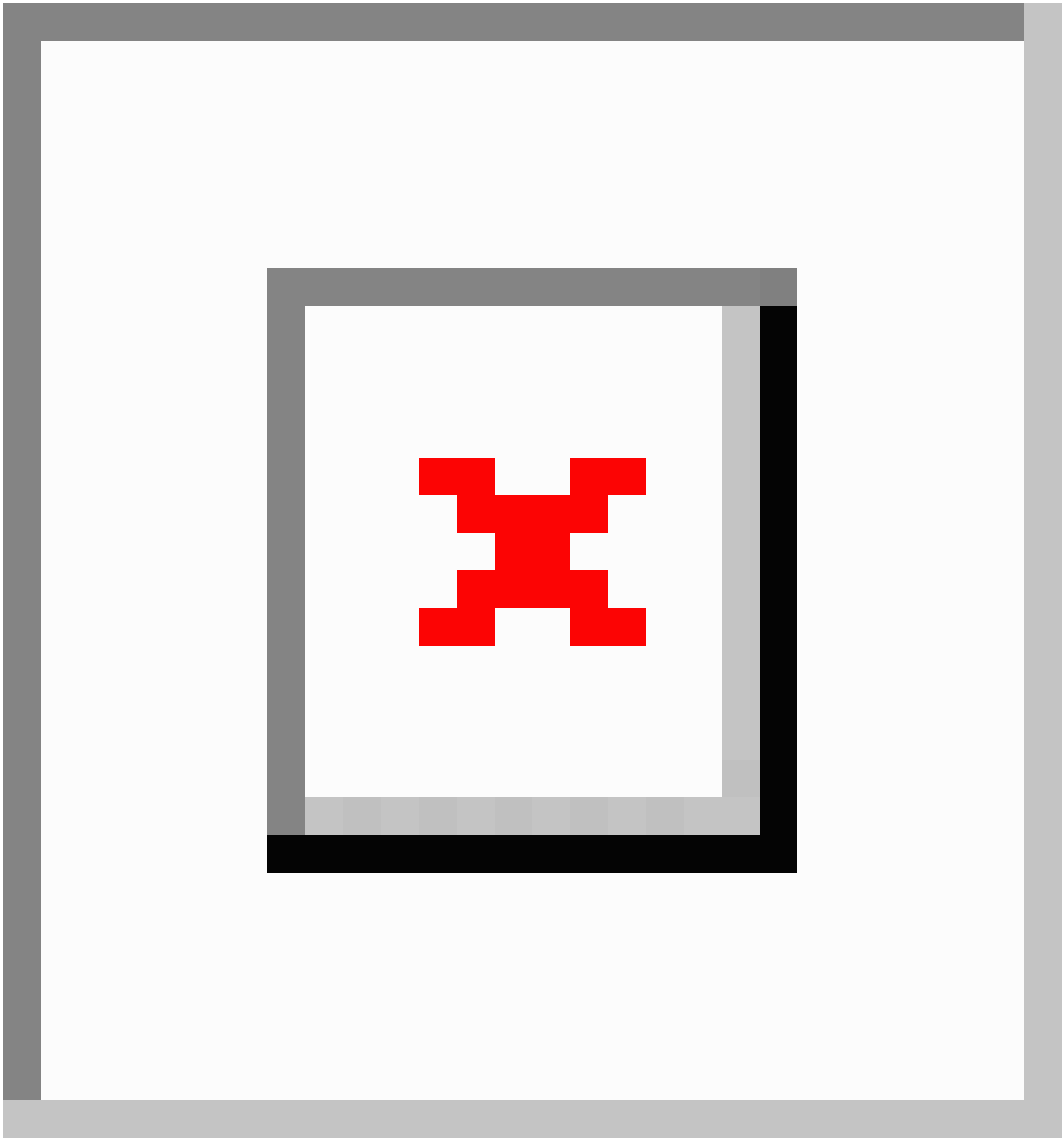
General Approach

The OPT has been designed to support general pseudonymization workflows that are needed in most biomedical research projects, as illustrated in Figure 1.

When a subject is admitted to the hospital, visits a study center, or has a follow-up visit, they are enrolled in the study. In this setting, the physicians or study nurses collect directly identifying and medical data and, according to the study protocol, the appropriate biosamples. The identifying attributes are entered into the OPT to create a unique pseudonym: the OPT Subject ID. During the follow-up visits, the study staff can use the OPT to retrieve an existing pseudonym from a subject that was already enrolled in the study. In all downstream data collection or processing, the OPT Subject ID can be used instead of identifying data so that the medical data are protected but still linked to the study subject and across visits. In addition, biosample data can also be entered into the OPT and linked to the appropriate subject to generate 1 or more additional pseudonyms: the OPT Biosample IDs. A label can then be generated for each biosample vial, containing the OPT Biosample ID, the OPT Subject ID, a DataMatrix Code, a QR code, or a barcode (containing the OPT Biosample ID) for tracking the biosample via scanners commonly used in laboratories. Study-specific information, for example, the exact information to capture for each study subject and biosample, the number and schedule of visits, and the types and schedules of biosample collections, can all be configured in the OPT. Moreover, in addition to its applicability in prospective studies, as described above, the software also supports importing existing

data about subjects and biosamples that can be used in retrospective study designs.

Figure 1. Basic concept of the OPT. IDAT: identifying data; MDAT: medical data; PID: patient ID; PSN: subject pseudonym; PSN-S: sample pseudonym; SDAT: sample data; SID: sample ID.



Implementation Details

To overcome challenges caused by the heterogeneity of IT infrastructures across different institutions and a potential lack of support by IT departments due to resource constraints, the OPT has been implemented based on programmable runtime environments that are available at practically any institution: office suites. These suites, especially the one by Microsoft, are among the most important and widely used applications around the world and still play a key role in many sectors today. The OPT is available for Microsoft Office as an Excel application and for LibreOffice as a Calc application. The application logic has been implemented in the embedded Basic scripting language using efficient algorithms for data management. Although Visual Basic for Applications is supported by Microsoft Office and LibreOffice Basic is supported by LibreOffice, they share similarities but are not fully compatible with each other. In the development process of the OPT, the Excel version serves as the primary implementation, and changes as well as additions are regularly ported to the LibreOffice version to achieve feature parity.

For generating the labels for the biosample vials, the OPT is delivered together with a single-page label printing application

that takes pseudonyms and metadata (eg, visit labels) as input and generates printable labels. Although this application is implemented using web technologies such as HTML, CSS, and JavaScript, it is delivered as files and can be executed locally without access to the internet. The label printing application works in any common web browser and can be called via the OPT. Properties of the labels to be printed can either be automatically transmitted via the URL for a single label or manually copied into the application via an input field for bulk printing of a larger number of labels. It is also possible to host the application on a web server. However, in this case, the URL function will be deactivated in the OPT to ensure that no data are sent to the server that hosts the application. It is important to note that the application still runs completely locally in the browser of the user, and no data ever leave the devices used to print labels. The pseudonyms and biosample metadata will be temporarily managed in the browser of the device.

Specific Functionalities

In addition to study subject and biosample management, the OPT also provides import and export functionalities, statistics, and a range of configuration options. In this section, we will briefly introduce each function, whereas a structured overview can be found in Multimedia Appendix 1. Regarding the subject-related functions, the OPT supports individual or bulk registration and a search function for finding pseudonyms for already registered subjects. An important feature of the software is a search function, required for any new patient or sample registration, which prevents multiple registrations of the same study participant. The search, to be performed as the first step of the registration, is linked to several data quality checks as well as a fuzzy record linkage process that prevents duplicate registrations. The bulk registration functionality enables the use of the OPT for retrospective pseudonymization of existing data sets. The search function supports wildcards and fuzzy matching across a configured set of master data attributes. Additional properties for the registered individuals can be documented to account for site-specific requirements.

Biosample-related functions are designed analogously to those for study subject management. In addition, labels can be generated and printed through the service described in the previous section.

Import and export functionalities are provided to enable the creation of backups (see the next section) and the migration from old versions of the OPT as part of update processes.

Finally, separate worksheets display statistical information about the data captured, such as the number of subjects registered or pseudonyms created for different study visits. Extensive configuration options are also available through a separate worksheet.

All functionalities of the OPT are described briefly in an integrated Quick User Guide and in detail in a comprehensive user manual [24].

Security Considerations and Features

The data collected during study subject and biosample registration, as well as the pseudonyms generated, are sensitive and a critical part of the data managed in any study. Hence, the confidentiality, integrity, and availability [25] of the data managed in the OPT must be ensured. In this context, the approach taken by the OPT clearly trades off some of the guarantees that could be provided by a client-server application against the possibility of rapid deployment and rollout. However, as described in the user manual, care has been taken to provide robust guarantees by specifying requirements on how the OPT should be deployed and used [24]. First, the OPT should not be placed on a local drive but on a network share that is integrated with the institution's Authentication and Authorization Infrastructure and, hence, provides means for controlling who is able to access the software in read or write mode and from which devices. Second, it is highly recommended that this share be backed up regularly so that data can be restored in case of problems. This should be complemented by regular, for example, daily, manual backups through the export functionality provided by the OPT and according to reminders that are displayed by the software. Finally, the office suites used as runtime environments do not provide multiuser support, and the application can only be opened by 1 user with write permission at any point in time. To enable parallel read access, the OPT comes with a script that opens a temporary read-only copy of the software. This allows, for example, laboratory technicians to use the OPT for generating biosample labels in parallel with ongoing registration processes. The measures described in this section have proven to be effective, and no problems have been encountered to date during extensive use of the software at many institutions (see the *Results* section).

Results

Overview of the Application

The graphical user interface of the OPT is divided into 10 different perspectives that provide access to the functionalities described in the previous sections. One of those sheets, the configuration sheet, is hidden from the users. All other sheets have write protection using the integrated protection functions of the spreadsheet software, except the input fields and the buttons, to ensure that data management is only performed through the specific functionalities provided by the software. A password is set by default for the write protection, which can be changed by the administrator at any time. However, it is important to keep the password safe. Figure 2 provides an overview of 4 important perspectives.

Figure 2. Perspectives of the OPT for (A) configuration, (B) registration and search, (C) data overview, and (D) statistics. OPT: ORCHESTRA Pseudonymization Tool.

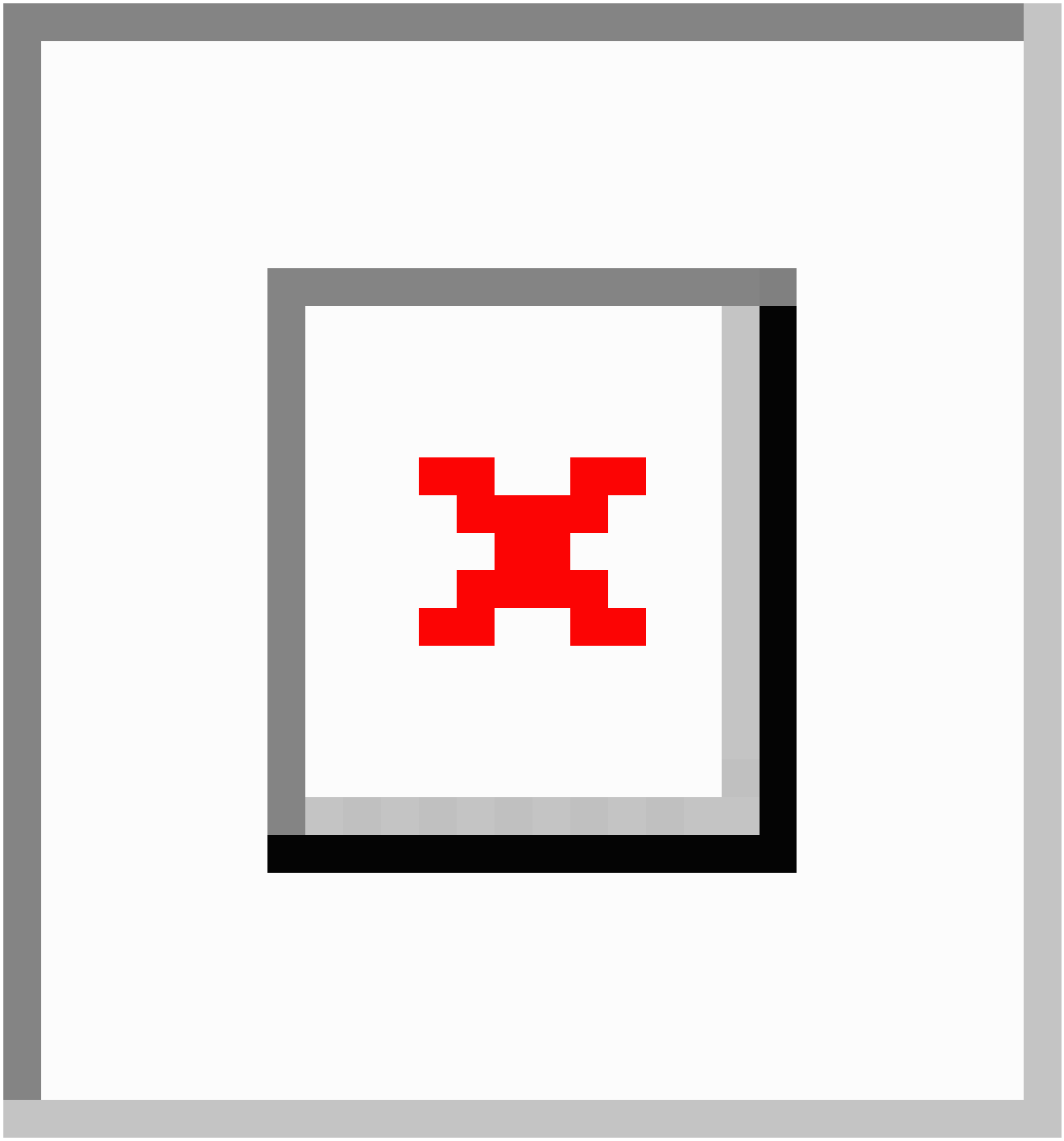
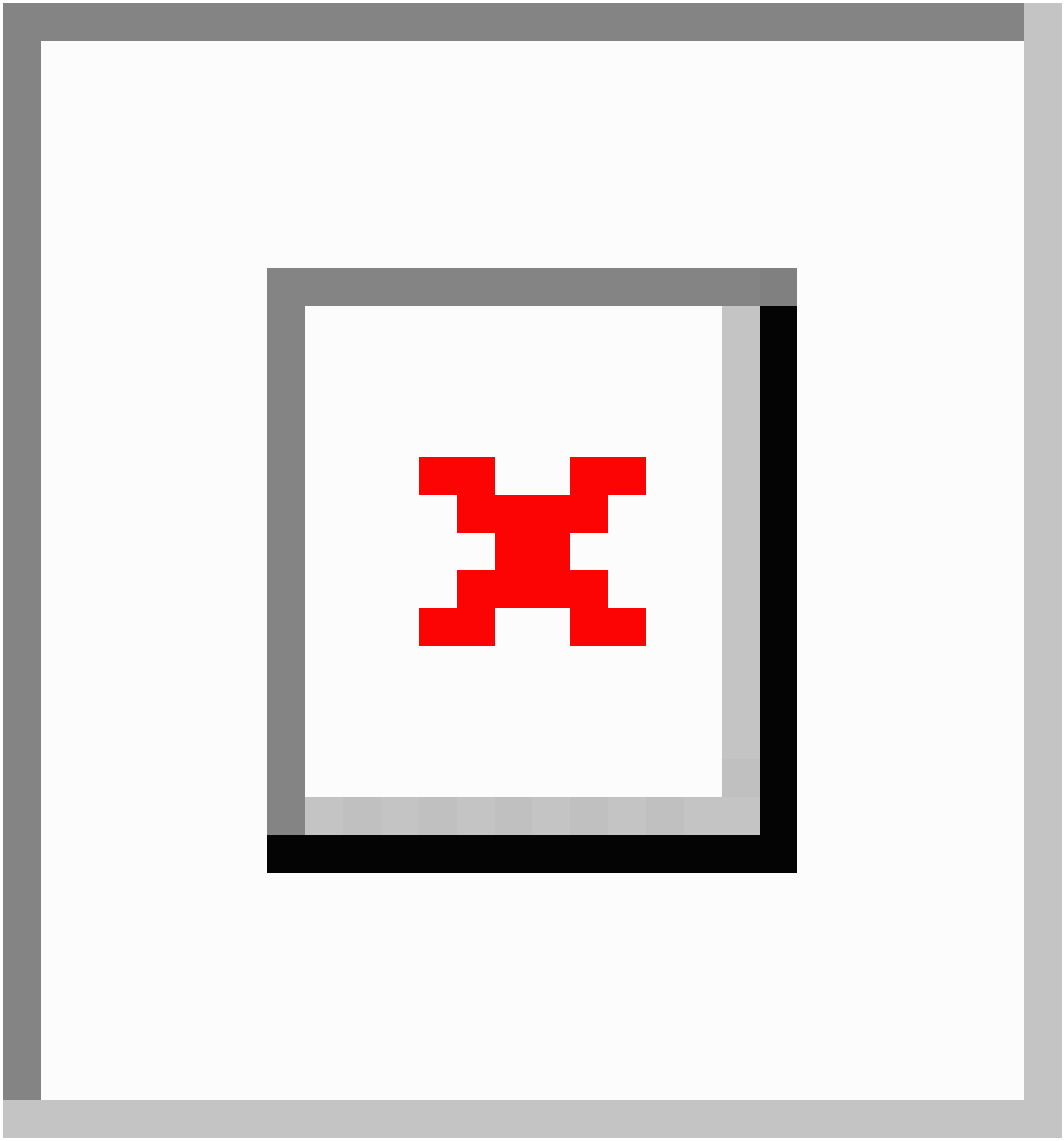


Figure 2A shows the configuration sheet, in which the specifics of the algorithm for generating pseudonyms, the study schedule, and the data fields to be documented can be specified. Figure 2B shows the interface provided for searching and registering subjects, with a search form on the left side of the sheet and a results list on the right side. All study subject data stored in the OPT are listed in the sheet shown in Figure 2C. This sheet also allows users to document any additional data that a site may require. Finally, Figure 2D shows a sheet providing statistical information on the number of subjects and biosamples registered, as well as insights into how these numbers have developed over time.

An overview of the label printing application is provided in Figure 3. As shown in the figure, the data that are to be printed on the labels are listed, and the number of rows and columns can be configured to support printing in bulk or for individual labels. The figure also shows an example of a sheet that can be printed and a detailed image of a single label. The data that are printed on those labels include the biosample and study subject IDs, the associated visit of the study schedule, and the biosample type.

Figure 3. Overview of the label printing application. OPT: ORCHESTRA Pseudonymization Tool.



Use of the OPT in the ORCHESTRA Project

ORCHESTRA is a 3-year international research project about the COVID-19 pandemic that was established in December 2020, involving 26 partners from 15 countries. The aim of ORCHESTRA is to share and analyze data from several retrospective and prospective studies to provide rigorous evidence for improving the prevention and treatment of COVID-19 and to better prepare for future pandemics [26,27].

The data management architecture in ORCHESTRA consists of 3 layers that build upon each other. The first layer is formed by "National Data Providers," which consist of the participating partners (universities, hospitals, and research networks). These provide the subject data and samples for joint analyses. On the second layer, "National Hubs" pool pseudonymized data in national instances of the Research Electronic Data Capture (REDCap) system [28]. Finally, the "ORCHESTRA Data Portal" forms the third layer, in which access to aggregated data and results is provided through a central repository.

In ORCHESTRA, the OPT was used for implementing pseudonymization at the data providers' sites. Each participating site named 1 or 2 persons responsible for technical aspects, such as setting up the required network share and installing updates, as well as several study nurses or clinicians, who would use the OPT. With these users, we performed regular training sessions and provided contact details in case of questions. As of June 2023, 19 instances of the OPT have been rolled out to 13 sites in 11 countries, including Germany, France, Italy, and Slovakia

in Europe; Congo in Africa; and Argentina in South America. A world map highlighting all the countries in which the OPT has been rolled out can be found in Multimedia Appendix 2.

On average, each instance of the OPT was used by up to 4 staff members. The OPT has been successfully rolled out, used, and maintained at large sites with committed IT departments, as well as at smaller, resource-constrained institutions. Overall, it has been in constant production use for more than 2 years. In the majority of the sites (10/13, 77%), the OPT Microsoft Excel version was used, whereas the remaining sites (3/13, 23%) used the LibreOffice release. In total, more than 10,000 study subjects and 15,000 samples have been registered in the OPT across all sites, and more than 10,000 labels have been printed. To evaluate the usability of the OPT, we conducted a survey among all active users, leveraging the widespread System Usability Scale [29] questionnaire, which includes 10 Likert-scale questions. During this survey, our system was designed to prevent multiple responses from individual participants and the submission of incomplete responses. We received 6 responses from 9 invited users, resulting in a score of 75 on a scale from 0 to 100, which adjectively translates to “good” [30].

Performance Evaluation

As mentioned, the OPT has been carefully designed to provide acceptable performance, even when large data sets are being processed or a large number of subjects or samples are being managed. In this section, we present the results of a brief performance evaluation. Our test environment consisted of an average office laptop, which was equipped with a quad-core 1.8 GHz Intel Core i7 CPU and a 64-bit Microsoft Windows 10 operating system. On top of it, Microsoft Excel 2016 (x32) and LibreOffice 7.0 (x64) were installed. Figure 4 provides an overview of the execution times of the most important functionalities of the OPT for different cohort sizes.

The numbers clearly show that the OPT works well and provides excellent performance for small or medium-sized data sets and acceptable performance for large data sets.

Figure 4. Execution times of the most important operations of the ORCHESTRA Pseudonymization Tool: (A) import, (B) registration, and (C) search.

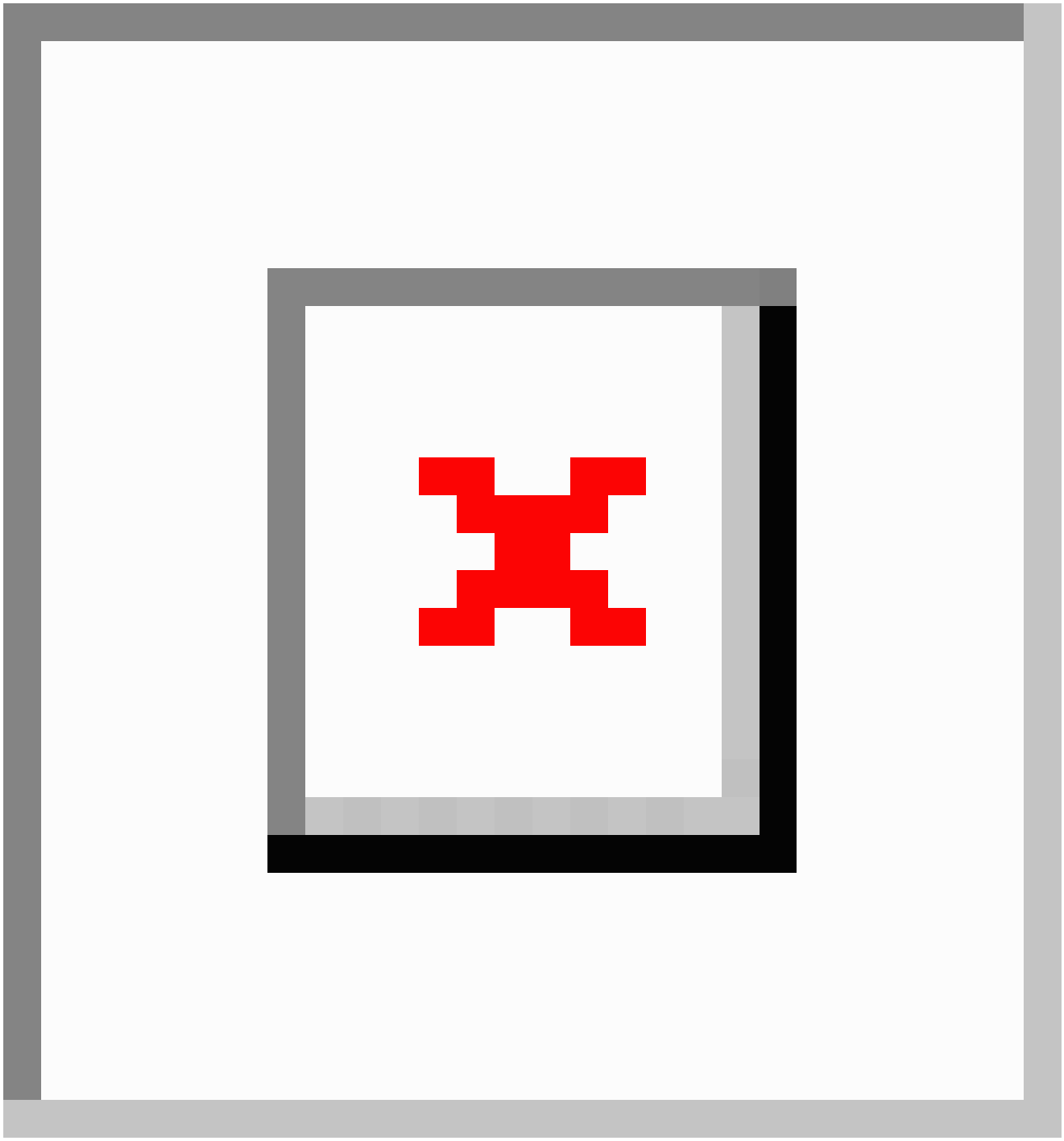


Figure 4A shows the average execution times for importing data about study subjects and samples. Data about subjects were imported into a completely empty OPT, whereas data about samples were imported into an OPT that already had the corresponding study subjects registered, so that each biosample was assigned to exactly 1 subject. For example, importing the data of 100,000 subjects took about 10 seconds in the Excel version and 54 seconds in the LibreOffice version. During the registration, the existence of the associated study subject in the OPT is checked, which makes the registration of samples slower compared to the registration of subjects. This is also noticeable in Figure 4B, which shows the average execution times for registering a single study subject or sample. As can be seen, using an OPT data set in which 100,000 entities were already registered, this took between 2 and 4 seconds in the Excel version and between 4 and 6 seconds in the LibreOffice version. Figure 4C shows the average execution times for searching for entities and obtaining their pseudonym, which is roughly twice as fast as the registration operation.

As performance is associated linearly with the number of entities already managed, subsecond response times can be expected for instances in which around 15,000 or fewer subjects or samples have been registered. This is consistent with our experiences from the deployments in the ORCHESTRA research network.

Discussion

Principal Findings

In this paper, we presented the OPT, a comprehensive, scalable, and pragmatic pseudonymization tool that can be rapidly rolled out across large research networks. To achieve this, the software has been implemented based on runtime environments that are available at practically any institution: office suites. The software supports a broad range of functionalities, from registering and pseudonymizing subject and biosample identities to search and depseudonymization functions, statistics about the data managed, as well as import and export features. We have described measures that are recommended to ensure the security of the data managed by the OPT and reported on our experiences gained after 2 years of successful operation in a large research network on COVID-19. Finally, we have also presented the results of a performance evaluation showing that the software provides excellent performance for small or medium-sized data sets and acceptable performance for large data sets. The OPT is available as open-source software [31] and can be configured to meet the needs of a wide range of biomedical research projects.

Limitations and Future Work

To achieve the design goals of the OPT, some compromises had to be made regarding data management. Compared to using client-server applications that use database management systems to store data, it is more difficult to ensure the confidentiality, integrity, and availability of the data managed with the OPT. There is also limited support for multiuser scenarios. However, we have developed and documented a set of measures that, if taken, help to still ensure a high level of data security. For this to work, it is important that users adhere to those recommendations. Therefore, all users of the OPT should familiarize themselves with the manual [24], and ideally, they should also be trained in the use and operation of the software. Despite these limitations, we strongly believe that our approach offers an innovative take on pseudonymization tools that can rapidly be rolled out across large research networks. Of course, it would be even more desirable if global standards for pseudonymization functions could be developed and agreed upon. Such global standards would ensure that solutions already existing at many research institutions are interoperable and can readily be used in joint research activities.

Comparison With Related Work

A range of pseudonymization tools has been described in the literature and are available as open-source software. However, they are either based on a client-server architecture and hence require quite some effort to be rolled out across sites, based on central services and hence not usable if consent is lacking for this type of processing, or offered as command-line utilities or programming libraries for IT experts.

Examples of client-server approaches include the work by Lablans et al [20] to provide a RESTful interface to pseudonymization services in modern web applications, which is based on a concept suggested by Pommerening et al [6] in 2006. Moreover, researchers from the University of Greifswald in Germany have designed and developed several client-server tools that can be used to manage subjects, samples, and other aspects of biomedical studies [32,33].

Examples of central services for pseudonymization include the EUPID, which was developed in 2014 by the Austrian Institute of Technology for the European Network for Cancer Research in Children and Adolescents project [21]. Another example is the Secure Privacy-preserving Identity management in Distributed Environments for Research (SPIDER) service, which was launched in May 2022 by the Joint Research Centre [34]. Both services support linking and transferring subject data across registries without revealing their identities. However, biosample data management is not possible with them. Further centralized concepts include the one described by Angelow et al [35].

Examples of command-line utilities, application programming interfaces, and programming libraries include the generic solution for record linkage of special categories of personal data developed by Fischer et al [36]; that by Preciado-Marquez et al [37]; and the PID (patient ID) generator developed by the TMF (Technologies, Methods and Infrastructure for Networked Medical Research e.V.), the German umbrella association for networked medical research [6].

Conclusion

Widely available office suites provide runtime environments that offer opportunities to rapidly roll out software components for biomedical studies across a wide range of large and resource-constrained research institutions. We have demonstrated this through the development, practical use, and evaluation of the OPT, which offers pseudonymization functionalities for study subjects and biosamples. As we believe that the software is of interest to the larger research community, it has been made available under a permissive open-source license [31].

Acknowledgments

This work has been funded by the European Union's Horizon 2020 research and innovation programme under the project ORCHESTRA (grant agreement 101016167).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of the ORCHESTRA Pseudonymization Tool functions.
[PNG File, 233 KB - medinform_v12i1e49646_app1.png]

Multimedia Appendix 2
Map of countries in which the ORCHESTRA Pseudonymization Tool has been rolled out.
[PNG File, 229 KB - medinform_v12i1e49646_app2.png]

References

1. Dron L, Dillman A, Zoratti MJ, Haggstrom J, Mills EJ, Park JJH. Clinical trial data sharing for COVID-19-related research. J Med Internet Res 2021 Mar 12;23(3):e26718. [doi: 10.2196/26718] [Medline: 33684053]
2. R&D Blueprint. A coordinated global research roadmap: 2019 novel coronavirus. : World Health Organization; 2020 Mar 12 URL: https://www.who.int/publications/m/item/a-coordinated-global-research-roadmap [accessed 2024-04-12]
3. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. Nat Biotechnol 2018 May 9;36(5):391-392. [doi: 10.1038/nbt.4128] [Medline: 29734317]
4. Walport M, Brest P. Sharing research data to improve public health. Lancet 2011 Feb 12;377(9765):537-539. [doi: 10.1016/S0140-6736(10)62234-9] [Medline: 21216456]
5. Mahmoud A, Ahlborn B, Mansmann U, Reinhardt I. Clientside pseudonymization with trusted third-party using modern web technology. Stud Health Technol Inform 2021 May 27;281:496-497. [doi: 10.3233/SHTI210212] [Medline: 34042618]
6. Pommerening K, Schröder M, Petrov D, Schlösser-Faßbender M, Semler SC, Drepper J. Pseudonymization service and data custodians in medical research networks and biobanks. In: INFORMATIK 2006 – INFORMATIK für Menschen: Gesellschaft für Informatik e.V; 2006, Vol. 1:715-721.
7. Tacconelli E, Gorska A, Carrara E, et al. Challenges of data sharing in European COVID-19 projects: a learning opportunity for advancing pandemic preparedness and response. Lancet Reg Health Eur 2022 Oct;21:100467. [doi: 10.1016/j.lanepe.2022.100467] [Medline: 35942201]
8. Rumbold J, Pierscionek B. Contextual anonymization for secondary use of big data in biomedical research: proposal for an anonymization matrix. JMIR Med Inform 2018 Nov 22;6(4):e47. [doi: 10.2196/medinform.7096] [Medline: 30467101]
9. Aamot H, Kohl CD, Richter D, Knaup-Gregori P. Pseudonymization of patient identifiers for translational research. BMC Med Inform Decis Mak 2013 Jul 24;13:75. [doi: 10.1186/1472-6947-13-75] [Medline: 23883409]
10. Wu X, Wang H, Zhang Y, Li R. A secure visual framework for multi-index protection evaluation in networks. Digit Commun Netw 2023 Apr;9(2):327-336. [doi: 10.1016/j.dcan.2022.05.007]
11. Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union. 2016 Apr 27. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679 [accessed 2024-04-12]
12. U.S. Department of Health and Human Services, Office for Civil Rights. HIPAA administrative simplification: regulation text: 45 CFR parts 160, 162, and 164 (unofficial version, as amended through March 26, 2013). U.S. Department of Health and Human Services. 2013 Mar 26. URL: https://www.hhs.gov/sites/default/files/hipaa-simplification-201303.pdf [accessed 2024-04-12]
13. Quinn P. Research under the GDPR - a level playing field for public and private sector research? Life Sci Soc Policy 2021 Mar 1;17(1):4. [doi: 10.1186/s40504-021-00111-z] [Medline: 33648586]
14. Rodriguez A, Tuck C, Dozier MF, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: a scoping review. Clin Trials 2022 Aug;19(4):452-463. [doi: 10.1177/17407745221087469] [Medline: 35730910]
15. Kohlmayer F, Lautenschläger R, Prasser F. Pseudonymization for research data collection: is the juice worth the squeeze? BMC Med Inform Decis Mak 2019 Sep 4;19(1):178. [doi: 10.1186/s12911-019-0905-x] [Medline: 31484555]
16. Gruschka N, Mavroeidis V, Vishi K, Jensen M. Privacy issues and data protection in big data: a case study analysis under GDPR. Presented at: 2018 IEEE International Conference on Big Data (Big Data); Dec 10 to 13, 2018; Seattle, WA p. 5027-5033. [doi: 10.1109/BigData.2018.8622621]
17. Lautenschläger R, Kohlmayer F, Prasser F, Kuhn KA. A generic solution for web-based management of pseudonymized data. BMC Med Inform Decis Mak 2015 Nov 30;15:100. [doi: 10.1186/s12911-015-0222-y] [Medline: 26621059]
18. European Union Agency for Cybersecurity, Drogkaris P, Bourka A. Recommendations on shaping technology according to GDPR provisions - an overview on data pseudonymisation. : European Network and Information Security Agency; 2018. [doi: 10.2824/74954]
19. Bialke M, Bahls T, Havemann C, et al. MOSAIC--a modular approach to data management in epidemiological studies. Methods Inf Med 2015;54(4):364-371. [doi: 10.3414/ME14-01-0133] [Medline: 26196494]
20. Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak 2015 Feb 7;15:2. [doi: 10.1186/s12911-014-0123-5] [Medline: 25656224]
21. Nitzlnader M, Schreier G. Patient identity management for secondary use of biomedical research data in a distributed computing environment. Stud Health Technol Inform 2014;198:211-218. [Medline: 24825705]

22. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015 Mar 20;350:h1139. [doi: 10.1136/bmj.h1139] [Medline: 25794882]
23. Connecting European cohorts to increase common and effective response to SARS-CoV-2 pandemic: ORCHESTRA. European Commission. 2022 Apr 21. URL: https://cordis.europa.eu/project/id/101016167/de [accessed 2023-06-02]
24. BIH-MI/opt: ORCHESTRA pseudonymization tool - user manual. GitHub. 2023 Sep 24. URL: https://github.com/BIH-MI/opt/blob/main/development/documentation/user-manual.pdf [accessed 2023-09-26]
25. ISO/IEC 27001:2022 information security, cybersecurity and privacy protection - information security management systems - requirements. : International Organization for Standardization; 2022 URL: https://www.iso.org/standard/27001 [accessed 2024-04-12]
26. Azzini AM, Canziani LM, Davis RJ, et al. How European research projects can support vaccination strategies: the case of the ORCHESTRA project for SARS-CoV-2. Vaccines (Basel) 2023 Aug 14;11(8):1361. [doi: 10.3390/vaccines11081361] [Medline: 37631929]
27. ORCHESTRA - EU horizon 2020 cohort to tackle COVID-19 internationally. ORCHESTRA. 2022 Sep 19. URL: https://orchestra-cohort.eu/ [accessed 2023-04-12]
28. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCAP)--a metadata-driven methodology and workflow process for providing translational research Informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
29. Brooke J. SUS: a quick and dirty usability scale. In: Usability Evaluation in Industry: CRC Press; 1996:189-194.
30. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009 May;4(3):114-123 [FREE Full text]
31. BIH-MI/opt: ORCHESTRA pseudonymization tool. GitHub. 2023 Jun 2. URL: https://github.com/BIH-MI/opt [accessed 2023-06-02]
32. Bialke M. Werkzeuggestützte Verfahren für die Realisierung einer Treuhandstelle im Rahmen des zentralen Datenmanagements in der epidemiologischen Forschung [Dissertation].: Universitätsmedizin der Ernst-Moritz-Arndt-Universität Greifswald; 2016 URL: https://d-nb.info/1124566945/34 [accessed 2024-04-12]
33. Bialke M, Penndorf P, Wegner T, et al. A workflow-driven approach to integrate generic software modules in a trusted third party. J Transl Med 2015 Jun 4;13:176. [doi: 10.1186/s12967-015-0545-6] [Medline: 26040848]
34. SPIDER pseudonymisation tool. European Commission. 2023 May 4. URL: https://eu-rd-platform.jrc.ec.europa.eu/spider/ [accessed 2023-06-02]
35. Angelow A, Schmidt M, Weitmann K, et al. Methods and implementation of a central biosample and data management in a three-centre clinical study. Comput Methods Programs Biomed 2008 Jul;91(1):82-90. [doi: 10.1016/j.cmpb.2008.02.002] [Medline: 18406002]
36. Fischer H, Röhrig R, Thiemann VS. Simple Batch Record Linkage System (SimBa) – a generic tool for record linkage of special categories of personal data in small networked research projects with distributed data sources: lessons learned from the Inno_RD project. In: Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. 64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS): German Medical Science GMS Publishing House; 2019. [doi: 10.3205/19gmds118]
37. Preciado-Marquez D, Becker L, Storck M, Greulich L, Dugas M, Brix TJ. MainzelHandler: a library for a simple integration and usage of the Mainzelliste. Stud Health Technol Inform 2021 May 27;281:233-237. [doi: 10.3233/SHTI210155] [Medline: 34042740]

Abbreviations

EUPID: European Unified Patient Identity Management
GDPR: General Data Protection Regulation
gPAS: Generic Pseudonym Administration Service
HIPAA: Health Insurance Portability and Accountability Act
OPT: ORCHESTRA Pseudonymization Tool
PID: patient ID
REDCap: Research Electronic Data Capture
SPIDER: Secure Privacy-preserving Identity management in Distributed Environments for Research
SUS: System Usability Scale
TMF: Technologies, Methods and Infrastructure for Networked Medical Research e.V.

Edited by C Lovis; submitted 06.06.23; peer-reviewed by J Scheibner, X Wu; revised version received 03.10.23; accepted 07.03.24; published 23.04.24.

Please cite as:
Abu Attieh H, Neves DT, Guedes M, Mirandola M, Dellacasa C, Rossi E, Prasser F
A Scalable Pseudonymization Tool for Rapid Deployment in Large Biomedical Research Networks: Development and Evaluation Study
JMIR Med Inform 2024;12:e49646
URL: https://medinform.jmir.org/2024/1/e49646
doi:10.2196/49646

Original Paper

User Preferences and Needs for Health Data Collection Using Research Electronic Data Capture: Survey Study

Hiral Soni[1], PhD; Julia Ivanova[1], PhD; Hattie Wilczewski[1], BS; Triton Ong[1], PhD; J Nalubega Ross[1], PhD; Alexandra Bailey[1], MS; Mollie Cummins[1,2], PhD; Janelle Barrera[1,3], MPH; Brian Bunnell[1,3], PhD; Brandon Welch[1,4], PhD

[1]Doxy.me Research, Doxy.me Inc, Charleston, SC, United States

[2]College of Nursing, University of Utah, Salt Lake City, UT, United States

[3]Department of Psychiatry and Behavioral Neurosciences, University of South Florida, Tampa, FL, United States

[4]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Hiral Soni, PhD
Doxy.me Research
Doxy.me Inc
18 Broad Street, 3rd Floor
Suite 6 and 7
Charleston, SC, 29401
United States
Phone: 1 8444369963
Email: sonihiralc@gmail.com

Abstract

Background: Self-administered web-based questionnaires are widely used to collect health data from patients and clinical research participants. REDCap (Research Electronic Data Capture; Vanderbilt University) is a global, secure web application for building and managing electronic data capture. Unfortunately, stakeholder needs and preferences of electronic data collection via REDCap have rarely been studied.

Objective: This study aims to survey REDCap researchers and administrators to assess their experience with REDCap, especially their perspectives on the advantages, challenges, and suggestions for the enhancement of REDCap as a data collection tool.

Methods: We conducted a web-based survey with representatives of REDCap member organizations in the United States. The survey captured information on respondent demographics, quality of patient-reported data collected via REDCap, patient experience of data collection with REDCap, and open-ended questions focusing on the advantages, challenges, and suggestions to enhance REDCap's data collection experience. Descriptive and inferential analysis measures were used to analyze quantitative data. Thematic analysis was used to analyze open-ended responses focusing on the advantages, disadvantages, and enhancements in data collection experience.

Results: A total of 207 respondents completed the survey. Respondents strongly agreed or agreed that the data collected via REDCap are accurate (188/207, 90.8%), reliable (182/207, 87.9%), and complete (166/207, 80.2%). More than half of respondents strongly agreed or agreed that patients find REDCap easy to use (165/207, 79.7%), could successfully complete tasks without help (151/207, 72.9%), and could do so in a timely manner (163/207, 78.7%). Thematic analysis of open-ended responses yielded 8 major themes: survey development, user experience, survey distribution, survey results, training and support, technology, security, and platform features. The user experience category included more than half of the advantage codes (307/594, 51.7% of codes); meanwhile, respondents reported higher challenges in survey development (169/516, 32.8% of codes), also suggesting the highest enhancement suggestions for the category (162/439, 36.9% of codes).

Conclusions: Respondents indicated that REDCap is a valued, low-cost, secure resource for clinical research data collection. REDCap's data collection experience was generally positive among clinical research and care staff members and patients. However, with the advancements in data collection technologies and the availability of modern, intuitive, and mobile-friendly data collection interfaces, there is a critical opportunity to enhance the REDCap experience to meet the needs of researchers and patients.

(JMIR Med Inform 2024;12:e49785) doi:10.2196/49785

KEYWORDS

Research Electronic Data Capture; REDCap; user experience; electronic data collection; health data; personal health information; clinical research; mobile phone

Introduction

Background

Accurate and complete health outcome data directly from patients or study participants (hereon referred to as *patients*) are critical for health care and research [1-3]. Unfortunately, it can be burdensome to extract patient-reported health data that researchers or providers need [4,5]. Collecting patient-reported outcomes data is becoming increasingly important in clinical research and care [6,7]. Self-administered web-based questionnaires, which patients can complete at a clinic or at home, are becoming a conventional approach to collect data for clinical research. Web-based questionnaires have advantages of being low-cost and easy to deploy at scale. A variety of clinical research electronic data capture (EDC) tools exist to streamline remote data collection and management. These systems comply with privacy regulations, integrate with different tools (such as electronic health records [EHRs]) for efficient data collection, and reduce the effort of sharing data [8]. However, user experience, cost, and maintenance of such commercial EDC systems are often prohibitive. An understanding of user experiences and preferences regarding EDC tools is critical in assessing stakeholder needs, satisfaction, and challenges in clinical and research settings.

REDCap (Research Electronic Data Capture; Vanderbilt University) is a global, secure web application for building and managing EDC for clinical research [9,10]. Developed by Vanderbilt University, REDCap is freely available for its consortium members (ie, network of nonprofit collaborators and supporters), who have an established agreement with the university. REDCap is compliant with global privacy regulations (such as the Health Insurance Portability and Accountability Act [HIPAA] of 1996) and used by more than 2.2 million researchers in more than 140 countries [9]. REDCap allows researchers to build and conduct electronic surveys, track and manage study information, schedule visits, and manage databases that are fully customizable and at no cost [11]. REDCap is designed to support data capture for research studies, providing (1) an intuitive interface for validated data capture, (2) audit trails for tracking data manipulation and export procedures, (3) automated export procedures for seamless data downloads to common statistical packages, and (4) procedures for data integration and interoperability with external sources.

Although REDCap is widely used, user needs and preferences of EDC via REDCap have rarely been studied [12,13]. For example, 1 usability study of a REDCap-based patient-facing intervention reported that patient participants found REDCap useful and easy to use but showed concerns about wordiness and inconsistent visual design [13]. Researchers have reported frequently on the implementation, use, and interventions using REDCap [10,14-20]. Understanding the preferences and needs of REDCap administrators and researchers using REDCap to capture data could help enhance existing features and EDC processes in general. While REDCap is a robust clinical research data management system, this study solely focuses on the experience of REDCap as an EDC tool. To the best of our knowledge, such preferences have not yet been studied.

Objective

The aim of this study was to survey REDCap administrators and researchers in the United States to assess their experience with REDCap, including perspectives on advantages, challenges, and suggestions for enhancement.

Methods

Study Settings and Respondents

We conducted a web-based survey with representatives of member organizations listed as REDCap Partners on the REDCap website [21]. The roles of the listed members were unclear at the time of invitation sent via email. The email communication included information related to the study goals, voluntary participation, and a link to the REDCap survey. Respondents were compensated with a US $10 electronic gift card for completing the survey.

Ethical Considerations

This study was reviewed and approved as exempt human subjects research by the Medical University of South Carolina Institutional Review Board (Pro00082875).

Survey Design

We developed a web-based survey with multiple-choice and free-response questions (Multimedia Appendix 1) to capture the perspectives of researchers and administrators from participating REDCap consortium organizations. Our research team includes experts in biomedical informatics, behavioral sciences, mixed methods research, and user experience. The survey included 4 sections, as follows:

- *Demographics*: multiple-choice questions capturing participant role in their respective organization (Q1) and organization use of REDCap (Q2)
- *Quality of patient-reported data collected via REDCap*: Likert-scale questions capturing perspectives (ranging from 1=strongly agree to 5=strongly disagree) on the accuracy, reliability and completeness of data reported using REDCap (Q3)
- *Patient experience with REDCap*: Likert-scale question focusing on perspectives (ranging from 1=strongly agree to 5=strongly disagree) on REDCap usability, including ease of use, success rate, and completion time (Q4).
- *Data collection experience*: Free-response questions asking about the advantages (Q5), challenges (Q6), and suggestions of enhancements related to data collection, patient experience, and engagement (Q7).

Data Collection and Analysis

We collected and managed study data using REDCap EDC tools hosted at the Medical University of South Carolina [22,23]. We generated plots and univariate statistics to summarize the data (eg, frequencies, means, SDs, and percentages). We conducted 1-way ANOVA tests to determine differences in data quality and patient experience variables by participant role and REDCap use duration. For the ANOVAs, the primary role variable was restructured to include "Educators" in the "Other" category due to the low sample size (n=1). Excel (Microsoft Corp) and SPSS (version 29; IBM Corp) were used for analyses. Free-response questions were qualitatively analyzed to identify emerging themes related to REDCap data collection experience [24]. We randomly selected 15% of the responses for initial coding and codebook development. The coding unit was done by the entirety of the participant entry. Thematic analysis of all qualitative data was done over 4 iterations using MAXQDA, during which emergent themes were identified. While the research team reviewed and honed the codes and codebook, 1 team member coded and finalized thematic coding. Discrepancies were resolved through consensus. Emergent themes were organized by frequency and topic, allowing for further qualitative analysis using complex coding query to determine concurrent themes. We reported the total frequencies per code, which may not align with the number of participants. For example, 1 participant may report a code multiple times throughout their response [25]. While thematic analysis allows us to identify principle emergent themes, it also can help identify uncommon trends that may be significant but would require further investigation in follow-up research [26]. Responses from incomplete surveys with missing quantitative or qualitative responses were excluded from the analysis

Results

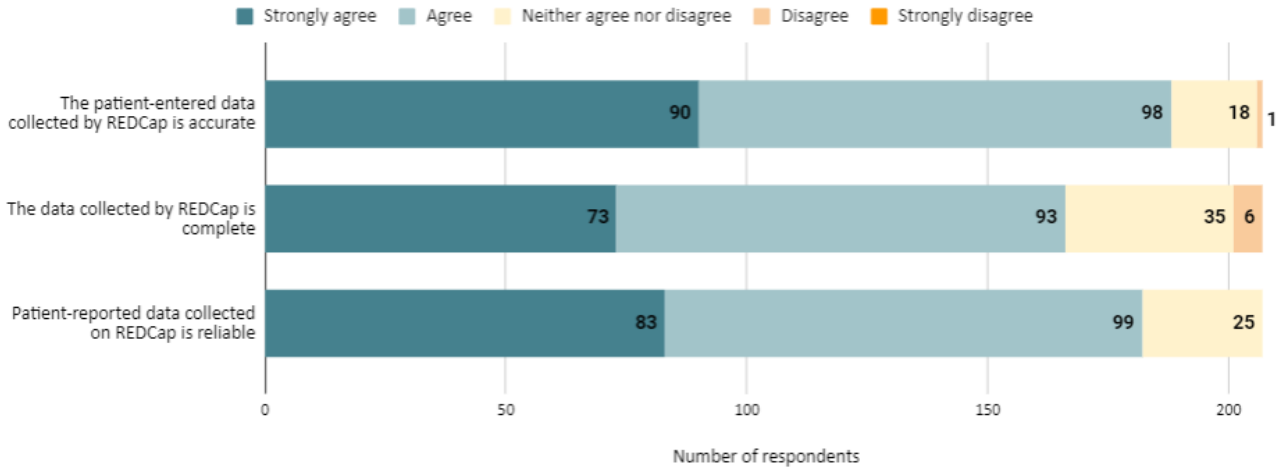
Demographics

Between October and November 2020, 3058 representatives from 1676 REDCap member organizations in the United States were invited to complete the survey. In total, 285 (9.3%) invitees started the survey, of which 207 completed the survey. Most (150/207, 72.5%) respondents were REDCap administrators, followed by researchers (25/207, 12.1%). Furthermore, 1 (0.5%) respondent was an educator and 31 (15%) respondents served in other roles, including IT directors and managers, research coordinators and managers, program managers, project managers, director of research, library directors, and data analysts. Respondents reported that their organization had used REDCap for <5 years (92/207, 44.4%), 5 to 10 years (83/207, 40.1%), or >10 years (32/207, 15.5%).

Quality of Patient-Reported Data Collected via REDCap

We asked respondents about their perspectives of the quality of the survey data, including the accuracy, reliability, and completeness of the data collected using REDCap (Figure 1). Most respondents strongly agreed or agreed that the data collected via REDCap are accurate (188/207, 90.8%), reliable (182/207, 87.9%), and complete (166/207, 80.2%). We observed no statistically significant group differences in accuracy ($F_{2,204}$=1.003; P=.37), completeness ($F_{2,204}$=0.243; P=.78), or reliability ($F_{2,204}$=0.245; P=.78) among respondent role groups. Furthermore, we observed no statistically significant group differences in accuracy ($F_{2,204}$=0.672; P=.51), completeness ($F_{2,204}$=0.045; P=.96), or reliability ($F_{2,204}$=1.712; P=.18) among REDCap use groups.

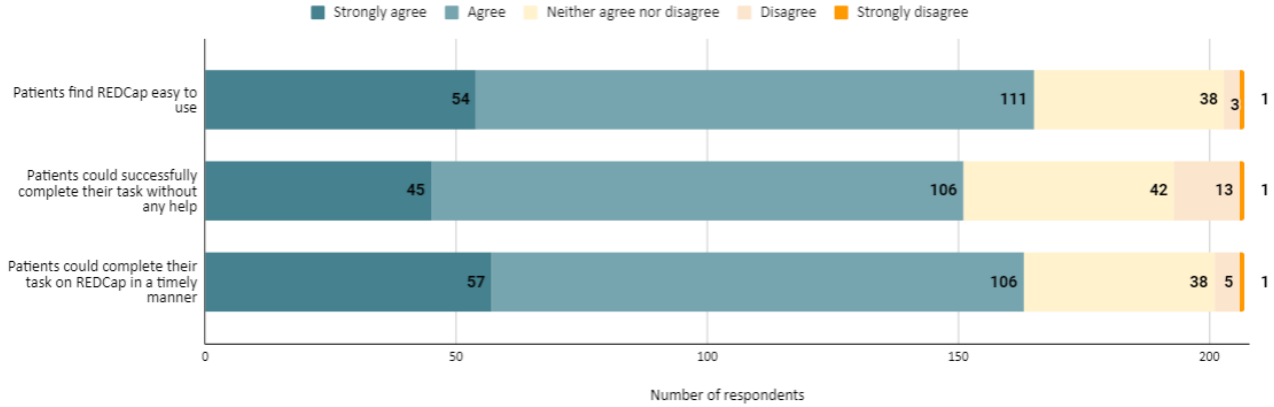
Figure 1. Quality of patient-reported data collected via REDCap (Research Electronic Data Capture).



Patient Experience With REDCap

We also asked respondents about their perspectives on patient experiences with completing surveys using REDCap. Figure 2 summarizes their responses. More than half of respondents strongly agreed or agreed that patients find REDCap easy to use (165/207, 79.7%), could successfully complete tasks without help (151/207, 72.9%), and could do so in a timely manner (163/207, 78.7%). We observed no statistically significant group differences in ease ($F_{2,204}$=2.025; P=.13), successful task completion ($F_{2,204}$=0.671; P=.51), or timely task completion ($F_{2,204}$=2.303; P=.10) among respondent role groups. Furthermore, we observed no statistically significant group differences in ease ($F_{2,204}$=0.711; P=.49), successful task completion ($F_{2,204}$=1.851; P=.16), or timely task completion ($F_{2,204}$=2.000; P=.13) among REDCap user groups.

Figure 2. Patient experience with REDCap (Research Electronic Data Capture).



REDCap Advantages, Challenges, and Enhancement Suggestions

We asked respondents about the advantages, challenges, and suggestions for future enhancements using free-response questions. The analysis yielded 8 primary codes: survey development, user experience, survey distribution, survey results, training and support, technology, security, and platform features. Within each of these themes, responses were further categorized at secondary and tertiary levels. Multimedia Appendix 2 shows the qualitative codebook with illustrative examples for each code. Table 1 shows the frequencies response classification of advantages, disadvantages, and enhancements for each code category based on respondents' responses.

Table 1. Counts and percentages of response classification of REDCap (Research Electronic Data Capture) users[a].

Code	Advantages, n (%)	Challenges, n (%)	Enhancements, n (%)
Survey development			
Design	52 (50.5)	59 (47.6)	40 (47.1)
Survey design	10 (9.7)	39 (31.5)	21 (24.7)
Response and logic	13 (12.6)	22 (17.7)	18 (21.2)
Survey setup	20 (19.4)	2 (1.6)	0 (0)
Flexibility	7 (6.8)	1 (0.8)	3 (3.5)
Organization	1 (1)	1 (0.8)	0 (0)
Testing	0 (0)	0 (0)	3 (3.5)
Customizations	7 (100)	17 (65.4)	15 (65.2)
Language support	0 (0)	9 (34.6)	8 (34.8)
Project interactions	0 (0)	3 (100)	6 (100)
Feature suggestions	1 (100)	16 (100)	49 (100)
User experience			
Usability	105 (55.9)	29 (61.7)	8 (53.3)
Ease of use	57 (30.3)	7 (14.9)	1 (6.7)
Accessibility	1 (0.5)	1 (2.1)	3 (20)
Intuitiveness	3 (1.6)	2 (4.3)	1 (6.7)
User-friendliness	11 (5.9)	4 (8.5)	2 (13.3)
Reliability	3 (1.6)	0 (0)	0 (0)
Simplicity	8 (4.3)	4 (8.5)	0 (0)
User interface	9 (52.9)	27 (50.9)	33 (52.4)
Visual interface	2 (11.8)	5 (9.4)	29 (46)
Devices	5 (29.4)	9 (17)	0 (0)
Functionality	1 (5.9)	8 (15.1)	0 (0)
Design configuration	0 (0)	4 (7.5)	1 (1.6)
Mobile experience	20 (60.6)	16 (64)	21 (50)
Ease of use	3 (9.1)	0 (0)	0 (0)
Interface	4 (12.1)	5 (20)	3 (7.1)
Mobile friendly	5 (15.2)	0 (0)	3 (7.1)
Mobile app	1 (3)	4 (16)	15 (35.7)
Patient experien ce	34 (55.7)	13 (48.1)	5 (62.5)
Convenience	10 (16.4)	0 (0)	0 (0)
Engagement	9 (14.8)	6 (22.2)	3 (37.5)
Patient input	3 (4.9)	8 (29.6)	0 (0)
Patient log-in	3 (4.9)	0 (0)	0 (0)
Efficiency	1 (1.6)	0 (0)	0 (0)
Empowerment	1 (1.6)	0 (0)	0 (0)
Researcher experience	8 (100)	0 (0)	0 (0)
Survey distribution and reminders			
Invitations and scheduling	20 (52.6)	25 (53.2)	25 (54.3)
Automated scheduling and messaging	5 (13.2)	1 (2.1)	2 (4.3)
Save and return	3 (7.9)	15 (31.9)	9 (19.6)

Code	Advantages, n (%)	Challenges, n (%)	Enhancements, n (%)
Invitation approaches	9 (23.7)	6 (12.8)	5 (10.9)
Calendar integration	1 (2.6)	0 (0)	3 (6.5)
Patient opt out	0 (0)	0 (0)	2 (4.3)
Reminders	7 (87.5)	7 (87.5)	7 (63.6)
Email text	0 (0)	0 (0)	1 (9.1)
Follow-up with patients	1 (12.5)	1 (12.5)	3 (27.3)
Easy distribution	11 (100)	0 (0)	3 (100)
Results and data			
Results view	5 (100)	0 (0)	4 (100)
Data sharing	8 (100)	0 (0)	3 (100)
Data quality	3 (100)	1 (100)	0 (0)
Training and support			
Education and training	7 (100)	7 (100)	8 (100)
Support	8 (100)	15 (100)	18 (100)
Patient support	2 (100)	12 (100)	16 (61.5)
Patient education and communication	0 (0)	0 (0)	10 (38.5)
Patient feedback	1 (100)	0 (0)	6 (100)
User misunderstanding and error	3 (100)	8 (100)	1 (100)
Technology and accessibility			
Consent	5 (100)	0 (0)	3 (100)
Technology integration	11 (100)	1 (100)	12 (100)
Technology access	17 (100)	51 (100)	0 (0)
Technology literacy	1 (100)	33 (100)	0 (0)
Security			
Privacy and compliance	15 (100)	2 (40)	2 (100)
Trust in technology	0 (0)	3 (60)	0 (0)
Platform features			
Data collection	14 (56)	4 (50)	5 (50)
Comprehensive	5 (20)	0 (0)	0 (0)
Data administration	3 (12)	1 (12.5)	3 (30)
Offline access	1 (4)	3 (37.5)	2 (20)
Familiarity	2 (8)	0 (0)	0 (0)
Cost	10 (100)	0 (0)	0 (0)
Comparison with other platforms	3 (100)	5 (100)	0 (0)
No input	7 (100)	22 (100)	52 (100)

[a]Due to the coding process (eg, double coding), the total number of secondary and tertiary codes may not add up to the primary code or 100%. The percentages are calculated based on the total number of codes in secondary and tertiary categories.

Survey Development and Customization

Respondents perceived that REDCap surveys were generally easy (20 codes) and quick (2 codes) to set up, build, organize, and maintain (2 codes). One participant commented on these topics, “Easy to build surveys Easy to make questions easy to answer Easy to build branching questions.”

However, respondents also noted that incorrect setup by the study staff and limited default formatting options and flexibility could be challenging in developing and completing surveys (3 codes).

While some respondents pointed out that REDCap provides continuous releases with new features (2 codes) and various design and automation options to ask a variety of questions for

efficient data collection (8 codes), respondents frequently pointed out the value of well-designed survey instruments in gathering high-quality information and engaging patients. They reported that complex, poorly designed surveys and ambiguous instructions (39 codes) could result in poor patient experience, potentially impacting the survey response rate and quality of data gathered. Respondents provided suggestions for enhancing survey design capabilities to streamline survey design and layout for the patients (including simplifying survey formatting, survey nesting abilities, and use of embedded fields). Respondents also suggested pilot testing of surveys before sending them out to patients (3 codes) and for study teams to follow best practices and guidelines to be more informed in survey methodologies and development. For example, 1 respondent commented:

> *Study teams following best practices with survey methodology and design, which can involve keeping surveys short & sweet, choosing appropriate field types for the question at hand, phrasing questions and response options well to reduce mental burden and make it easier for patients to answer questions.*

Respondents also reported that the availability of various response types, data validation, and branching logic ensure high-quality data collection (13 codes). One respondent commented on this advantage, "The wide array of validations can help patients enter data correctly."

Another respondent noted similarly, "Data validation and branching logic make participants conform to data standards and allows researchers to obtain higher quality data."

While data validation was discussed positively, respondents more frequently noted the challenges with response and logic types (22 codes), often pointing out that the actual response and logic types available from REDCap are not conducive to good survey design. One respondent made a clear reference to this issue saying, "It all depends on who sets up the survey, but until recently it has been a challenge to create grids of disparate data entry fields."

In addition, some respondents noted that due to the logic types, patients can make critical mistakes affecting the completeness of the data:

> *...branching logic at a very question to determine if they qualify or not. Sometimes, they accidently select different value in a hurry, and the survey gets completed. It is hard for them to change the response or refill the survey without admin help.*

Respondents noted many enhancement potentials within this category, such as voice input (4 codes), superior data entry experience (5 codes), use of a more conversational approach in response types (1 code), more effective multimedia (5 codes), and gamification of survey (2 codes). While REDCap offers multimedia options, respondents often suggested that options become more interactive and effective:

> *...more visual aids in questions, and the ability to answer with images. For example, by painting the areas afflicted on an image.*

One respondent explained how multimedia may be further useful:

> *...ability to add images to response options. Especially when working with minorities (traffic lights, or smiley faces).*

In addition to the design of the surveys, respondents noted that while REDCap surveys are readily customizable (7 codes), there are far more reported challenges (17 codes) and need for enhancements (15 codes). Respondents noted customization was not possible in some cases: "Default formatting options are limited."

However, many respondents focused on the lack of multi-language support (9 codes) as the critical challenge:

> *...multi-linguistic support. This is always a challenge for any software system/platform, and REDCap is no different...*

They frequently suggested enhancements to include multi-language support (8 codes) and customizations in forms' appearance (6 codes). For example, 1 respondent mentioned, "Allow for some more customization of the overall look/feel of surveys."

With respect to challenges with survey interactions, respondents reported that REDCap capabilities at the time did not send new surveys or allow patients to complete future surveys if previous surveys were incomplete (2 codes). One respondent mentioned the following:

> *...[t]he longitudinal design functionality in REDCap requires a participant to take each form before moving to the next, but our experiment design does not require this, and sometimes people will miss sessions and need to move on to the form for the next one. But if we stack all of the forms in one event, we cannot direct people to an individual form, only to the queue.*

One participant commented on REDCap's "inability to provide staff log-in status." (1 code). Respondents requested features for internal messaging or chat between study staff (2 codes), enhancing flow and cross-linking between projects (2 codes), ability to easily add study staff members outside of the organization (1 code), and ability for patients to skip longitudinal surveys (1 code).

User Experience

Respondents perceived REDCap to be easy to use for both patients (ie, to take surveys) and the study staff (ie, to build and distribute surveys; 57 codes). One respondent commented as follows:

> *REDCap is the easiest way to survey patients, families, and staff who are not part of our study team. We would not be able to conduct these surveys without it!*

They also perceived REDCap to be user-friendly (11 codes), simple (8 codes), intuitive (3 codes), timely (2 codes), and reliable (3 codes). Although some respondents reported REDCap allows for quick data collection (7 codes), they perceived that

lengthy or poorly designed surveys (eg, too many clicks and not enough instructions) could lead to fatigue and poor participation (15 codes). While the usability perceptions were generally positive, respondents reported that the platform was not as user-friendly or outdated as other commercial data collection platforms (4 codes), unintuitive (2 codes), and clunky for study staff (4 codes). They reported that "REDCap is not the simplest tool to learn how to use" for study staff (4 codes) and patients (3 codes). Respondents suggested the need to enhance accessibility features, such as the ability to change font size, screen reader view, and text-to-voice, among others (3 codes). In total, 8 (3.9%) of 207 respondents reported that the REDCap interface was advantageous for study staff considering its consistent interface and automated features, which reduce burden.

Respondents generally reported REDCap's visual user interface as challenging to use. Although some respondents perceived the interface to be clean or simple looking (9 codes) and optimized for various devices (5 codes), other respondents perceived that REDCap's interface was not modern looking (7 codes) or appealing (5 codes). One respondent mentioned, "The web interface of our survey pages are very basic, and narrow," whereas another respondent said, "[REDCap has] Very set layout of each item, can't make it look more 'modern' like other websites are at this time."

Respondents considered REDCap as not having a configurable design (4 codes) and some noted the user interface's poor functionality (8 codes). One respondent described both issues when explaining the challenges of the user interface:

> *REDCap is simply not user friendly in any way. The data structures are often too rigid and frankly outdated in being an effective tool for data collection.*

Respondents suggested the redesign of the REDCap user interface to be consistent with modern data collection platforms (27 codes), options to change the visual appearance and formatting of the surveys (3 codes), adding progress tracking aids (such as an automatic progress bar) for patients (2 codes), and a more flexible interface (1 code).

Some respondents appreciated REDCap's mobile access (4 codes), availability of mobile apps for study staff (REDCap mobile app; 2 codes) and patients (MyCap; 6 codes) supporting offline data collection, and perceived REDCap to be easy to use on mobile devices (3 codes) and mobile friendly (5 codes). While respondents appreciated the mobile interface, they reported that the mobile experience is affected by poor and suboptimal mobile user interface and scaling on smaller screens (5 codes). One participant reported the following:

> *We design our surveys on a computer, but many of our participants use their phones. We try to check how answers scale when the screen size changes, but some phones rescale to a different aspect ratio leading to challenges.*

They also reported that although the REDCap mobile app is available for study staff, it is not ideal and is difficult for study staff to set up the app (4 codes). One respondent mentioned the following:

> *I think that the REDCap mobile app is a bit too far separated from the web version, in as much as there is no access to external modules and other important features.*

Respondents suggested a need for an enhanced mobile app and interface (21 codes), including advanced capabilities for the study staff to view study records and perform analysis (2 codes) and push notifications (2 codes). One respondent mentioned the following:

> *[They need] better workflows with mobile phones, like notifications instead of just text messages. Something like an App except not the current one which is focus on asymmetric internet access.*

Respondents also commented on patient experience with REDCap. Overall, respondents noted that REDCap makes it easier for patients to complete the surveys at their convenience (10 codes), all while increasing engagement levels (9 codes). They saw REDCap as a way to make data collection more efficient and empowered (2 codes), especially as patients did not need to register or remember usernames or passwords to use the platform (3 codes). One participant said, "[Survey] Can be done at the patient's convenience from any digital device." A common challenge reported was the patient's desire and motivation to complete the surveys, being able to use the platform, and fatigue with lengthy surveys (13 codes). Suggestions for improving patient experience included maintaining engagement using visual aids and gamification (3 codes), a patient dashboard to keep them up to date on status of longitudinal studies (1 code) and making the platform more patient friendly (1 code). One respondent commented as follows:

> *For longer surveys, having a way of maintaining engagement by making the surveys more interactive (e.g. fun feedback to participants as they progress) would be nice. Some periodic messages of encouragement like "Great job!" "Keep it up!"*

Survey Distribution and Reminders

Respondents found it advantageous that REDCap included multiple ways to invite patients, such as emails or embedded links (4 codes). REDCap surveys were easy to distribute (11 codes) and could be automated and scheduled on a timeline easily. One participant commented on this aspect, "It can send surveys to participants directly, and on a schedule when the project is longitudinal." REDCap's ability to send patients custom links was an advantage respondents liked (3 codes): "For online surveys: able [to] send individualized email links...automated email with message that has piping upon completion." One respondent pointed out that there was "no scheduling component for visits" and suggested this feature. One respondent suggested the ability to send attachments with automatic notifications.

In addition, the ability to send completion reminder emails to patients was reported to reduce the burden on clinic staff while engaging patients (7 codes). Reminders also allowed the study staff members to follow up on incomplete surveys but 1 respondent mentioned that this was challenging while respondents suggested for improvements in customizing

reminders and enhanced tracking for incomplete surveys longitudinally (3 codes):

> *If there was a more efficient way to upload and manage patient invitations, as well as identify which patients have completed the survey within previous xx months therefore a new survey invitation does not need to be sent.*

Respondents noted patients sometimes missed invitations and reminders because email service providers blocked the emails (2 codes): "We have had email providers block REDCap emails, specifically Yahoo.com email." There was also confusion about the email sender as the emails were "from" REDCap instead of the study staff (2 codes):

> *From my experience... The emails that are sent out to respondents are not user friendly. The 'From' text box comes from REDCap, not from my email address.*

In addition, this respondent noted the emails were not user-friendly, sometimes arriving with broken links going to patient's junk mail, and requiring patients create a completely new log-in to complete a survey. One respondent suggested REDCap may "make it easier to send mass emails that are individually linked with the patient's profile; create a prettier or more visually appealing interface for patients." Furthermore, integrations to link communications to personal calendars were thought to be beneficial (3 codes). Respondents wanted a way to automatically opt out patients from surveys that were being distributed over a period (2 codes). One participant stated they, "would really like to be able to set a flag for opt-out subject [s] when distributing surveys over a period of time. We currently have to remove their emails to prevent future distribution."

Respondents commented on the "Save and Return" feature (25 codes), which allows patients to leave and return using a unique code to complete the survey at a later time. Although REDCap's Save and Return feature existed, respondents noted that this feature was often difficult to use (15 codes). They reported that patients may forget or not save their return code or may not know how to return to the survey, resulting in incomplete data or delay in data collection. One respondent commented, "It is not always obvious how to 'save and return later' if that is an option or even be aware that that is an option." Respondents suggested improvements (9 codes) to send the unique save and return code via emails, with reminders and save in invitation logs such that the study staff could provide it to patients if needed. In addition, respondents suggested that improving user-friendliness and patient awareness of this feature could increase response rates and data completion. A participant noted the following:

> *If they [patients] don't complete the survey the first time they often forget their return code and lose it. It would really help if the reminder emails had the return code, of if it could be included on the survery [sic] invitation log page that would make it much easier to find and give to the patient.*

Results and Data

Respondents liked that REDCap made data exportation easy for storage and analysis purposes (8 codes). Not only was it easy to export data out of the REDCap survey tool, it also made the analysis of the data much easier for the study staff, even those with minimal statistics training. As 1 respondent put it, "[REDCap has a] good translation into a dataset [and] easy statistics for those with minimal statistical training." Respondents (4 codes) pointed out the need for improving data exports and seamless communication with third-party solutions to send and receive information:

> *Being able to send to communicate and receive information from other software programs like Clinical Conductor for Demographic information and seamless data uploads.*

Respondents perceived that it is easy to create reports and monitor patient responses on REDCap and review specific data points (5 codes). Some respondents provided suggestions to edit charts and graphics as well as being able to share user- or survey-specific data (3 codes). For example, 1 respondent mentioned the following:

> *Ability for researchers to edit/modify graphics that can be automatically displayed with reports within redcap. This would facilitate researchers' ability to use those charts."*

Another respondent mentioned, "built in tools to share summary-level data (you vs the whole study) or findings." While respondents perceived that REDCap allows capturing accurate and complete high-quality data (3 codes), 1 respondent mentioned the following:

> *As with every self-service data entry portal accuracy of self-service data entry is wildly unreliable. There is real value to having a trained rep assisting the client enter information, when possible.*

Training and Support

Respondents reported that REDCap's active online community and support allowed REDCap users (including administrators and researchers) to find information and answers on how to manage, design, and conduct surveys (8 codes): "...it has a huge user base and a great consortium full of all the information you need to begin administering [surveys]." Respondents mentioned needing REDCap or IT support for patients to complete consent forms or surveys (12 codes). Although support existed for survey designers and administrators, it did not extend to patients completing surveys. Respondents suggested REDCap needed a way to educate or support patients in completing surveys (10 codes) and obtain help via on-demand messaging to study staff members (2 codes). As 1 participant suggested, REDCap should allow patients to "Click icon and get video explaining any information on a field." Another participant asked that REDCap have the following:

> *Dedicated instrument defined support button at the top that takes participants to a page made by the study team where we can put in a zoom room link monitored by study staff, phone numbers, or some pointers on definitions/examples on the instrument.*

Respondents also suggested a need for obtaining standardized patient feedback surveys to better engage them and understand their experience (6 codes).

Although some respondents mentioned REDCap required minimal training to get started (7 codes), some respondents (especially REDCap administrators) mentioned the need for training survey designers to set up REDCap tools and surveys to design high-quality surveys (7 codes). When asked about challenges, 1 participant mentioned the following:

> *Lack of resources for support (in person- phone) and functionality. It is not always easy and takes a lot of time to build tools. Not able to use to its fullest capacity or correctly—basically training ourselves. Library or community network does not help either. Not knowing how to set up properly more complicating functions inhibits usage.*

Respondents suggested more information and mandatory training for survey builders, including better guidelines and training videos to enhance builder and patient experience (8 codes). Respondents also perceived that patients taking surveys often do not understand how to fill out surveys or certain questions (8 codes) and having expert survey designers and well-designed surveys could alleviate these concerns (1 code).

Technology

Respondents often noted challenges of access to the internet and devices (51 codes) as well as technology literacy (33 codes):

> *Patients [without] a computer, device, or smart phone may not be able to use REDCap.*

As REDCap is web based, data collection could be difficult in rural and low-resource areas due to lack of access to technology (4 codes), such as a computer or reliable internet connection. Another participant noted, "I do work in global health, so our colleagues in resource-limited settings have challenges with the internet connection."

They also noted REDCap's ability to integrate with other technologies, such as messaging tools (eg, Twilio) as well as open application programmable interface to be beneficial (11 codes). In comparison, 1 participant noted as a challenge that, "integrating the ReCap [sic] extract with Epic [EHR] data. But once the system is setup it's easy to maintain." Respondents suggested integrations with other clinical trial management systems for seamless data transfers and EHRs to conduct surveys or autopopulate patient medical information:

> *The only other thing that would be super cool is if it could blow surveys into EPIC for documentation when needed.*

Respondents also referred to the informed consent capabilities of REDCap (8 codes). Even though they noted the consent module to be advantageous to obtain remote informed consents especially after the COVID-19 pandemic (5 codes), respondents suggested more enhancements, such as a 1-step consent process (3 codes).

Security

Respondents commented positively on the security and compliance of REDCap (15 codes). They reported that HIPAA compliance and the ability to store patient data securely are important advantages of REDCap. One participant commented that "all client data can be stored in one HIPAA compliant platform."

Respondents mentioned mistrust of technology (3 codes) could make patients feel uncomfortable sharing medical information on web-based platforms. One respondent commented that surveys requiring password protection are difficult for patients. They also provided enhancement suggestions (2 codes) related to maintaining HIPAA compliance, enhancing security, and assuring patients that their health information is safe and secure with REDCap.

Platform Features

Respondents found REDCap advantageous in enabling researchers to collect and patients to provide health data remotely (23 codes):"It has made it much easier for patients to submit their questionnaires and information using an online platform," especially during and after the COVID-19 pandemic.

Respondents perceived REDCap as a comprehensive or versatile (5 codes) data collection solution noting the following: "It provides us a comprehensive tool for collecting, tracking, and managing patient data and outreach." They also noted administration and maintenance (3 codes) to be advantageous as REDCap allows "being able to maintain administrative research tasks together with the data collection." They noted REDCap's offline data collection (using REDCap mobile apps) to be challenging (3 codes) and suggested that the offline feature should be improved for better data collection experience (2 codes). In addition, respondents noted the familiarity with REDCap among researchers (2 codes) and seamlessness (1 code) for the study personnel to be advantageous.

In addition, REDCap being available for free to REDCap consortium members was sought to be beneficial (10 codes). While some respondents noted REDCap being simpler and easier than other commercial platforms and paper forms (3 codes), some also noted that REDCap's interface was not easy to use or user-friendly compared to modern data collection tools (5 codes).

No Input

Respondents did not provide inputs with respect to advantages, disadvantages, and enhancement suggestions stating lack of experience or ability to provide inputs or not using REDCap for patient data collection (81 codes). Some nonsensical or unrelated comments lacking information context or irrelevant responses were excluded from the analysis. For example, when asked about enhancement suggestions for REDCap, 1 participant responded, "To REDCap or??."

Discussion

Overview

This study aimed to identify the advantages, challenges, and future opportunities for enhancements from the perspectives of REDCap administrators and researchers. To the best of our knowledge, this is one of the early studies of user perspectives on REDCap services and features. We believe that the findings of this study will aid REDCap developers and consortium users in better understanding stakeholder needs to enhance and customize REDCap features as well as researchers in improved survey development and data collection.

Principal Findings

Respondents had overwhelmingly positive perceptions of REDCap's survey design and data collection interface. The vast majority of respondents agreed or strongly agreed that data collected via REDCap were accurate (188/207, 90.8%), reliable (182/207, 87.9%), and complete (166/207, 80.2%). They found REDCap advantageous as it is free for its consortium members, secure, and easy to use. Respondents also perceived REDCap as easy and flexible to create and customize surveys including a variety of response and validation options, which make data collection easier for survey takers. However, respondents pointed out that poor survey design—often attributed to human factors (eg, lengthy forms and lack of knowledge among study staff) or technology limitations (eg, restrictions in survey and visual formatting in REDCap)—could result in poor patient experience and, ultimately, response and completion rates. Optimal design of survey forms is critical for assuring patient comprehension of the forms and accurate data collection [27,28]. Furthermore, direct investigations of REDCap user experiences and preferences could allow better understanding of the need for study staff and patient education. In addition, further research related to user needs for survey development and optimization can lead to enhancing their experience of developing high-quality surveys. One respondent pointed out the following:

> *It [REDCap] needs a much better understanding of how users engage the questions on a form (e.g., sit and watch users and staff try to figure out acceptable data type entries!). Needs a solid revamping in how it works "out front" and to run a series of user groups—patient and staff.*

Although respondents appreciated the availability of REDCap's community support for administrators and study staff, they pointed out that REDCap has room for improvement in this realm: the tool is not simple to learn, and there is a need for more training of study staff to help develop efficient, unambiguous survey instruments that can enhance patient experience. Poorly designed surveys and questions could potentially lead to incomplete responses and inaccurate data. Respondents pointed out the need for supporting patients, especially to ensure they understand the questions and can obtain help when needed in filling out surveys. Direct help from study staff members to fill out surveys or having the ability to directly send a message to study staff could alleviate misunderstandings and errors in completing surveys. Previous research suggests that the ability to obtain clarifications about survey questions can enhance response accuracy [28]. Further research and availability of resources are necessary to guide study staff members in creating well-designed instruments. In addition, understanding the factors affecting patients' experience in completing REDCap surveys and reasons for misunderstanding and errors could also enhance the REDCap experience and health data collection processes.

Opinions on patient experience and usability were more mixed. Most respondents agreed or strongly agreed that patients found REDCap easy to use (90.4%), able to be completed without assistance (79.8%), and able to be completed in a timely manner (87.5%). These strongly positive perceptions of REDCap usability are consistent with a prior study in which 6 out of 7 participants needed no help using REDCap, achieved 71% to 100% task completion, and provided 89% positive reaction words [13]. Qualitative outcomes showed that respondents perceived REDCap made it convenient for patients to provide data remotely without having to log in or remember credentials. Although they commented that patients can complete REDCap surveys using a device of choice (such as a laptop or mobile), technology access and technology literacy appeared to be a concern. Living in rural or low-income areas also presented issues for survey access. Respondents noted low-resource areas without stable internet access meant data collection was not reliable. Lack of internet access not only meant surveys could not be accessed but also meant the data collection process could be interrupted. REDCap's MyCap and REDCap Mobile app can allow study staff and patients to complete the collection of data offline, but they were also deemed challenging due to the lack of features compared with the web interface. In a study by Doyle et al [19], the REDCap mobile interface was less favorably received by participants. Similarly, REDCap's *Save and Return* feature allows users to complete surveys at a later time, which could be helpful during poor internet access; however, respondents recommended enhancements in the feature to improve patient experience, specifically an easier way for patients to remember and retrieve the return code. One participant noted this difficulty that patients face in attempting to use the feature:

> *If they don't complete the survey the first time, they often forget their return code and lose it. It would really help if the reminder emails had the return code, or if it could be included on the survey invitation [log-in] page...*

It is imperative to better understand patient and research participant experience with REDCap in completing surveys via larger and direct studies.

This study identified opportunities to improve the usability of REDCap. Respondents suggested enhancements in the patient-facing survey user interface to be in line with present EDC tools on the market, wanting a sleeker, modern, and cleaner looking interface. A variety of EDC tools are available for health care and non–health care data collection providing modern, device-friendly, and intuitive user interfaces to promote patient engagement [29-32]. In recent years, virtual conversational agents or chatbots have emerged as intuitive and engaging mediums for data collection. Modern data collection tools allow

survey designers to develop chatbot-based interactions to collect health data mimicking human-to-human conversations. Studies have shown that individuals prefer chatbot-based conversational data collection experience in comparison to traditional web-based forms [33,34]. Visual and graphical enhancements in REDCap appearance of surveys, patient communication, and researcher interface could support modernization of REDCap-based surveys, thus providing study staff and patients with clear and effective experience of health data collection.

Respondents wanted the mobile interface updated to look more like other commercial products, such as Qualtrics or SurveyMonkey. As more individuals are using mobile devices to obtain health information, it is of great importance to enhance their experience with mobile data collection [35]. They also suggested that the mobile apps have similar features as the web-based REDCap. Other requests included REDCap to support more languages or a translation service, where surveys could be translated to patients' preferred languages. Though it has some language capabilities, including Spanish, respondents wanted more language options built into REDCap. In addition, there was concern about the literacy of patients leading to suggestions for REDCap to include tools allowing patients with various literacy levels to access surveys. Respondents suggested inclusion of voice capabilities and more multimedia and gamification features in response options, such as a picture interface where patients could locate their pain visually for researchers. Inclusion of these features could further enhance the experience among patients with higher accessibility needs and low literacy. We also noted that some respondents suggested features that were available within REDCap at the time of conducting the survey. Suggestions included availability of REDCap's mobile version, embedded fields for responses, and integrations with messaging services such as Twilio. This again points out the need for education among study staff and organizational administrators to enable the optimal and effective use of REDCap features.

Limitations

This study is not without limitations. Although we recruited over 200 respondents, the sample size is small in comparison with the existing user base. We recruited fewer researchers (25/207, 12.1%) than administrators (150/207, 72.5%) who may be more directly involved in survey design and data collection. We also did not ask for participants' training and experience with REDCap. Future studies should focus on better understanding user perspectives (especially researchers) while also considering the type and amount of REDCap training received by the user. We asked individuals' opinions that are valuable but may be subject to bias, incomplete recall, or lack of information. For example, we asked information about their institution's REDCap use, but we did not include a response option or decline responding if they did not have accurate information. We also used REDCap as the platform to conduct our survey, which may have potentially biased responses by familiarizing participants with REDCap more than necessary. Participants' free-ended responses may have been influenced by how our study was designed or how the features were used. Future, more direct studies are warranted to better understand preferences. We recruited respondents from current REDCap consortium members, who may be more likely to believe REDCap is highly usable, as they may act as REDCap champions within institutions. We may be missing critical information by not capturing the perspectives of people who are not frequent users or consortium members. Future research should capture opinions of novice or past REDCap users. We also did not ask for information about participants' institutions, REDCap versions and plug-ins used, or institutional policies and customizations. It is possible that participant feedback may be related to institutional requirements or policies. Furthermore, we asked researcher and administrator opinions on patient experience. However, we did not directly assess the patient or research participant experience. Understanding patient experience is important to study in future research. In addition, a comparison of the REDCap experience with other EDC platforms could provide a better understanding of study staff and patient needs. A recent study compared individuals' experience in completing health forms using REDCap versus a chatbot platform. The results revealed that over 69% of participants preferred a chatbot for data collection with higher usability and net promotor scores for the chatbot [33]. The chatbot provided superior engagement and interactivity and was perceived as more intuitive than a standard, web-based REDCap interface. Future studies should look into better understanding study staff and patient needs to optimize survey development and data collection experience.

Conclusions

This pilot study aimed to assess stakeholder perspectives on experience with REDCap as an electronic health data collection tool. The findings revealed researchers and administrators perceive REDCap as a valued, low-cost resource that enables them to remotely collect and report health data in a secure and easy way. They also indicated a generally positive health data collection experience by clinical research and care staff members and patients. Although, with the advancements in data collection technologies and availability of interactive and intuitive user interfaces, there is a critical opportunity to enhance the REDCap experience to meet the needs of its vast user base of researchers and patients.

Acknowledgments

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health (award number 1R41LM013419-01). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. BB was funded by the National Institute of Mental Health (grant K23MH118482). BW was funded by the National Library of Medicine (grant R41LM013419).

Conflicts of Interest

BW is a shareholder of Doxy.me Inc and Dokbot LLC. All other authors are employees of Doxy.me Inc, a commercial telemedicine company. All authors declare no other conflicts of interest.

Multimedia Appendix 1
Survey.
[DOCX File , 16 KB - medinform_v12i1e49785_app1.docx]

Multimedia Appendix 2
Qualitative codebook.
[DOCX File , 26 KB - medinform_v12i1e49785_app2.docx]

References

1. Caron-Flinterman JF, Broerse JE, Bunders JF. The experiential knowledge of patients: a new resource for biomedical research? Soc Sci Med 2005 Jun;60(11):2575-2584. [doi: 10.1016/j.socscimed.2004.11.023] [Medline: 15814182]
2. Hanley B, Truesdale A, King A, Elbourne D, Chalmers I. Involving consumers in designing, conducting, and interpreting randomised controlled trials: questionnaire survey. BMJ 2001 Mar 03;322(7285):519-523 [FREE Full text] [doi: 10.1136/bmj.322.7285.519] [Medline: 11230065]
3. Saczynski JS, McManus DD, Goldberg RJ. Commonly used data-collection approaches in clinical research. Am J Med 2013 Nov;126(11):946-950 [FREE Full text] [doi: 10.1016/j.amjmed.2013.04.016] [Medline: 24050485]
4. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. Ann Transl Med 2018 Feb;6(3):42 [FREE Full text] [doi: 10.21037/atm.2018.01.13] [Medline: 29610734]
5. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med Inform Decis Mak 2016 Jul 13;16(1):90 [FREE Full text] [doi: 10.1186/s12911-016-0332-1] [Medline: 27411943]
6. Sacristan JA, Aguaron A, Avendaño C, Garrido P, Carrion J, Gutierrez A, et al. Patient involvement in clinical research: why, when, and how. Patient Prefer Adherence 2016 Apr;10:631-640. [doi: 10.2147/ppa.s104259]
7. van der Scheer L, Garcia E, van der Laan AL, van der Burg S, Boenink M. The benefits of patient involvement for translational research. Health Care Anal 2017 Sep 24;25(3):225-241 [FREE Full text] [doi: 10.1007/s10728-014-0289-0] [Medline: 25537464]
8. Summary of the HIPAA privacy rule. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html [accessed 2021-12-09]
9. REDCap homepage. REDCap. URL: https://www.project-redcap.org/ [accessed 2022-05-09]
10. Patridge EF, Bardyn TP. Research electronic data capture (REDCap). J Med Libr Assoc 2018 Jan 12;106(1) [FREE Full text] [doi: 10.5195/jmla.2018.319]
11. REDCap (Research Electronic Data Capture) - Harvard Catalyst. Harvard Catalyst. URL: https://catalyst.harvard.edu/redcap/ [accessed 2024-06-05]
12. Reichold M, Heß M, Kolominsky-Rabas P, Gräßel E, Prokosch HU. Usability evaluation of an offline electronic data capture app in a prospective multicenter dementia registry (digiDEM Bayern): mixed method study. JMIR Form Res 2021 Nov 03;5(11):e31649 [FREE Full text] [doi: 10.2196/31649] [Medline: 34730543]
13. Stambler DM, Feddema E, Riggins O, Campeau K, Breuch LA, Kessler MM, et al. REDCap delivery of a web-based intervention for patients with voice disorders: usability study. JMIR Hum Factors 2022 Mar 25;9(1):e26461 [FREE Full text] [doi: 10.2196/26461] [Medline: 35333191]
14. Kianersi S, Luetke M, Ludema C, Valenzuela A, Rosenberg M. Use of research electronic data capture (REDCap) in a COVID-19 randomized controlled trial: a practical example. BMC Med Res Methodol 2021 Aug 21;21(1):175 [FREE Full text] [doi: 10.1186/s12874-021-01362-2] [Medline: 34418958]
15. Tamuhla T, Tiffin N, Allie T. An e-consent framework for tiered informed consent for human genomic research in the global south, implemented as a REDCap template. BMC Med Ethics 2022 Nov 24;23(1):119 [FREE Full text] [doi: 10.1186/s12910-022-00860-2] [Medline: 36434585]
16. Wong TC, Captur G, Valeti U, Moon J, Schelbert EB. Feasibility of the REDCap platform for single center and collaborative multicenter CMR research. J Cardiovasc Magn Reson 2014 Jan 16;16(Supplement 1):P89. [doi: 10.1186/1532-429x-16-s1-p89]
17. Chen C, Turner SP, Sholle ET, Brown SW, Blau VL, Brouwer JP, et al. Evaluation of a REDCap-based workflow for supporting federal guidance for electronic informed consent. AMIA Jt Summits Transl Sci Proc 2019;2019:163-172 [FREE Full text] [Medline: 31258968]
18. Lee CA, Gamino D, Lore M, Donelson C, Windsor LC. Use of research electronic data capture (REDCap) in a sequential multiple assignment randomized trial (SMART): a practical example of automating double randomization. BMC Med Res Methodol 2023 Jul 06;23(1):162 [FREE Full text] [doi: 10.1186/s12874-023-01986-6] [Medline: 37415099]

19. Doyle S, Pavlos R, Carlson SJ, Barton K, Bhuiyan M, Boeing B, et al. Efficacy of digital health tools for a pediatric patient registry: semistructured interviews and interface usability testing with parents and clinicians. JMIR Form Res 2022 Jan 17;6(1):e29889 [FREE Full text] [doi: 10.2196/29889] [Medline: 35037889]
20. Garcia KK, Abrahão AA. Research development using REDCap software. Healthc Inform Res 2021 Oct;27(4):341-349 [FREE Full text] [doi: 10.4258/hir.2021.27.4.341] [Medline: 34788915]
21. Partners - REDCap. REDCap. URL: https://projectredcap.org/partners/ [accessed 2023-04-21]
22. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
23. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
24. Braun V, Clarke V. Thematic analysis. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA Handbook of Research Methods in Psychology, Vol. 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological. Washington, DC: American Psychological Association; 2012:57-71.
25. Bernard HR. Research Methods in Anthropology: Qualitative and Quantitative Approaches. Lanham, MD: Rowman & Littlefield Publishers; 2018.
26. Bernard HR, Wutich A, Ryan GW. Analyzing Qualitative Data: Systematic Approaches. Thousand Oaks, CA: SAGE Publications; 2016.
27. Bargas-Avila JA, Brenzikofer O, Roth SP, Tuch AN, Orsini S, Opwis K. Simple but crucial user interfaces in the world wide web: introducing 20 guidelines for usable web form design. In: Mátrai R, editor. User Interfaces. London, UK: IntechOpen; May 1, 2010.
28. Conrad FG, Schober MF. Clarifying survey questions when respondents don't know they need clarification. National Center for Education Statistics. 2001. URL: https://nces.ed.gov/FCSM/pdf/2001FCSM_Conrad.pdf [accessed 2023-04-24]
29. ClinCapture homepage. ClinCapture. URL: https://www.clincapture.com/ [accessed 2023-05-09]
30. Pawelek J, Baca-Motes K, Pandit JA, Berk BB, Ramos E. The power of patient engagement with electronic health records as research participants. JMIR Med Inform 2022 Jul 08;10(7):e39145 [FREE Full text] [doi: 10.2196/39145] [Medline: 35802410]
31. SurveyMonkey homepage. SurveyMonkey. URL: https://www.surveymonkey.com/ [accessed 2023-05-09]
32. Qualtrics XM: the leading experience management software. Qualtrics XM. URL: https://www.qualtrics.com/ [accessed 2023-05-09]
33. Soni H, Ivanova J, Wilczewski H, Bailey A, Ong T, Narma A, et al. Virtual conversational agents versus online forms: patient experience and preferences for health data collection. Front Digit Health 2022 Oct 13;4:954069 [FREE Full text] [doi: 10.3389/fdgth.2022.954069] [Medline: 36310920]
34. Ponathil A, Ozkan F, Welch B, Bertrand J, Chalil Madathil K. Family health history collected by virtual conversational agents: an empirical study to investigate the efficacy of this approach. J Genet Couns 2020 Dec 03;29(6):1081-1092. [doi: 10.1002/jgc4.1239] [Medline: 32125052]
35. Heimlich R. More use cell phones to get health information. Pew Research Center. 2012 Nov 14. URL: https://www.pewresearch.org/fact-tank/2012/11/14/more-use-cell-phones-to-get-health-information/ [accessed 2022-04-18]

Abbreviations

EDC: electronic data capture
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
REDCap: Research Electronic Data Capture

Edited by C Lovis; submitted 09.06.23; peer-reviewed by S Wang, C Chen; comments to author 12.03.24; revised version received 10.04.24; accepted 04.05.24; published 25.06.24.

Please cite as:
Soni H, Ivanova J, Wilczewski H, Ong T, Ross JN, Bailey A, Cummins M, Barrera J, Bunnell B, Welch B
User Preferences and Needs for Health Data Collection Using Research Electronic Data Capture: Survey Study
JMIR Med Inform 2024;12:e49785
URL: https://medinform.jmir.org/2024/1/e49785
doi:10.2196/49785
PMID:

Original Paper

Introducing Attribute Association Graphs to Facilitate Medical Data Exploration: Development and Evaluation Using Epidemiological Study Data

Louis Bellmann[1], PhD; Alexander Johannes Wiederhold[1], MD; Leona Trübe[1], PhD; Raphael Twerenbold[2,3,4], Dr Med; Frank Ückert[1], Dr Med; Karl Gottfried[1]

[1]Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[2]Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[3]German Center for Cardiovascular Research (DZHK) Partner Site Hamburg-Kiel-Lübeck, Hamburg, Germany

[4]University Center of Cardiovascular Science, University Heart & Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Corresponding Author:
Louis Bellmann, PhD
Institute for Applied Medical Informatics
University Medical Center Hamburg-Eppendorf
Martinistr. 52
Hamburg, 20246
Germany
Phone: 49 15228842404
Email: l.bellmann@uke.de

Abstract

Background: Interpretability and intuitive visualization facilitate medical knowledge generation through big data. In addition, robustness to high-dimensional and missing data is a requirement for statistical approaches in the medical domain. A method tailored to the needs of physicians must meet all the abovementioned criteria.

Objective: This study aims to develop an accessible tool for visual data exploration without the need for programming knowledge, adjusting complex parameterizations, or handling missing data. We sought to use statistical analysis using the setting of disease and control cohorts familiar to clinical researchers. We aimed to guide the user by identifying and highlighting data patterns associated with disease and reveal relations between attributes within the data set.

Methods: We introduce the attribute association graph, a novel graph structure designed for visual data exploration using robust statistical metrics. The nodes capture frequencies of participant attributes in disease and control cohorts as well as deviations between groups. The edges represent conditional relations between attributes. The graph is visualized using the Neo4j (Neo4j, Inc) data platform and can be interactively explored without the need for technical knowledge. Nodes with high deviations between cohorts and edges of noticeable conditional relationship are highlighted to guide the user during the exploration. The graph is accompanied by a dashboard visualizing variable distributions. For evaluation, we applied the graph and dashboard to the Hamburg City Health Study data set, a large cohort study conducted in the city of Hamburg, Germany. All data structures can be accessed freely by researchers, physicians, and patients. In addition, we developed a user test conducted with physicians incorporating the System Usability Scale, individual questions, and user tasks.

Results: We evaluated the attribute association graph and dashboard through an exemplary data analysis of participants with a general cardiovascular disease in the Hamburg City Health Study data set. All results extracted from the graph structure and dashboard are in accordance with findings from the literature, except for unusually low cholesterol levels in participants with cardiovascular disease, which could be induced by medication. In addition, 95% CIs of Pearson correlation coefficients were calculated for all associations identified during the data analysis, confirming the results. In addition, a user test with 10 physicians assessing the usability of the proposed methods was conducted. A System Usability Scale score of 70.5% and average successful task completion of 81.4% were reported.

Conclusions: The proposed attribute association graph and dashboard enable intuitive visual data exploration. They are robust to high-dimensional as well as missing data and require no parameterization. The usability for clinicians was confirmed via a

user test, and the validity of the statistical results was confirmed by associations known from literature and standard statistical inference.

(JMIR Med Inform 2024;12:e49865) doi:10.2196/49865

KEYWORDS

data exploration; cohort studies; data visualization; big data; statistical models; medical knowledge; data analysis; cardiovascular diseases; usability

Introduction

The amount and availability of data around us are constantly increasing. Researchers are increasingly using statistical models to guide their data-driven scientific work. However, as the relationships discovered increase in complexity, the models themselves are becoming gradually less transparent. In high-stake decision fields, such as health care, data explanation and justification of decision-making are essential for the applicability and distribution of novel technologies. Here, we present new methods for extracting statistical insights from large data sources and visualizing the results based on graph structures. The methods balance complexity and comprehensive description of the results on the one hand and clarity and interpretability for clinicians and patients on the other hand.

The availability of large quantities of medical data is growing [1,2] and thus enabling machine learning methods to play an ever-increasing role in medical research [3-5]. With the undoubtedly numerous advantages of “big data” in medicine arises the problem of increasing complexity and lack of transparency for clinicians [6,7]. In this context, the call for more interpretable statistical models is gaining more attention [8,9]. In addition to the interpretability of the applied models and results, good data visualization methods are key for the knowledge communication with clinicians and patients. Many methods have been developed over the years [10-12].

For data-driven analysis, approaches originating from the mathematical field of graph theory gain an increasing amount of attention for health care applications [13]. A graph consists of nodes representing arbitrary objects and edges each connecting 2 nodes corresponding to some form of relation between them. Graph-based database technologies, such as Neo4j (Neo4j, Inc) [14], allow more efficient retrieval of large amounts of data compared to traditional relational database systems [15,16], and many software tools for interactive, graphical user interfaces are available [14,17-20].

Knowledge graphs are a form of data representation capturing large quantities of data from potentially multiple sources in a graph structure. Existing data are usually processed and jointly represented to enable accessible, often visual, exploration of condensed knowledge across different data modalities and sources. Owing to their intuitive and versatile character, knowledge graphs have many applications in the medical domain [21]. Examples are the representation of biomolecular pathways [22], research related to COVID-19 or diabetes [23,24], knowledge about dietary supplement [25], and networks of complex disease interactions [26].

Statistical analysis discovering relations between variables within a medical data set can be captured within a graph structure. In this context, Bayesian networks are of increasing interest in the medical domain [27,28]. They represent conditional dependencies as edges and the absence of an edge as probabilistic independence [29]. Using these conditional dependencies, Bayesian networks can be used for inferring neural networks [30] or diagnosis prediction [31]. However, they are sensitive to missing data during the model training process [32]. Markov models describe states, for example, events during a patient’s hospital stay, as nodes and transition probabilities between states as edges [33]. As a result, Markov models are applied for the analysis of time-dependent dynamic processes in health care [33-35]. In association rule learning, relations between variables are extracted from a data set based on different measurements of interest, for example, conditional probability [36]. This concept is applied to extract patterns from clinical databases [37] or find suitable drug treatments [38]. All 3 approaches capture variable relations across a complete data set.

In this work, we developed the attribute association graph (AAG), a new graph structure capturing statistical knowledge extracted from a data set. We aimed to combine the focus of knowledge graphs on interpretability, accessibility, and visual exploration with graph-based statistical methods. We sought to develop a novel and robust tool for statistical analysis that is intuitively usable by physicians. We tailored our approach specifically to the needs of data-driven analysis in the medical domain by incorporating disease and control cohorts and aiming for robustness to high-dimensional or not normally distributed data, small sample sizes, and missing values. The graph is visualized, and nodes and edges representing variable relations of interest are highlighted to attract the attention of the user and facilitate the data analysis. We complemented the AAG with a dashboard for further data exploration. Only mouse clicking and search bar prompting in English are required for the navigation of the graph and dashboard. We aimed to evaluate the validity of the statistical analysis represented by the graph structure and dashboard. Therefore, we conducted an exemplary data analysis based on a large epidemiological study. The results of the analysis were compared with findings from literature and standard statistical inference using CIs of Pearson correlation coefficients. In addition, we assessed the usability of the visualization for medical researchers. We conducted user tests with physicians using standardized usability tests, user tasks, open feedback questions, and a free data exploration. The generated graph structure and dashboard are freely available to clinical researchers for exploration on their own computers.

Methods

AAG Definition

Our goal is to visualize participant attributes and the statistical traits and relationships between them in a compact, interpretable, and intuitive way. As a participant attribute, we consider a singular value or semantically meaningful value group for a variable, for example, "the participant was diagnosed with hypertension" or "participant has total cholesterol level above 200 mg/dL." For the statistical analysis, we use simple metrics, which were found to be intuitive for clinicians [39]. The metrics are calculated for a disease and control group and compared to identify attributes with a large deviation. Thus, in contrast to traditional association rule mining [36], Bayesian networks [29], or Markov models [34], attributes can be selected, which appear more often in the disease group compared to the control group. As we analyze relations of singular attributes instead of associations between variables, our results are methodologically different from correlation analysis, such as chi-square tests [40] or Pearson correlation coefficients [41].

In the AAG, single attributes are captured as nodes and visualized as colored spheres of different sizes. Each node has parameters for the name of the attribute's variable, its value, and a short description including units of measurement for metric variables. In addition, we assigned labels to each node depending on the broad categories of the represented attribute, for example, *Cardiac*, *Condition*, or *Medical History*.

For metric variables, we calculated reference ranges based on their value distribution within the whole data set. We defined the reference range as all values within SD around the mean. On the basis of reference ranges, we derived 3 additional nodes for the attribute associated with values below, within, and above the reference range. The 3 nodes inherit the parameter's *name* and *description* from the original nodes. They have the value *low*, *normal*, or *high*. In addition, they contain the lower and upper bound of the reference range. All participants are assigned to 1 of the 3 nodes based on their attribute value. Thus, metric values, for example, patient laboratory results, are labeled in comparison to the whole data set and enriched with semantics.

In addition, we enriched the nodes with several statistical measurements of the described participant attribute within the data set. The resulting parameters are given in Table 1. Note that the relative attribute share accounts for the common problem of missing data [42,43] and is an upper bound to the relative total share. By measuring the difference and quotient of relative attribute shares, the distinction in attribute distribution between the 2 groups is expressed. The size and color of the node visualization capture parts of these measurements to support the data exploration with visual highlights.

Table 1. Statistical parameters for a node describing attribute *a* together with a short description and formula[a].

Parameter	Description	Formula
Absolute count	Number of group members having attribute *a*	c_i
Relative total share	Fraction of group members have attribute *a*	
Relative attribute share	Relative total share, missing value adjusted	
Relative attribute share difference	Absolute difference of relative attribute shares	
Relative attribute share quotient	Fraction of maximum and minimum relative attribute share	

[a]Parameters with subscript *d* refer to the disease group. Parameters with subscript *c* refer to the control group. Subscript *i* refers to a definition for both groups, that is, $i \in \{d,c\}$. Let g_i be the group size, and  be the number of group members having a valid value for the attribute *a*, that is, not a missing value.

We assigned a frequency label impacting the node's size based on the maximum relative attribute share. Therefore, a node's size indicates how common an attribute is within one of the groups. Let p be the maximum relative attribute share of a node. The node is assigned to 1 of the following 3 frequency label types:

- $p \geq 0.5$: labeled as *highly frequent*; the node has the largest size.
- $0.1 \leq p < 0.5$: labeled as *frequent*; the node has a medium size.
- $p < 0.1$: labeled as *infrequent*; the node has the smallest size.

In addition, we assigned a distinction label to each node from which its color is derived. The distinction label, and thus the node color, indicates how much the attribute distribution differs between groups. Here, brighter colors signal a larger distinction. We reuse the symbols in Table 1. Each node is assigned 1 of 5 colors and distinction label types:

- $\delta \geq 0.2$ or $\gamma \geq 2.0$:
  - $p_d > p_c$: labeled as *highly related*; the node is colored in red.
  - $p_d < p_c$: labeled as *highly inverse*; the node is colored in blue.
- ($\delta \geq 0.1$ or $\gamma \geq 1.5$) and $\delta < 0.2$ and $\gamma < 2.0$:
  - $p_d > p_c$: labeled as *related*; the node is colored in orange.
  - $p_d < p_c$: labeled as *inverse*; the node is colored in turquoise.
- $\delta < 0.1$ and $\gamma < 1.5$: labeled as *unrelated*; the node is colored in beige.

Combining size and color, nodes that are displayed largest and brightest represent attributes with high frequency and large distinction between groups. As all parameters calculated for an individual node depend only on data for a single variable, the computation time needed for the calculation of all nodes of the graph scales linearly with the number of variables and linear with the sample size.

In the AAG, edges point from a source node to a target node, indicating the conditional dependence of the target attribute on the source attribute. The edges are displayed as lines with arrows pointing from the source node sphere to the target node sphere. The calculated statistical parameters for the conditional dependence are presented in Table 2. Note that the relative conditional share is conceptually equivalent to confidence in association rule learning [36]. By measuring the difference and quotient of the relative conditional share and the unconditional relative attribute share of the target node, the impact of the added condition is expressed. This impact can be negative if the unconditional relative attribute share is larger than the relative conditional share. We assign a type to each edge to capture the impact of the added condition. In the visualization, the line thickness of the edge is given by its type. We reuse the symbols in Table 2. Each node is assigned to 1 of the following 3 types:

- $\delta' \geq 0.2$ or $\gamma \geq 2.0$: assigned to the *high conditional difference* type; the edge has the thickest line.
- ($\delta' \geq 0.1$ or $\gamma \geq 1.5$) and $\delta' < 0.2$ and $\gamma < 2.0$: assigned to the *medium conditional difference* type; the edge has a thinner line.
- $\delta' < 0.1$ and $\gamma < 1.5$: assigned to the *low conditional difference* type; the edge has the thinnest line.

Table 2. Statistical parameters for an edge pointing from a source node *x* to a target node y[a].

Parameter	Description	Formula
Absolute cooccurrence	Number of group members having both attributes of *x* and *y*	o_i
Relative conditional share	Fraction of group members with attribute of *x*, also having attribute of *y*	
Conditional and unconditional target share difference	Absolute increase of relative conditional share compared to relative attribute share of *y*	
Conditional and unconditional target share quotient	Quotient of relative conditional share and relative attribute share of *y*	

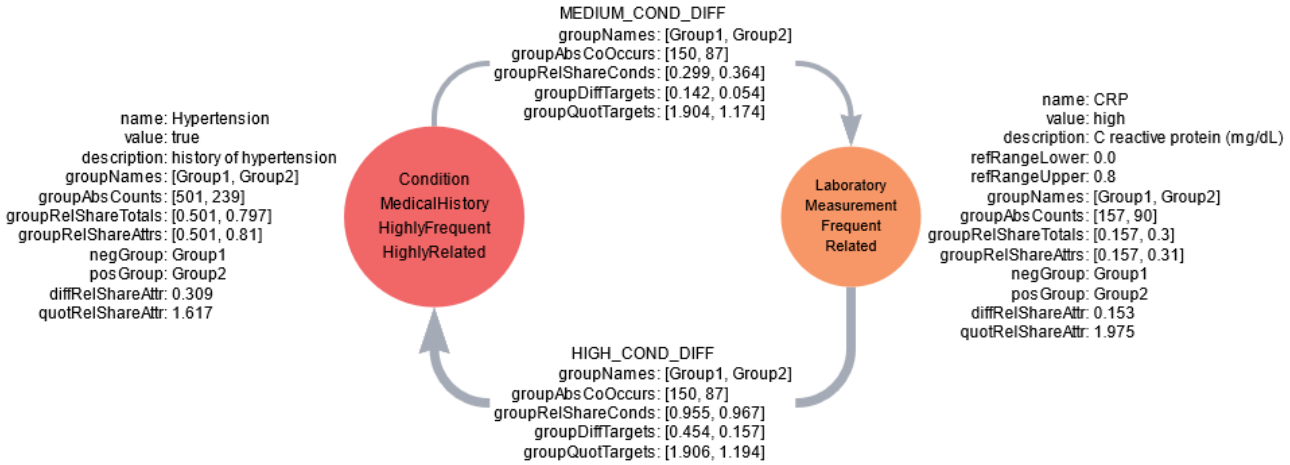
[a]Subscript *i* refers to a definition for both groups. Let be the absolute count of *x* and be the relative attribute share of *y*.

The computation time for the generation of all the AAG's edges scales quadratically with the number of variables in the data set and linear with the sample size.

In the last step, the nodes and edges are filtered by their statistical parameters to highlight the most relevant attributes and conditional dependencies. A detailed description of the filtering procedure is provided in Multimedia Appendix 1 [41,44,45]. We represented the extracted data in a graph structure using the graph data platform Neo4j [14] and the graphical user interface Neo4j Bloom (Neo4j, Inc) [19]. The graph structure can be navigated by mouse clicking and via a search bar typing prompts in English.

Figure 1 [46] shows a minimal fictional example of an AAG with 2 nodes capturing fictional data about history of hypertension and high C-reactive protein (CRP) measurements as well as their relationship in participant group 1 (control group) and 2 (disease group). We conducted a hypothetical data analysis, as we intend the AAG to be used. For CRP measurements (mg/dL), a fictional reference range of 0.0-0.8 was derived. From the difference of the relative total share and relative attribute share, we can infer existing missing values on group 2 for both attributes. In group 1, no missing values exist because relative total share and relative attribute share do not differ. Regarding the quotient of relative attribute shares, we can infer group 2's participants being almost twice as likely to show a high CRP value. Thus, a CRP measurement >0.8 mg/dL might be highly related to the condition or property of group 2 compared to participants of group 1. A history of hypertension appears approximately 30% more often in group 2, giving a 60% proportional increase. As a result, its node is labeled as highly related to the condition or property of group 2. Viewing the data of the edges, we find that almost all participants with a high CRP measurement also have a history of hypertension in both groups. Therefore, high CRP values could be an indicator for hypertension in both fictional groups. Conversely, only approximately one-third of participants with a history of hypertension also show high measurements of CRP. This pattern of conditional relationship is similar between groups and could thus be independent of the group definitions, for example, medical condition and control group.

Figure 1. An attribute association graph with 2 nodes represented as spheres and 2 edges represented as lines with arrows. The arrow indicates the target node of the edge. Node parameters are depicted next to the spheres. Labels are shown inside the spheres with one label per line. The edge's parameters are depicted on top of the edge. The heading above the edge's parameters specifies the edge type (MEDIUM_COND_DIFF for medium conditional difference, HIGH_COND_DIFF for high conditional difference). Absolute counts (groupAbsCounts), relative total shares (groupRelShareTotals), relative attribute shares (groupRelShareAttrs), difference between relative attribute shares (diffRelShareAttr), quotient between relative attribute shares (quotRelShareAttr), absolute cooccurrence (groupAbsCoOccurs), relative conditional share (groupRelShareConds), difference to target relative attribute share (groupDiffTargets), and quotient to target relative attribute share (groupQuotTargets) are depicted as lists with the score for group 1, followed by the score for group 2. Group 2 is the disease group (posGroup), and group 1 is the control group (negGroup). The color of the sphere indicates the deviation label of the node: orange (related) and red (highly related). The size of the sphere indicates the frequency label from medium (frequent) to the largest size (highly frequent). The line thickness indicates the type of edge from medium (medium conditional difference) to thickest (high conditional difference). Descriptions of all parameter names, edge types, labels as well as color, size and thickness encoding can be found in the ZFDM repository. CRP: C-reactive protein.



Dashboard

To complement the AAG, we generated a dashboard using the NeoDash (Neo4j, Inc) [17] toolkit. With the dashboard, users can investigate the average and distribution of metric variables across participant groups in more detail. In addition to the cardiovascular disease and control cohorts, the group of all participants contained in the Hamburg City Health Study (HCHS) data set was included. We developed 2 different tabs. The first tab allows for comparison of participant groups. We included the sizes of disease and control group. In addition, variable distributions can be compared between groups. For this purpose, we applied the following workflow to all metric variables and participant groups. First, we measured the variable average within the group. Second, we generated a binned distribution by rounding the measurements to multiples of 0.1, 0.5, 1, 5, 10, or 50 depending on the SD within the group. Bins containing <0.5% of the participants or <3 participants are summarized. We removed distributions without any bins fulfilling these requirements. The user can select 2 groups and variables for the distributions shown in the first tab of the dashboard. The averages of all metric variables for all 3 groups are shown in the first tab as well. To make them comparable in a figure, the averages of each variable are normalized by the maximum average of that variable. In the second tab, the user can investigate the relationship between 2 variables within a participant group. For the first variable, the generated binned distribution across the group is shown. For the second variable, we use precalculated averages of participants within a bin. The x-axis of the resulting figure shows the bin values of the first variable, and the y-axis shows the average value of the second variable for participants of that bin.

HCHS Data Set and Cohort Selection

To evaluate the AAG and dashboard, we used an exemplary data exploration workflow of a large epidemiological cohort study. We compared the results with findings from literature and standard statistical analysis. The HCHS is a single-center, prospective, observational, population-based cohort study of 45,000 randomly selected residents of the metropolitan region of Hamburg, Germany, aged between 45 and 74 years. The study design has been published [47], and the study is registered [48]. The study focuses on major chronic diseases, causes for their development, as well as factors for survival and support for life in survivorship. The study considers >6000 properties per participant. The data are raised from 18 examinations, primarily targeting major organ systems, as well as questionnaires about medical and family history, physical condition, dietary habits, lifestyle, and various other topics. The examinations will be repeated after 6 years to obtain large-scale, long-term assessments. For this analysis, the HCHS committee provided a subset of the whole HCHS data set focusing on cardiovascular and cancer diseases. The subset consists of 524 selected attributes for the first 10,000 participants enrolled in HCHS, including information about laboratory analyses; electrocardiography (ECG); magnetic resonance imaging; vascular ultrasound examinations; blood pressure measurements; cardiovascular and cancer medical history questionnaires; as well as dietary, lifestyle and sleeping habits. We selected 131 (25%) of these 524 attributes, translated their descriptions to English, assigned labels to each variable to broad variable groups, and added Systematized Nomenclature of Medicine Clinical Terms [49] or Logical Observation Identifier Names and Codes [50] codes. When no directly fitting code was found, we chose the code of a related term. A full list of all variables, descriptions, labels, vocabulary codes, and data types can be

freely accessed [46]. In some cases, the reference ranges calculated for the AAG deviated from the usual reference ranges known from the literature because of a different value distribution in the HCHS data set. In these cases, we manually adjusted the reference intervals according to the Merck Manual of Diagnosis and Therapy manual [44]. A full list of the adjusted reference ranges can be found in Table S1 in Multimedia Appendix 1. In this work, we focused on participants with a general cardiovascular condition. We included participants in this cohort who met any of the following criteria: showed any pathological cardiovascular findings during the cardiac magnetic resonance imaging examination; had a missing sinus rhythm; had a finding of atrial fibrillation or flutter in the ECG check; or reported a medical history of cardiac infarction, coronary artery disease, angina pectoris, congestive heart failure, myocarditis, or valvular endocarditis in the questionnaire. As a result, the disease cohort contained 1917 participants. In addition, we derived the control group of 8083 participants not exhibiting any of the conditions and findings.

User Tests

Study Design

We conducted a user test using a mixed methods approach to evaluate the usability of the AAG. The associated questionnaire can be found in Multimedia Appendix 2. We did not consider the proposed dashboard in the user test, as dashboards are widely used in the medical domain [11,51-53]. The usability testing consisted of 3 main parts in the following order: (1) in a 30-minute preparation phase, participants independently worked through the AAG user manual and the Neo4j Bloom overview website [19]; (2) a semistructured interview with open feedback questions and user tasks was conducted; and (3) participants completed the System Usability Scale (SUS) [54]. The SUS is a standardized and validated instrument for usability testing of systems, frequently used in this context [25,52,55,56]. The SUS comprises 10 questions rated on a 5-point Likert scale. The total score, ranging from 0 to 100, is calculated from all questions to ensure comparability. With the addition of user tasks and feedback questions tailored to the AAG, we aimed to create additional insights on the usability of the specific parts of the graph as well as observe the data exploration conducted by the users. The user tasks can be grouped into three categories: (1) reproducing the introduced labels and metric parameters; (2) using the application functionalities necessary for exploration; and (3) conducting a free exploration of 2 AAG subgraphs of the HCHS data set: first, the 10 nodes with the highest quotient of relative attribute shares related to the cardiovascular disease group; and second, the subgraph of nodes regarding laboratory measurements. The user results for tasks of categories 1 and 2 were evaluated as correct or incorrect by the authors. During the exploration of the 2 subgraphs, the users were asked to verbalize their findings, and the results were recorded and categorized by the authors. The participant answers to the open feedback questions were also broadly categorized by the authors.

Participant Recruitment

The study participants for the user tests included 10 physicians from various specialties and fields of activity. This group comprised 2 anesthetists, 2 cardiologists, 1 neurologist, 1 radiologist, 2 resident doctors in the field of child and adolescent psychiatry, 1 medical student in the final year, and 1 physician working in the public health sector. With this heterogeneous group composition, we aimed for a comprehensive usability assessment of the presented methods across the clinical field. The recruitment of participants was conducted on a voluntary basis, supported by the research team's network. It was assumed that the participants had no bias regarding the AAGs, as the methodology and visualization had not been officially released and were therefore not used by the participants at the time of the user test.

Ethical Considerations

The HCHS study was approved by the Ethics Committee of the Hamburg chamber of physicians (PV5131) and has been registered at ClinicalTrial.gov (NCT03934957).

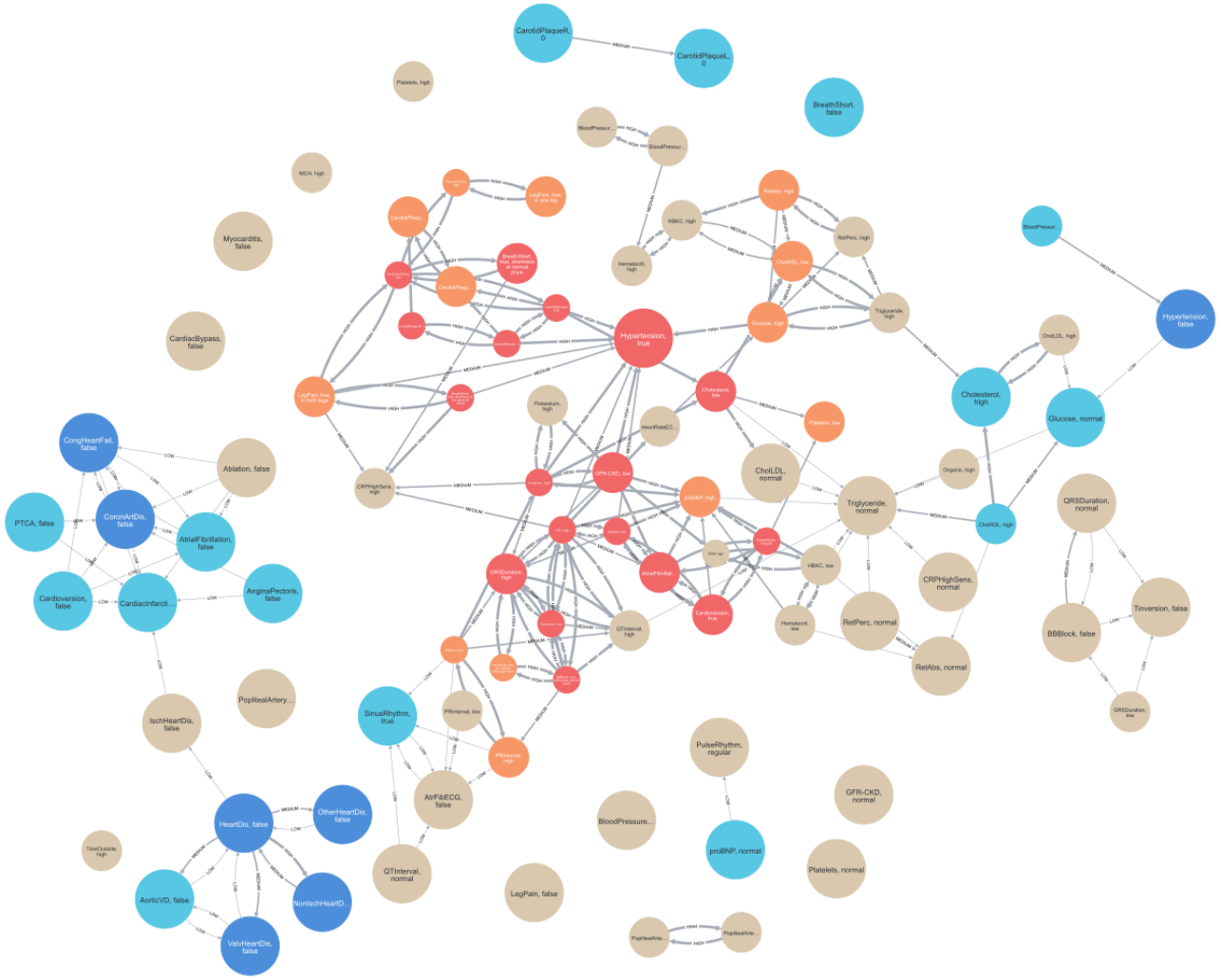
Results

Exemplary Data Analysis

We have generated the AAG for the disease and control group within the HCHS data set based on our definition of a general cardiovascular disease. In this paragraph, we give an exemplary data analysis using the graph and some aspects of the dashboard. This analysis was conducted by the authors of this work independently of the exploration of users during the usability test. The analysis is meant to showcase the usability of the graph representations and is by no means exhaustive. The Neo4j database dumps, configuration files, and user guide can be freely accessed [46]. In addition, the software tool used to generate AAGs was made publicly available [57] and will be presented in an upcoming publication. To assess the compatibility of the presented methods with standard statistical inference, we calculated Pearson correlation coefficients [41], 1-tailed CIs at the confidence level of 95% using the Fisher transformation [45], and P values for 1-tailed null hypothesis testing of statistical independence for all associations discussed in the following data analysis. The results can be found in Table S2 in Multimedia Appendix 1.

For brevity, we define the cardiovascular disease group as group A and its control group as group B. Group A contains 1917 participants, and group B contains 8083 participants. The generated AAG is shown in Figure 2 [14,46]. The nodes labeled as related or highly related form a cluster in the middle of the graph with the highest density of edges between them. Most of the inverse and highly inverse labeled nodes are primarily located on the periphery of the graph with many interconnections but few connections to the inner cluster. This observation indicates a clear distinction highlighted by the graph between the attributes based on their cooccurrence with cardiovascular disease within the HCHS data set.

Figure 2. The attribute association graph describing the cardiovascular disease cohort and control group extracted from the Hamburg City Health Study data set. Screenshot taken from the Neo4j Browser. Nodes are depicted as spheres, and edges are depicted as lines between spheres. The color of the sphere indicates the deviation label of the node: vanilla (unrelated), orange (related), red (highly related), turquoise (inverse), and blue (highly inverse). The size of the sphere indicates the frequency label from the smallest (infrequent) to the largest size (highly frequent). The line thickness indicates the type of edge from thinnest (low conditional difference) to thickest (high conditional difference). The text inside the node spheres states the variable name, followed by the value of the attribute. Data and variable descriptions can be found in the ZFDM repository. For a higher-resolution version of this figure, see Multimedia Appendix 3. Variable descriptions are found in Multimedia Appendix 4.



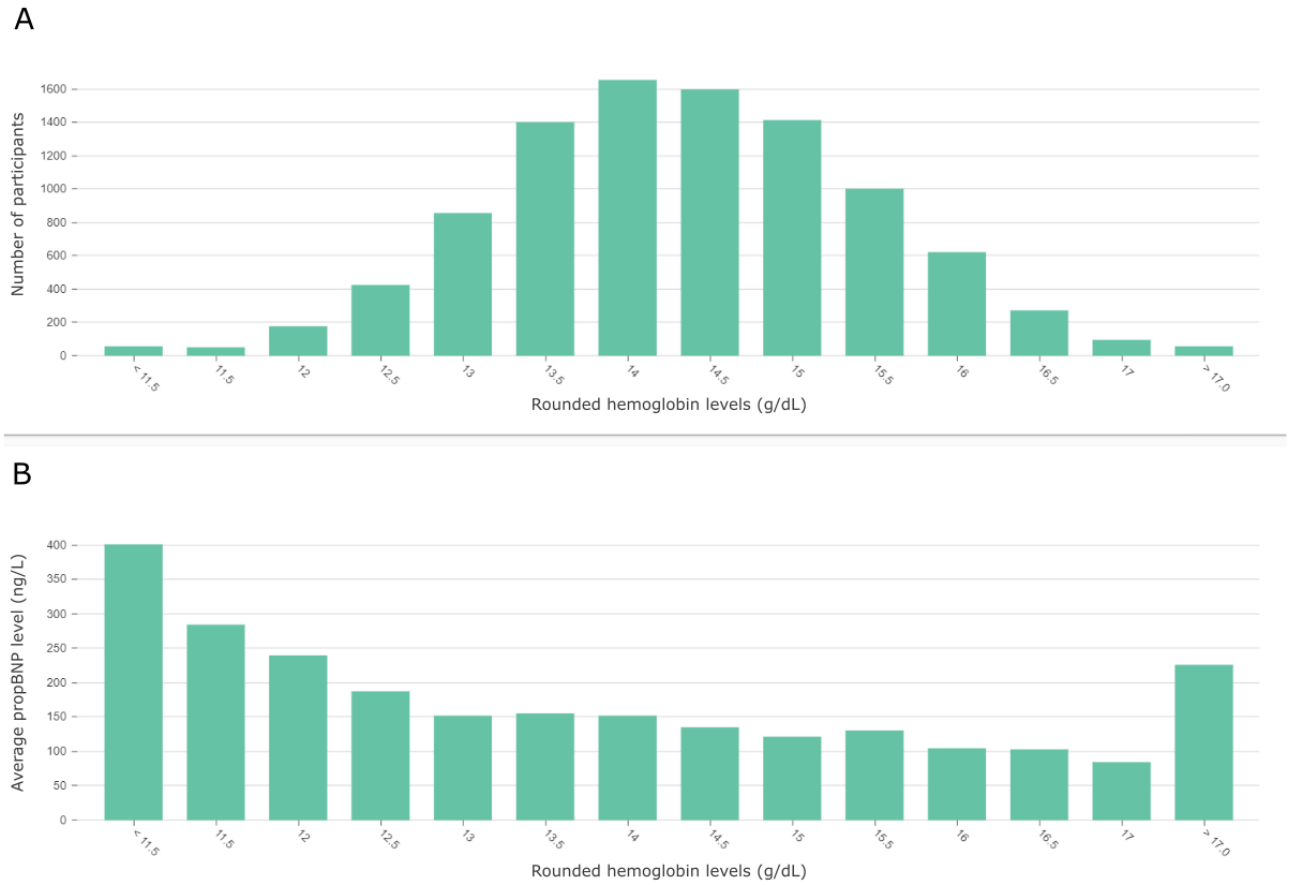
For a more detailed analysis of this AAG, we focused on the laboratory results data shown in Figure 3 [14,46]. Within the graph, 3 nodes are labeled as highly related, along with several adjacent nodes labeled as related. The nodes representing glomerular filtration rate <60 mL/min/1.73 m^2 ("GFR-CKD, low") and creatinine levels >1.2 mg/dL ("creatine, high") are identified as highly related and are interconnected. Furthermore, they are also connected to the node representing elevated potassium levels >4.15 mmol ("potassium, high") through high conditional difference relationships. The presence of a low glomerular filtration rate, high creatinine, and elevated potassium levels are all correlated with chronic kidney disease [58], which in turn is a risk factor for the development of cardiovascular conditions [58,59]. Thus, all 3 laboratory results are associated with heart disease in clinical settings [60], which coincides with the findings presented in this graph. The respective 95% CIs lie fully above 0 for creatine and potassium levels and fully below 0 for the glomerular filtration rate. The relative attribute share of the nodes for glomerular filtration rate <60 mL/min/1.73 m^2 ("GFR-CKD, low") in group A is, with 12%, more than twice as high as the relative total share. This indicates missing values for glomerular filtration rate measurements in participant with a cardiovascular condition. The related node in the center of Figure 3 ("proBNP, high") represents elevated N-terminal prohormone of B-type natriuretic peptide (proBNP) levels >125 ng/L, which were identified as a biomarker for cardiac diseases [61]. With 47%, group A has a 1.7-fold increased relative attribute share for this attribute compared to group B. The associated CI for the Pearson correlation coefficient is strictly positive. The node has 3 incoming edges of high conditional difference. Of these 3 edges, 2 describe the relationship between low glomerular filtration rate and high creatinine levels to elevated proBNP levels. Participants of group B with 1 of these properties are at least 1.6-fold more likely to show elevated proBNP levels >125 ng/L compared to general patients of group B. The same pattern can be observed in group A, which is consistent with the impact of worsening kidney function on proBNP concentration [62,63]. The CIs of the Pearson correlation coefficient of proBNP and glomerular filtration rate is strictly negative, and the CI for

creatinine and proBNP levels is fully positive. The third incoming edge is of type high conditional difference. It indicates a relationship between hemoglobin levels <13 g/dL ("HBKC, low") and elevated proBNP measurements. Although the node for low hemoglobin levels is labeled as unrelated, measurements <13 g/dL appear with a 1.4-fold increase in group B compared to group A. The associated CI is close to, but fully above, 0. Interestingly, participants of both groups with low hemoglobin levels are approximately 1.5-fold more likely to exhibit high proBNP measurements compared to general participants of their group, a phenomenon observed in other studies [64-66]. The Pearson correlation coefficient CI for proBNP and hemoglobin levels are close to, but fully below, 0. Overall, these 3 relationships confirm that while elevated proBNP levels serve as a biomarker for cardiac conditions, other factors may also contribute to its elevation.

Figure 4 was extracted from the dashboard and discloses the relationship of hemoglobin and proBNP levels across the whole data set in more detail. Average proBNP values increase for participants with hemoglobin levels <13 g/dL. Interestingly, proBNP levels also increase in participants with high hemoglobin values >17 g/dL. For further investigation, we returned to the graph and inspected the node ("HBKC, high") for high hemoglobin levels >15.5 g/dL. This threshold is exceeded by 21.5% of the participants in group A and by only 15.3% of the participants in group B. These observations align with the calculated, strictly positive CI and findings of other studies associating high hemoglobin concentrations with cardiovascular disease [67,68]. The third node ("cholesterol, low"), which is labeled as highly related, can be seen in the lower center of Figure 3. It represents total cholesterol levels <150 mg/dL, which is exhibited by 16.3% of group A and only 5.5% of group B. Conversely, total cholesterol levels >200 mg/dL are observed in 47.3% of group A and 61.2% of group B. As a result, the corresponding node ("cholesterol, high") is labeled as inversely related.

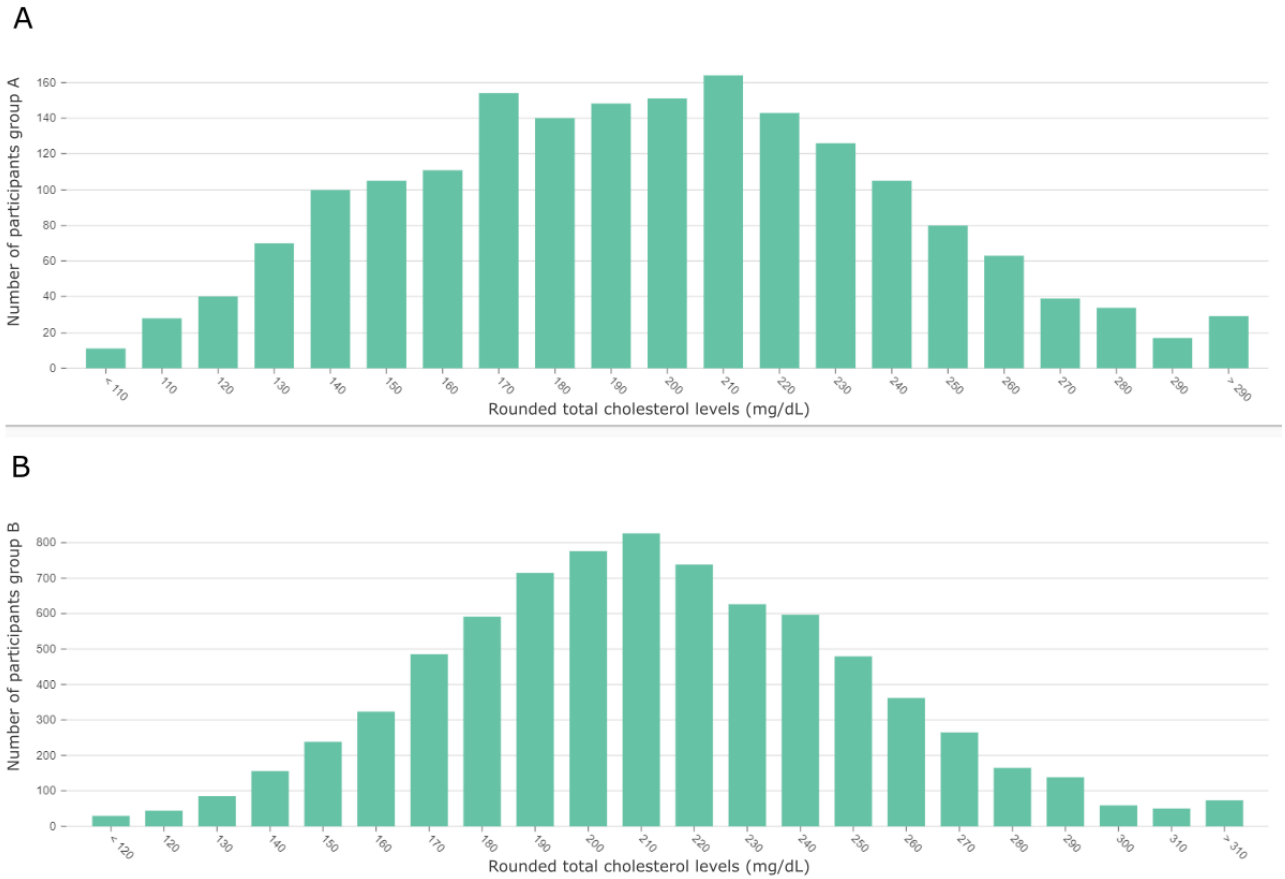
Figure 3. A subgraph of the full attribute association graph describing the cardiovascular disease cohort and control group extracted from the Hamburg City Health Study data set. Screenshot taken from the Neo4j Browser. Only nodes representing laboratory measurements and edges between them are shown. The color of the sphere indicates the deviation label of the node: vanilla (unrelated), orange (related), red (highly related), turquoise (inverse), and blue (highly inverse). The size of the sphere indicates the frequency label from the smallest (infrequent) to the largest size (highly frequent). The line thickness indicates the type of edge from thinnest (low conditional difference) to thickest (high conditional difference). The text inside the node spheres states the variable name, followed by the value of the attribute. Data and variable descriptions can be found in the ZFDM repository. CKD: chronic kidney disease; CRP: C-reactive protein; GFR: glomerular filtration rate; HBKC: hemoglobin level; HDL: high-density lipoprotein; LDL: low-density lipoprotein; proBNP: prohormone of B-type natriuretic peptide. For a higher-resolution version of this figure, see Multimedia Appendix 5. Variable descriptions are found in Multimedia Appendix 4.

Figure 4. (A) Distribution of hemoglobin levels (g/dL) across all participants of the Hamburg City Health Study data set. (B) The average N-terminal prohormone of B-type natriuretic peptide (proBNP) level (ng/L) per participant of the data set with a rounded hemoglobin level specified on the x-axis. This figure is a screenshot from the dashboard.



However, in Figure 5, we can observe that the highest number of participants in both groups exhibit a slightly elevated total cholesterol level of 210 mg/dL. Next, we inspected the 2 nodes (“CholLDL, normal” and “CholLDL, high”) for low-density lipoprotein (LDL) cholesterol levels. Measurements >130 mg/dL (“CholLDL, high”) appear with a 1.3-fold increase in group B. LDL cholesterol levels <130 mg/dL (“CholLDL, normal”) appear in 68.1% of group A and 59.7% of group B. These observations are peculiar because elevated total and LDL cholesterol are commonly recognized as important risk factors for cardiovascular diseases [69-73]. A similar pattern can be inferred from the 2 nodes (“CholHDL, low” and “CholHDL, high”) for measurements of high-density lipoprotein (HDL) cholesterol. Levels <45 mg/dL appear with a 1.7-fold increase in group A, whereas measurements >83 mg/dL showed a 1.8-fold increase in group B. This observation coincides with the widely accepted inverse association of HDL levels with cardiovascular diseases [74,75]. It is noteworthy that the nodes for high LDL and HDL cholesterol levels share an edge with the node for high total cholesterol levels. The same holds true for low HDL, normal LDL, and low total cholesterol measurements. These edges are all labeled with “high conditional difference.” The CIs for all 3 cholesterol measurements and the membership to group A are strictly negative. The CIs for total cholesterol levels and HDL as well as LDL cholesterol measurements are strictly positive, with the correlation coefficient of LDL and total cholesterol being close to 1. In summary, reduced overall cholesterol, LDL cholesterol, and HDL cholesterol levels appear more often in the cardiovascular disease group compared to the control group and are associated with each other. As stated earlier, this observation contradicts the commonly accepted association of elevated overall and LDL cholesterol with cardiovascular diseases. It could be attributed to the widely used therapy with statins [76], which mainly targets the reduction of LDL and overall cholesterol [77]. On the basis of this idea, the high conditional difference relation between elevated creatinine levels and low total cholesterol measurements found in Figure 3 and the associated strictly negative CI for the Pearson correlation coefficient could be explained by statin-associated muscle symptoms [78]. However, additional information about the medication history of the participants would be required and could be a starting point for further investigation.

Figure 5. (A) Distribution of total cholesterol levels (mg/dL) for the cardiovascular disease group (group A) and (B) its control group (group B) derived from the Hamburg City Health Study data set. This figure is a screenshot taken from the dashboard.



User Tests

The participants indicated a work experience in the current field ranging from 1 to 10 years, with an average of 5.8 years. The data exploration tools mostly used by the participants were SPSS (IBM) [79], R [80], and Microsoft Excel (Microsoft) [81]. No users mentioned any prior experience with graph-based statistical analysis tools. The results of the user test can be found in Multimedia Appendix 3.

In Figure 6, the results of the SUS questionnaire are shown and range from 62.5 to 85.0. The mean of 70.5 indicates the passing of usability criteria [82] and a rating of "good" usability [83]. In addition, physicians rated the user-friendliness on a scale from 1 (very bad) to 10 (very good), with a mean of 7.0 in accordance with the SUS results.

In Figure 7, the percentage of the 10 participants with successful completion is shown for each user task. The average score across all tasks is 81.4%, with 6 (86%) of 7 navigation tasks being correctly completed by all participants. However, only 20% (2/10) of participants queried successfully for the 10 nodes most statistically associated with the disease group by the quotient of relative attribute shares. Regarding the description tasks of category 1, all but 1 task of reproducing label and parameter meaning was completed by at least 70% (7/10) of users. An exception was task C3.2 where participants should describe the meaning of the edge parameter for the difference of relative conditional share and relative attribute share of the target node. This task was only completed correctly by 30% (3/10) of the participants. In addition, only 30% (3/10) of the participants found the parameter names for nodes understandable, and only 10% (1/10) of the participants classified the edge parameter names as clear.

During the free data exploration, all participants noticed the unusually low levels of total and LDL cholesterol in the cardiovascular disease group compared to the control group, which is also discussed during the exemplary data analysis conducted by the authors. In addition, 40% (4/10) of the participants suspected this association to be caused by medication not represented in the data set. Overall, 60% (6/10) of the participants discussed ECG signals, and 60% (6/10) of the participants discussed kidney metabolism. Moreover, 70% (7/10) of the physicians mentioned the results of their data exploration to be plausible, except for total and LDL cholesterol unprompted. Regarding the answers to the open feedback questions, 80% (8/10) of the participants mentioned the colors and sizes of nodes to be helpful, and 40% (4/10) of the participants referred to the display of attribute connections as edges becoming apparent. Moreover, 30% (3/10) of the participants mentioned the benefit of initial data exploration without the need for numerical values. As to disadvantages of the AAG, 30% (3/10) of the users mentioned the edge definitions being hard to understand, 20% (2/10) assessed the graphs to be too crowded to get a good overview, and 20% (2/10) stated that they would need more practice to use the tool efficiently.

Figure 6. The System Usability Scale (SUS) score for each of the 10 participants of the user test. In addition, the average score is represented by a horizontal dashed line in red.

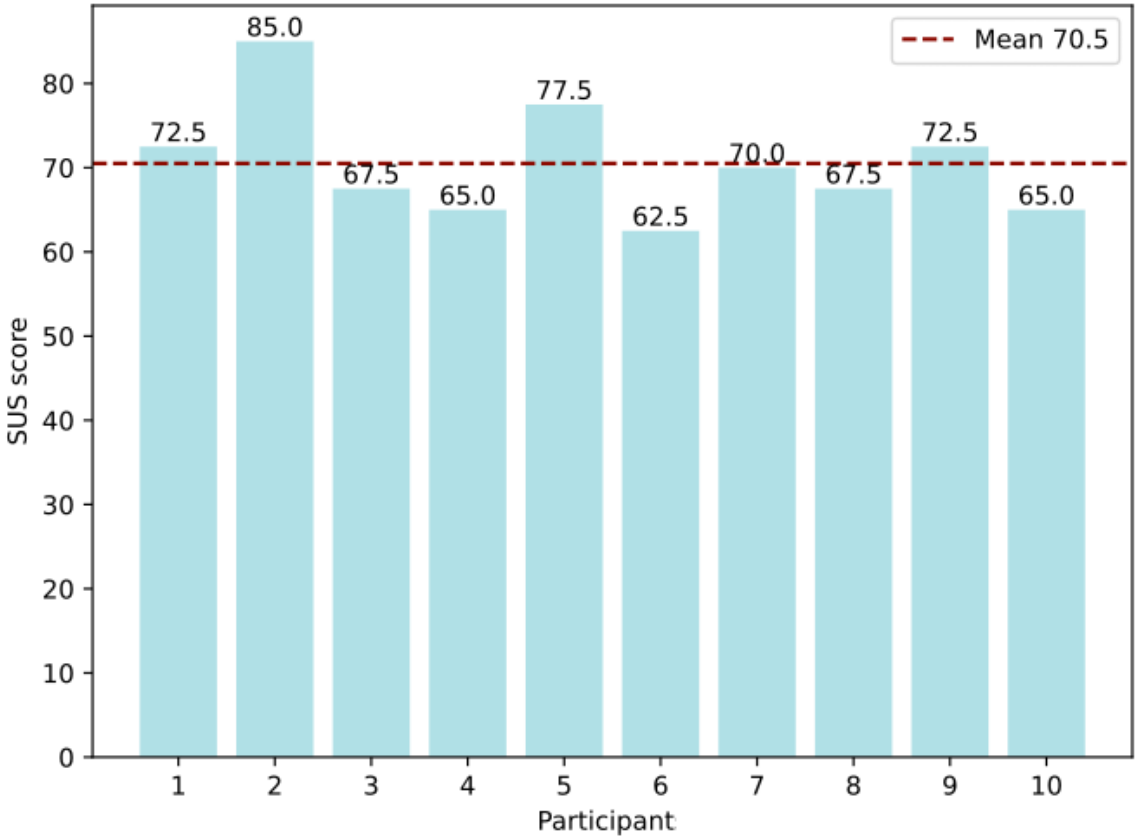
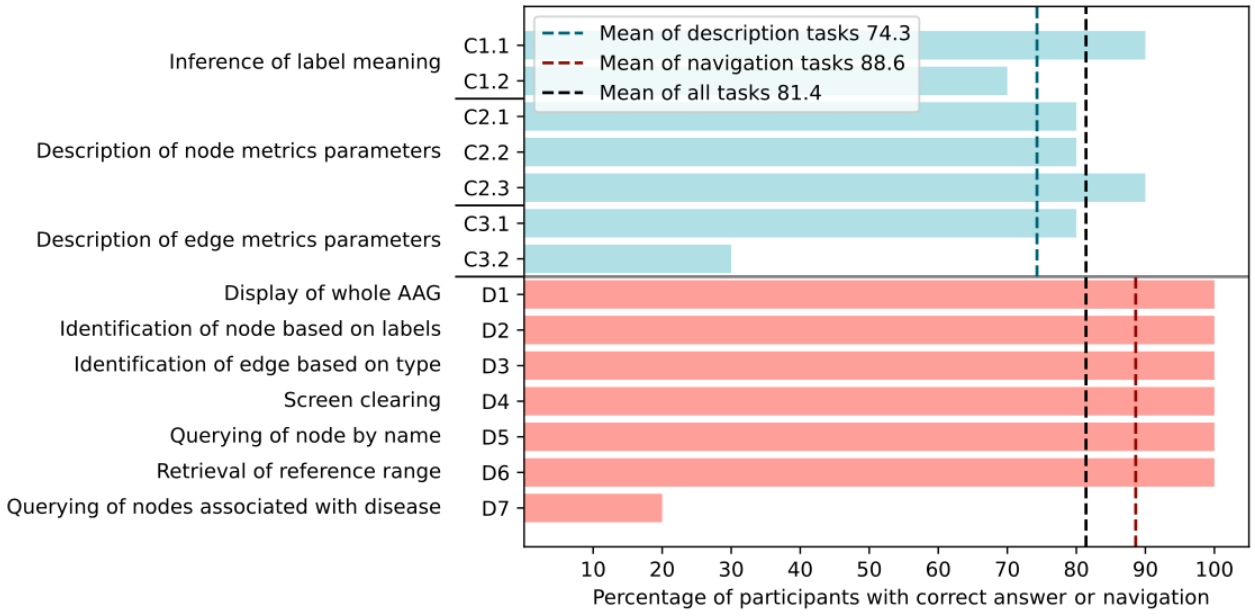


Figure 7. Correct task completion by participants during the user test in percentage. Task numbering is taken from the questionnaire. A short description of the tasks is given on the left. Bars for description and reproduction of labels and metrics (task category 1) are depicted in turquoise. Bars for graph navigation tasks (task category 2) are depicted in pink. Average percentages of correct tasks are plotted as dashed lines for description, navigation, and all tasks.



Discussion

Principal Findings

In this work, we presented the AAG for visual exploration of medical data sets using disease and control cohorts. The graph structure represents attributes as nodes and identifies as well as visually highlights attributes, which are linked to the observed disease by robust statistical metrics. Relations between attributes are captured as edges by conditional frequencies. As a result, attributes associated with the observed disease are visually clustered and clearly separated from attributes, which are associated with the control group. The graph structure detects and handles missing values without the need for data deletion.

The usability of the AAG and dashboard was assessed using an exemplary data analysis. All but 1 association of laboratory measurements and cardiovascular diseases extracted from the HCHS data set are in line with findings from the literature. The exceptions are unusually low total and LDL cholesterol levels in participants with cardiovascular disease, which might be caused by lipid-lowering therapy. All results extracted from the AAG were confirmed by standard statistical inference using null hypothesis testing and CI for the Pearson correlation coefficient. In addition, a user test with physicians was conducted using the standardized SUS questionnaire, nonstandardized open feedback questions as well as user tasks, and a free data exploration. The SUS score of 70.5% and average successful task completion of 81.4% show a general acceptance and good usability of the AAG. After the initial 30-minute preparation period, all users were able to navigate the graph and could extract medical knowledge that they considered plausible and meaningful. In addition, all participants identified the unusual lipid levels in participants of the cardiovascular disease group and some suspected medication not represented in the data set to be the cause. The encoding of statistical results by color, size, and clustering of nodes as well as thickness of edges was seen as helpful by the users. The users regarded the tool as useful for accessible hypothesis formation during the initial research phase.

Comparison With Prior Work

Other existing data-driven approaches based on graph structures focus mainly on the connection of different data sources as knowledge graphs [23,24,26] or direct clinical decision support through outcome prediction [27,31,35,38,84-86]. To our knowledge, a graph structure capturing statistical measurements of a medical data set using disease and control cohorts with a clear focus on interpretability and visualization is a novel approach. In addition, as our proposed methods consider single attributes and pairs of attributes, they are robust to high-dimensional data, which pose a problem for many other statistical models applied to the medical domain [87]. We believe that the usability of graph-based visualizations in the medical field is rarely assessed using standardized tests such as the SUS questionnaire. The only other results known to the authors reported a slightly lower SUS score of 64.4 [25].

Regarding the graph-based statistical framework, we see our work closest related to Bayesian networks [29] and association rule learning [36]. While Bayesian networks can hold strong predictive power [88], the choice of prior distribution and sensitivity to data quality can be challenging for clinicians [89]. In association rule learning, conditional relationships between attributes are partially expressed through the confidence parameter, which is quite similar to our methodology in that regard. However, we enrich the added condition with semantics by calculating difference and quotient to the unconditional relative frequencies. Finally, none of the 2 methods measure statistical differences between disease and control cohorts. We believe this to be vital in our approach for generation of insight and adoption in the medical domain.

Limitations

We intended the AAG and dashboard as a compact visualization for data exploration in the initial phase of research projects. We aimed to incorporate easily interpretable, robust metrics in the form of conditional and unconditional absolute and relative frequencies as well as their deviations between disease and control cohorts. However, because of this choice of metrics, the accuracy could be lower when used in prediction tasks compared to, for example, Bayesian networks or other nonlinear models. In addition, CIs and null hypothesis significance testing play a key role in statistical inference of medical data [90]. They are not incorporated into the methods presented here but could be a follow-up to the initial exploration using the AAG. Finally, temporal data cannot be handled with the proposed methodology in the current form, and Markov models [33] could be applied instead.

Regarding the usability of the visualization, the results of the user test indicate a need for simplification of the parameter names regarding the statistical measurements. In addition, the comparison of conditional and unconditional frequencies captured in the edges of the graph structure was not accessible enough for the users. Moreover, the prompt for retrieval of nodes most associated with one of the groups was considered too lengthy by the users. The authors will incorporate this valuable feedback in the next update iteration of the presented methods.

Conclusions

In this work, we introduced the AAG, a novel graph-based representation of statistical data combined with a dashboard. These structures can be visually explored and allow for data analysis tailored to the needs of the medical domain. The usability of the graph structure and dashboard was confirmed by user tests conducted with physicians. In addition, the validity of the incorporated statistical analysis was assessed through an exemplary data analysis of a large epidemiological study, and its compatibility with standard statistical methodology and findings from the literature was established. For the future, it might be of interest to enable clinicians in generating their own AAGs without the need for programming experience as an extension to their existing data analysis workflow. To achieve this, we developed a software package [57], which will be presented in an upcoming publication. We think that accessible data analysis and intuitive presentation for clinicians and patients is the way forward in a world of ever-growing data availability and complexity.

Acknowledgments

The authors would like to thank the Hamburg City Health Study (HCHS) committee for granting access to the HCHS cohort study data set. The authors received no specific funding for this work. In terms of overall funding for the underlying HCHS, various institutes and departments at the University Medical Center Hamburg-Eppendorf contribute with their own individual and scaled budgets. The HCHS is additionally funded by the Joachim Herz Foundation, the Leducq Foundation (grant 16 CVD 03), the euCanSHare grant agreement (grant 825903-euCanSHare H2020), and the Innovative Medicine Initiative (grant 116074). The HCHS is further supported by Deutsche Gesetzliche Unfallversicherung (DGUV), Deutsches Krebsforschungszentrum (DKFZ), Deutsches Zentrum für Herz-Kreislauf-Forschung (DZHK), Deutsche Stiftung für Herzforschung, Seefried Stiftung, Bayer, Amgen, Novartis, Schiller, Siemens, Topcon, and Unilever and by donations from the "Förderverein zur Förderung der HCHS e.V." and TePe (2014). Sponsor funding has in no way influenced the content or management of this study. RT reports research support from the German Center for Cardiovascular Research (DZHK), the Kühne Foundation, the Joachim Herz Foundation, the Swiss National Science Foundation (Grant NoP300PB_167803) and the Swiss Heart Foundation.

Data Availability

The data sets generated and analyzed during this study are available in the ZFDM repository [46]. The HCHS data set itself is not publicly available due participant data privacy. The attribute association graph (AAG) and dashboard data as Neo4j dumps, Neo4j Bloom configuration as JSON files, as well as a detailed installation and user guide as PDF file and descriptions for all variables of the Hamburg City Health Study data subset can be found in the repository [46]. Adjusted reference ranges and filter criteria for the AAGs, Pearson correlation coefficients, as well as the user test questionnaire and results can be found in Multimedia Appendices 1, 2 and 6. Code repository, Python package, and software tool to create custom AAGs [57] will be described in an upcoming publication.

Conflicts of Interest

RT reports speaker honoraria/consulting honoraria from Abbott, Amgen, Astra Zeneca, Psyros, Roche, Siemens, Singulex and Thermo Scientific BRAHMS. RT is co-founder and shareholder of the ART-EMIS Hamburg GmbH, which holds an international patent application on a computing device for estimating the probability of myocardial infarction (International Publication Numbers WO2022043229A1, TW202219980A).

Multimedia Appendix 1
Attribute association graph filter criteria, manually adjusted reference ranges, Pearson correlation coefficients, CIs at the confidence level of 95% using the Fisher transformation, and *P* values for 1-tailed null hypothesis testing of statistical independence.
[DOCX File , 34 KB - medinform_v12i1e49865_app1.docx]

Multimedia Appendix 2
User test questionnaire.
[PDF File (Adobe PDF File), 161 KB - medinform_v12i1e49865_app2.pdf]

Multimedia Appendix 3
The attribute association graph describing the cardiovascular disease cohort and control group extracted from the Hamburg City Health Study data set. Screenshot taken from the Neo4j Browser. Nodes are depicted as spheres, and edges are depicted as lines between spheres. The color of the sphere indicates the deviation label of the node: vanilla (unrelated), orange (related), red (highly related), turquoise (inverse), and blue (highly inverse). The size of the sphere indicates the frequency label from the smallest (infrequent) to the largest size (highly frequent). The line thickness indicates the type of edge from thinnest (low conditional difference) to thickest (high conditional difference). The text inside the node spheres states the variable name, followed by the value of the attribute. Data and variable descriptions can be found in the ZFDM repository.
[PNG File , 4960 KB - medinform_v12i1e49865_app3.png]

Multimedia Appendix 4
Variable descriptions for Figures 2 and 3.
[XLSX File (Microsoft Excel File), 24 KB - medinform_v12i1e49865_app4.xlsx]

Multimedia Appendix 5
A subgraph of the full attribute association graph describing the cardiovascular disease cohort and control group extracted from the Hamburg City Health Study data set. Screenshot taken from the Neo4j Browser. Only nodes representing laboratory

measurements and edges between them are shown. The color of the sphere indicates the deviation label of the node: vanilla (unrelated), orange (related), red (highly related), turquoise (inverse), and blue (highly inverse). The size of the sphere indicates the frequency label from the smallest (infrequent) to the largest size (highly frequent). The line thickness indicates the type of edge from thinnest (low conditional difference) to thickest (high conditional difference). The text inside the node spheres states the variable name, followed by the value of the attribute. Data and variable descriptions can be found in the ZFDM repository. CKD: chronic kidney disease; CRP: C-reactive protein; GFR: glomerular filtration rate; HBKC: hemoglobin level; HDL: high-density lipoprotein; LDL: low-density lipoprotein; proBNP: prohormone of B-type natriuretic peptide.
[PNG File , 1350 KB - medinform_v12i1e49865_app5.png]

Multimedia Appendix 6
Results of the user test.
[XLSX File (Microsoft Excel File), 25 KB - medinform_v12i1e49865_app6.xlsx]

References

1. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearb Med Inform 2014 Aug 15;9(1):14-20 [FREE Full text] [doi: 10.15265/IY-2014-0020] [Medline: 25123716]
2. Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. SAGE Open Med 2020;8:2050312120934839 [FREE Full text] [doi: 10.1177/2050312120934839] [Medline: 32637104]
3. Egger J, Gsaxner C, Pepe A, Pomykala KL, Jonske F, Kurz M, et al. Medical deep learning-a systematic meta-review. Comput Methods Programs Biomed 2022 Jun;221:106874 [FREE Full text] [doi: 10.1016/j.cmpb.2022.106874] [Medline: 35588660]
4. Baldi P. Deep learning in biomedical data science. Annu Rev Biomed Data Sci 2018 Jul 20;1(1):181-205. [doi: 10.1146/annurev-biodatasci-080917-013343]
5. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017 Jun 21;19:221-248 [FREE Full text] [doi: 10.1146/annurev-bioeng-071516-044442] [Medline: 28301734]
6. Price WN. Big data and black-box medical algorithms. Sci Transl Med 2018 Dec 12;10(471):eaao5333 [FREE Full text] [doi: 10.1126/scitranslmed.aao5333] [Medline: 30541791]
7. Poon AI, Sung JJ. Opening the black box of AI-medicine. J Gastroenterol Hepatol 2021 Mar;36(3):581-584. [doi: 10.1111/jgh.15384] [Medline: 33709609]
8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May 13;1(5):206-215 [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
9. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
10. Wanderer JP, Nelson SE, Ehrenfeld JM, Monahan S, Park S. Clinical data visualization: the current state and future needs. J Med Syst 2016 Dec;40(12):275. [doi: 10.1007/s10916-016-0643-x] [Medline: 27787779]
11. Badgeley MA, Shameer K, Glicksberg BS, Tomlinson MS, Levin MA, McCormick PJ, et al. EHDViz: clinical dashboard development using open-source technologies. BMJ Open 2016 Mar 24;6(3):e010579 [FREE Full text] [doi: 10.1136/bmjopen-2015-010579] [Medline: 27013597]
12. Torsvik T, Lillebo B, Mikkelsen G. Presentation of clinical laboratory results: an experimental comparison of four visualization techniques. J Am Med Inform Assoc 2013;20(2):325-331 [FREE Full text] [doi: 10.1136/amiajnl-2012-001147] [Medline: 23043123]
13. Schrodt J, Dudchenko A, Knaup-Gregori P, Ganzinger M. Graph-representation of patient data: a systematic literature review. J Med Syst 2020 Mar 12;44(4):86 [FREE Full text] [doi: 10.1007/s10916-020-1538-4] [Medline: 32166501]
14. GenAI apps, grounded in your data. Neo4j Graph Data Platform: The Leader in Graph Databases. URL: https://neo4j.com/ [accessed 2022-11-09]
15. Almabdy S. Comparative analysis of relational and graph databases for social networks. In: Proceedings of the 1st International Conference on Computer Applications & Information Security. 2018 Presented at: ICCAIS '18; April 4-6, 2018; Riyadh, Saudi Arabia p. 1-4 URL: https://ieeexplore.ieee.org/document/8441982 [doi: 10.1109/cais.2018.8441982]
16. Khan W, ahmed E, Shahzad W. Predictive performance comparison analysis of relational and NoSQL graph databases. Int J Adv Comput Sci Appl 2017;8(5). [doi: 10.14569/ijacsa.2017.080564]
17. NeoDash - dashboard builder for Neo4j. Neo4j. URL: https://neo4j.com/labs/neodash/ [accessed 2022-12-14]
18. GraphXR: visual analytics, graph BI, and more. Kineviz. URL: https://www.kineviz.com [accessed 2022-12-16]
19. Neo4j bloom. Neo4j Graph Data Platform. URL: https://neo4j.com/product/bloom/ [accessed 2022-12-16]
20. Home page. Graphviz. URL: https://graphviz.org/ [accessed 2022-12-16]
21. Rajabi E, Kafaie S. Knowledge graphs and explainable AI in healthcare. Information 2022 Sep 28;13(10):459. [doi: 10.3390/info13100459]

22. Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, et al. Reactome graph database: efficient access to complex pathway data. PLoS Comput Biol 2018 Jan;14(1):e1005968 [FREE Full text] [doi: 10.1371/journal.pcbi.1005968] [Medline: 29377902]
23. Gütebier L, Bleimehl T, Henkel R, Munro J, Müller S, Morgner A, et al. CovidGraph: a graph to fight COVID-19. Bioinformatics 2022 Oct 14;38(20):4843-4845 [FREE Full text] [doi: 10.1093/bioinformatics/btac592] [Medline: 36040169]
24. Dedié A, Bleimehl T, Täger J, Preusse M, de Angelis MH, Jarasch A. DZDconnect: mit vernetzten daten gegen diabetes. Diabetologe 2021 Sep 28;17(8):780-787. [doi: 10.1007/s11428-021-00807-y]
25. He X, Zhang R, Rizvi R, Vasilakes J, Yang X, Guo Y, et al. ALOHA: developing an interactive graph-based visualization for dietary supplement knowledge graph through user-centered design. BMC Med Inform Decis Mak 2019 Aug 08;19(Suppl 4):150 [FREE Full text] [doi: 10.1186/s12911-019-0857-1] [Medline: 31391091]
26. Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Min 2016 Jul 25;9(1):23 [FREE Full text] [doi: 10.1186/s13040-016-0102-8] [Medline: 27462371]
27. McLachlan S, Dube K, Hitman GA, Fenton NE, Kyrimi E. Bayesian networks in healthcare: distribution by medical condition. Artif Intell Med 2020 Jul;107:101912 [FREE Full text] [doi: 10.1016/j.artmed.2020.101912] [Medline: 32828451]
28. Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. Proc AMIA Symp 2000:106-110 [FREE Full text] [Medline: 11079854]
29. Kitson NK, Constantinou AC, Guo Z, Liu Y, Chobtham K. A survey of Bayesian network structure learning. Artif Intell Rev 2023 Jan 17;56(8):8721-8814. [doi: 10.1007/s10462-022-10351-w]
30. Smith VA, Yu J, Smulders TV, Hartemink AJ, Jarvis ED. Computational inference of neural information flow networks. PLoS Comput Biol 2006 Nov 24;2(11):e161 [FREE Full text] [doi: 10.1371/journal.pcbi.0020161] [Medline: 17121460]
31. Burnside ES. Bayesian networks: computer-assisted diagnosis support in radiology. Acad Radiol 2005 Apr;12(4):422-430. [doi: 10.1016/j.acra.2004.11.030] [Medline: 15831415]
32. Ke X, Keenan K, Smith VA. Treatment of missing data in Bayesian network structure learning: an application to linked biomedical and social survey data. BMC Med Res Methodol 2022 Dec 19;22(1):326 [FREE Full text] [doi: 10.1186/s12874-022-01781-9] [Medline: 36536286]
33. Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. Med Decis Making 1993 Jul 02;13(4):322-338. [doi: 10.1177/0272989X9301300409] [Medline: 8246705]
34. Beck JR, Pauker SG. The Markov process in medical prognosis. Med Decis Making 1983;3(4):419-458. [doi: 10.1177/0272989X8300300403] [Medline: 6668990]
35. Hogendoorn W, Moll FL, Sumpio BE, Hunink MG. Clinical decision analysis and Markov modeling for surgeons: an introductory overview. Ann Surg 2016 Aug;264(2):268-274. [doi: 10.1097/SLA.0000000000001569] [Medline: 26756750]
36. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. SIGMOD Rec 1993 Jun 01;22(2):207-216. [doi: 10.1145/170036.170072]
37. Stilou S, Bamidis PD, Maglaveras N, Pappas C. Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. Stud Health Technol Inform 2001;84(Pt 2):1399-1403. [Medline: 11604957]
38. Harahap M, Husein AM, Aisyah S, Lubis FR, Wijaya BA. Mining association rule based on the diseases population for recommendation of medicine need. J Phys Conf Ser 2018 Apr 30;1007:012017. [doi: 10.1088/1742-6596/1007/1/012017]
39. Johnston BC, Alonso-Coello P, Friedrich JO, Mustafa RA, Tikkinen KA, Neumann I, et al. Do clinicians understand the size of treatment effects? a randomized survey across 8 countries. CMAJ 2016 Jan 05;188(1):25-32 [FREE Full text] [doi: 10.1503/cmaj.150430] [Medline: 26504102]
40. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinb Dubl Phil Mag J 2009 Apr 21;50(302):157-175. [doi: 10.1080/14786440009463897]
41. Pearson K. Note on regression and inheritance in the case of two parents. Proc R Soc Lond 1997 Jan 31;58(347-352):240-242 [FREE Full text] [doi: 10.1098/rspl.1895.0041]
42. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021;8(1):140 [FREE Full text] [doi: 10.1186/s40537-021-00516-9] [Medline: 34722113]
43. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. Clin Trials 2004;1(4):368-376. [doi: 10.1191/1740774504cn032oa] [Medline: 16279275]
44. Padilla O, Abadie J. Normal laboratory values. MSD Manual Professional Version. URL: https://www.msdmanuals.com/professional/resources/normal-laboratory-values/normal-laboratory-values [accessed 2023-01-23]
45. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 1915 May;10(4):507. [doi: 10.2307/2331838]
46. Bellmann L, Wiederhold AJ, Trübe L, Twerenbold R, Ückert F, Gottfried K. Introducing attribute association graphs to facilitate medical data exploration: development and evaluation using epidemiological study data. Universität Hamburg. URL: https://www.fdr.uni-hamburg.de/record/13421 [accessed 2023-09-18]

47. Jagodzinski A, Johansen C, Koch-Gromus U, Aarabi G, Adam G, Anders S, et al. Rationale and design of the Hamburg city health study. Eur J Epidemiol 2020 Feb 08;35(2):169-181 [FREE Full text] [doi: 10.1007/s10654-019-00577-4] [Medline: 31705407]
48. HCHS study record. ClinicalTrials. URL: https://beta.clinicaltrials.gov/study/NCT03934957 [accessed 2023-09-18]
49. Home page. Systematized Nomenclature of Medicine (SNOMED). URL: https://www.snomed.org/ [accessed 2022-12-14]
50. Huff SM, Rocha RA, McDonald CJ, de Moor GJ, Fiers T, Bidgood WDJ, et al. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc 1998 May 01;5(3):276-292 [FREE Full text] [doi: 10.1136/jamia.1998.0050276] [Medline: 9609498]
51. Helminski D, Kurlander JE, Renji AD, Sussman JB, Pfeiffer PN, Conte ML, et al. Dashboards in health care settings: protocol for a scoping review. JMIR Res Protoc 2022 Mar 02;11(3):e34894 [FREE Full text] [doi: 10.2196/34894] [Medline: 35234650]
52. Wu DT, Vennemeyer S, Brown K, Revalee J, Murdock P, Salomone S, et al. Usability testing of an interactive dashboard for surgical quality improvement in a large congenital heart center. Appl Clin Inform 2019 Oct;10(5):859-869 [FREE Full text] [doi: 10.1055/s-0039-1698466] [Medline: 31724143]
53. Elm JJ, Daeschler M, Bataille L, Schneider R, Amara A, Espay AJ, et al. Feasibility and utility of a clinician dashboard from wearable and mobile application Parkinson's disease data. NPJ Digit Med 2019;2:95 [FREE Full text] [doi: 10.1038/s41746-019-0169-y] [Medline: 31583283]
54. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996:189-194.
55. Coe AM, Ueng W, Vargas JM, David R, Vanegas A, Infante K, et al. Usability testing of a web-based decision aid for breast cancer risk assessment among multi-ethnic women. AMIA Annu Symp Proc 2016;2016:411-420 [FREE Full text] [Medline: 28269836]
56. Hirschmann J, Sedlmayr B, Zierk J, Rauh M, Metzler M, Prokosch HU, et al. Evaluation of an interactive visualization tool for the interpretation of pediatric laboratory test results. Stud Health Technol Inform 2017;243:207-211. [Medline: 28883202]
57. Bellmann L. GraphXplore code repository. GitHub. URL: https://github.com/UKEIAM/graphxplore [accessed 2024-04-23]
58. Metra M, Cotter G, Gheorghiade M, Dei Cas L, Voors AA. The role of the kidney in heart failure. Eur Heart J 2012 Sep 10;33(17):2135-2142. [doi: 10.1093/eurheartj/ehs205] [Medline: 22888113]
59. Vindhyal MR, Khayyat S, Shaaban A, Duran BA, Kallail KJ. Decreased renal function is associated with heart failure readmissions. Cureus 2018 Aug 09;10(8):e3122 [FREE Full text] [doi: 10.7759/cureus.3122] [Medline: 30338197]
60. Wannamethee SG, Shaper AG, Perry IJ. Serum creatinine concentration and risk of cardiovascular disease: a possible marker for increased risk of stroke. Stroke 1997 Mar;28(3):557-563. [doi: 10.1161/01.str.28.3.557] [Medline: 9056611]
61. Panagopoulou V, Deftereos S, Kossyvakis C, Raisakis K, Giannopoulos G, Bouras G, et al. NTproBNP: an important biomarker in cardiac diseases. Curr Top Med Chem 2013;13(2):82-94. [doi: 10.2174/1568026611313020002] [Medline: 23470072]
62. Srisawasdi P, Vanavanan S, Charoenpanichkit C, Kroll MH. The effect of renal dysfunction on BNP, NT-proBNP, and their ratio. Am J Clin Pathol 2010 Jan 01;133(1):14-23. [doi: 10.1309/ajcp60htpgigfcnk]
63. Takase H, Dohi Y. Kidney function crucially affects B-type natriuretic peptide (BNP), N-terminal proBNP and their relationship. Eur J Clin Invest 2014;44(3):303-308. [doi: 10.1111/eci.12234] [Medline: 24372567]
64. Willis MS, Lee ES, Grenache DG. Effect of anemia on plasma concentrations of NT-proBNP. Clin Chim Acta 2005 Aug;358(1-2):175-181. [doi: 10.1016/j.cccn.2005.03.009] [Medline: 15878465]
65. Karakoyun I, Colak A, Arslan FD, Hasturk AG, Duman C. Anemia considerations when assessing natriuretic peptide levels in ED patients. Am J Emerg Med 2017 Nov;35(11):1677-1681. [doi: 10.1016/j.ajem.2017.05.048] [Medline: 28587950]
66. Hogenhuis J, Voors AA, Jaarsma T, Hoes AW, Hillege HL, Kragten JA, et al. Anaemia and renal dysfunction are independently associated with BNP and NT-proBNP levels in patients with heart failure. Eur J Heart Fail 2007 Aug;9(8):787-794 [FREE Full text] [doi: 10.1016/j.ejheart.2007.04.001] [Medline: 17532262]
67. Chonchol M, Nielson C. Hemoglobin levels and coronary artery disease. Am Heart J 2008 Mar;155(3):494-498. [doi: 10.1016/j.ahj.2007.10.031] [Medline: 18294483]
68. Lee G, Choi S, Kim K, Yun J, Son JS, Jeong S, et al. Association of hemoglobin concentration and its change with cardiovascular and all-cause mortality. J Am Heart Assoc 2018 Jan 29;7(3):e007723 [FREE Full text] [doi: 10.1161/JAHA.117.007723] [Medline: 29378732]
69. Verschuren WM, Jacobs DR, Bloemberg BP, Kromhout D, Menotti A, Aravanis C, et al. Serum total cholesterol and long-term coronary heart disease mortality in different cultures. Twenty-five-year follow-up of the seven countries study. JAMA 1995 Jul 12;274(2):131-136. [Medline: 7596000]
70. Anderson KM, Castelli WP, Levy D. Cholesterol and mortality. 30 years of follow-up from the Framingham study. JAMA 1987 Apr 24;257(16):2176-2180. [doi: 10.1001/jama.257.16.2176] [Medline: 3560398]
71. Houterman S, Verschuren WM, Hofman A, Witteman JC. Serum cholesterol is a risk factor for myocardial infarction in elderly men and women: the Rotterdam study. J Intern Med 1999 Jul;246(1):25-33 [FREE Full text] [doi: 10.1046/j.1365-2796.1999.00525.x] [Medline: 10447222]

72. Ference BA, Ginsberg HN, Graham I, Ray KK, Packard CJ, Bruckert E, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. Eur Heart J 2017 Aug 21;38(32):2459-2472 [FREE Full text] [doi: 10.1093/eurheartj/ehx144] [Medline: 28444290]
73. Abdullah SM, Defina LF, Leonard D, Barlow CE, Radford NB, Willis BL, et al. Long-term association of low-density lipoprotein cholesterol with cardiovascular mortality in individuals at low 10-year risk of atherosclerotic cardiovascular disease. Circulation 2018 Nov 20;138(21):2315-2325. [doi: 10.1161/circulationaha.118.034273]
74. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. High density lipoprotein as a protective factor against coronary heart disease. The Framingham study. Am J Med 1977 May;62(5):707-714. [doi: 10.1016/0002-9343(77)90874-9] [Medline: 193398]
75. Emerging Risk Factors Collaboration, Di Angelantonio E, Sarwar N, Perry P, Kaptoge S, Ray KK, et al. Major lipids, apolipoproteins, and risk of vascular disease. JAMA 2009 Nov 11;302(18):1993-2000 [FREE Full text] [doi: 10.1001/jama.2009.1619] [Medline: 19903920]
76. Gu Q, Paulose-Ram R, Burt VL, Kit BK. Prescription cholesterol-lowering medication use in adults aged 40 and over: United States, 2003-2012. NCHS Data Brief 2014 Dec(177):1-8. [Medline: 25536410]
77. Jafri H, Alsheikh-Ali AA, Karas RH. Meta-analysis: statin therapy does not alter the association between low levels of high-density lipoprotein cholesterol and increased cardiovascular risk. Ann Intern Med 2010 Dec 21;153(12):800-808. [doi: 10.7326/0003-4819-153-12-201012210-00006] [Medline: 21173414]
78. Ward NC, Watts GF, Eckel RH. Statin toxicity. Circ Res 2019 Jan 18;124(2):328-350. [doi: 10.1161/circresaha.118.312782]
79. SPSS software. IBM Corp. URL: https://www.ibm.com/de-de/spss [accessed 2023-06-12]
80. R Core Team. The R project for statistical computing. R Foundation. URL: https://www.r-project.org/ [accessed 2023-06-12]
81. Microsoft Excel spreadsheet software. Microsoft 365. URL: https://www.microsoft.com/en-us/microsoft-365/excel [accessed 2023-10-10]
82. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Int J Hum Comput Interact 2008 Jul 30;24(6):574-594. [doi: 10.1080/10447310802205776]
83. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Use Exp 2009;4(3):114-123 [FREE Full text]
84. Liu C, Wang F, Hu J, Xiong H. Temporal phenotyping from longitudinal electronic health records: a graph based framework. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD '15; August 10-13, 2015; Sydney, Australia p. 705-714 URL: https://dl.acm.org/doi/10.1145/2783258.2783352 [doi: 10.1145/2783258.2783352]
85. Zhang S, Liu L, Li H, Xiao Z, Cui L. MTPGraph: a data-driven approach to predict medical risk based on temporal profile graph. In: Proceedings of the 2016 IEEE International Conference on Trust, Security and Privacy in Computing and Communications. 2016 Presented at: TrustCom '16; August 23-26, 2016; Tianjin, China p. 1174-1181 URL: https://ieeexplore.ieee.org/document/7847074 [doi: 10.1109/trustcom.2016.0191]
86. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep 2017 Jul 20;7(1):5994 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
87. Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021 Oct 28;4(1):153 [FREE Full text] [doi: 10.1038/s41746-021-00521-5] [Medline: 34711924]
88. Park E, Chang H, Nam HS. A Bayesian network model for predicting post-stroke outcomes with available risk factors. Front Neurol 2018;9:699 [FREE Full text] [doi: 10.3389/fneur.2018.00699] [Medline: 30245663]
89. Kyrimi E, Dube K, Fenton N, Fahmi A, Neves MR, Marsh W, et al. Bayesian networks in healthcare: what is preventing their adoption? Artif Intell Med 2021 Jun;116:102079. [doi: 10.1016/j.artmed.2021.102079] [Medline: 34020755]
90. Ioannidis JP. What have we (not) learnt from millions of scientific papers with P values? Am Stat 2019 Mar 20;73(sup1):20-25. [doi: 10.1080/00031305.2018.1447512]

Abbreviations

AAG: attribute association graph
CRP: C-reactive protein
ECG: electrocardiography
HCHS: Hamburg City Health Study
HDL: high-density lipoprotein
LDL: low-density lipoprotein
proBNP: prohormone of B-type natriuretic peptide
SUS: System Usability Scale

Edited by C Lovis; submitted 12.06.23; peer-reviewed by A Scherag, L Loeb, M Bjelogrlic, C Gaudet-Blavignac; comments to author 28.08.23; revised version received 11.10.23; accepted 04.05.24; published 24.07.24.

Please cite as:
Bellmann L, Wiederhold AJ, Trübe L, Twerenbold R, Ückert F, Gottfried K
Introducing Attribute Association Graphs to Facilitate Medical Data Exploration: Development and Evaluation Using Epidemiological Study Data
JMIR Med Inform 2024;12:e49865
URL: https://medinform.jmir.org/2024/1/e49865
doi:10.2196/49865
PMID:

Original Paper

Evaluating and Enhancing the Fitness-for-Purpose of Electronic Health Record Data: Qualitative Study on Current Practices and Pathway to an Automated Approach Within the Medical Informatics for Research and Care in University Medicine Consortium

Gaetan Kamdje Wabo[1], MSc; Preetha Moorthy[1], MSc; Fabian Siegel[1,2], MD; Susanne A Seuchter[3], BSc; Thomas Ganslandt[3,4], MD

[1]Center for Preventive Medicine and Digital Health Baden-Wuerttemberg, Department of Biomedical Informatics, Medical Faculty of Mannheim, University of Heidelberg, Mannheim, Germany

[2]Department of Urology and Urosurgery, University Medical Center of Mannheim, Medical Faculty of Mannheim, University of Heidelberg, Mannheim, Germany

[3]Medical Center for Information and Communication Technology, Erlangen University Hospital, Erlangen, Germany

[4]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Gaetan Kamdje Wabo, MSc
Center for Preventive Medicine and Digital Health Baden-Wuerttemberg
Department of Biomedical Informatics
Medical Faculty of Mannheim, University of Heidelberg
Building 3, Level 4
Theodor-Kutzer-Ufer 1-3
Mannheim, 68167
Germany
Phone: 49 621 383 8088
Email: gaetankamdje.wabo@medma.uni-heidelberg.de

Abstract

Background: Leveraging electronic health record (EHR) data for clinical or research purposes heavily depends on data fitness. However, there is a lack of standardized frameworks to evaluate EHR data suitability, leading to inconsistent quality in data use projects (DUPs). This research focuses on the Medical Informatics for Research and Care in University Medicine (MIRACUM) Data Integration Centers (DICs) and examines empirical practices on assessing and automating the fitness-for-purpose of clinical data in German DIC settings.

Objective: The study aims (1) to capture and discuss how MIRACUM DICs evaluate and enhance the fitness-for-purpose of observational health care data and examine the alignment with existing recommendations and (2) to identify the requirements for designing and implementing a computer-assisted solution to evaluate EHR data fitness within MIRACUM DICs.

Methods: A qualitative approach was followed using an open-ended survey across DICs of 10 German university hospitals affiliated with MIRACUM. Data were analyzed using thematic analysis following an inductive qualitative method.

Results: All 10 MIRACUM DICs participated, with 17 participants revealing various approaches to assessing data fitness, including the 4-eyes principle and data consistency checks such as cross-system data value comparison. Common practices included a DUP-related feedback loop on data fitness and using self-designed dashboards for monitoring. Most experts had a computer science background and a master's degree, suggesting strong technological proficiency but potentially lacking clinical or statistical expertise. Nine key requirements for a computer-assisted solution were identified, including flexibility, understandability, extendibility, and practicability. Participants used heterogeneous data repositories for evaluating data quality criteria and practical strategies to communicate with research and clinical teams.

Conclusions: The study identifies gaps between current practices in MIRACUM DICs and existing recommendations, offering insights into the complexities of assessing and reporting clinical data fitness. Additionally, a tripartite modular framework for fitness-for-purpose assessment was introduced to streamline the forthcoming implementation. It provides valuable input for developing and integrating an automated solution across multiple locations. This may include statistical comparisons to advanced

machine learning algorithms for operationalizing frameworks such as the 3×3 data quality assessment framework. These findings provide foundational evidence for future design and implementation studies to enhance data quality assessments for specific DUPs in observational health care settings.

(JMIR Med Inform 2024;12:e57153) doi:10.2196/57153

KEYWORDS

data quality; fitness-for-purpose; secondary use; thematic analysis; EHR data; electronic health record; data integration center; Medical Informatics Initiative; MIRACUM consortium; Medical Informatics for Research and Care in University Medicine; data science; integration; data use; visualization; visualizations; record; records; EHR; EHRs; survey; surveys; medical informatics

Introduction

Insight Into Medical Informatics in Research and Care in University Medicine Data Integration Centers and Data Use Projects

The German Medical Informatics Initiative (MI-I) [1] was launched by the German Federal Ministry of Education and Research to enhance digital health and clinical research infrastructure advancements in Germany. This initiative comprises multiple large consortia. The Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium [2] is among the MI-I consortia that focus on integrating clinical and research data to enhance patient care and facilitate data-driven medical research at German university hospitals. However, MIRACUM includes 10 university hospitals and further medical research institutions across Germany, all of which instantiate medical Data Integration Centers (DICs). The DICs are crucial in gathering, harmonizing, and integrating clinical data from various source systems, including electronic health records (EHRs), clinical imaging systems, and other health-related databases. Additionally, the DICs' efficient data pipelines support uniform and secure data storage, enabling significant privacy-preserved sharing and analysis of patient data.

Among others, a cornerstone of the MIRACUM consortium's mission is also to foster data-driven medical research and the improvement of clinical patient care through the implementation of data use projects (DUPs). The DUPs use integrated and harmonized clinical data to answer pertinent research questions, such as identifying patterns, creating predictive models, or supporting evidence-based decision-making in the clinical field. Some of the key investigation areas of focus in the MIRACUM DUPs' applications so far include personalized medicine [3-5], clinical decision support [6-10], disease monitoring and surveillance [11], drug safety and pharmacovigilance [12], population health management [11], and translational research [13]. Ensuring an appropriate level of data quality (DQ) is imperative for the successful execution of DUPs that use MIRACUM DICs' clinical data, particularly during the development and implementation of data extraction-transformation-loading [14] processes.

Evaluating the DQ Impact on Secondary Use: Emphasis on Fitness-for-Purpose

Despite the establishment of robust data integration pipelines within each of the MIRACUM DICs [2,15], improving the fitness-for-purpose of generated data requires overcoming the challenges related to DQ for specific DUP purposes. As a crucial determinant for generating credible evidence [16], high-quality data are essential for drawing sound conclusions. This ensures that a larger target group, for instance, clinicians, health care providers, or givers, may rely on research findings. Conversely, compromised DQ [17] can lead to erroneous results, harmful treatment decisions, or a loss of public trust in the scientific community. However, several studies [18-22] have presented and discussed frameworks and methodologies about how to assess and ensure the quality of EHR data. It remains crucial to gauge the fitness of clinical data to enable achieving designated medical DUP objectives. In the context of observational DIC data-driven studies, for instance, ensuring DQ is of high importance to ensure the validity of the study results, which can emerge as a paramount challenge to overcome. Weiskopf and Weng [20] and Kahn et al [22] emphasized the relevance of adhering to the recommended references, such as the 3×3 data quality assessment (DQA) framework [21], when using EHR data to pursue specific research inquiries. Nevertheless, these current approaches might not comprehensively address the subjective understanding of data fitness that may arise from the diverse backgrounds of secondary data users.

The definition of fitness-for-purpose of clinical data might emphasize greater complexity depending on the requirements of the intended data use or research question to be investigated. In certain DUP contexts, clinical data may be considered as "fit for purpose," when all eligible patients simply present complete information about specific treatments. In other DUP scenarios, a plausible correlation between these treatments and certain specific diagnostic indicators, laboratory outcomes, or caregiver-related metadata may be additionally required. Girman et al [23] proposed a definition for EHR data fitness-for-purpose, which can be summarized in two dimensions: (1) relevance and (2) reliability. The first dimension, relevance, ensures that the target data elements are available, and a sufficient number of representative patients are present for the study. The second dimension, reliability, verifies whether the data to be used are sufficiently accurate, complete, and traceable (provenance).

However, the interpretation of what constitutes "fitness," and how to assess it, can substantially vary across different professional domains. For example, a computer scientist might prioritize the algorithmic reliability of data, whereas health care providers might pay attention to the clinical relevance of the data to patient care. These differing perspectives can profoundly affect the ways of assessing data fitness, highlighting the need for more empirical investigation in this domain, especially in multidisciplinary settings such as those of the MIRACUM DICs.

Prior Works, Contemporary Overview of DQ, and Existing Research Gaps

Against the backdrop described in the previous section, a dynamic DQA tool was designed and implemented [14,24], during the earlier stages of the MIRACUM project, based on the DQA framework suggested by Kahn et al [22]. This tool allows for capturing and assessing DQA metrics, including data completeness, conformance, and plausibility checks, between data sources and target systems to validate DICs' extraction-transformation-loading processes. Additionally, it facilitates cross-location comparisons of DQ distributions. Furthermore, conducting a DQA study that included comorbidity analysis [25] enriched our initial understanding of strategies to assess the DQ for specific research purposes. However, these significant efforts have yet to comprehensively provide a workable solution for automated assistance for assessing the fitness of DIC data to complete the ongoing or upcoming DUPs.

To bridge the divide between current practices and required improvements, it is imperative to initially capture and understand the activities applied by the MIRACUM DICs to assess and report about the data fitness-for-purpose throughout the data delivery process. Therefore, Reynolds et al [26] introduced objective considerations for evaluating the data fitness-for-purpose, but the authors did not delve deeper into how the DQ checks could be performed based on the research question criteria set by the researchers in an automated way. In contrast, Cho et al [27] developed a fitness-for-purpose tool that assesses data completeness, which may be more practical than intrinsic DQA tools. The predictive data completeness, as proposed by Weiskopf et al [21], which aims to assess, for instance, the impact of missing documentation of data elements such as chronic diseases on the prediction of clinical events, was scarcely addressed. Furthermore, Raman et al [28] suggested conducting study-specific fitness-for-purpose assessment in the early stage of trials and sharing the results in a transparent manner. Nevertheless, knowing whether and how these assessments correlate with successful trials could be valuable. Despite these insights, there is a considerable benefit in exploring common practices and prerequisites for creating a universally deployable solution, spanning multiple sites and systems, and automating the assessment of data fitness-for-purpose.

This research addresses several benefits. On the one hand, it is the first qualitative empirical investigation, gathering and analyzing the experiences in the assessment of data fitness-for-purpose in German medical DICs. This offers the opportunity to gain valuable information about possible existing solutions, challenges, and observable deficits, in comparison with internationally established evidence [21,27,28]. On the other hand, this allows for the investigation of relevant practice-oriented requirements for the development of an evidence-based tool that would enable an automated holistic assessment for DIC data fitness-for-purpose.

Research Aims

In light of the aforementioned research landscape, this investigation seeks to identify and describe (1) to what extent do the MIRACUM DICs address the fitness-for-purpose of observational health care data and how do the applied approaches contrast with existing recommendations and (2) what key requirements are necessary to develop an automated system within the MIRACUM DICs for assessing the fitness-for-purpose of EHR data for research or clinical applications.

Methods

Approach Overview

In this section, we detail the methods used to address the research questions described previously. This encompasses outlining the ethical considerations, the study design, the study sample, and the procedures used to gather and qualitatively analyze the data.

Ethical Considerations

The study did not require formal ethics approval from the Ethics Committee II of the Medical Faculty Mannheim at the University of Heidelberg, as it involved anonymized data collection that complies with the requirements of the professional code for physicians and the General Data Protection Regulation. Informed consent was obtained from all participants for the participation in the survey on assessing the quality of observational data for secondary use. All data were deidentified to ensure privacy and confidentiality, with no personal information collected or published, and participants received no compensation.

Study Design

In the course of this investigation, we performed a qualitative study. Therefore, we meticulously implemented a survey across all 10 medical DICs participating in the MIRACUM consortium. Adhering to the directives outlined in the GESIS (Society of Social Science Infrastructure Institutions) survey guideline version 2.0 [29], we developed a survey instrument consisting of 6 open-ended questions (refer to Textbox 1). Such a format facilitates respondents' ability to articulate their perspectives unrestrictedly, thereby fostering the acquisition of valuable insights and preventing the proliferation of superfluous response alternatives [29]. The instrument was formatted into a questionnaire that was distributed to all participating DIC locations. An accompanying guide was additionally provided during the survey dissemination, which outlined the expected time for the survey completion, the preference for keyword-oriented responses, the intended deadline for response submission, and background information regarding the survey (Multimedia Appendix 1).

Textbox 1. Overview of the survey questionnaire.

Metadata related to the location and survey

- Location name
- Data Integration Center (DIC) data quality (DQ) officer
- Deadline
- Survey feedback completion date

Metadata related to the survey respondents

- Gender
- Educational background
- Highest degree
- Years of experience with data quality assessment (DQA)
- Years of experience with observational data fitness-for-purpose assessment

Survey questions

- Question 1: On average, how many data requests or local data use projects are handled at your site DIC per quarter?
- Question 2: Which contents are mostly in focus in your local data use projects (eg, care evaluation in transfusion medicine), and which data repositories are most frequently queried in this context (Informatics for Integrating Biology and Bedside, Observational Medical Outcomes Partnership, Fast Healthcare Interoperability Resources, etc)?
- Question 3: How are data use project–specific DQ requirements collected from the perspective of data requesters at their DIC?
- Question 4: In addition to the current Medical Informatics for Research and Care in University Medicine DQA tool, what tools or technical approaches do you use for data use project–specific DQA?
- Question 5: What measures are taken at your location to communicate with data requesting sites about the quality of provided data for the intended purpose, so that data requesters have opportunities to estimate the fitness of the data to complete the intended project?
- Question 6: What would be their expectations or requirements for a fitness-for-purpose cross-site DQ framework that you could adopt in the future to measure DQ related to their data use projects?

Sample

All medical DICs from the 10 university hospitals affiliated with the MIRACUM consortium were invited to participate in the survey. For each DIC (MIRACUM site), the responsibility of completing the survey was delegated to potential participants. The participants were specialized professionals with expertise in DQ, responsible for evaluating and enhancing the quality of observational data for secondary use within their respective DIC in the context of intern- or cross-location DUPs. While each DIC submitted a single survey response detailing their practice related to assessing data's fitness-for-purpose, they had the flexibility to involve 1 or multiple participants based on their availability during the data collection period.

Data Collection

We conducted the data collection from April 15 to June 15, 2022, using a survey questionnaire comprised of 6 open-ended questions. To streamline the survey process for all participating sites and to maintain clear documentation of the participant's responses, we used Atlassian Confluence (version 7.13.11) [30] as our documentation software.

The data collection process was initiated by extending formal invitations to the DICs through email. These invitations included a link directing to a confluence main page, which outlined the objectives of the survey and provided instructions on how to respond to the survey questions. Additionally, each participating location had access to a distinct embedded confluence-landing subpage, structured with 2 information entry areas. The first input area enabled the entry of meta-information, collecting demographic data on each participant involved in the survey. These included gender, educational background, highest degree obtained, and years of experience in DQA and fitness-for-purpose evaluation. The second area was devoted to gathering specific responses to the 6 open-ended survey questions. Upon completing the meta-information provision and responding to the survey questions, the involved participants completed the survey by submitting their site response and documenting the date of completion.

The survey specifically inquired about the quarterly frequency and current objectives driving observational DUPs within the DICs. Our inquiry was guided by the 3×3 EHR DQA guideline by Weiskopf et al [21], which illustrates the project-specific nature of data fitness evaluations for particular uses. This framework served as a preliminary theoretical basis for our study, allowing us to explore how DICs manage and align data requester expectations concerning observational DQ throughout the data lifecycle—from the initial request to the final delivery. In addition to examining the procedural aspects of DUP-related DQ management, our survey aimed to uncover the underlying mechanisms through which DICs implement and communicate the fitness-for-purpose of data within research and clinical teams. Additionally, we sought to identify the standards required

by each DIC for a scalable, automated solution to assess and report on data fitness-for-purpose in the MIRACUM DUP context.

Data Analysis

We analyzed the data using the thematic analysis (TA) method suggested by Braun and Clarke [31,32]. This approach was strategically chosen to identify and rectify any potential errors in the coding process. This methodology aligns with best practices recommended for qualitative research and is supported by precedents established in similar studies [31-33]. The TA framework offers practical and meaningful steps to deeply understand the common thoughts, experiences, or behaviors [34] among the specific cohort of participants.

Following the TA framework of Braun and Clarke [32], our analysis progressed through 6 structured stages, including data familiarization, initial codes generation, theme identification, theme review, theme definition and naming, and report writing [32,34]. First, we familiarized ourselves with the material through multiple, exhaustive readings of each location's feedback and took targeted notes. This facilitated the jotting down of early impressions. Second, we generated initial codes by manually classifying each relevant data segment. In this context, we proposed an initial coding concept (Table 1), which we created based on the vivo coding [34] (verbatim coding) method. This consisted of codes derived from the data by mainly using the language and terminology used by the study participants. This approach helped encapsulate codes reflecting the perspectives and actions expressed by the study participants. Throughout this process, there were numerous iterative discussions within the research team. Third, we used a dynamic approach to combine, compare, and analyze each of the generated code. This interpretive process enabled to inductively derive appropriate themes that have a concise and meaningful connection to the survey data. However, we ensured the themes also accurately reflected the entire data set. In the fourth step, we thoroughly reviewed these themes, ensuring that each theme has sufficient commonality, coherence, and distinctiveness to each other. In the fifth step, we assigned descriptive and accurate titles to each theme, enabling a comprehensive illustration of the key information of participants' responses. Finally, in the sixth step, we created this manuscript as part of a TA-guided qualitative data analysis process. The condensed analytical process documentation is included in Multimedia Appendix 2. The main objective of the initial coding was to ascertain the reliability and accuracy of the framework prior to its application to a larger data set, thereby enhancing the overall integrity and validity of the research findings. Furthermore, the research team collaboratively discussed and refined the initial coding to avoid overlapping with other codes. This supported the refinement of the initial coding to the final code (see Results section). This increased both the relevance and the representativeness of the inductively generated themes and codes in light of the large information corpus provided by the participating DIC locations. The final codes are presented in the *Results* section of this paper.

We migrated the aggregated data from the confluence platform into a comprehensive Microsoft Word document for in-depth analysis (Multimedia Appendices 2 and 3). The collected data from the first survey question were analyzed using the open-source software RStudio (Posit) [35]. The scripts used for this analysis are presented in Multimedia Appendix 4.

Table 1. Initial coding.

Selected preliminary codes	Illustrating examples of key terms
Clinical research purposes as primary emphasis of data use projects	Clinical research, clinical trials, and observational studies
Applying the 4-eyes principle	Mutual control and the involvement of an independent person
Using overview dashboard	Data portal, integration portal, and information dash
Check for data plausibility	Plausibility checks, data verification, and data items comparisons
System comparisons	Data quality comparison between data sources and target systems
Data consistency checks	In-depth examination of data completeness, data conformity, and data correctness (including plausibility) with regard to an intended data use
Data provenance collection	Collection and documentation about where the data came from

Results

Overview

This study involved all 10 MIRACUM DICs including 17 participants. The greatest proportion of the participants were male (11/17, 65%), had backgrounds in computer sciences (8/17, 47%), and held a master's degree (6/17, 35%). On average, the participants had 3.7 (SD 5.3) years of experience in assessing DQ and 1.3 (SD 1.5) years in evaluating data fitness-for-purpose. The study also revealed that most MIRACUM-affiliated DICs conduct 2 to 5 DUPs quarterly, with a single location handling up to 20 DUPs in the same timeframe. Additionally, the analysis identified 27 codes grouped into 6 themes.

Scope of Participants

Of the 10 MIRACUM DIC sites solicited for participation, all accepted to engage in the study. In total, 17 participants agreed to respond to the survey questionnaire. Table 2 presents the demographic characteristics of the participants and shows the distribution of their years of experience in assessing both general DQ and data fitness-for-purpose.

Table 3 shows an overview of the DUP frequencies across the MIRACUM DICs. The MIRACUM-affiliated sites have

undertaken an average of 5.8 (SD 6.5) DUPs on a quarterly basis, with a few sites executing up to 15 or 20 DUPs within the same timeframe. One DIC did not conduct any DUP.

Table 2. Overview of respondent's metadata (N=17).

Features	Values
Sex, n (%)	
Male	11 (65)
Female	4 (23)
Others	0 (0)
Missing	2 (12)
Educational background, n (%)	
Informatics related	8 (47)
Statistics related	1 (6)
Health related	2 (12)
Others	1 (6)
Missing	5 (29)
Highest degree, n (%)	
Doctoral grade	1 (6)
Master	6 (35)
German "Diplom"	3 (18)
Bachelor	1 (6)
Vocational training	1 (6)
Missing	5 (29)
Years of experience with data quality assessment	
Mean (SD)	3.7 (5.3)
Range	0-20
Years of experience with observational data fitness-for-purpose assessment	
Mean (SD)	1.3 (1.5)
Range	0-4

Table 3. Quarterly distribution of data use project frequencies across the Medical Informatics for Research and Care in University Medicine (MIRACUM) Data Integration Centers as of May 2022.

DUP[a] frequencies	Values (n=58)
MIRCAUM sites, n (%)	
MIRACUM site 1	20 (34)
MIRACUM site 2	3 (5)
MIRACUM site 3	15 (26)
MIRACUM site 4	4 (7)
MIRACUM site 5	2 (4)
MIRACUM site 6	3 (5)
MIRACUM site 7	4 (7)
MIRACUM site 8	0 (0)
MIRACUM site 9	6 (10)
MIRACUM site 10	1 (2)
Summary statistics	
Total of DUP frequencies, n (%)	58 (100)
Mean (SD)	5.8 (6.5)
Frequency range	0-20

[a]DUP: data use project.

Thematic Representation of Data

Overview

This subsection presents a thematic representation of findings derived from responses in the MIRACUM DICs. Through a comprehensive coding process, we identified 27 distinct codes, which were subsequently organized into 6 key themes. These themes approach various relevant aspects of assessing the fitness for use of EHR data within the context of DUPs. Below is an overview of the identified themes: objectives of DUPs in MIRACUM DICs, use of heterogeneous types of data repositories, strategies for gathering DUP-specific DQ criteria, methods for evaluating the data fitness-for-purpose, existing implementations and reporting mechanisms for data fitness-for-purpose, and requirements for a scalable data fitness-for-purpose assessment solution. Results are presented below, summarized and illustrated with quotes from site feedback, aligned with the COREQ (Consolidated Criteria for Reporting Qualitative Research).

Objectives of DUPs in MIRACUM DICs

Most MIRACUM sites (8/10, 80%) reported that DUPs primarily engaged in clinical research, with a focus on the analysis of clinical events, the assessment of health care quality, and the development of prediction models for clinical associations. The participating sites provided detailed accounts of their research activities.

The MIRACUM site 4 underscored, for instance, the relevance of qualifying research questions and quality assessment, including “Qualifying research questions (doctoral dissertations etc.), quality assessment, proof of qualification, where so far mostly the mirror system of ORBIS serves as data repository.”

Another participating site illustrated the type of clinical questions they address, by focusing on “Clinical research questions e.g. number and context data on splenectomies, context data on urological sepsis.” This indicates an in-depth examination of specific medical procedures and conditions, as observed at MIRACUM site 4.

A further narrative provided a broader perspective on the research activities, describing a research focus on

> *Department- and unit-specific clinical questions e.g., prediction of departmental sepsis and associations with specific treatment procedures/ICD diagnoses. Other example: patient case-based analysis of multiple clinical complications associated with specific clinical and demographic characteristics.*

This comprehensive approach at MIRACUM site 9 exemplifies the depth and complexity of the clinical research being conducted, which would aim to link various clinical and demographic factors with health outcomes. These narratives collectively describe the diverse and detailed nature of the clinical research efforts within DUPs, thereby demonstrating an important commitment to improving health care quality and developing predictive models based on extensive clinical data.

Use of Heterogeneous Types of Data Repositories

Overview

In the analysis of clinical data repositories used for DUP execution, 5 prominent repositories were identified from site feedback, revealing a diverse and dynamic landscape of data management systems. These systems range from broadly used ones such as the Clinical Data Warehouse (DWH) to specialized frameworks designed to meet specific research needs or privacy considerations. This variety reflects the differing technological

preferences across sites and underscores the complexity of DQ and data management in clinical research settings. The key repositories are as follows:

Use of Clinical DWH

The Clinical DWH emerges as the most commonly used repository for executing DUPs, as evidenced by its application in DUP-related DQA and reporting activities. In total, 3 (30%) of the 10 surveyed sites indicated that they primarily rely on the DWH for their internal research queries, quality assurance, and reporting needs. The respondents at MIRACUM site 2 and site 9 indicated that while the DWH is the primary repository for general queries, other specialized repositories, such as Informatics for Integrating Biology and Bedside (i2b2), Observational Medical Outcomes Partnership (OMOP), and Fast Healthcare Interoperability Resources (FHIR), are reserved for specific project needs, such as MI-I or MIRACUM requests. The following statements can show this:

> *Internal research queries, quality ensuring and reporting are mainly performed using the DWH. The i2b2/OMOP/FHIR repositories are mainly used for MI-I/MIRACUM specific requests.* [Survey question 2, MIRACUM site 2]

> *Most queries through the cDWH and i2b2 repo.* [Survey question 2, MIRACUM site 9]

Data Representation Model (i2b2, OMOP, and FHIR)

These models are notably prevalent, being used in 7 (70%) of the 10 sites. They are particularly suited to internal and cross-location DUP projects that require specific data handling or analysis frameworks. This pervasive use is corroborated by feedback from MIRACUM site 2, where i2b2 and OMOP are used for internal projects: "The i2b2/OMOP/FHIR repositories are mainly used for MI-I/MIRACUM specific requests." This indicates the existence of a relevant and versatile framework capable of supporting a multitude of research needs.

ORBIS System Use

At 1 (10%) surveyed site, the ORBIS system, particularly its mirror component, is exclusively used as the primary data repository. The system supports a range of academic and quality assurance activities, such as doctoral dissertations and qualification proofs, thereby emphasizing its specialized application in academic and clinical research environments. This is presented in the following statement: "Qualifying research questions (doctoral dissertations etc.), quality assessment, proof of qualification, where so far mostly the mirror system of ORBIS serves as data repository" (Survey question 2, MIRACUM site 3).

OPAL DataSHIELD Framework

In total, 2 (20%) of the 10 sites surveyed indicated a preference for the OPAL DataSHIELD framework for data storage and analysis. This preference may indicate a strategic choice for environments where data privacy and security are of paramount importance. This is evidenced by the use of DataSHIELD for analysis without direct querying of data repositories, as observed at MIRACUM site 10: "Analyzing is performed via DataSHIELD, therefore no direct query in data repositories" (Indirect i2b2).

CentraXX Repository

The local research repository CentraXX is referenced by 1 (10%) site for the purpose of storing frequently requested data items. The respondents from MIRACUM site 6 reported that "The most frequently requested data items are stored in the local research repository CentraXX" (Survey question 2). This repository's use serves to illustrate its role in facilitating rapid access to high-demand data, which could be of paramount importance for the efficient conduct of DUPs at the local level.

Strategies for Gathering DUP-Specific DQ Criteria

Overview

The analysis of data collection methods used by the DICs for the DUPs reveals a multifaceted approach aimed at enhancing the fitness of data through rigorous validation processes and strategic requester-provider interactions. In this subsection, we present the reported practices to ensure the alignment and quality of data for DUPs.

Detailed Request Submission

A noteworthy observation is that a considerable proportion of DICs (4/10, 40%) emphasize the importance of detailed and precise data requests. This approach serves as a preemptive measure to ensure that the data delivered are aligned with the needs of the requester. In the event of discrepancies between the requested and provided data, the aforementioned sites use a postprocessing step to rectify any misalignments. This process is captured by a statement from one of the survey respondents:

> *During the data request, we advise that the requested data should be described as fine-grained and exact as possible. If the data provided does not match the request, a "post-processing" process will be initiated.* [Survey question 3, MIRACUM site 1]

Participatory Discussion for Data Validation

Half of the surveyed sites (5/10, 50%) use participatory discussions between data providers and requesters as a core strategy for data validation. These discussions aim to identify and mitigate any factors that might reduce the quality of the data, such as issues with free text information or the specific documentation practices of the data-providing institution. The respondents noted that:

> *The requested data are usually discussed at least once with the requester and quality-reducing aspects are worked out together, e.g., free text information, documentation practice in the respective data-providing institution (usually the requester comes from the same institution and knows it very well).* [Survey question 3, MIRACUM site 3]

Quality Requirements Gathering

A collection of DQ requirements is conducted at a minority of the sites (1/10, 10%), with the use of a feasibility or data request form. The form is completed by both data requesters and internal data request administrators in order to gather specific DQ

requirements. The aforementioned process is described as follows:

> *Project-related data quality requirements are gathered using a Feasibility Request (FR) form completed by the data requester & internal data request administrator.* [Survey question 3, MIRACUM site 9]

However, 2 (20%) of the 10 sites did not apply any approach to gather the DUP-related DQ requirements.

Methods for Evaluating the Data Fitness-for-Purpose

Overview

In terms of applied methods to assess whether the DIC data suite is for the carrying out of various DUPs, 3 approaches were revealed.

Four-Eyes Principle

This method, which was observed at 3 (30%) of the 10 sites, places an emphasis on ensuring that the data are fit for the intended purposes through mutual control and content validation. The fundamental principle is straightforward: before any data are provided or issued. It is subjected to scrutiny by at least 2 individuals, thereby enhancing both accuracy and reliability.

The MIRACUM site 3 highlighted the development of a metadata repository–supported DQA tool, indicating that this approach is being further systematized: "Mutual control before issue/provision (4-eyes principle), an MDR-supported DQA tool is under development."

Similarly, MIRACUM site 9 provided further insight into the practical application of this principle, describing a 2-tier validation process as follows:

> *4-eyes principle: Content validation of the queries by a second data scientist (possibly also with a separate query), so that it is ensured that the query actually does what it is supposed to do. Content-related plausibility control of the results from the query through medical colleagues.*

Comparison of Data Values Distribution

This approach, used by 2 (20%) of the 10 sites, validates data by contrasting the distribution of data values from different systems. The MIRACUM site 2 provided an illustrative example of this method by comparing the hit ratios generated by independent systems: "For project-specific validation, comparison of hit ratio from different systems created by an independent person." This process not only identifies discrepancies or anomalies but also serves to reinforce the integrity of the data through independent verification.

Data Consistency Checks

The consistency checks are used by 2 (20%) of the 10 sites to verify the appropriateness of data formats, types, and variable associations. The MIRACUM site 10 exemplifies this process by using DataSHIELD to verify data formats and the number of variables prior to initiating analyses: "Verification of the data format or type, the number of variables via DataSHIELD before the analyses." This step could be of paramount importance in ensuring that the data meet the requisite standards for subsequent analytical procedures. It might serve as a foundational check that prevents the propagation of errors in data handling and analysis.

Existing Implementations and Reporting Mechanisms for Data Fitness-for-Purpose

Overview

The following measures were identified to technically implement and report fitness-for-purpose of clinical data.

Data Requester Feedback and Adjustment Process

At a majority of the study sites (6/10, 60%), the implementation of a feedback mechanism that involves data requesters is of high importance. Here, data requester feedback is systematically gathered and analyzed. In light of this input, the queries used to select data undergo adjustments and validations in accordance with the 4-eyes principle. This process is illustrated by a representative from MIRACUM site 9:

> *Then the data request administrator, who goes through the data to be delivered together with the data requester, delivers the data. In case of change requests/incorrect quality in the data, the data selection queries are adjusted and validated again via the 4-eyes principle, and documented.*

Feedback Loop and Quality Control

In 2 (20%) of the 10 surveyed sites, a continuous feedback loop is established among the data requester, data request administrator, internal data scientists, and the data transfer office. This iterative process is highly relevant for the refinement of DQ prior to final delivery. As described by MIRACUM site 9, the cycle involves a series of checks and rechecks:

> *This results in the feedback cycle: data requester => data request administrator => internal data scientists => data request administrator => data requester. Only in case of a complete match (from the data requester's perspective) the final data delivery takes place.*

Reporting Quality Using Dashboard

The use of technological tools to report on DQ was observed in 2 (20%) sites. Specifically, these sites use an overview dashboard within a self-developed data integration portal. This tool provides a streamlined and transparent view of DQ metrics, which can be essential for ongoing assessment and improvement. A representative from MIRACUM site 7 described the utility of the dashboard as follows: "Overview dashboard in the self-developed data integration portal," indicating a technology-driven approach to quality reporting.

Requirements for a Scalable Solution for the Data Fitness-for-Purpose Assessment

Table 4 summarizes the requirements gathered during the survey including some exemplifying quotes. The MIRACUM DICs emphasized diverse key attributes for the development of a data fitness-for-purpose assessment tool. These include flexibility, understandability, practicability, and extendibility in terms of usability. From a technical perspective, there was a preference

for dashboard implementation, system comparison features, data consistency checks, and ensuring uniformity in the FHIR profiles. Additionally, the consideration of data provenance was mentioned as an important feature in assessing fitness-for-purpose.

Table 4. Summary of requirements for implementing a data fitness-for-purpose assessment tool.

Requirement dimension and key attributes	DICs[a] (n=10), n (%)	Illustrating quotes
Usability		
Flexibility	1 (10)	"Flexible organization of the DQ system" [Survey question 6, MIRACUM[b] site 1].
Understandability	2 (20)	"Understandability for the clinician and data scientist/statistician Fitness-for-use dashboard" [Survey question 6, MIRACUM site 2].
Practicability	1 (10)	"Generally enough that it can be used in every DIZ and for every request. It should be pragmatic and easy to understand, so that it can always be used as a basic tool and its benefits are seen equally by all parties (data provider, data supplier, data requester)" [Survey question 6, MIRACUM site 3].
Extendibility	1 (10)	"In the short term, it is limited to the essentials to be able to use it and gain experience. In the long term, it may even be possible to modularize it and thus use it only in parts" [Survey question 6, MIRACUM site 3].
Technical and functionalities		
Dashboard implementation	3 (30)	"Implementation of a dashboard" [Survey question 6, MIRACUM site 1].
System comparison	3 (30)	"Complete non-interactive integration of the DQ process as an operation within the data pipelines for complete monitoring of the mapping of source and target systems with automatic machine-readable report generation (no PDF)" [Survey question 6, MIRACUM site 7].
Data consistency checks	3 (30)	"Mapping and automation of DQ checks based on the specific data quality metrics • Data completeness: are there enough patients at the DIZ site to carry out the planned projects • Data plausibility: formulation & automation of general-transferable plausibility checks (e.g., no readmission after a death, ...) that could affect the outcomes of most DRs • Data conformity: uniform mapping and verification of conformity of ICD, OPS, LOINC codes, and adequate reporting in the systematics" [Survey question 6, MIRACUM site 9].
FHIR[c] profiles uniformity	3 (30)	"Uniform FHIR profiles across MIRACUM partners" [Survey question 6, MIRACUM site 10].
Data provenance collection	1 (10)	"Structured Provenance Documentation: • where did the data come from, • what processing steps were performed on the data up to the time of data delivery, • Are there changes to the data that may represent a potential impact on the planned data use project?" [Survey question 6, MIRACUM site 9].

[a]DIC: Data Integration Center.

[b]MIRACUM: Medical Informatics in Research and Care in University Medicine.

[c]FHIR: Fast Healthcare Interoperability Resources.

Discussion

Principal Findings

Overview

This investigation examined the practices surrounding the fitness-for-purpose of observational health care data within MIRACUM DICs. It revealed both strengths and areas needing improvement in current approaches compared to established evidence. It also highlighted the active engagement of MIRACUM DICs in assessing data suitability for clinical research, using a diverse array of data repository infrastructures. However, the findings also underscored an important variation in methods for DQA and a general lack of standardization, suggesting the need for a more harmonized approach to enhance the accuracy and reliability of clinical data for specific secondary use purposes. Practical requirements were identified to support future implementation studies toward automating the observed processes.

Addressing Fitness-for-Purpose of Observational Health Care Data

Our study set out to examine the extent to which MIRACUM DICs address the fitness-for-purpose of observational health care data and the ways in which these approaches differ from

existing benchmarks. On the one hand, the demographic data suggest that the majority of participating MIRACUM DIC DQ experts have a computer science background and the prevalent academic qualification was a master's degree. This is indicative of the technological proficiency of the participants about data handling and engineering, but the participants may lack specific clinical or statistical expertise. This could be required for accurately assessing the suitability of clinical data for secondary use purposes. On the other hand, it is apparent that the MIRACUM DICs are actively engaged in fitness-for-purpose assessment of data, particularly for clinical research questions ranging from clinical event investigations to complex prediction models. This reflects the substantial amount of evidence [3,5,8,9,12] achieved since the inception of the German MI-I, implying a potential of fitness-for-purpose in DIC data, which should be further investigated based on the DIC data–driven research question. Interestingly, various repositories are used, including Clinical DWH, i2b2, OMOP, FHIR, ORBIS system, and OPAL DataSHIELD framework, indicating the use of diversified data infrastructure. However, there is a lack of harmonization and standardization in terms of strategies for assessing the suitability of DIC data with methods including the 4-eyes principle, system comparisons, and data consistency checks.

Compared to existing evidence, carrying out data consistency checks by the DICs nuancedly aligns with the recommendations of Kahn et al [22]. This process helps the MIRACUM sites in using the existing MIRACUM DQA tool to determine whether the integrated DIC data across the source and target databases [14,24] adhere to specified standards (conformance), whether certain data elements or values are sufficient (completeness), or if they are accurately reported (plausibility). In contrast, implementing the items from the 3×3 DQA guideline proposed by Weiskopf et al [21] was barely observable. This approach recommends specific DUPs by using particular methods (eg, rate of data regularity as proposed by Sperrin et al [36]) to assess DQ. For instance, this involves examining whether information regarding patients, data variables, and timeframe are completely, correctly, and currently integrated. This enables the research or clinical team to transparently determine the fitness-for-purpose of the data delivered by DICs. Automating such a process would enhance the integration of much more evidence, ensuring DQ for executing DUPs within the MIRACUM DICs. Therefore, the investigation of Razzaghi et al [37] could be seen as an initial roadmap toward a technical implementation. This describes a framework operationalizing such a fitness-for-purpose assessment process, with a particular focus on assessing clinical diagnosis or care. Furthermore, this should be evaluated in a closed association with DQ principles (outlier detection, plausibility, etc), as reported in various existing recommendations [20-22,38].

Key Requirements for Automating the Assessment Process

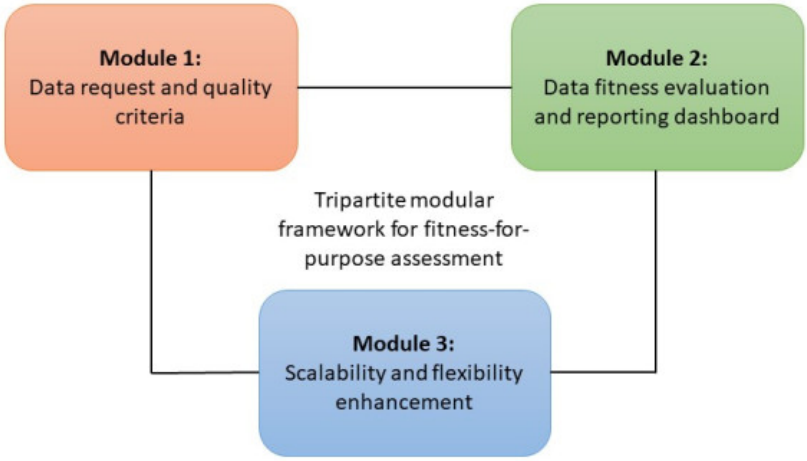
Our second research question focused on identifying the essential requirements for automating the assessment of fitness-for-purpose of EHR data within MIRACUM DICs. The TA revealed that from the perspective of the MIRACUM DICs, a scalable system for assessing the suitability of clinical data should be flexible, easily understandable, practicable, and modularizable (able to be extended in modules). However, the DICs expressed a need for user-friendly dashboards, facilitating an automated performance of system comparisons, data consistency checks, and data provenance collection in DUPs. These practice-oriented expectations adhere to the definition of fitness-for-purpose proposed by Girman et al [23], which considers assessing the availability of relevant data and the data provenance as key points for capturing fitness-for-purpose. A prior study by Gierend et al [15] explored the current status of provenance collection, which may present straightforward support in implementing this approach at the MIRACUM DIC level. However, the heterogeneity in the used data repositories at the DICs revealed that any automated solution would need to be repository agnostic or interoperable with multiple type of databases. In addition, the strategies for gathering of DUP-specific DQ criteria included consistent data description by requesters and participatory discussion between data providers and requesters. This suggests that an automated solution would require incorporating a flexible but stringent set of DQ criteria or functions, such as suggested by Weiskopf et al [20,21], Razzaghi et al [37], or Schmidt et al [38], that can be adjusted based on dynamic inputs. Finally, common approaches for assessing the fitness-for-purpose included the 4-eyes principle, comparison of data value distribution, and data consistency checks. These DQ verification processes suggest the need to incorporate robust DQA mechanisms into automation. These mechanisms could potentially use machine learning or rule-guided algorithms similar to the machine learning–supported DQA framework developed by Mark et al [39]. By considering these diverse aspects, we can consistently streamline the assessment of the fitness-for-purpose of clinical data, benefiting both research and clinical applications.

Derivation of a Roadmap Framework for Fitness-for-Purpose Assessment

Overview

Based on the findings discussed earlier, a guiding framework for an automated solution for assessing the fitness-for-purpose of EHR at the MIRACUM DICs was elaborated and structured into 3 modules (Figure 1). These modules have been designed to be pragmatically implementable and closely aligned with the survey insights to facilitate a smooth transition into practical application.

Figure 1. Tripartite modular framework for fitness-for-purpose assessment.



Data Request and Quality Criteria Module

This module should centralize and streamline the data requests and collection of DUP-specific DQ criteria. This module should include a user interface for data requesters to input their data requirements (eg, study fit criteria) and fitness-for-purpose criteria (eg, based on DQ rules from the 3×3 framework [21]).

Data Fitness and Reporting Dashboard Module

This module incorporates mechanisms for data fitness assessment reflecting the need for evaluating the data fitness for specific DUPs. It should use the fitness-for-purpose criteria collected from module 1 to perform data accuracy evaluation, alongside functionalities for comparing data values across different systems (eg, i2b2, OMOP, and FHIR). Furthermore, this module should feature a decision-supporting dashboard for visualizing these assessments by providing an overview of the data's fitness-for-purpose, which may facilitate structured documentation of DQ assurance tasks into the data management plans of DUPs.

Scalability and Flexibility Enhancement Module

This module is dedicated to the technical aspects of items from modules 1 and 2, addressing the requirements for a scalable solution, including the implementation of a user interface and dashboard, system comparison features, data consistency and provenance checks, and uniformity in FHIR profiles. Moreover, it should be implemented under consideration of usability requirements expressed by the DICs.

Each module was elaborated to address specific aspects identified in the survey to ensure a comprehensive and user-friendly solution for the automated DQ or fitness-for-purpose tool. This approach should not only facilitate the implementation process but also improve the efficiency and effectiveness of DQ and fitness-for-purpose checks across all MIRACUM DICs.

Strengths and Limitations

The strengths of this investigation lie in its comprehensive data collection and analysis. Using established approaches, such as the TA-based framework of Braun and Clarke [31,32], we provided a holistic view of the current state of the assessment of fitness-for-purpose observational EHR data at the MIRACUM DICs. Given the importance of reflexivity in qualitative research, as discussed by Braun and Clarke [32], we acknowledge our theoretical stance toward a constructivist approach, where knowledge is viewed as a construct rather than discovery. This perspective informed our analysis to interpret the data through a lens that considers both the factual and the contextual dimensions conveyed by the participants. As such, our survey questions, detailed in Textbox 1, were crafted not only to gather empirical information but also to probe deeper into the implications and the perceived effectiveness of DQ practices. Furthermore, the comprehensive examination of all MIRACUM DICs, illustrated diversified ways of conducting secondary use of EHR data and fitness-for-purpose assessments, highlighting additional proficiency.

However, our study also presents some limitations. The study cohort was restricted to the MIRACUM consortium. Expanding the research to include other MI-I consortia and DICs not affiliated with the MI-I project would enhance the sample's representativeness and the generalizability of the earned findings.

Implications for Future Research and Practice

To ensure that DIC data are suitable for research and clinical applications, it is crucial for MIRACUM DICs to align their DQA processes with established best practices and recommendations. Future research should aim to investigate the correlation adhering to the recommendations with the successful conduction of DUPs by evaluating whether clinical research questions (eg, cross-location analysis of comorbidities [25]) were more consistently addressable. For research practice, there is an accentuated need to develop automated and scalable solutions to assess the fitness-for-purpose of EHR DIC data, and our study provides the foundational insights to drive this technological advancement. Importantly, the solution could be based on the proposed tripartite modular framework. This could ease the integration into existing infrastructures of MIRACUM and eventually MI-I DICs, taking into account the key requirements identified by this investigation.

Conclusions

While the MIRACUM DICs focus on assessing the quality of clinical data for DUP conduction, there is a relevant variation

in the used methods and requirements to automate these processes. With the increasing volume and complexity of health data, having a standardized and scalable solution to automate the assessment of fitness-for-purpose is essential to maintain the integrity of research and clinical practice. Follow-up investigations should be directed toward building systems guided by evidence-based practices that should be customized to the specific needs and circumstances of MIRACUM and MI-I DICs.

Acknowledgments

This research was funded in part by the German Federal Ministry of Education and Research through the Medical Informatics in Research and Care in University Medicine consortium (grants 01ZZ1801E and 01ZZ1801A). For the publication fee, we acknowledge financial support from Heidelberg University. This work was performed in (partial) fulfillment of the requirements for obtaining the degree “Dr sc hum” of the Medical Faculty Mannheim of the University of Heidelberg (GKW).

Data Availability

The qualitative data collected and analyzed during this study are available in Multimedia Appendices 2 and 3. Furthermore, the R-based script used to plot the data use project execution frequency is available as well in Multimedia Appendix 4.

Authors' Contributions

GKW developed the research concept and methodology, collected and analyzed data, generated and reviewed themes and codes, visualized data, programmed in R, wrote the original draft, and reviewed and edited the manuscript. PM contributed to the methodology, reviewed themes and codes, and edited and reviewed the manuscript. SAS collected data and also edited and reviewed the manuscript. FS edited and reviewed the manuscript, provided supervision, and secured funding. TG contributed to the conceptualization and methodology of the study, edited and reviewed the manuscript, provided supervision, managed the project, and acquired funding. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Instructions for sites participation at the survey.
[PDF File (Adobe PDF File), 20 KB - medinform_v12i1e57153_app1.pdf]

Multimedia Appendix 2
Documentation of qualitative analysis of survey feedback data.
[PDF File (Adobe PDF File), 504 KB - medinform_v12i1e57153_app2.pdf]

Multimedia Appendix 3
Survey feedbacks from the participating Medical Informatics for Research and Care in University Medicine sites (deidentified).
[PDF File (Adobe PDF File), 435 KB - medinform_v12i1e57153_app3.pdf]

Multimedia Appendix 4
R-Project to perform the visualization of data use project frequencies.
[ZIP File (Zip Archive), 1024 KB - medinform_v12i1e57153_app4.zip]

References

1. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
2. Prokosch H, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med 2018;57(S 01):e82-e91 [FREE Full text] [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
3. Metzger P, Hess ME, Blaumeiser A, Pauli T, Schipperges V, Mertes R, et al. MIRACUM-pipe: an adaptable pipeline for next-generation sequencing analysis, reporting, and visualization for clinical decision making. Cancers (Basel) 2023;15(13):3456 [FREE Full text] [doi: 10.3390/cancers15133456] [Medline: 37444566]
4. Lauk K, Peters M, Velthaus J, Nürnberg S, Ueckert F. Use of process modelling for optimization of molecular tumor boards. Appl Sci 2022;12(7):3485. [doi: 10.3390/app12073485]
5. Peng Y, Nassirian A, Ahmadi N, Sedlmayr M, Bathelt F. Towards the representation of genomic data in HL7 FHIR and OMOP CDM. Stud Health Technol Inform 2021;283:86-94. [doi: 10.3233/SHTI210545] [Medline: 34545823]

6. Pape L, Schneider N, Schleef T, Junius-Walker U, Haller H, Brunkhorst R, et al. The nephrology eHealth-system of the metropolitan region of Hannover for digitalization of care, establishment of decision support systems and analysis of health care quality. BMC Med Inform Decis Mak 2019;19(1):176 [FREE Full text] [doi: 10.1186/s12911-019-0902-0] [Medline: 31477119]
7. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, et al. KETOS: Clinical decision support and machine learning as a service—a training and deployment platform based on docker, OMOP-CDM, and FHIR web services. PLoS One 2019;14(10):e0223010 [FREE Full text] [doi: 10.1371/journal.pone.0223010] [Medline: 31581246]
8. Schaaf J, Prokosch H, Boeker M, Schaefer J, Vasseur J, Storf H, et al. Interviews with experts in rare diseases for the development of clinical decision support system software—a qualitative study. BMC Med Inform Decis Mak 2020;20(1):230 [FREE Full text] [doi: 10.1186/s12911-020-01254-3] [Medline: 32938448]
9. Schaaf J, Sedlmayr M, Prokosch H, Ganslandt T, Schade-Brittinger C, von Wagner M, et al. The status quo of rare diseases centres for the development of a clinical decision support system—a cross-sectional study. In: dHealth 2020—Biomedical Informatics for Health and Care. Amsterdam, Netherlands: IOS Press; 2020:176-183.
10. Schaaf J, Sedlmayr M, Sedlmayr B, Prokosch H, Storf H. Evaluation of a clinical decision support system for rare diseases: a qualitative study. BMC Med Inform Decis Mak 2021;21(1):65 [FREE Full text] [doi: 10.1186/s12911-021-01435-8] [Medline: 33602191]
11. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: two national projects to promote data sharing in healthcare. Yearb Med Inform 2019;28(1):195-202 [FREE Full text] [doi: 10.1055/s-0039-1677917] [Medline: 31419832]
12. Reinecke I, Siebel J, Fuhrmann S, Fischer A, Sedlmayr M, Weidner J, et al. Assessment and improvement of drug data structuredness from electronic health records: algorithm development and validation. JMIR Med Inform 2023;11:e40312 [FREE Full text] [doi: 10.2196/40312] [Medline: 36696159]
13. Unberath P, Knell C, Prokosch H, Christoph J. Developing new analysis functions for a translational research platform: extending the cBioPortal for cancer genomics. Stud Health Technol Inform 2019;258:46-50. [Medline: 30942711]
14. Kapsner LA, Kampf MO, Seuchter S, Kamdje-Wabo G, Gradinger T, Ganslandt T, et al. Moving towards an EHR data quality framework: the MIRACUM approach. In: German Medical Data Sciences: Shaping Change–Creative Solutions for Innovative Medicine. Amsterdam, Netherlands: IOS Press; 2019:247-253.
15. Gierend K, Freiesleben S, Kadioglu D, Siegel F, Ganslandt T, Waltemath D. The status of data management practices across German medical data integration centers: mixed methods study. J Med Internet Res 2023 Nov 08;25:e48809 [FREE Full text] [doi: 10.2196/48809] [Medline: 37938878]
16. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc 2021;28(10):2251-2257 [FREE Full text] [doi: 10.1093/jamia/ocab132] [Medline: 34313749]
17. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet 2011;Chapter 1:Unit1.19 [FREE Full text] [doi: 10.1002/0471142905.hg0119s68] [Medline: 21234875]
18. Arts DGT, de Keizer NF, Scheffer G. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002;9(6):600-611 [FREE Full text] [doi: 10.1197/jamia.m1087] [Medline: 12386111]
19. Stausberg J, Nasseh D, Nonnemacher M. Measuring data quality: a review of the literature between 2005 and 2013. Stud Health Technol Inform 2015;210:712-716. [Medline: 25991245]
20. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
21. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A data quality assessment guideline for electronic health record data reuse. EGEMS (Wash DC) 2017;5(1):14 [FREE Full text] [doi: 10.5334/egems.218] [Medline: 29881734]
22. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
23. Girman CJ, Ritchey ME, Lo Re V. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. Pharmacoepidemiol Drug Saf 2022;31(7):717-720 [FREE Full text] [doi: 10.1002/pds.5444] [Medline: 35471704]
24. Kapsner LA, Mang JM, Mate S, Seuchter SA, Vengadeswaran A, Bathelt F, et al. Linking a consortium-wide data quality assessment tool with the MIRACUM metadata repository. Appl Clin Inform 2021;12(4):826-835. [doi: 10.1055/s-0041-1733847] [Medline: 34433217]
25. Kamdje-Wabo G, Gradinger T, Löbe M, Lodahl R, Seuchter S, Sax U, et al. Towards structured data quality assessment in the German Medical Informatics Initiative: initial approach in the MII demonstrator study. In: MEDINFO 2019: Health and Wellbeing e-Networks for All. Amsterdam, Netherlands: IOS Press; 2019:1508-1509.

26. Reynolds MW, Bourke A, Dreyer NA. Considerations when evaluating real-world data quality in the context of fitness for purpose. Pharmacoepidemiol Drug Saf 2020;29(10):1316-1318 [FREE Full text] [doi: 10.1002/pds.5010] [Medline: 32374042]
27. Cho S, Ensari I, Elhadad N, Weng C, Radin J, Bent B, et al. An interactive fitness-for-use data completeness tool to assess activity tracker data. J Am Med Inform Assoc 2022;29(12):2032-2040 [FREE Full text] [doi: 10.1093/jamia/ocac166] [Medline: 36173371]
28. Raman SR, O'Brien EC, Hammill BG, Nelson AJ, Fish LJ, Curtis LH, et al. Evaluating fitness-for-use of electronic health records in pragmatic clinical trials: reported practices and recommendations. J Am Med Inform Assoc 2022;29(5):798-804 [FREE Full text] [doi: 10.1093/jamia/ocac004] [Medline: 35171985]
29. Züll C. Open-ended questions (version 2.0). GESIS survey guidelines. Leibniz Institute for the Social Sciences. 2016. URL: https://d-nb.info/1191070131/34 [accessed 2024-08-02]
30. Kohler S. Atlassian Confluence 5 Essentials. Birmingham: Packt Publishing Ltd; 2013:334.
31. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101.
32. Braun V, Clarke V. Thematic analysis. United States: American Psychological Association; 2012:57-71.
33. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach 2020;42(8):846-854. [doi: 10.1080/0142159X.2020.1755030] [Medline: 32356468]
34. Manning J. In vivo coding. In: The International Encyclopedia of Communication Research Methods. New York, NY: Wiley; 2017:18.
35. Gandrud C. Reproducible Research with R and R Studio. New York, NY: CRC Press; 2013:323.
36. Sperrin M, Thew S, Weatherall J, Dixon W, Buchan I. Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. AMIA Annu Symp Proc 2011;2011:1318-1325 [FREE Full text] [Medline: 22195193]
37. Razzaghi H, Greenberg J, Bailey LC. Developing a systematic approach to assessing data quality in secondary use of clinical data based on intended use. New York, NY: Wiley; 2022:2379-6146.
38. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021;21(1):63 [FREE Full text] [doi: 10.1186/s12874-021-01252-7] [Medline: 33810787]
39. Mark S, Gaurav S, Timothy O, Hayley P, Linda T, Kira N, et al. Development and validation of ML-DQA—a machine learning data quality assurance framework for healthcare. 2022 Presented at: 7th Machine Learning for Healthcare Conference; August 5-6, 2022; Durham, NC p. 741-759 URL: https://arxiv.org/abs/2208.02670 [doi: 10.48550/arXiv.2208.02670]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
DIC: Data Integration Center
DQ: data quality
DQA: data quality assessment
DUP: data use project
DWH: Data Warehouse
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
GESIS: Society of Social Science Infrastructure Institutions
i2b2: Informatics for Integrating Biology and Bedside
MI-I: Medical Informatics Initiative
MIRACUM: Medical Informatics in Research and Care in University Medicine
OMOP: Observational Medical Outcomes Partnership
TA: thematic analysis

Edited by G Eysenbach, J Klann; submitted 06.02.24; peer-reviewed by A Hassan, J Aarts, R Bidkar; comments to author 15.04.24; revised version received 31.05.24; accepted 22.07.24; published 19.08.24.

Please cite as:
Kamdje Wabo G, Moorthy P, Siegel F, Seuchter SA, Ganslandt T
Evaluating and Enhancing the Fitness-for-Purpose of Electronic Health Record Data: Qualitative Study on Current Practices and Pathway to an Automated Approach Within the Medical Informatics for Research and Care in University Medicine Consortium
JMIR Med Inform 2024;12:e57153
URL: https://medinform.jmir.org/2024/1/e57153
doi:10.2196/57153
PMID:39158950

Multifaceted Natural Language Processing Task–Based Evaluation of Bidirectional Encoder Representations From Transformers Models for Bilingual (Korean and English) Clinical Notes: Algorithm Development and Validation

Kyungmo Kim[1], MS; Seongkeun Park[2], MD, PhD; Jeongwon Min[1], MS; Sumin Park[3], PhD; Ju Yeon Kim[4], MD, PhD; Jinsu Eun[5], MS; Kyuha Jung[5], MS; Yoobin Elyson Park[5], MS; Esther Kim[5], BS; Eun Young Lee[4], MD, PhD; Joonhwan Lee[5], PhD; Jinwook Choi[3,6], MD, PhD

1

2

3

4

5

6

Corresponding Author:
Jinwook Choi, MD, PhD

Abstract

Background: The bidirectional encoder representations from transformers (BERT) model has attracted considerable attention in clinical applications, such as patient classification and disease prediction. However, current studies have typically progressed to application development without a thorough assessment of the model's comprehension of clinical context. Furthermore, limited comparative studies have been conducted on BERT models using medical documents from non–English-speaking countries. Therefore, the applicability of BERT models trained on English clinical notes to non-English contexts is yet to be confirmed. To address these gaps in literature, this study focused on identifying the most effective BERT model for non-English clinical notes.

Objective: In this study, we evaluated the contextual understanding abilities of various BERT models applied to mixed Korean and English clinical notes. The objective of this study was to identify the BERT model that excels in understanding the context of such documents.

Methods: Using data from 164,460 patients in a South Korean tertiary hospital, we pretrained BERT-base, BERT for Biomedical Text Mining (BioBERT), Korean BERT (KoBERT), and Multilingual BERT (M-BERT) to improve their contextual comprehension capabilities and subsequently compared their performances in 7 fine-tuning tasks.

Results: The model performance varied based on the task and token usage. First, BERT-base and BioBERT excelled in tasks using classification ([CLS]) token embeddings, such as document classification. BioBERT achieved the highest F_1-score of 89.32. Both BERT-base and BioBERT demonstrated their effectiveness in document pattern recognition, even with limited Korean tokens in the dictionary. Second, M-BERT exhibited a superior performance in reading comprehension tasks, achieving an F_1-score of 93.77. Better results were obtained when fewer words were replaced with unknown ([UNK]) tokens. Third, M-BERT excelled in the knowledge inference task in which correct disease names were inferred from 63 candidate disease names in a document with disease names replaced with [MASK] tokens. M-BERT achieved the highest hit@10 score of 95.41.

Conclusions: This study highlighted the effectiveness of various BERT models in a multilingual clinical domain. The findings can be used as a reference in clinical and language-based applications.

(JMIR Med Inform 2024;12:e52897) doi:10.2196/52897

KEYWORDS

natural language processing; NLP; natural language inference; reading comprehension; large language models; transformer

Introduction

Since 2015, deep learning is increasingly being used in clinical natural language processing (NLP) [1]. Large language models (LLMs) based on deep learning technology are widely used in numerous clinical NLP domains [2]. Because contextual comprehension is critical for the overall performances of NLP models, studies have focused on the development of models

that excel in conveying contextual information. Conventional approaches of NLP involve crafting word-to-word sequence models such as the hidden Markov model and using limited datasets annotated with labels such as disease and medication names [3-5]. However, studies are increasingly focusing on fine-tuning LLMs that have been pretrained on massive unlabeled biomedical literature sources, such as Medical Information Mart for Intensive Care (MIMIC-III) [6] and PubMed [7,8]. This shift in the NLP research direction has substantially elevated the contextual understanding capabilities of models and inspired studies on clinical NLP that focus on LLM utilization. For example, studies on automated summarization [9-11] have effectively extracted critical phrases from diverse sources, including biomedical papers and patient records. In addition, studies on entity extraction [12-15] have identified major entities such as disease names and drug names. However, these studies have focused exclusively on English-language corpora.

In the multilingual clinical domain, we proposed a set of contextual understanding conditions, with a comprehensive suite of clinical NLP evaluations specifically for these conditions. The proposed approach involves comparatively assessing bidirectional encoder representations from transformers (BERT) models [16] to provide guidelines for selecting the most suitable BERT model for a particular condition.

We proposed 2 hypotheses to examine 4 BERT models. First, we assumed that within the multilingual clinical domain, a language model capable of comprehending multiple languages would achieve superior performance. Second, models with the capacity to comprehend medical contexts would demonstrate superior efficacies. We selected BERT-base [16], Korean BERT (KoBERT) [17], Multilingual BERT (M-BERT) [18], and BERT for Biomedical Text Mining (BioBERT) [7] for the study. We pretrained these models on visit records on 160,000 patients. Subsequently, we introduced a series of comprehensive downstream tasks to learn the conditions required for these models to achieve effective contextual understanding. We assumed that an effective language model thrives in contextual comprehension under the following conditions:

- The model can determine whether the provided documents pertain to the same patient (tasks 1 and 2).
- The model is proficient in identifying the department associated with a given document (task 3).
- The model can discern the descriptions within medical records for the conditions of different patients (tasks 4 and 5).
- The model can ascertain the connection among sentences (task 6).
- The model can competently deduce disease names based on existing knowledge (task 7).

The rationale of the proposed conditions is the widespread adoption of BERT models in the medical domain for various applications.

BERT has been applied in medical natural language inference research to assess the relationship between 2 sequences (premise and hypothesis) with entailment, contradiction, or neutrality labels. Percha et al [19] used a fine-tuned BERT to locate clinical notes relevant to query sentences. Romanov and Shivade [20] created the MedNLI clinical dataset for natural language inference. They used several models and methodologies, such as bag-of-words, InferSent, and enhanced sequential inference models, to confirm the efficacy and validity of their datasets. Boukkour et al [21] introduced an alternative approach to BERT tokenization, proposing a convolutional neural network–based character-based tokenizer as a replacement for WordPiece Tokenizer, which is used to pretrain BERT, to improve BERT performance on the MedNLI dataset. Kanakarajan et al [22] pretrained the ELECTRA model, which is named “efficiently learning an encoder that classifies token replacements accurately,” [23] using abstracts from PubMed, and evaluated its performance on the MedNLI dataset.

BERT was applied to categorize clinical notes. Rasmy et al [24] introduced Med-BERT, which pretrained BERT using electronic health record data to classify diabetes and pancreatic cancer datasets. This model exceeded gated recurrent units by 2 - 4 in terms of area under the receiver operating characteristic score. Zhang and Jankowski [25] proposed average pooling transformer layers handling token-, sentence-, and document-level embeddings for classifying *International Classification of Diseases* codes. Their model outperformed the BERT-base model by 11 points.

For the reading comprehension task, BERT can be used to determine the answer span within a given text. Pampari et al [26] proposed the electronic medical record question answering (emrQA) dataset to determine the answer span to a question in a clinical context. Yue et al [27] compared the performances of BERT-base, BioBERT, and ClinicalBERT [8] on the emrQA dataset and additional test datasets to address the problems of the emrQA dataset. Rawat et al [28] used 30 logical forms to express questions in semistructured texts and identified the correct responses in the emrQA dataset. They entered clinical notes and questions and used multitask training to simultaneously predict the logical structure of the question and the text span of the answer in a clinical note. Savery et al [29] introduced the MEDIQA-AnS dataset, which contains questions and corresponding answers regarding the health care concerns of patients. The correct answers to these questions, which contain valuable information about the patients, are used as summaries.

BERT can be used to extract information from clinical notes. Yang et al [15] used the 2010 i2b2 [30], 2012 i2b2 [31], and 2018 national NLP clinical challenges (n2c2) [32] datasets to compare the information extraction performances of BERT models, namely, BERT-base, ELECTRA, A Lite BERT (ALBERT) [33], and Robustly Optimized BERT Pretraining Approach (RoBERTa) [34]. The test results revealed that RoBERTa outperformed the other models. Richie et al [35] used Clinical BERT [8] to extract the social determinants of patient health, namely, employment, living tobacco, alcohol, drug use, and their attributes, from the n2c2 2022 Track 2 dataset [36]; for instance, texts such as “works” and “unemployed” were extracted for detailing employment information.

Although studies have extensively examined BERT versatility, they have focused only on English corpora. To address this limitation, we comprehensively analyzed the efficacies of BERT models in various tasks involving medical documents in both Korean and English.

The rest of the manuscript is organized as follows. The *Methods* section outlines the diverse tests used for BERT analysis and their application procedures. The *Results* section presents a summary of the outcomes of each test. The *Discussion* section outlines the distinctive characteristics of each BERT model and presents a thorough analysis for understanding the reasons for these characteristics. Finally, the *Conclusion* section summarizes the study and emphasizes its significance.

The aim of this study was to identify the BERT models that perform optimally in the bilingual (Korean and English) clinical domain. To achieve this objective, we designed 7 tasks, evaluated the performance of 4 BERT variants (BERT-base, BioBERT, KoBERT, and M-BERT) across these tasks, and assessed their relative significance.

Methods

Dataset

We obtained outpatient records from 8 departments, namely, endocrinology, respiratory, cardiovascular, gastroenterology, rheumatology, nephrology, allergy medicine, and infectious medicine departments, at Seoul National University Hospital in South Korea. We collected the records of 164,460 outpatients between 2010 and 2019. The dataset comprised 2,453,934 documents, with 412,499,140 tokens generated after tokenization using white space. The distribution of tokens and documents for various departments was as follows: endocrinology (tokens: 91,352,271; docs: 496,938), respiratory (tokens: 31,556,578; docs: 195,048), cardiovascular (tokens: 114,978,554; docs: 696,061), gastroenterology (tokens: 57,755,571; docs: 416,062), rheumatology (tokens: 24,857,675; docs: 204,600), nephrology (tokens: 70,865,514; docs: 322,629), allergy medicine (tokens: 17,024,481; docs: 92,041), and infectious medicine departments (tokens: 4,108,496; docs: 30,555). Table 1 provides statistical data for the corpus. Table 2 presents the clinical note of a patient experiencing rheumatoid arthritis.

Table . Statistical data of clinical notes in Seoul National University Hospital between 2010 and 2019.

Department	Tokens, n	Documents, n
Endocrinology	91,352,271	496,938
Respiratory	31,556,578	195,048
Cardiovascular	114,978,554	696,061
Gastroenterology	57,755,571	416,062
Rheumatology	24,857,675	204,600
Nephrology	70,865,514	322,629
Allergy medicine	17,024,481	92,041
Infectious medicine	4,108,496	30,555
Sum	412,499,140	2,453,934

Table . The example of a clinical note that was used for training bidirectional encoder representations from transformers models (for better understanding, an English translation has been added).

Section	Contents
History	• Korean: *3117.2.1 arthralgia r/o d/t letrozole* 로 병원 방문; English (translated): *3117.2.1 arthralgia, rule out (r/o) due to letrozole. Visited hospital* • Korean: *meloxicam 7.5 mg bid* 복용한 *hx* 있다.; English (translated): *Has a history of taking meloxicam 7.5 mg twice daily.* • Korean: *f/u loss* 마지막 방문 때 *RF*[a], *ACCP*[b], *ANA*[c]등 처방했다.; English (translated): *Prescribed RF, ACCP, ANA, etc, during the last visit.* • Korean: *Arthralgia , neutropenia* 가 있다. 손이 붓고 마디가 아프다. 약먹지만 붓기가 빠지지 않는 것 같다.; English (translated): *Experiencing arthralgia and neutropenia. Hands are swollen and joints are painful. Although taking medication, the swelling does not seem to be going down.*
P/E & Lab[d]	• Korean: *PIP S −/+ T −/− wrist S −/+ , T −/+ Toe s −/− T −/− 2120 . 3 lab bone scan: normal CBC*[e]*, WNL*[f]*, and CRP*[g]*0.10*; English (translated): *PIP S −/+ T −/− wrist S −/+ , T −/+ Toe s −/− T −/− 2120 . 3 lab bone scan: normal CBC, WNL, and CRP 0.10*
Assessment	• Korean: *Arthralgia r/o d/t letrozole* 장상피화생; English (translated): *Arthralgia r/o d/t letrozole. Intestinal metaplasia*
Plan	• Korean: *RF, ACCP, ANA , x-ray Celebrex 50 mg tid -->Celebrex 100 mg tid*; English (translated): *RF, ACCP, ANA, x-ray Celebrex 50 mg tid -->Celebrex 100 mg tid*

[a]RF: rheumatoid factor.

[b]ACCP: anticitrullinated protein antibody.

[c]ANA: antinuclear antibody.

[d]P/E & Lab: physical examination and laboratory.

[e]CBC: complete blood count.

[f]WNL: within normal limits.

[g]CRP: C-reactive protein.

Ethical Considerations

We obtained approval to use the original data collection for research purposes from the institutional review board (IRB) at Seoul National University Hospital (IRB no. C-2108-008-1242). According to the institution's IRB policy, the data cannot be publicly disclosed due to patient privacy concerns. Instead, we provide an overview of the data in Table 2.

BERT Models

The BERT-base model is a precursor in pretrained transformer encoders [37]. Vast open-domain data sources, including Wikipedia and BooksCorpus, are used to train the model [38]. The model is primarily focused on English text. The configuration of this dataset facilitates the expression of contextual representations of English sequences.

The BioBERT model is an evolution of BERT and is pretrained on PubMed data and enriched with biomedical entities, rendering BioBERT proficient in comprehending terminologies such as disease and drug names. In this study, we used the latest iteration of BioBERT, that is, BioBERT version 1.1.

The SKT Corporation in South Korea devised the KoBERT model to enhance the comprehension and processing of the Korean language. Data from Korean Wikipedia and news articles were used to pretrain the model.

The M-BERT model was obtained from a richly varied corpus of 104 languages, enabling a contextual representation that spans both English and Korean sequences.

Pretraining

To enhance the bilingual clinical contextual understanding capabilities of BERT models, we conducted additional pretraining using an extensive dataset comprising 159,460 out of 164,460 outpatient records from Seoul National University Hospital, employing masked language modeling. The data were preprocessed meticulously using this strategy. WordPiece Tokenizer was used by BERT-base, BioBERT, and M-BERT; SentencePiece Tokenizer [39] was used by KoBERT. All tokenizers were case-sensitive. Subsequently, random tokens within the input sequence were replaced with [MASK] tokens. This process was reiterated 10 times to yield the data required for pretraining. The pretraining task of the model involved reinstating the [MASK] token to its original token, drawing on the data crafted through this preprocessing procedure.

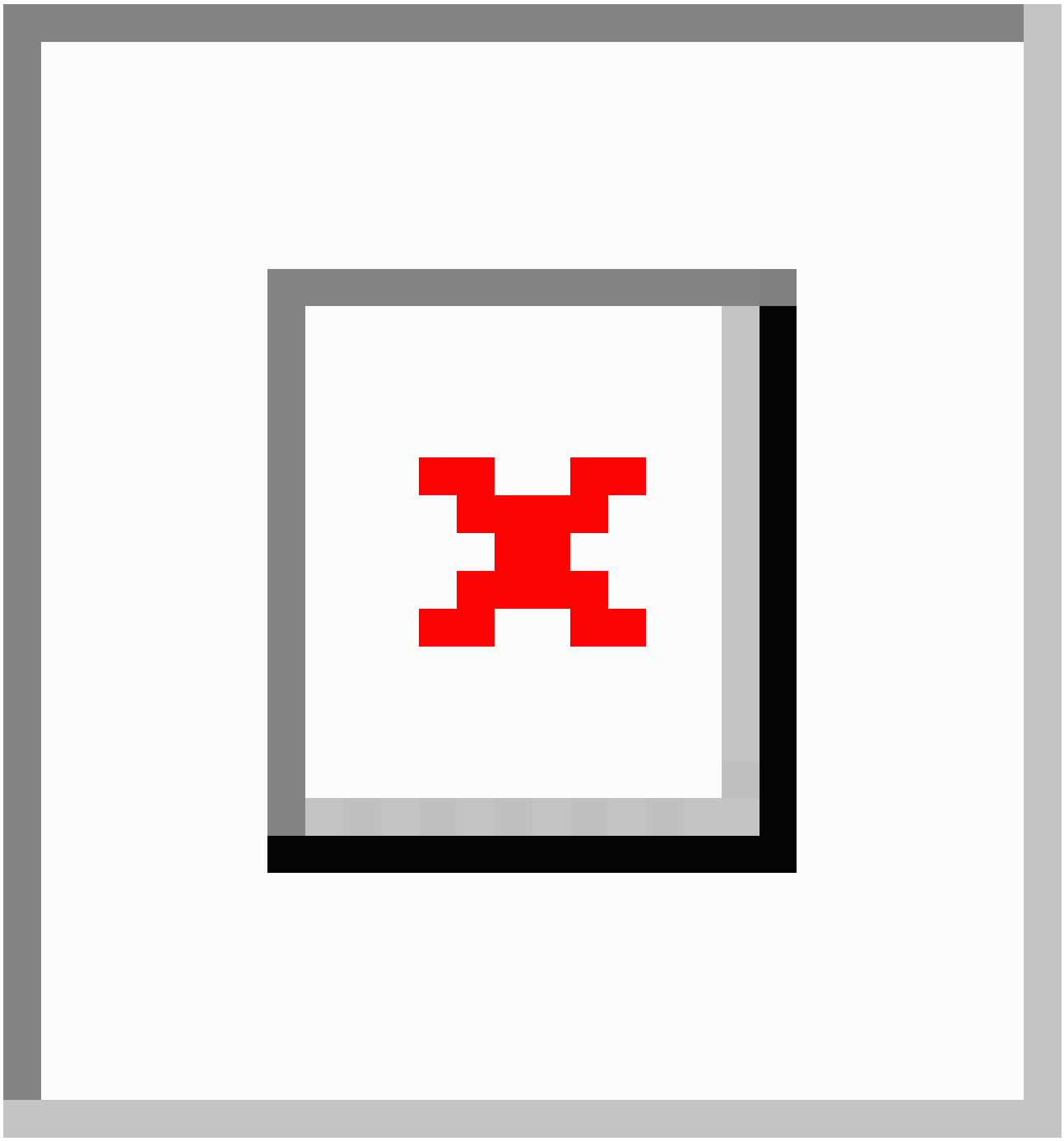
Multifaceted Clinical NLP Tasks

The evaluation framework encompassed 5 characteristics. Each characteristic was examined through 7 distinct downstream tasks that were designed to assess the clinical contextual comprehension capabilities of various BERT models.

Homogeneity Determination

As seen in Figure 1, we used 2 single outpatient records per input sequence to determine document homogeneity. Each model performed binary classification, discerning whether the records corresponded to those of the same patient (task 1). We extended this examination to the section level, tasking each model with predicting homogeneity based on a smaller segment of a page (task 2). In task 2, the objective was to determine whether 2 sequences originated from the same patient record, with 1 sequence containing an assessment section and the other section containing a randomized section.

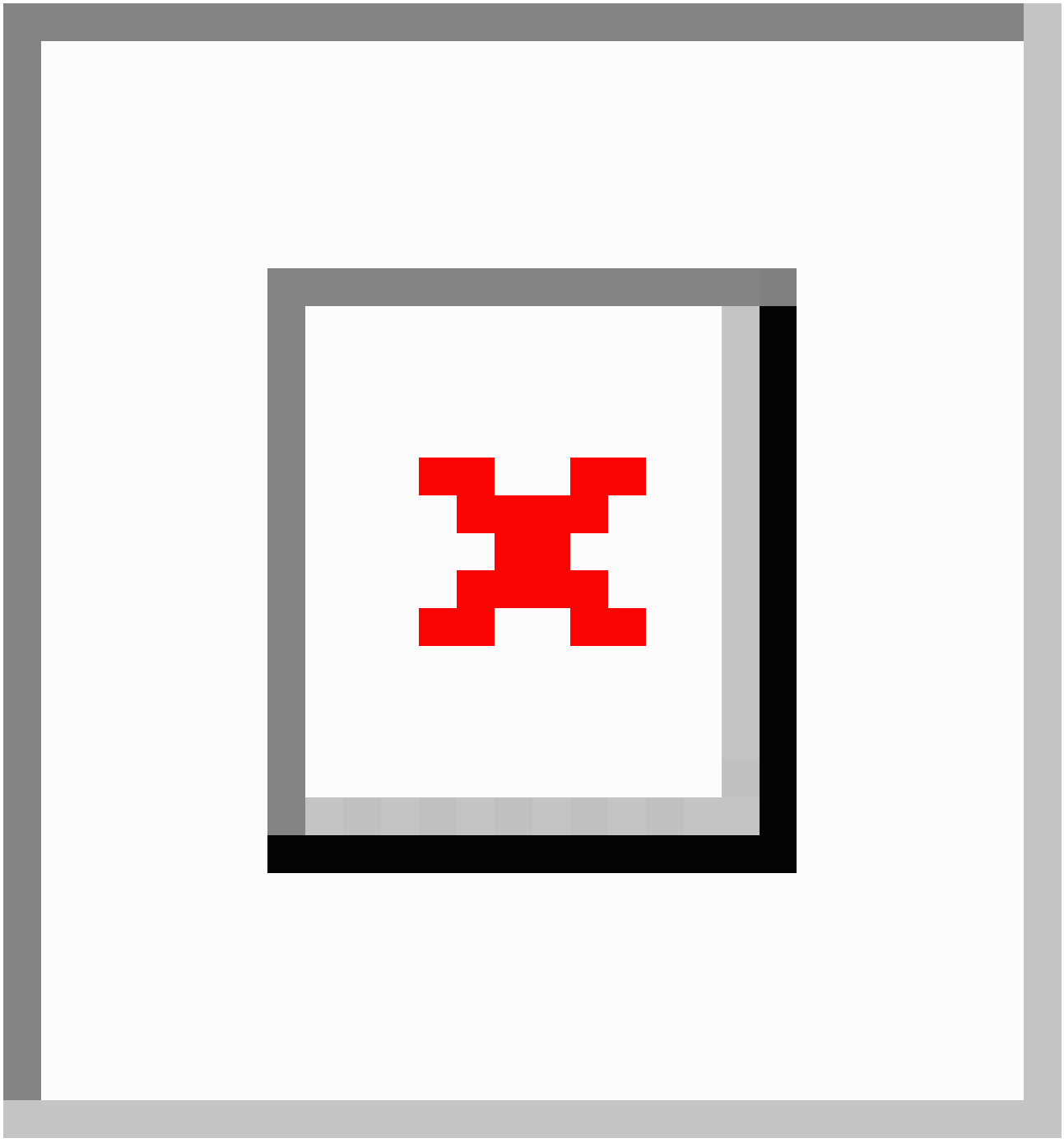
Figure 1. Document homogeneousness test (tasks 1 and 2). BERT: bidirectional encoder representations from transformers; CLS: classification.



Document Representativeness

As seen in Figure 2, to assess document representativeness, we devised a task that focused on department identification by using individual visit records (task 3).

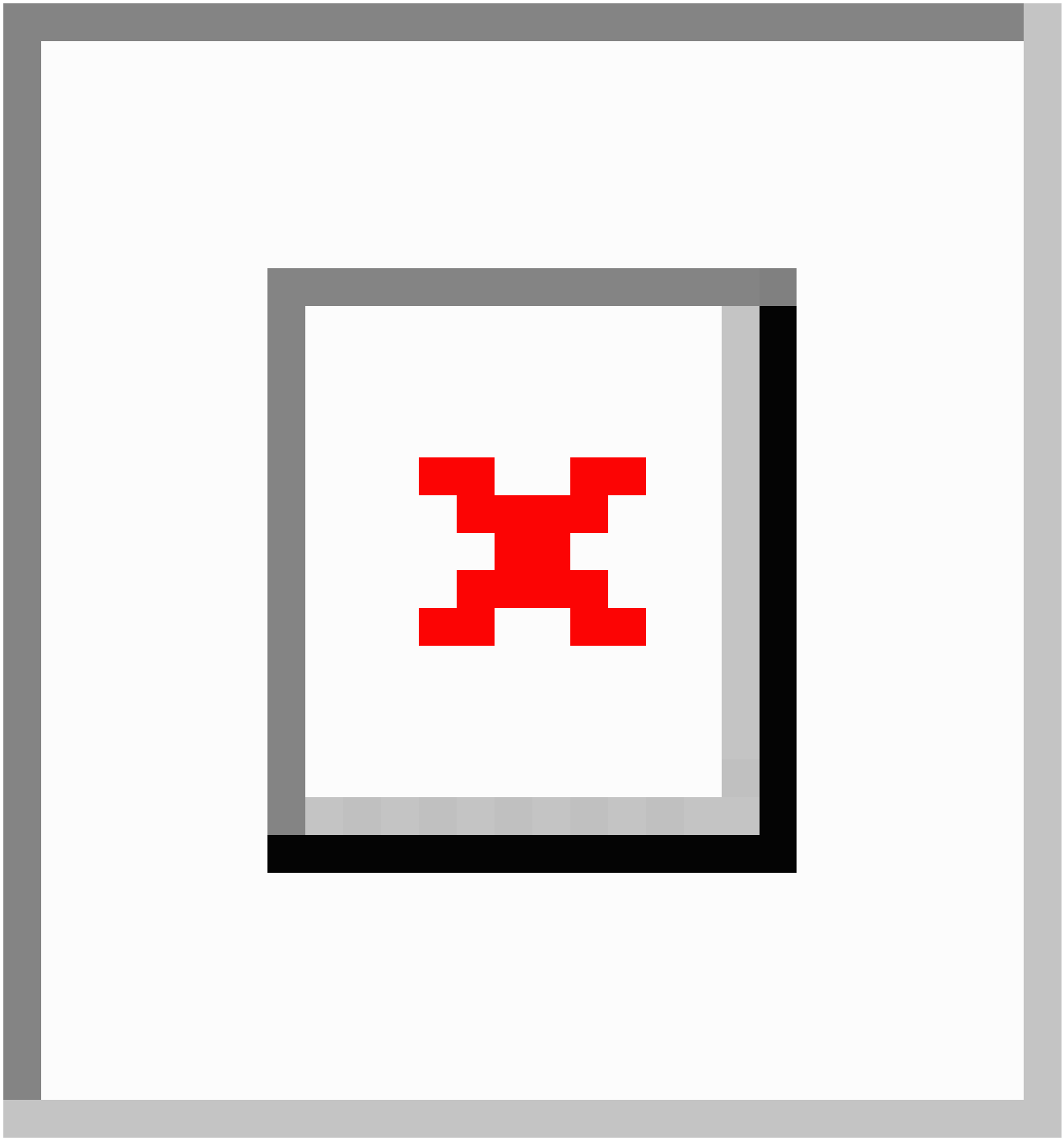
Figure 2. Document representativeness test: classifying documents (task 3). BERT: bidirectional encoder representations from transformers; CLS: classification; SEP: separator.



Reading Comprehension Test

The reading comprehension test (Figure 3) test extracted summarized content from a visit record. We focused on extracting the assessment section from the Subjective, Objective, Assessment, Plan (SOAP) or the history, physical examination, laboratory, assessment, and plan sections. The experiments encompassed 2 setups, namely, 1 setup with section-shuffled documents (task 4) and 1 setup with maintained section-order documents (task 5).

Figure 3. Reading comprehension test: identifying the department associated with a given document (with section shuffling: task 4; w/o section shuffling: task 5). BERT: bidirectional encoder representations from transformers; CLS: classification; SEP: separator; w/o: without.

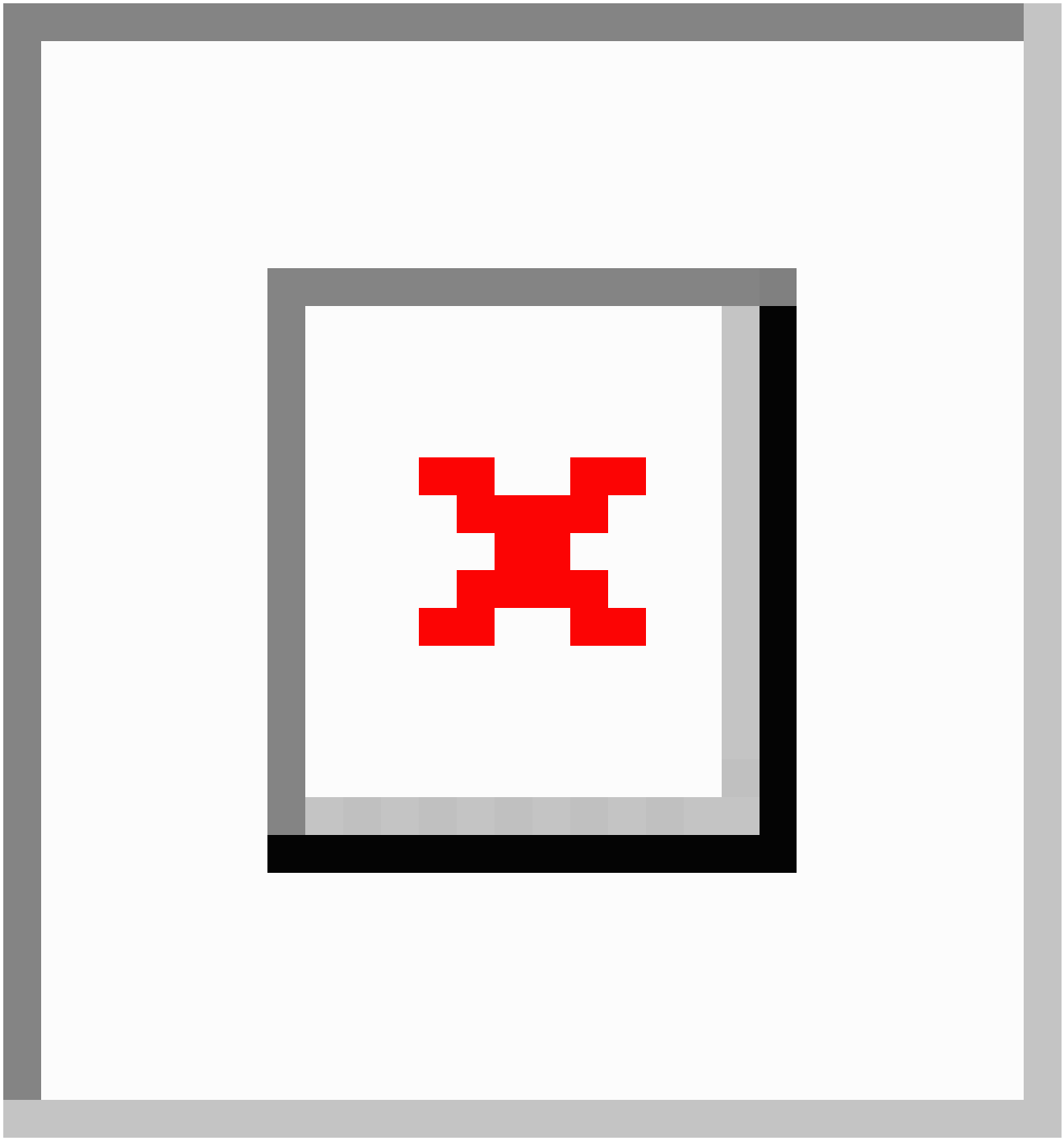


Contextual Connections

As seen in Figure 4, we introduced a task that required the model to differentiate the most recent visit record from a set of 4 candidate documents when given a query document representing the oldest visit record (task 6). The limitation of BERT models regarding the amount of the input length they can handle necessitates a workaround because simultaneously inputting both the query document and 4 candidate documents is not feasible. To address this problem, we adopted a 2-step approach. First, each individual document was independently inputted into BERT to acquire document embeddings. Subsequently, these document embeddings, forming a pair comprising the query document and the kth candidate document embeddings, were introduced into a feedforward neural network (FFNN) [40]. For example, if the query and document embedding pair for the most recent visit were inputted into the FFNN, the model was trained to output a prediction value of 1; this value was assigned based on our assumptions. We postulated that the query document, which corresponded to the earliest visit among the 5 documents, and the last document, which denoted the most recent visit, encompassed the most distinct narrative. Consequently, we measured the cosine distances between these 2 embeddings and directed the model to output a prediction value of 1, which indicated the greatest distance in terms of cosine similarity. By contrast, if the query and nonanswer

document embedding pairs were presented to the FFNN, the model was trained to output a prediction value of zero.

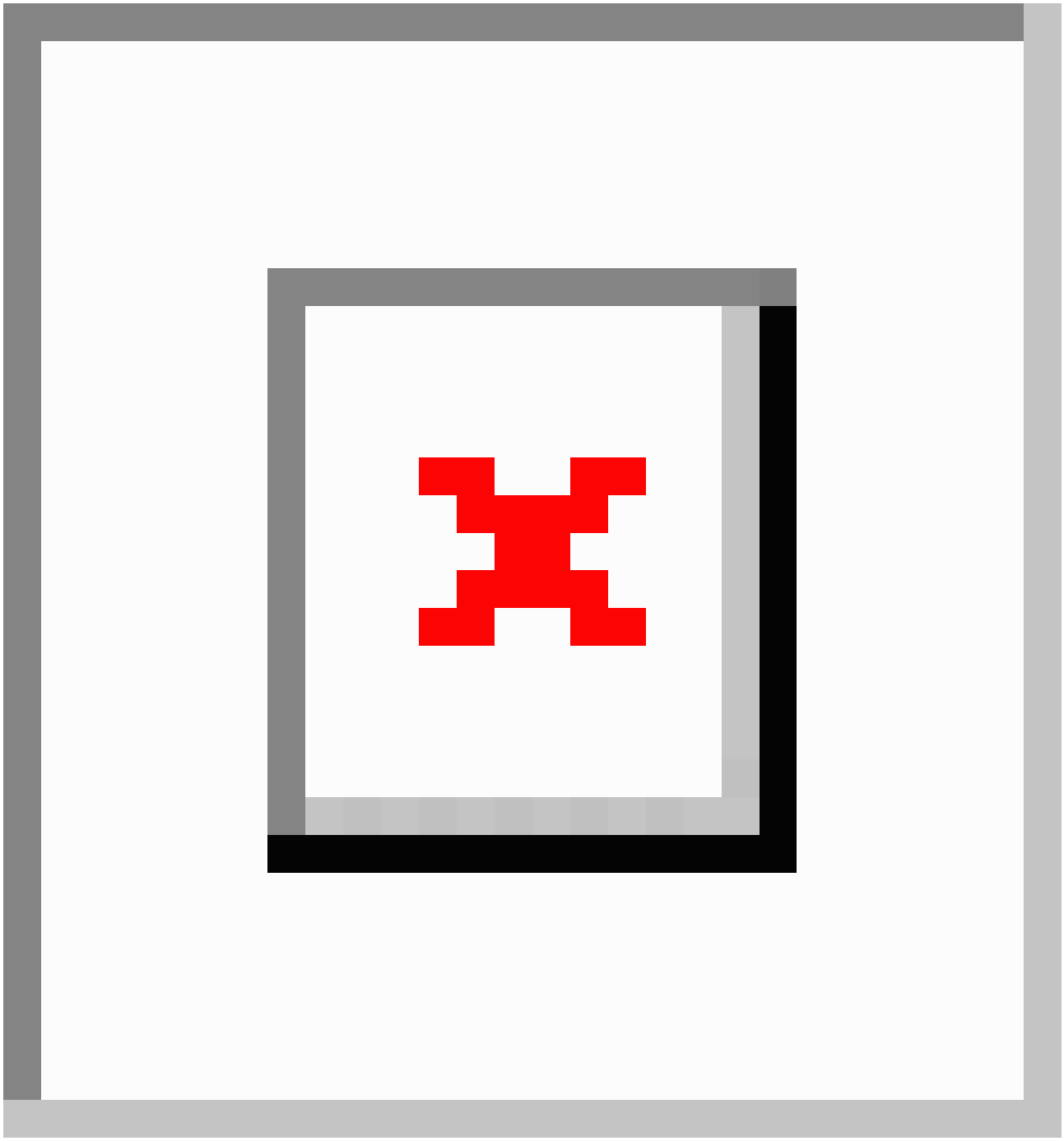
Figure 4. Document connectivity test: finding the last visited document (task 6). BERT: bidirectional encoder representations from transformers.



Knowledge Reasoning

The knowledge reasoning characteristic (Figure 5) evaluated the capacity of a model to deduce entities from masked text (task 7). Each model was tasked with deducing disease names from masked visit records in which the disease names had been replaced with [MASK] tokens. We used MetaMap [41] to create a dataset by identifying diagnostic names. Each model, when presented with the [MASK] token and context, selected the correct disease name from 63 disease names. A comprehensive list of the entities is shown in Table S1 in Multimedia Appendix 1.

Figure 5. Knowledge reasoning test: finding the disease name (task 7). BERT: bidirectional encoder representations from transformers; CLS: classification; SEP: separator.



Experimental Settings

We trained and evaluated 4 types of publicly available BERT models through the following process. We used records of 159,460 patients out of 164,460 patients for pretraining. In the pretraining procedure, 15% of random tokens from the 159,460 patient records were masked. Among them, 80% of the masked tokens were replaced with [MASK] tokens, 10% were replaced with random tokens, and the remaining 10% retained their original tokens. We trained the BERT models to restore [MASK] tokens to their original tokens.

After pretraining, the 4 BERT models were fine-tuned for tasks 1 - 7. For fine-tuning, we used 5000 patient records that were not used in pretraining. We assigned 4000 patients to the training set and 1000 patients to the test set and then created training and evaluation data specific for each task. In each task, the 4 pretrained BERT models were trained using the training set and evaluated on the test set.

In the pretraining step, 4 NVIDIA 3090 graphics processing units (GPUs) were used in parallel for 3 epochs. After pretraining, all the models were fine-tuned using a 1080ti GPU except for task 6, in which 3090 GPU were used, because this task required more calculation procedures and memory. The detailed hyperparameter settings are described in Table S2 in Multimedia Appendix 1. The detailed experimental settings and analysis code used in this study are available on GitHub [42].

Results

Results of Tasks 1-3

In tasks 1 - 3, BERT-base and BioBERT exhibited the best scores; Tables 3 and 4 present the corresponding results.

Table . Results of various BERT[a] models in tasks 1 and 2.

Model	Task 1: Determination of whether 2 documents are from the same patients			Task 2: Determination of whether 2 sections are from the same patients		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
BERT-base	84.44	94.19	89.05	89.28	87.60	88.43
BioBERT[b]	83.36	96.21	89.32	92.92	82.73	87.53
KoBERT[c]	83.95	74.05	78.69	90.68	75.78	82.56
M-BERT[d]	83.22	94.02	88.29	83.56	93.38	88.19

[a]BERT: bidirectional encoder representations from transformers.

[b]BioBERT: BERT for Biomedical Text Mining.

[c]KoBERT: Korean BERT.

[d]M-BERT: Multilingual BERT

Table . Results of various BERT[a] models in task 3.

Model	Task 3: Identification of the department associated with a given document accuracy
BERT-base	96.75
BioBERT[b]	97.44
KoBERT[c]	95.38
M-BERT[d]	96.06

[a]BERT: bidirectional encoder representations from transformers.

[b]BioBERT: BERT for Biomedical Text Mining.

[c]KoBERT: Korean BERT.

[d]M-BERT: Multilingual BERT.

In the homogeneity test conducted on document-level inputs (task 1), BioBERT achieved the highest F_1-score, whereas in the test conducted on the section-level inputs (task 2), BERT-base achieved the highest F_1-score. Comparing the scores under tasks 1 and 2 revealed that BioBERT exhibited a more substantial drop in performance than those of other models. By contrast, KoBERT consistently demonstrated a diminished performance compared with that exhibited by other BERT models. In the document representativeness test, which entailed the selection of a single department from a set of 8 department candidates, BioBERT exhibited superior performance in terms of accuracy, which was the evaluation metric.

Results of Tasks 4-7

In tasks 4 - 7, M-BERT achieved the best scores (Tables 5 and 6).

Table . Results of various BERT[a] models in tasks 4 and 5.

Model	Task 4: Finding the assessment section with inputs that are section-shuffled			Task 5: Finding the assessment section with inputs that are not section-shuffled		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
BERT-base	71.03	61.74	60.83	74.59	59.14	56.69
BioBERT[b]	72.16	56.31	51.64	74.71	55.99	51.17
KoBERT[c]	76.57	77.41	76.88	92.61	93.88	93.15
M-BERT[d]	93.15	94.61	93.77	96.52	96.37	96.44

[a]BERT: bidirectional encoder representations from transformers.

[b]BioBERT: BERT for Biomedical Text Mining.

[c]KoBERT: Korean BERT.

[d]M-BERT: Multilingual BERT.

Table . Results of various BERT[a] models in task 7.

Model	Task 7: Determination of disease names based on existing knowledge		
	hit@1	hit@3	hit@10
BERT-base[a]	60.26	77.47	93.40
BioBERT[b]	59.40	80.20	95.12
KoBERT[c]	46.20	72.02	91.54
M-BERT[d]	61.12	81.64	95.41

[a]BERT: bidirectional encoder representations from transformers.

[b]BioBERT: BERT for Biomedical Text Mining.

[c]KoBERT: Korean BERT.

[d]M-BERT: Multilingual BERT.

In the reading comprehension tests (tasks 4 and 5), the performances of the models were evaluated in terms of the F_1-score, which was calculated by measuring the proportion of tokens within the predicted interval that correctly overlapped with the actual interval. M-BERT achieved the highest performance in reading comprehension tests. In addition, the models exhibited the largest performance differences in these tests. In the context connectivity test (task 6), M-BERT exhibited the highest performance with an F_1-score of 64.75, whereas all the other models achieved a score lower than 60 (BERT-base: 59.78; BioBERT: 58.39; KoBERT: 25.62; and M-BERT: 64.75). In the knowledge-reasoning test (task 7), the M-BERT model exhibited the best performance. The primary objective of this test was to accurately prognosticate 63 potential candidate diagnoses, as extracted from clinical documents, in which the diagnosis name was substituted with [MASK]. In our assessment, we used hit@k (where k=1, 3, or 10). For instance, in task 7, BERT computes probabilities for 63 diseases based on a provided context. In this context, hit@k is a true positive if k diseases with the highest probability encompass the correct disease. The final evaluation score is then determined by dividing the number of true positives by the total number of sequences under assessment.

Discussion

Suitability of BERT-Base and BioBERT for English [CLS] Embedding (Tasks 1-3)

In tasks 1 - 3, the BERT classification ([CLS]) embedding was the input for the FFNN. The [CLS] token, positioned at the far-left side of the input sequence, is a classification token. The embedding of this token is commonly used as a feature for classification tasks, indicating the model's comprehension of segment-level or document-level context. In tasks 1 and 2, homogeneity was assessed at the document and section levels, respectively, and BioBERT and BERT-base demonstrated the highest performances, respectively. In task 3, BioBERT achieved the highest score. Based on these observations, we inferred that BERT-base and BioBERT would be suitable for tasks involving [CLS] embedding.

Generally, a model's ability to understand context diminishes as the number of tokens absent from its dictionary increases. Unknown ([UNK]) tokens represent tokens absent from the model's dictionary, and the presence of these tokens correlates with lower model performance. The higher the frequency of [UNK] tokens, the greater the challenge for the model to accurately comprehend the context. Notably, despite the limited inclusion of Korean tokens, these models excelled in tasks 1 - 3 (Table S4 in Multimedia Appendix 1). BERT-base and

BioBERT, which were pretrained on English sentence patterns, exhibited improved performances because of the prevalence of English sentences in outpatient visit records, which typically detailed their diseases.

Influence of Multilingual Capabilities in Reading Comprehension Tasks on Outcomes (Tasks 4 and 5)

In tasks 4 and 5, the reading comprehension ability of the model was assessed by determining the scope of the assessment section. Among models, M-BERT demonstrated the highest performance, whereas BERT-base and BioBERT exhibited the lowest test scores. The presence of extensive multilingual capabilities in the reading comprehension tests was the predominant factor influencing these outcomes.

To comprehend why BERT-base and BioBERT exhibit markedly inferior performance compared with M-BERT in tasks 4 and 5, understanding the composition of the BERT model dictionaries and the function of the [UNK] token is crucial. In BERT models, a dedicated tokenizer is used to segment text into tokens. These tokens are retained if present in the model's dictionary; otherwise, the tokens are substituted with [UNK] tokens, representing unknown entities. Consequently, a higher prevalence of [UNK] tokens indicates a diminished ability of the model to comprehend the semantic nuances of the sequence. In tasks 4 and 5, where each token's semantic relevance determines its association with an assessment section, models with inadequate knowledge of individual tokens exhibit poor performance. The dictionaries of BERT-base and BioBERT contain minimal Korean characters, resulting in the majority of Korean tokens being replaced with [UNK] tokens. By contrast, M-BERT encompasses a comprehensive range of Korean characters in its dictionary. Therefore, BERT-base and BioBERT exhibit notably inferior performance in tasks 4 and 5 compared with M-BERT.

Relationship Between Multilingual Capability and Task Complexity (Task 7)

Task 7, which was focused at evaluating the aptitude of a model for knowledge inference, was more complex than other tasks. Notably, M-BERT outperformed the other models in task 7, securing hit@1, hit@3, and hit@10 scores of 61.12, 81.64, and 95.41, respectively. These results highlighted the pivotal role of the dictionary in knowledge inference. Furthermore, when processing documents in multiple languages, M-BERT outperformed BERT-base, which had been exclusively trained in a single language.

For task 6, the test results were poor. An analysis indicated that BERT models did not excel in this task because of the prevalence of outpatient medical records in the copy-and-paste format (Table S6 in Multimedia Appendix 1). Consequently, the significance of task 6 in this study was low.

Contributions to the Clinical Text Processing and Medical Fields

Importance of Multilingual Models

The experiment highlights the significance of using multilingual language models in processing bilingual clinical notes. The findings demonstrated that using a model capable of handling 2 languages yields superior performance compared with relying solely on a single language model. This insight is particularly relevant for countries such as Korea and Japan, where clinical documentation typically involves a mixture of languages.

Base for Model Selection

Furthermore, this study provides empirical evidence for choosing a proper BERT model, a factor not substantiated in existing NLP research. For instance, in previous studies, such as that conducted by Kim and Lee [43], M-BERT was used for tasks such as extracting disease names, symptoms, and body parts from Korean text without providing explicit justification. The experimental results satisfied this gap by showcasing the superiority of M-BERT in understanding bilingual clinical text and supporting appropriate BERT selection in future studies.

Limitations and Future Works

Limited Scope of Clinical Notes

This analysis primarily focused on outpatient visit records. Future studies should encompass a broad range of clinical notes, including surgical notes, hospitalization records, and discharge summaries. Comparing and validating the performance of BERT models across various types of clinical documentation provides a comprehensive understanding of their effectiveness.

Single-Institution Data

This study exclusively used data from Seoul National University Hospital, which can limit the generalizability of the findings. Clinical notes can vary considerably in style and content across various health care institutions. Therefore, future studies should involve data from multiple hospitals to validate BERT model performance in various clinical settings.

More Tasks Should Be Verified

The BERT model requires further validation in bilingual clinical text. Oh et al [44] conducted a study to recognize protected health information in the publicly available i2b2 2014 dataset. However, we could not perform this task because manual labeled annotations are required to extract non-English entities in bilingual clinical notes. In future studies, various tasks using bilingual clinical notes should be proposed.

Conclusions

In this study, we comprehensively compared 4 BERT models, encompassing text in both English and Korean, within the multilingual clinical domain. We pretrained these models with approximately 160,000 patient records and evaluated their performances for 7 diverse downstream tasks. The experimental findings are summarized as follows.

First, the BERT-base and BioBERT models excelled in document classification tasks using [CLS] tokens. These results highlighted their superiority over M-BERT in tasks involving simple pattern recognition in word sequences. Second, the significance of having a comprehensive dictionary was evident in the reading comprehension task in which comprehensive token usage was required. The exceptional performance of M-BERT, which encompassed a broad range of Korean and English tokens, clearly confirmed the importance of the dictionary. Third, multilingual proficiency was pivotal for tasks

that demanded complex reasoning. Both M-BERT and BioBERT excelled in task 7, which focused on diagnosing a multitude of candidates, and notably, M-BERT consistently outperformed BioBERT.

Our findings highlighted the suitability of BioBERT and BERT-base for tasks that relied on sequence patterns in multilingual clinical domains. In addition, M-BERT, which had an expansive dictionary and aptitude for leveraging Korean and English clinical contexts, was highly suitable for tasks involving textual content comprehension. The experimental results of the BERT models in mixed-language clinical documents provide valuable insights for future medical NLP research and appropriate BERT model selection for different types of tasks.

Acknowledgments

This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (no. NRF-2021R1I1A4A01042182).

Data Availability

The datasets generated and/or analyzed during the current study are not publicly available due to patient privacy concerns. Patient records contain personal information, and, as such, Seoul National University's institutional review board does not permit public disclosure of the data.

Authors' Contributions

KK, SP, JM, and SP conceptualized the paper, developed the methodology, and prepared the original draft of the manuscript. KK contributed to software implementation and validated the findings. JYK and EYL curated the data, conducted investigations, and contributed to data analysis and interpretation. JE, KJ, YEP, EK, and JL contributed to methodology development, conducted formal analysis, and provided insights throughout the research process. JC supervised the study, managed the project administration, and contributed to reviewing and editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials on disease entities, hyperparameter settings, number of documents in the pretraining dataset, tokens, tokenization, and masked language modeling loss.
[DOCX File, 125 KB - medinform_v12i1e52897_app1.docx]

References

1. Wu H, Wang M, Wu J, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. NPJ Digit Med 2022 Dec 21;5(1):186. [doi: 10.1038/s41746-022-00730-6] [Medline: 36544046]
2. Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges. Cureus 2023 May;15(5):e39305. [doi: 10.7759/cureus.39305] [Medline: 37378099]
3. Zhang J, Shen D, Zhou G, Su J, Tan CL. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. J Biomed Inform 2004 Dec;37(6):411-422. [doi: 10.1016/j.jbi.2004.08.005] [Medline: 15542015]
4. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557-562. [doi: 10.1136/amiajnl-2011-000150] [Medline: 21565856]
5. Torii M, Wagholikar K, Liu H. Detecting concept mentions in biomedical text using hidden Markov model: multiple concept types at once or one at a time? J Biomed Semantics 2014 Jan 17;5(1):3. [doi: 10.1186/2041-1480-5-3] [Medline: 24438362]
6. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
7. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
8. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; Jun 7, 2019; Minneapolis, MN p. 72-78. [doi: 10.18653/v1/W19-1909]
9. Krishna K, Khosla S, Bigham J, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Aug 1-6, 2021; Online p. 4958-4972. [doi: 10.18653/v1/2021.acl-long.384]

10. Hu J, Li Z, Chen Z, Li Z, Wan X, Chang TH. Graph enhanced contrastive learning for radiology findings summarization. arXiv. Preprint posted online on Jun 8, 2022 URL: https://arxiv.org/abs/2204.00203 [accessed 2022-04-01] [doi: 10.48550/arXiv.2204.00203]
11. Kanwal N, Rizzo G. Attention-based clinical note summarization. arXiv. Preprint posted online on Apr 18, 2021 URL: https://arxiv.org/abs/2104.08942 [accessed 2021-04-18] [doi: 10.48550/arXiv.2104.08942]
12. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. J Am Med Inform Assoc 2021 Aug 13;28(9):1892-1899. [doi: 10.1093/jamia/ocab090] [Medline: 34157094]
13. Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. J Am Med Inform Assoc 2020 Jan 1;27(1):13-21. [doi: 10.1093/jamia/ocz063] [Medline: 31135882]
14. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S111-S119. [doi: 10.1016/j.jbi.2015.06.010] [Medline: 26122527]
15. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. J Am Med Inform Assoc 2020 Dec 9;27(12):1935-1942. [doi: 10.1093/jamia/ocaa189] [Medline: 33120431]
16. Devlin J, Chang MW, Lee K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Jun 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
17. SKTBrain/KoBERT: Korean BERT pre-trained cased (KoBERT). GitHub. 2019. URL: https://github.com/SKTBrain/KoBERT.git [accessed 2022-05-02]
18. Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT? Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Jul 28 to Aug 2, 2019; Florence, Italy p. 4996-5001. [doi: 10.18653/v1/P19-1493]
19. Percha B, Pisapati K, Gao C, Schmidt H. Natural language inference for curation of structured clinical registries from unstructured text. J Am Med Inform Assoc 2021 Dec 28;29(1):97-108. [doi: 10.1093/jamia/ocab243] [Medline: 34791282]
20. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018; Brussels, Belgium p. 1586-1596. [doi: 10.18653/v1/D18-1187]
21. El Boukkouri H, Ferret O, Lavergne T, Noji H, Zweigenbaum P, Tsujii J. CharacterBERT: reconciling ELMo and BERT for word-level open-vocabulary representations from characters. Presented at: Proceedings of the 28th International Conference on Computational Linguistics; Dec 8-13, 2020; Barcelona, Spain p. 6903-6915. [doi: 10.18653/v1/2020.coling-main.609]
22. Kanakarajan KR, Kundumani B, Sankarasubbu M. BioELECTRA: pretrained biomedical text encoder using discriminators. Presented at: Proceedings of the 20th Workshop on Biomedical Language Processing; Jun 11, 2021; Online p. 143-154. [doi: 10.18653/v1/2021.bionlp-1.16]
23. Clark K, Luong MT, Le QV, Manning CD. Electra: pre-training text encoders as discriminators rather than generators. arXiv. Preprint posted online on Mar 23, 2020 URL: https://arxiv.org/abs/2003.10555 [accessed 2023-09-15] [doi: 10.48550/arXiv.2003.10555]
24. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021 May 20;4(1):86. [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
25. Zhang N, Jankowski M. Hierarchical BERT for medical document understanding. arXiv. Preprint posted online on Mar 11, 2022 URL: https://arxiv.org/abs/2204.09600 [accessed 2023-09-15] [doi: 10.48550/arXiv.2204.09600]
26. Pampari A, Raghavan P, Liang J, Peng J. EmrQA: a large corpus for question answering on electronic medical records. Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018; Brussels, Belgium p. 2357-2368. [doi: 10.18653/v1/D18-1258]
27. Yue X, Gutierrez BJ, Sun H. Clinical reading comprehension: a thorough analysis of the emrAQ dataset. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020; Online p. 4474-4486. [doi: 10.18653/v1/2020.acl-main.410]
28. Rawat BPS, Weng WH, Min SY, Raghavan P, Szolovits P. Entity-enriched neural models for clinical question answering. Presented at: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing; Jul 9, 2020; Online p. 112-122. [doi: 10.18653/v1/2020.bionlp-1.12]
29. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. Sci Data 2020 Oct 2;7(1):322. [doi: 10.1038/s41597-020-00667-z] [Medline: 33009402]
30. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011 Sep 1;18(5):552-556. [doi: 10.1136/amiajnl-2011-000203]
31. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc 2013;20(5):806-813. [doi: 10.1136/amiajnl-2013-001628] [Medline: 23564629]

32. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 2020 Jan 1;27(1):3-12. [doi: 10.1093/jamia/ocz166]
33. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv. Preprint posted online on Sep 26, 2019 URL: https://arxiv.org/abs/1909.11942 [accessed 2023-09-15] [doi: 10.48550/arXiv.1909.11942]
34. Liu Z, Lin W, Shi Y, Zhao J. A robustly optimized BERT pre-training approach with post-training. Presented at: Proceedings of the 20th Chinese National Conference on Computational Linguistics; Aug 13-15, 2021; Huhhot, China p. 1218-1227.
35. Richie R, Ruiz VM, Han S, Shi L, Tsui FR. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. J Am Med Inform Assoc 2023 Jul 19;30(8):1379-1388. [doi: 10.1093/jamia/ocad046] [Medline: 37002953]
36. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. J Am Med Inform Assoc 2023 Jul 19;30(8):1367-1378. [doi: 10.1093/jamia/ocad012] [Medline: 36795066]
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); Dec 4-7, 2017; Long Beach, CA, USA p. 5998-6008.
38. Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. Presented at: 2015 IEEE International Conference on Computer Vision (ICCV); Dec 7-13, 2015; Santiago, Chile p. 19-27. [doi: 10.1109/ICCV.2015.11]
39. Kudo T, Richardson J. SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Oct 31 to Nov 4, 2018; Brussels, Belgium p. 66-71. [doi: 10.18653/v1/D18-2012]
40. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019; Hong Kong, China p. 3982-3992. [doi: 10.18653/v1/D19-1410]
41. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236. [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
42. medinfoman/multifaceted-berts: a study that verified the performance of BERT models in clinical text from various perspectives. GitHub. URL: https://github.com/medinfoman/multifaceted-berts.git [accessed 2024-03-09]
43. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak 2020 Sep 30;20(Suppl 7):242. [doi: 10.1186/s12911-020-01241-8] [Medline: 32998724]
44. Oh SH, Kang M, Lee YH. Protected health information recognition by fine-tuning a pre-training transformer model. Healthc Inform Res 2022 Jan;28(1):16-24. [doi: 10.4258/hir.2022.28.1.16]

Abbreviations

[CLS]: classification
[UNK]: unknown
ALBERT: A Lite BERT
BERT: bidirectional encoder representations from transformers
BioBERT: BERT for Biomedical Text Mining
ELECTRA: efficiently learning an encoder that classifies token replacements accurately
emrQA: electronic medical record question answering
FFNN: feedforward neural network
GPU: graphics processing unit
HMM: hidden Markov model
IRB: institutional review board
KoBERT: Korean BERT
LLM: large language model
M-BERT: Multilingual BERT
MIMIC-III: Medical Information Mart for Intensive Care
NLP: natural language processing
RoBERTa: Robustly Optimized BERT Pretraining Approach
SOAP: Subjective, Objective, Assessment, Plan

Edited by C Lovis; submitted 19.09.23; peer-reviewed by C Haag, D Chrimes, M Chatzimina; revised version received 08.07.24; accepted 17.08.24; published 30.10.24.

Please cite as:
Kim K, Park S, Min J, Park S, Kim JY, Eun J, Jung K, Park YE, Kim E, Lee EY, Lee J, Choi J
Multifaceted Natural Language Processing Task–Based Evaluation of Bidirectional Encoder Representations From Transformers Models for Bilingual (Korean and English) Clinical Notes: Algorithm Development and Validation
JMIR Med Inform 2024;12:e52897
URL: https://medinform.jmir.org/2024/1/e52897
doi:10.2196/52897

Unsupervised Feature Selection to Identify Important ICD-10 and ATC Codes for Machine Learning on a Cohort of Patients With Coronary Heart Disease: Retrospective Study

Peyman Ghasemi[1,2], MSc; Joon Lee[1,3,4,5], PhD

1

2

3

4

5

Corresponding Author:
Joon Lee, PhD

Abstract

Background: The application of machine learning in health care often necessitates the use of hierarchical codes such as the International Classification of Diseases (ICD) and Anatomical Therapeutic Chemical (ATC) systems. These codes classify diseases and medications, respectively, thereby forming extensive data dimensions. Unsupervised feature selection tackles the "curse of dimensionality" and helps to improve the accuracy and performance of supervised learning models by reducing the number of irrelevant or redundant features and avoiding overfitting. Techniques for unsupervised feature selection, such as filter, wrapper, and embedded methods, are implemented to select the most important features with the most intrinsic information. However, they face challenges due to the sheer volume of ICD and ATC codes and the hierarchical structures of these systems.

Objective: The objective of this study was to compare several unsupervised feature selection methods for ICD and ATC code databases of patients with coronary artery disease in different aspects of performance and complexity and select the best set of features representing these patients.

Methods: We compared several unsupervised feature selection methods for 2 ICD and 1 ATC code databases of 51,506 patients with coronary artery disease in Alberta, Canada. Specifically, we used the Laplacian score, unsupervised feature selection for multicluster data, autoencoder-inspired unsupervised feature selection, principal feature analysis, and concrete autoencoders with and without ICD or ATC tree weight adjustment to select the 100 best features from over 9000 ICD and 2000 ATC codes. We assessed the selected features based on their ability to reconstruct the initial feature space and predict 90-day mortality following discharge. We also compared the complexity of the selected features by mean code level in the ICD or ATC tree and the interpretability of the features in the mortality prediction task using Shapley analysis.

Results: In feature space reconstruction and mortality prediction, the concrete autoencoder–based methods outperformed other techniques. Particularly, a weight-adjusted concrete autoencoder variant demonstrated improved reconstruction accuracy and significant predictive performance enhancement, confirmed by DeLong and McNemar tests (P<.05). Concrete autoencoders preferred more general codes, and they consistently reconstructed all features accurately. Additionally, features selected by weight-adjusted concrete autoencoders yielded higher Shapley values in mortality prediction than most alternatives.

Conclusions: This study scrutinized 5 feature selection methods in ICD and ATC code data sets in an unsupervised context. Our findings underscore the superiority of the concrete autoencoder method in selecting salient features that represent the entire data set, offering a potential asset for subsequent machine learning research. We also present a novel weight adjustment approach for the concrete autoencoders specifically tailored for ICD and ATC code data sets to enhance the generalizability and interpretability of the selected features.

(JMIR Med Inform 2024;12:e52896) doi:10.2196/52896

KEYWORDS

unsupervised feature selection; ICD-10; International Classification of Diseases; ATC; Anatomical Therapeutic Chemical; concrete autoencoder; Laplacian score; unsupervised feature selection for multicluster data; autoencoder-inspired unsupervised feature selection; principal feature analysis; machine learning; artificial intelligence; case study; coronary artery disease; artery disease; patient cohort; artery; mortality prediction; mortality; data set; interpretability; International Classification of Diseases, Tenth Revision

Introduction

Machine learning is increasingly being used in health care to analyze patient data and provide insights on improving health outcomes and the quality of care [1]. With the rise of electronic health data (EHD) and entering a large amount of data per patient in hospitals, there are big opportunities to train machine learning models for a variety of applications, such as the prediction or diagnosis of diseases, outcome prediction, and treatment planning [1,2]. EHD are a valuable source of information on a patient, containing details on their demographics, hospital visits, medical diagnoses, physiological measurements, and treatments received [3]. However, despite the opportunities offered by these large data sets, there are challenges in terms of data quality, privacy, and the complexity of medical conditions [1]. In terms of machine learning, EHD can include many irrelevant and redundant features, where their direct use can lead to the "curse of dimensionality" as the high dimensionality of the data can make it more difficult to extract meaningful patterns and relationships [4]. Therefore, it is important to apply appropriate techniques for dimensionality reduction and feature engineering to address this challenge and improve the effectiveness of predictive models built from EHD.

Feature selection is one of the critical aspects of machine learning. It involves selecting a subset of relevant features that are the most useful for predicting a target variable. In the case of medical data, these features could include patient demographics, medical history, laboratory test results, and diagnosis codes [3]. Feature selection is essential because it can help improve the accuracy and performance of machine learning models by reducing the number of irrelevant or redundant features and avoiding overfitting [4]. Unsupervised feature selection is a type of feature selection method that is used when there is no target variable available to guide the selection of features. Unlike supervised feature selection, which chooses features that better predict a certain target variable, unsupervised feature selection methods rely on the intrinsic structure of the data to identify the most important features. This behavior helps the selected features to be unbiased and perform well when there are no labeled data. It can also reduce the risk of overfitting to a certain target variable and ensure robustness to new target variables [5]. This is an important advantage in health care, where collecting labeled data is usually difficult and the same data are often used to predict multiple target variables.

Generally, there are 3 main categories of feature selection methods: filter, wrapper, and embedded methods. Filter methods use statistical tests such as variance to rank individual features within a data set and select the features that maximize the desired criteria. However, they usually lack the ability to consider the interactions between features [6]. Wrapper methods, on the other hand, select features that optimize an objective function for a clustering algorithm. Therefore, these methods are generally specific to particular clustering algorithms and may not be suitable for use with other algorithms. Wrapper methods can detect potential relationships between features, but this often results in increased computational complexity [5]. Embedded methods also take into account feature relationships but generally do so more efficiently by incorporating feature selection into the learning phase of another algorithm. Lasso regularization is one of the well-known embedded methods that can be applied to a variety of machine learning models [6].

The *International Classification of Diseases, Tenth Revision* (*ICD-10*) is a method of classifying diseases that was created by the World Health Organization and is used internationally [7]. It categorizes diseases based on their underlying cause, characteristics, symptoms, and location in the body and uses codes to represent each disease. The *ICD-10* system organizes thousands of codes in a hierarchical structure that includes chapters, sections, categories, and expansion codes. Within this structure, section codes and their corresponding chapter codes can be thought of as child-parent relationships, with each *ICD-10* code serving as a node in the classification system. The same relationship applies to categories and sections, as well as expansion codes and categories. The high number of codes in this system is one of the major challenges of using them in machine learning applications [8]. It is worth noting that Canada has added or changed some codes in the lower levels according to their health care system requirements (*ICD-10, Canada* [*ICD-10-CA*]) [9].

Similar to International Classification of Diseases (ICD) codes, the Anatomical Therapeutic Chemical (ATC) classification system, developed by the World Health Organization Collaborating Centre for Drug Statistics Methodology, is an international tool for the active and systematic categorization of active pharmaceutical ingredients [10]. ATC codes are also structured hierarchically and are assigned based on the organ or system they impact, as well as their therapeutic, pharmacological, and chemical properties. This hierarchical system comprises 5 distinct levels, with the lower levels providing detailed information about the pharmacological subgroup and chemical substance, and the highest level representing the anatomical main group. As in the *ICD-10*, the ATC's hierarchy introduces child-parent relationships at each level.

In this research, we used 3 administrative databases comprising *ICD-10* and ATC codes pertaining to patients with coronary artery disease (CAD). These databases, relevant to acute care, ambulatory care, and pharmacy facilities, were used to select the most insightful codes characterizing this cohort.

Methods

Data Set and Preprocessing

The Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease (APPROACH) registry [11] is one of the most comprehensive data repositories of CAD management in the world, matching unique disease phenotypes with rich clinical information and relevant outcomes for patients in Alberta, Canada, who have undergone diagnostic cardiac catheterization or revascularization procedures. Our cohort's patients were selected from the APPROACH registry. These patients underwent diagnostic angiography between January 2009 and March 2019 at 1 of the following 3 hospitals in Alberta: Foothills Medical Centre, University of Alberta Hospital, and Royal Alexandra Hospital. We excluded patients

with ST elevation myocardial infarction from the study to focus on nonemergency CAD.

Discharge Abstract Database (DAD), National Ambulatory Care Reporting System (NACRS), and Pharmaceutical Information Network (PIN) data for the abovementioned patients were extracted from Alberta provincial health records. The DAD contains summary hospitalization information from all acute care facilities in Alberta. The NACRS includes all visits to ambulatory care facilities (ie, emergency department, urgent care, and day surgery visits) in the province as well as some nonabstracted data from other specialty clinics. The PIN is based on a system that collects all prescription medicine dispensations from pharmacies all over Alberta.

In the DAD and NACRS, for each patient, we aggregated all *ICD-10-CA* codes of hospital admissions or physician visits every 3 months following the first admission date (all codes in that period are treated as 1 record's codes). This helps us to make sure that chronic diseases are captured more comprehensively in fewer records and reduce the effect of noisy records. We did a similar procedure for ATC codes in the PIN data set and aggregated the codes every 6 months, since most medications prescription refills did not extend beyond 6 months. We one-hot encoded the *ICD-10-CA* and ATC codes and their parent nodes for each record. For example, if the *ICD-10-CA* code "I251" was present, "I25," "I20-I25," and "Chapter IX" were also encoded in the one-hot table. Similarly, if the ATC code "C07AB02" was present, "C07AB," "C07A," "C07," and "C" were also encoded. We show the number of all unique *ICD-10-CA* or ATC codes in the data set with N_{All}.

To validate the performance of the selected features in a real clinical problem, we pulled the mortality data of the patients enrolled in the cohort from the Vital Statistics Database and matched them with the aggregated records to determine 90-day mortality following the end of the last procedure.

Feature Selection

The following unsupervised algorithms were used for feature selection:

- Concrete autoencoder (CAE) [6]: In this method, continuous relaxation of discrete random (concrete) variables [12] and the Gumbel-Softmax reparameterization trick are used to construct a special layer in the neural network to transform discrete random variables into continuous ones, which allows for efficient computation and optimization using gradient-based methods. The reparameterization trick allows the use of a softmax function in this layer, which is differentiable, unlike the argmax function. This characteristic is useful for designing an autoencoder, in which features are selected in the concrete layer (as the encoder) through the softmax operation and a common neural network (as the decoder) is used to reconstruct the main feature space out of the selected features. During the training, a temperature parameter can be gradually decreased, allowing the concrete selector layer to try different features in the initial epochs and behave more similar to an argmax function in the last epochs to keep the best features. After training, we can use an argmax function on the weights to find the features passed to the neurons of the encoder layer. One of the major problems of this method is that it may converge to a solution where some duplicate features are selected in some neurons (ie, fewer than the desired number of features are selected).
- Autoencoder-inspired unsupervised feature selection (AEFS) [13]: This method combines autoencoder and group lasso tasks by applying an $L_{1,2}$ regularization on the weights of the autoencoder. The autoencoder in this method tries to map the inputs to a latent space and then reconstruct the inputs from that space. The $L_{1,2}$ regularization will optimize the weights (change them toward 0) to select a smaller number of features. The neural network structure of the autoencoder will enable the model to incorporate both linear and nonlinear behavior of the data in the results. After training this neural network, the features with higher weight values in the first layer can be selected as the most informative features. The authors claimed that this algorithm showed promising results in computer vision tasks.
- Principal feature analysis (PFA) [14]: This method selects features based on principal component analysis (PCA). The most important features are selected by applying a *k*-means clustering algorithm to the components of PCA and finding the features dominating each cluster (closest to the mean of the cluster). This algorithm is primarily designed for computer vision.
- Unsupervised feature selection for multicluster data (MCFS) [15]: This approach prioritizes the preservation of the multicluster structure of data. The algorithm involves constructing a nearest neighbor graph of the features, solving a sparse eigen-problem to find top eigenvectors with regard to the smallest eigenvalues. Then, an L_1-regularized least squares problem is optimized to find the linear weights between the features and the eigenvectors. This allows us to define the MCFS score as the maximum weight of each feature across different clusters and select the highest scores as the best features.
- Laplacian score (LS) [16]: The LS algorithm uses the nearest neighbor graph to capture the local structure of the data in the affinity matrix. For each feature, its adjusted variation is calculated by removing the feature's mean, normalized by a degree matrix, which itself is derived from the sum of similarities in the affinity matrix. The Laplacian matrix, essential for this calculation, is formed by subtracting the affinity matrix from the degree matrix. The significance of each feature is then assessed by the LS, which is the ratio of the feature's ability to preserve local information (captured by its adjusted variation's alignment with the Laplacian matrix) to its overall variance (measured by its alignment with the degree matrix). The lower the LS, the more relevant the feature for representing the intrinsic geometry of the data set.

We applied the LS, AEFS, PFA, MCFS, and CAE algorithms to a 67% training data set (split based on patients) of one-hot encoded features to select the best 100 features (N_{Best}=100) with the following specifications (we chose N_{Best} based on preliminary experimentations).

For the AEFS method, we used a single hidden-layer autoencoder and optimized the loss function as described in Han et al [13], with α=0.001 as the trade-off parameter of the reconstruction loss and the regularization term and β=0.1 as the penalty parameter for the weight decay regularization. The choice of these parameters was based on preliminary experimentations on a small set of data and exploring α and β of {0.001, 0.1, 1, 1000}.

For the PFA method, we used incremental PCA instead of the normal PCA in the original paper [14], with a batch size of $2N_{All}$ due to the high computational cost. We decomposed the data to NAll2 components and then applied *k*-means clustering to find N_{Best} clusters. We also tried { NAll5,NAll3,NAll2 } as the number of components of the PCA in the preliminary experiments.

To use the LS and MCFS methods for feature selection, we used the Euclidean distances between features to construct a nearest neighbor graph G based on the 5 nearest neighbors. For the LS method, we set the weights of the connected nodes of G to 1, assuming a large *t* in the LS formulation. Then, we computed the LS for each feature and selected the top features with higher scores. Due to the high computational resources required for the LS and MCFS methods, we did not explore different parameters and used the same settings suggested by the implementation codes of these algorithms.

As the structure of the loss function allows us to prioritize some target variables, the CAE method was applied in 2 different ways—with and without adjusting weights for features. The reason for adjusting the weights is that since there are many correlated features in the *ICD-10-CA* and ATC code data sets, the model may choose one of them randomly [3]. Therefore, we applied the function in equation 1 as the class weights of the features to the loss function of the model:

(1)WF=11+dF

where W_F is the weight for feature *F* and *d*(*F*) is the depth of feature *F* as a node of the *ICD-10-CA* or ATC tree. This weight adjustment will force the model to give more importance to the features at the top of the tree and to generalize more in clinical settings. In the rest of the paper, this variant of the CAE model will be referred to as the CAE with weight adjustment (CAEWW) and the regular CAE model will be referred to as the CAE with no weight adjustment (CAENW).

We defined N_{Best} neurons in the concrete selector layer and used a learning rate of 0.001, a batch size of 64, and 1000 epochs. We also controlled the learning of the concrete selector layer by the temperature parameter that started from 20 and decreased to 0.01 exponentially (this annealing strategy was suggested by Abid et al [6] for better convergence). The decoder of the CAE was a feed-forward neural network with 2 hidden layers with 64 neurons and used a sigmoid activation function for the output layer and a leaky rectified linear unit activation function for the other layers. The learning rate, number of neurons, and the layers were determined based on preliminary experiments for the fastest convergence of the autoencoder.

Evaluation of Selected Features: Reconstruction of Initial Feature Space

To evaluate the effectiveness of the selected features, we trained a simple feed-forward neural network model using the chosen features to reconstruct the original feature space for each data set separately. The neural network consisted of 2 hidden layers, each with 64 neurons, and used leaky rectified linear unit activation functions, with a 10% dropout rate, in the hidden layers and a sigmoid activation function in the output layer. We trained the model using the same training set used in the feature selection step and evaluated its performance on the remaining 33% test set using binary cross entropy. We also calculated the accuracy of each feature selection method to determine which method produced the most accurate results. One of the challenges in comparing models with a large number of targets is that the accuracy values are inflated, because most of the targets are heavily imbalanced (ie, most of them were 0s) and the models were able to predict them easily. To circumvent this issue, we used a 2-tailed *t* test analysis and compared the accuracy values of the classes with the accuracy of a baseline model that simply outputs the mode of the training data for each class regardless of the input.

Evaluation of Selected Features: Prediction of 90-Day Mortality

To demonstrate the utility of using unsupervised feature selection methods in a supervised setting, we conducted a case study where we used the selected features from each method to predict 90-day mortality following the end of the last procedure for each data set separately. Since our data sets were highly imbalanced, with only ~6%, ~2%, and ~1% of the aggregated records for the DAD, NACRS, and PIN data set, respectively, leading to 90-day mortality, we upsampled the minority class using random sampling to balance the training sets. We then trained extreme gradient boosting (XGBoost) models using the training sets with 5-fold cross-validation to tune the hyperparameters for each model. We used the best models to predict the binary outcome variables on the test sets and measured their performances. XGBoost was selected for its efficiency with sparse data, which was crucial for our data sets. XGBoost's regularization features help prevent overfitting [17]. Additionally, its ability to provide interpretable models through tree-based Shapley values aligns with our objective to not only predict mortality but also understand the contributing factors [18]. XGBoost's scalability on multiple processors and speed (for both training [17] and Shapley analysis [18]) are also beneficial for processing large volumes of data and complex model tuning. After training the mortality prediction models for each method and data set, we calculated tree-based Shapley values corresponding to the features. This allowed us to rank the importance of each feature and explain their roles in predicting mortality.

We have made the implementation code for the methods discussed available at our GitHub repository [19].

Ethical Considerations

This study received ethics approval from the Conjoint Health Research Ethics Board at the University of Calgary

(REB20-1879). Informed consent was waived due to the retrospective nature of the data and the large number of patients involved, making it impractical to seek consent from each patient. All data were deidentified. No compensation was provided to the participants as the study did not involve direct participant interaction.

Results

Data Set Description

Table 1 summarizes the characteristics of the patients in the cohort at the time of their initial catheterization. The total numbers of patients with at least 1 record in the respective data sets, as well as the time ranges for each data set, are provided in Table 2. The aggregation procedure described in the *Methods* section reduced the number of records to the values listed in the "Aggregated Records" row, and the table also includes the total number of codes (unique *ICD-10-CA* or ATC codes and their parent codes) in a data set, along with the average number of codes per record. Multimedia Appendix 1 illustrates the percentages of the 20 most common *ICD-10-CA* and ATC codes within each processed data set. Within the data set, there were 9942 cases corresponding to a 90-day mortality, resulting in a 20% mortality rate in the cohort. The final aggregated data for each data set were split into 67% for the training sets and 33% for the test sets at the patient level.

Table . Key characteristics of the patients with CAD[a] enrolled in the cohort.

Variable		Overall (N=51,506)
Total population, n (%)		51,506 (100)
Sex, n (%)		
	Female	12,875 (25)
	Male	38,631 (75)
Age (years), mean (SD)		66.09 (11.41)
BMI (kg/m ^{2}), mean (SD)		29.51 (7.45)
	Missing data, n (%)	8449 (16.4)
CAD type, n (%)		
	Non-ST elevation myocardial infarction	24,119 (46.83)
	Unstable angina	10,671 (20.72)
	Stable angina	9832 (19.09)
	Missing data	6884.0 (13.37)
Canadian Cardiovascular Society angina grade, n (%)		
	II (slight limit)	4688 (9.1)
	IVb	7513 (14.59)
	IVa (hospitalized with acute coronary syndrome)	21,117 (41)
	III (marked limit)	2581 (5.01)
	IVc	1627 (3.16)
	I (strenuous)	1309 (2.54)
	Atypical	698 (1.36)
	Other or missing data	11,973 (23.25)
Diabetes, n (%)		
	No diabetes	37,544 (72.89)
	Type II	12,067 (23.43)
	Type I	806 (1.56)
	Other	1089 (2.11)
Dyslipidemia, n (%)		32,967 (64.01)
Heart failure, n (%)		3689 (7.16)
Atrial fibrillation or flutter, n (%)		1220 (2.37)
Hypertension, n (%)		32,264 (62.64)
Angina, n (%)		2559 (4.97)
Family history of CAD, n (%)		15,209 (29.53)
Smoking, n (%)		
	Never	25,822 (50.13)
	Current	11,196 (21.74)
	Past	14,488 (28.13)
Chronic lung disease, n (%)		5318 (10.33)
Cerebrovascular disease, n (%)		2040 (3.96)
Psychiatric history, n (%)		1097 (2.13)
Venous insufficiency, n (%)		476 (0.92)
Alcohol consumption, n (%)		599 (1.16)

Variable		Overall (N=51,506)
Extent of CAD, n (%)		
	3 VDs[b]	247 (0.48)
	3 VDs (one >75%)	7765 (15.08)
	3 VDs (>75% proximal LAD[c])	5704 (11.07)
	3 VDs (proximal LAD)	3318 (6.44)
	2 VDs	5392 (10.47)
	2 VDs (>75% LAD)	569 (1.1)
	2 VDs (both >75%)	5215 (10.13)
	2 VDs (>75% proximal LAD)	2819 (5.47)
	1 VD (>75% proximal LAD)	2299 (4.46)
	1 VD (>75%)	8504 (16.51)
	1 VD (50% - 75%)	4032 (7.83)
	Severe left main disease	3058 (5.94)
	Left main disease	2584 (5.02)

[a]CAD: coronary artery disease.

[b]VD: vessel disease.

[c]LAD: left anterior descending.

Table . Summary statistics of the DAD[a], NACRS[b], and PIN[c] data sets.

Summary statistics	Data set		
	DAD	NACRS	PIN
Patients with at least 1 record, n	49,075	50,628	49,052
Records, n	273,910	3,974,403	28,807,136
Aggregated records, n	166,083	173,507	997,997
Unique *ICD-10-CA*[d] or ATC[e] codes and their parent codes, n	9651	7803	2315
Codes per aggregated record, mean (SD)	24.90 (16.55)	15.27 (12.55)	33.31 (18.95)
Time range	2004 - 2022	2010 - 2022	2004 - 2022

[a]DAD: Discharge Abstract Database.

[b]NACRS: National Ambulatory Care Reporting System.

[c]PIN: Pharmaceutical Information Network.

[d]*ICD-10-CA*: *International Classification of Diseases, Tenth Revision, Canada*.

[e]ATC: Anatomical Therapeutic Chemical.

Performances of the Feature Selection Methods

Table 3 shows the accuracies and binary cross entropies of the models based on the selected features from each method. Table 4 shows the accuracy, F_1-score, and area under the receiver operating characteristic curve (AUC-ROC) metrics of the XGBoost models to predict 90-day mortality.

Table . Average accuracy and binary cross entropy (BCE) loss of different sets of selected features in reconstructing the original feature space in a neural network structure.

Feature selection method	DAD[a]		NACRS[b]		PIN[c]	
	Accuracy, mean (95% CI)	BCE, mean (95% CI)	Accuracy, mean (95% CI)	BCE, mean (95% CI)	Accuracy, mean (95% CI)	BCE, mean (95% CI)
CAEWW[d]	0.9992[e] (0.9992-0.9993)	0.0121[e] (0.0121-0.0121)	0.9994[e] (0.9994-0.9995)	0.0091[e] (0.0091-0.0091)	0.9972[e] (0.9969-0.9975)	0.0432[e] (0.0432-0.0432)
CAENW[f]	0.9992[e] (0.9991-0.9993)	0.0121[e] (0.0121-0.0121)	0.9994[e] (0.9993-0.9994)	0.0094[e] (0.0094-0.0094)	0.9972[e] (0.9969-0.9974)	0.0438[e] (0.0438-0.0438)
AEFS[g]	0.9976 (0.9972-0.9980)	0.0370 (0.0370-0.0370)	0.9982 (0.9979-0.9985)	0.0274 (0.0274-0.0274)	0.9884 (0.9867-0.9901)	0.1794 (0.1794-0.1794)
MCFS[h]	0.9991[e] (0.9990-0.9991)	0.0145[e] (0.0145-0.0145)	0.9992[e] (0.9992-0.9993)	0.0117[e] (0.0117-0.0117)	0.9956[e] (0.9951-0.9962)	0.0677[e] (0.0677-0.0677)
PFA[i]	0.9975 (0.9971-0.9979)	0.0382 (0.0382-0.0382)	0.9981 (0.9978-0.9985)	0.0286 (0.0286-0.0286)	0.9871 (0.9852-0.9891)	0.1982 (0.1982-0.1982)
LS[j]	0.9989[e] (0.9988-0.9990)	0.0165[e] (0.0165-0.0165)	0.9991[e] (0.9990-0.9992)	0.0136[e] (0.0136-0.0136)	0.9945[e] (0.9938-0.9952)	0.0850[e] (0.0850-0.0850)
Mode of the training set (baseline model)	0.9975 (0.9971-0.9979)	0.0384 (0.0322-0.0447)	0.9981 (0.9978-0.9984)	0.0294 (0.0245-0.0342)	0.9870 (0.9850-0.9889)	0.2012 (0.1712-0.2312)

[a]DAD: Discharge Abstract Database.

[b]NACRS: National Ambulatory Care Reporting System.

[c]PIN: Pharmaceutical Information Network.

[d]CAEWW: concrete autoencoder with weight adjustment.

[e]Significantly different from the baseline model that outputs the mode of each class ($P<.05$). The P values are presented in Table S1 of Multimedia Appendix 2.

[f]CAENW: concrete autoencoder with no weight adjustment.

[g]AEFS: autoencoder-inspired unsupervised feature selection.

[h]MCFS: unsupervised feature selection for multicluster data.

[i]PFA: principal feature analysis.

[j]LS: Laplacian score.

Table . Performance of the extreme gradient boosting (XGBoost) model in predicting 90-day mortality using different sets of selected features.

Feature selection method	DAD[a]			NACRS[b]			PIN[c]		
	Accuracy	F_1-score	AUC-ROC[d]	Accuracy	F_1-score	AUC-ROC	Accuracy	F_1-score	AUC-ROC
CAEWW[e]	0.86	0.37	0.87	0.85	0.15	0.75	0.84	0.1	0.82
CAENW[f]	0.86	0.36	0.87[g]	0.86	0.15	0.75	0.83	0.1	0.82
AEFS[h]	0.88	0.21	0.61[g]	0.9	0.09	0.56[g]	0.85	0.08	0.69[g]
MCFS[i]	0.86	0.37	0.89[g]	0.84	0.13	0.74	0.81	0.09	0.84[g]
PFA[j]	0.92	0.05	0.5[g]	0.97	0.02	0.5[g]	0.93	0.05	0.54[g]
LS[k]	0.77	0.26	0.81[g]	0.8	0.11	0.73[g]	0.76	0.08	0.82

[a]DAD: Discharge Abstract Database.

[b]NACRS: National Ambulatory Care Reporting System.

[c]PIN: Pharmaceutical Information Network.

[d]AUC-ROC: area under the receiver operating characteristic curve.

[e]CAEWW: concrete autoencoder with weight adjustment.

[f]CAENW: concrete autoencoder with no weight adjustment.

[g]Significantly different from the AUC-ROC of the model trained on CAEWW features in their corresponding data set (P<.05) using the DeLong test [20]. The P values are presented in Table S2 in Multimedia Appendix 2.

[h]AEFS: autoencoder-inspired unsupervised feature selection.

[i]MCFS: unsupervised feature selection for multicluster data.

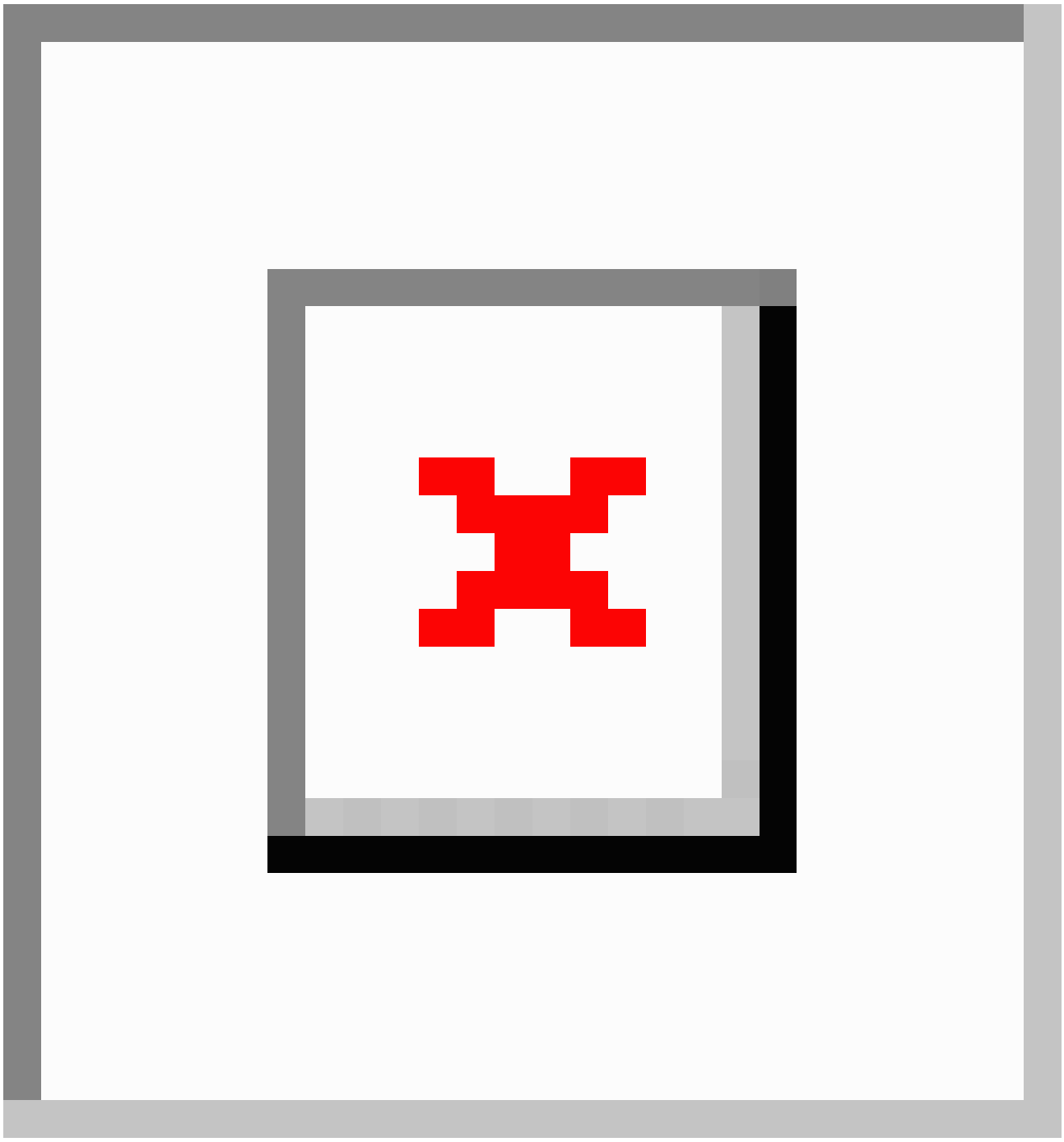
[j]PFA: principal feature analysis.

[k]LS: Laplacian score.

Both tables indicate that the CAE methods generally selected superior features compared to the other algorithms. Adjusting the weights within the CAE improved the performance of feature space reconstruction slightly. In terms of predicting 90-day mortality, the CAE methods again performed better than the other methods, as evidenced by the AUC-ROC. This superior performance was statistically significant in most instances (P<.05), according to the DeLong test for AUC-ROC. Furthermore, the McNemar test revealed a significant difference between the overall performance of the mortality prediction models trained on the features of the CAEWW method and those based on the other methods (P<.05). The P values of the McNemar and DeLong tests can be found in Table S2 of Multimedia Appendix 2.

Figure 1 shows the log-scale histograms of the original feature space reconstruction accuracy in each *ICD-10-CA* or ATC code for different feature selection methods. It shows that CAEWW and CAENW were the best methods in terms of reconstructing the majority of the features with high accuracy. The other methods, despite having high average accuracy, performed poorly in reconstructing some of the features.

Figure 1. Log-scale histograms of the initial feature space reconstruction accuracy in each International Classification of Diseases (ICD) code for different feature selection methods and different data sets: (A) concrete autoencoder with weight adjustment (CAEWW), (B) concrete autoencoder with no weight adjustment (CAENW), (C) autoencoder-inspired unsupervised feature selection (AEFS), (D) unsupervised feature selection for multicluster data (MCFS), (E) principal feature analysis (PFA), and (F) Laplacian score (LS). DAD: Discharge Abstract Database; NACRS: National Ambulatory Care Reporting System; PIN: Pharmaceutical Information Network.



Characteristics of the Selected Features

We also calculated the average depths of the codes selected by each method and compared them (using 2-tailed *t* test analysis) against the CAEWW method that is intended to select more general codes (ie, smaller depths). The CAEWW method selected codes with average depths of 1.38, 1.42, and 1.99 for the DAD, NACRS, and PIN data sets, respectively. Although CAEWW's code depths were significantly lower than the depths of the codes selected by the other methods (all $P<.05$), there was no significant difference between the CAEWW and CAENW methods (with average depths of 1.48, 1.45, 2.03; all $P>.05$). The *P* values can be found in Table S3 of Multimedia Appendix 2. Figure 2 illustrates the difference in average code depth among the different methods.

We used the average of the mean absolute Shapley values from each mortality prediction model as an index for the importance of the features selected by each method. To compare the CAEWW method with the other methods across the 3 different data sets, we conducted a 2-tailed *t* test analysis. The CAEWW method did not show any significant difference from the

CAENW methods across all data sets (all $P>.05$). However, the CAEWW method did yield significantly higher mean absolute Shapley values compared to the other methods (AEFS and PFA; all $P<.001$). The AEFS and PFA methods had the lowest Shapley values compared to all the other methods, indicating that they selected lower-quality features for this task. The P values are available in Table S4 in Multimedia Appendix 2. Figure 3 illustrates the aforementioned differences in mean absolute Shapley values. Figure 4 shows the Shapley plots of the 20 most important features selected by the CAEWW method across different data sets. The corresponding Shapley plots for the other methods can be found in Multimedia Appendix 3 (Figures S1-S5). Additionally, Multimedia Appendix 4 (Tables S1-S18) includes all chosen features, detailed descriptions, and the average absolute Shapley values across all data sets and methods.

Figure 2. Average depths of the selected codes by each method in the *ICD-10-CA* or ATC tree. Methods with average depths significantly ($P<.05$) larger than the CAEWW method in their corresponding data set are marked with asterisks (*). AEFS: autoencoder-inspired unsupervised feature selection; ATC: Anatomical Therapeutic Chemical CAENW: concrete autoencoder with no weight adjustment; CAEWW: concrete autoencoder with weight adjustment; DAD: Discharge Abstract Database; *ICD-10-CA*: *International Classification of Diseases, Tenth Revision, Canada*; LS: Laplacian score; MCFS: unsupervised feature selection for multicluster data; NACRS: National Ambulatory Care Reporting System; PFA: principal feature analysis; PIN: Pharmaceutical Information Network.

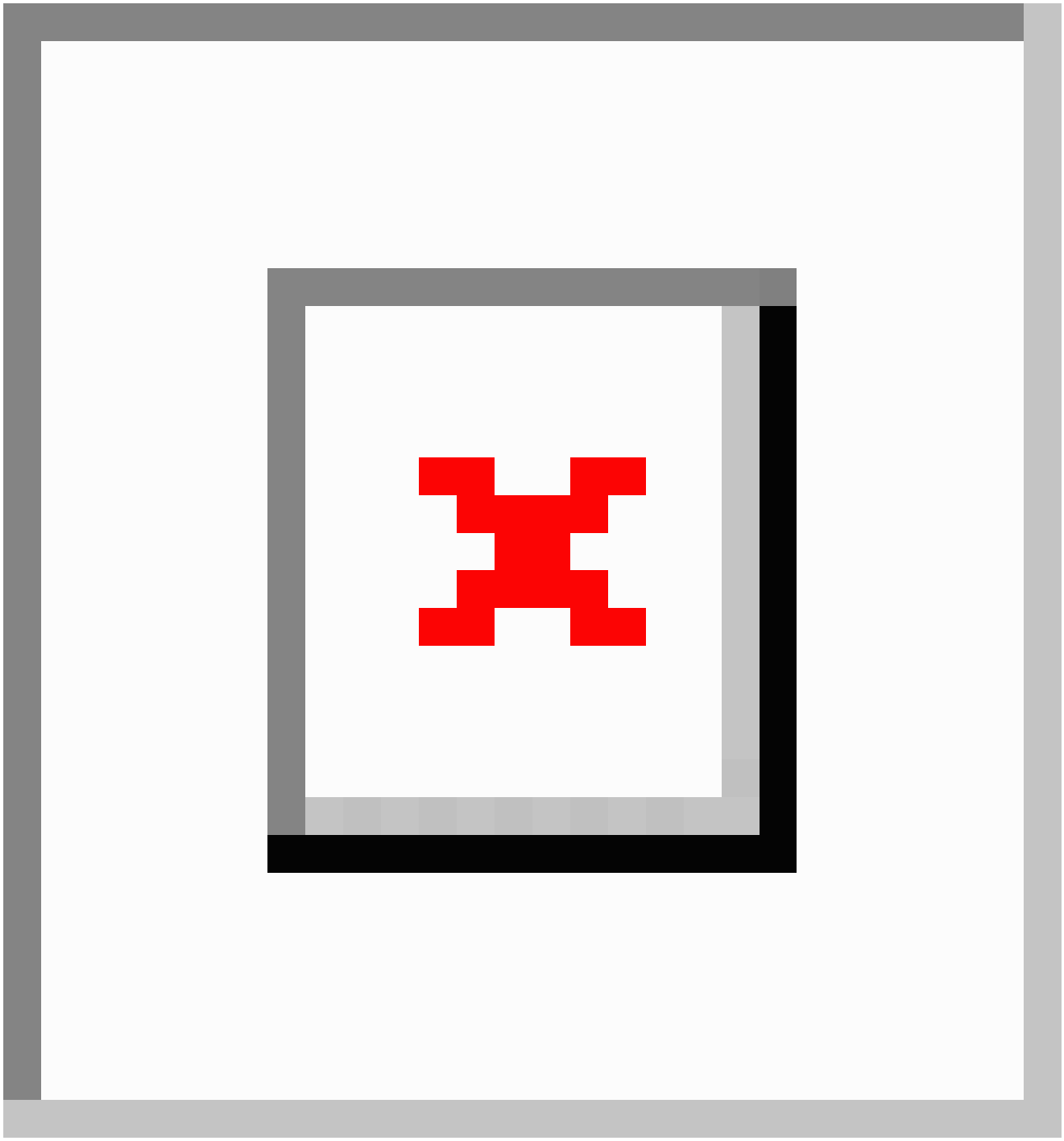


Figure 3. Average of mean absolute Shapley values of features in each mortality prediction model. Methods with average values significantly (P<.05) smaller than the CAEWW method in their corresponding data set are marked with asterisks (*). AEFS: autoencoder-inspired unsupervised feature selection; CAENW: concrete autoencoder with no weight adjustment; CAEWW: concrete autoencoder with weight adjustment; DAD: Discharge Abstract Database; LS: Laplacian score; MCFS: unsupervised feature selection for multicluster data; NACRS: National Ambulatory Care Reporting System; PFA: principal feature analysis; PIN: Pharmaceutical Information Network.

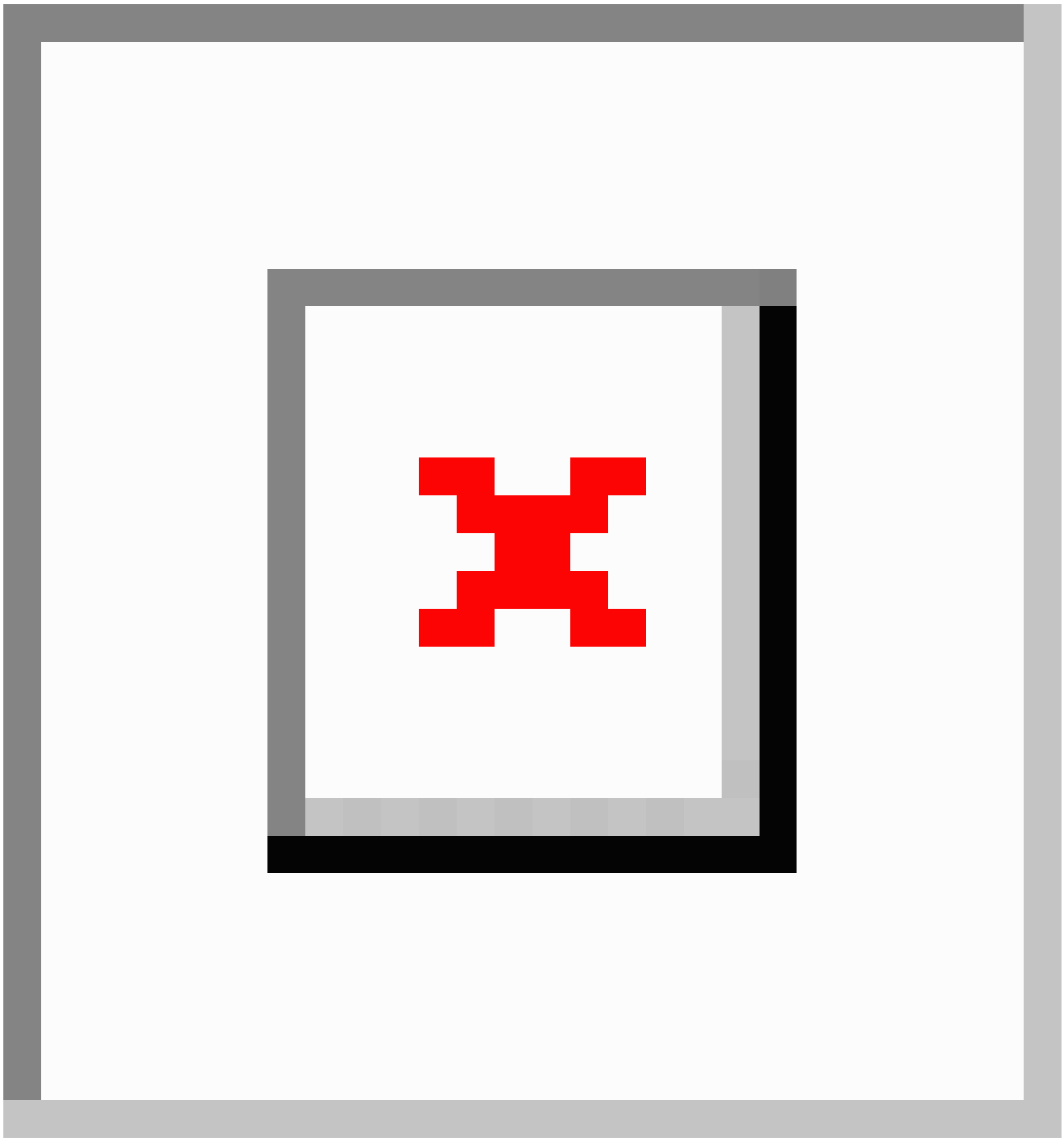
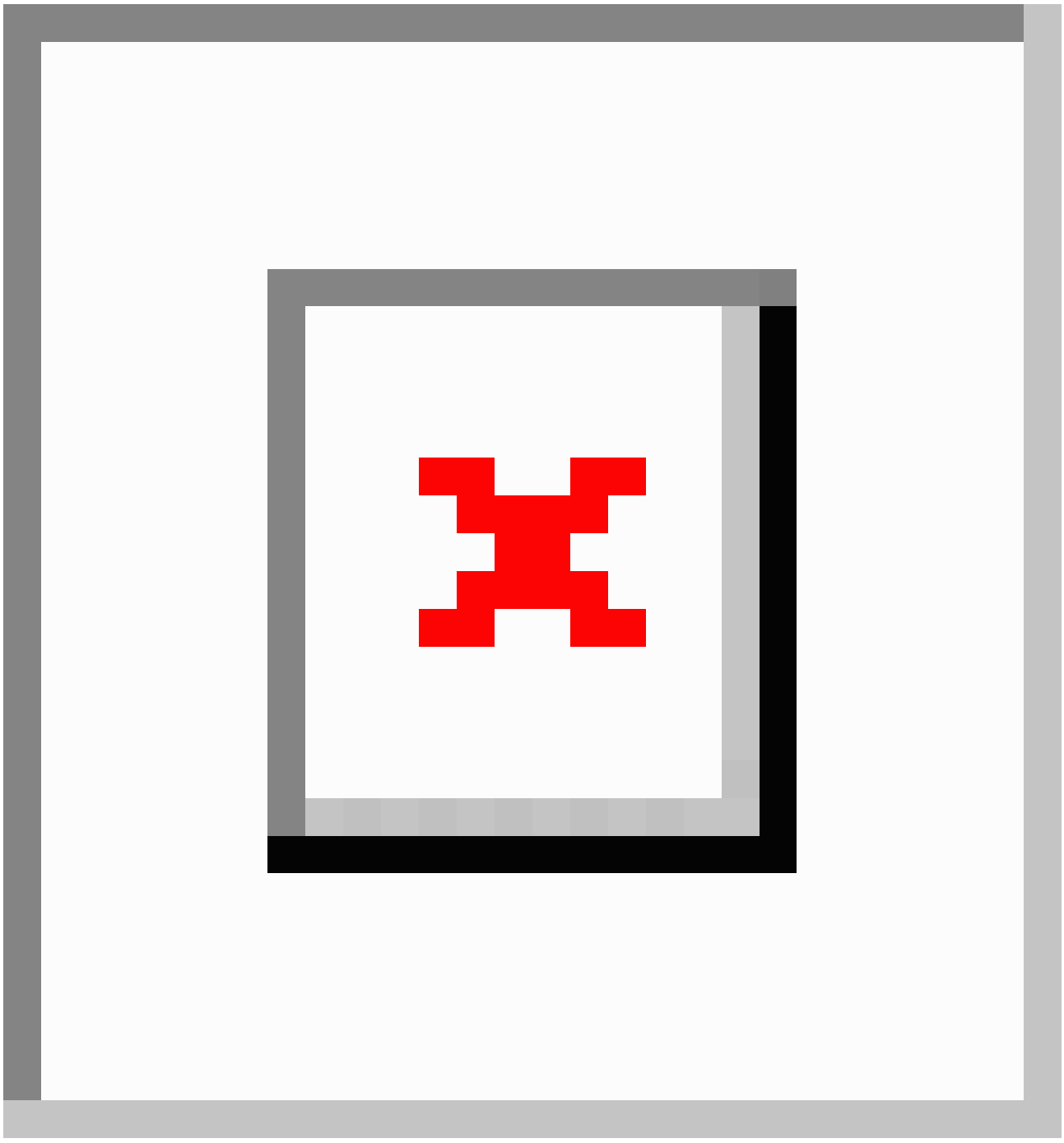


Figure 4. SHAP values of the features selected by the CAEWW method across different data sets (20 most important features): (A) Discharge Abstract Database (DAD), (B) National Ambulatory Care Reporting System (NACRS), and (C) Pharmaceutical Information Network (PIN). CAEWW: concrete autoencoder with weight adjustment; SHAP: Shapley additive explanations.



Discussion

Principal Findings

The high dimensionality of the ICD and ATC code databases necessitates the use of dimensionality reduction techniques to feed the data into machine learning models. Due to interpretability concerns in the health domain, selecting from original features, rather than transforming them into new features, is an essential step in reducing dimensionality. In this study, we demonstrated that the CAE methods performed the best in selecting the most informative ICD and ATC codes in an unsupervised setting. Using a clinical outcome as a case study, we also demonstrated that ICD and ATC code features selected by the CAE methods were able to predict the outcome variable with better accuracy than the other methods in the study, even though they were derived from an unsupervised setting in the absence of the target variable. This indicates that the selected features can be considered unbiased toward a specific target variable and explain the phenomenon appropriately. We also showed that the AEFS and PFA methods did not select high-quality features in our data set and were not suitable for both tasks of reconstructing the feature space and predicting 90-day mortality. The LS and MCFS methods, however, showed better performance in both tasks (slightly lower than the CAE methods). Furthermore, the features selected

by the CAE methods (especially CAEWW) were generally higher-level codes (ie, lower depth in the hierarchical structure), which helps the study to find generalized solutions.

It is worth mentioning that our methodology code is publicly shared, allowing other researchers to use the desired methods for selecting the most informative features within cohorts with large ICD, ATC, or other hierarchical-coded health databases [19].

Computational Cost

The MCFS, PFA, and LS methods had multiple special matrix operations that made them computationally expensive. Considering our large-scale, high-dimensional data set, these algorithms were not possible to run on a personal computer and we had to optimize the operations for an advanced computing cluster with 40 Intel Xeon Gold 6342 2.80 GHz CPUs and 2048 GB RAM. The MCFS and LS feature selection experiments had some shared operations and together took over 2 days to complete. The PFA method also needed less than a day for the entire feature selection experiments. The AEFS and CAE methods, however, had the advantage of using GPUs and optimized deep learning libraries for training the neural networks and were faster. Each of these algorithms took less than 4 hours on an Nvidia A100 80 GB GPU.

Selected Features

The Shapely analysis of the 20 most important features selected in each data set using the CAEWW method for predicting mortality revealed the multidimensional capabilities of this method in identifying relevant information. In the DAD and NACRS data sets, it selected disease codes relevant to mortality among patients with CAD. In both data sets, diseases related to cardiovascular conditions, hypertensive and circulatory disorders, metabolic disorders, renal failures [21], and cancer [22] were selected, which are important factors in the outcomes of patients with CAD. Furthermore, DAD-based features included accidents, arthropathies, and hospitalization-specific conditions, whereas the NACRS data set resulted in falls [23]; digestive disorders [24]; and codes related to the rehabilitations, management, or complexity of the diseases. In the PIN data set, direct interventions for CAD and related risk factors were mainly selected, including high-ceiling diuretics, statins, ace inhibitors, angiotensin II receptor blockers (plain and combinations), direct factor Xa inhibitors, vasodilators, antianemic preparations, antithrombotic agents, and other lipid-modifying agents, addressing heart failure, cholesterol management, blood pressure control, anticoagulation, anemia, and blood flow. Also, drugs related to accompanying diseases or conditions with CAD were selected: gastrointestinal issues (pantoprazole and general drugs for acid-related disorders [25,26] and drugs for constipation [27]), pain management (opioids, other analgesics, and antipyretics) [28], inflammatory conditions and immune responses (anti-inflammatory and antirheumatic products and corticosteroids for systemic use) [29,30], mental and behavioral health (antipsychotics) [31], respiratory conditions (adrenergic inhalants [32]), and urological issues [33].

Previous studies typically selected codes as machine learning model features based on expert opinions, the presence of high-level codes (eg, categories or chapters), or a combination of both [34,35]. To the best of our knowledge, only 1 study [3] attempted to offer a sophisticated feature selection method using tree-lasso regularization for ICD code data sets, but it was in a supervised setting that required an outcome variable. Our study provides a general tool for health researchers to select the most informative ICD and ATC codes without biasing the study toward a specific outcome variable. We also introduced a unique target weight adjustment function to the CAE model to guide the model to select higher levels of the ICD table compared to the model without adjustment.

Limitations and Future Work

One of the limitations of this study was the incapability of the CAE method to select an exact number of desired features. Since the neurons in the concrete selector layer work independently, there is a possibility of selecting duplicate features. Therefore, the number of final selected features can be fewer than the desired number. Although it indicates that the decoder model is still capable of reconstructing the initial feature space with a smaller number of features, some researchers may prefer to have an exact number of features they desire for their models. One previous study [36] has used a special regularization term in the training step to enforce the model not to select duplicate features. This method can be investigated for the ICD and ATC codes in the future.

The aggregation of codes should be viewed as a trade-off in this study. We needed to select a reasonable aggregation period that covers both long-term and short-term diseases. A shorter period could skew the results by including multiple correlated records from the same patient. Conversely, longer periods could weigh short-term diseases equally with long-term ones, and the codes of the patients with fewer records (eg, recent patients in the cohort) would have a lower chance of selection.

Another limitation was that we only used 3 data sets of a specific disease cohort to choose the features. Therefore, the selected features in this study may not generalize to other patient cohorts or diseases. Furthermore, we selected the 100 best features, but other data sets or patient cohorts may require a different number of features. Future studies may investigate the impact of the number of features on the results. Moreover, our hyperparameter analysis was conducted within a constrained scope due to limited computational resources. Future studies could further explore the impact of a broader range of hyperparameter values. We anticipate that the CAEs hold potential for this area due to their flexible neural network structure and optimized algorithms. A similar limitation also applies to the mortality prediction case study, where we only trained XGBoost and did not explore other model types.

Conclusions

In this study, we investigated 5 different methods for selecting the best features in ICD and ATC code data sets in an unsupervised setting. We demonstrated that the CAE method can select better features representing the whole data set that can be useful in further machine learning studies. We also introduced weight adjustment of the CAE method for ICD and ATC code data sets that can be useful in the generalizability

and interpretability of the models, given that it prioritizes selecting high-level definitions of diseases.

The CAEWW method outperformed all other methods in reconstructing the initial feature space across all data sets. We validated the selected features through a supervised learning task, predicting 90-day mortality after discharge using 3 distinct data sets. Features selected via the CAEWW method demonstrated significantly improved performance on this task, as evidenced by the DeLong and McNemar tests. Given the advantages of the CAE method, we recommend its use in the feature selection phase of EHD analysis with ICD or ATC codes.

Acknowledgments

This study was supported by the Libin Cardiovascular Institute PhD Graduate Scholarship, the Alberta Innovates Graduate Scholarship, and a Project Grant from the Canadian Institutes of Health Research (PJT 178027).

Authors' Contributions

PG led the conceptualization of the study, designed the research methodology, executed the study, performed the data analysis, and drafted the original manuscript. JL provided the data set essential for the research, contributed to study supervision, and critically reviewed and revised the manuscript. Both authors read and approved the final manuscript.

Conflicts of Interest

JL is the chief technology officer, a co-founder, and a major shareholder of Symbiotic AI, Inc. The authors have no further conflicts of interest to declare.

Multimedia Appendix 1
The percentages of the 20 most common *ICD-10-CA* and ATC codes present in the processed data sets. ATC: Anatomical Therapeutic Chemical; *ICD-10-CA*: *International Classification of Diseases, Tenth Revision, Canada*.
[PNG File, 70 KB - medinform_v12i1e52896_app1.png]

Multimedia Appendix 2
P values associated with Tables 3 and 4, and characteristics of the selected features.
[DOCX File, 21 KB - medinform_v12i1e52896_app2.docx]

Multimedia Appendix 3
Shapley value plots of the features selected by our methods across the different data sets (20 most important features).
[DOCX File, 925 KB - medinform_v12i1e52896_app3.docx]

Multimedia Appendix 4
Tables of the features selected by our methods across the different data sets with each feature’s description, rank, and average absolute Shapley score.
[DOCX File, 187 KB - medinform_v12i1e52896_app4.docx]

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 2;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
2. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv 2023 Nov 23;55(1):1-36. [doi: 10.1145/3477600]
3. Kamkar I, Gupta SK, Phung D, Venkatesh S. Stable feature selection for clinical prediction: exploiting icd tree structure using Tree-Lasso. J Biomed Inform 2015 Feb;53:277-290. [doi: 10.1016/j.jbi.2014.11.013] [Medline: 25500636]
4. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med 2021 Oct 28;4(1):153. [doi: 10.1038/s41746-021-00521-5] [Medline: 34711924]
5. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. A review of unsupervised feature selection methods. Artif Intell Rev 2020 Feb;53(2):907-948. [doi: 10.1007/s10462-019-09682-y]
6. Abid A, Balin MF, Zou J. Concrete autoencoders for differentiable feature selection and reconstruction. arXiv. Preprint posted online on Jan 27, 2019. [doi: 10.48550/arXiv.1901.09346]
7. World Health Organization. International Statistical Classification of Diseases and Related Health Problems: Alphabetical Index: World Health Organization; 2004.
8. Yan C, Fu X, Liu X, et al. A survey of automated international classification of diseases coding: development, challenges, and applications. Intell Med 2022 Aug;2(3):161-173. [doi: 10.1016/j.imed.2022.03.003]

9. World Health Organization, Canadian Institute for Health Information. International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Canada (ICD-10-CA): Tabular List: Canadian Institute for Health Information; 2015.
10. Structure and principles. WHO Collaborating Centre for Drug Statistics Methodology. URL: https://www.whocc.no/atc/structure_and_principles/ [accessed 2023-07-30]
11. Ghali WA, Knudtson ML. Overview of the alberta provincial project for outcome assessment in coronary heart disease. on behalf of the APPROACH investigators. Can J Cardiol 2000 Oct;16(10):1225-1230. [Medline: 11064296]
12. Maddison CJ, Mnih A, Teh YW. The concrete distribution: a continuous relaxation of discrete random variables. arXiv. Preprint posted online on Nov 2, 2017. [doi: 10.48550/arXiv.1611.00712]
13. Han K, Wang Y, Zhang C, Li C, Xu C. Autoencoder inspired unsupervised feature selection. Presented at: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Apr 15 to 20, 2018:; Calgary, AB p. 2941-2945. [doi: 10.1109/ICASSP.2018.8462261]
14. Lu Y, Cohen I, Zhou XS, Tian Q. Feature selection using principal feature analysis. In: MM '07: Proceedings of the 15th ACM International Conference on Multimedia: Association for Computing Machinery; 2007:301-304. [doi: 10.1145/1291233.1291297]
15. Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data. In: KDD '10: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Association for Computing Machinery; 2010:333-342. [doi: /10.1145/1835804.1835848]
16. He X, Cai D, Niyogi P. Laplacian score for feature selection. In: Weiss Y, Schölkopf B, Platt J, editors. Advances in Neural Information Processing Systems 18 (NIPS 2005): MIT Press; 2005. URL: https://papers.nips.cc/paper_files/paper/2005/hash/b5b03f06271f8917685d14cea7c6c50a-Abstract.html [accessed 2024-07-15]
17. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Association for Computing Machinery; 2016:785-794. [doi: 10.1145/2939672.2939785]
18. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
19. Ghasemi P. Unsupervised feature selection to identify important ICD-10 and ATC codes for machine learning. : GitHub URL: https://github.com/data-intelligence-for-health-lab/ICD10-Unsupervised-Feature-Selection [accessed 2023-09-16]
20. Sun X, Xu W. Fast implementation of Delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett 2014 Nov;21(11):1389-1393. [doi: 10.1109/LSP.2014.2337313]
21. Hajar R. Risk factors for coronary artery disease: historical perspectives. Heart Views 2017;18(3):109-114. [doi: 10.4103/HEARTVIEWS.HEARTVIEWS_106_17] [Medline: 29184622]
22. Mamas MA, Brown SA, Sun LY. Coronary artery disease in patients with cancer: it's always the small pieces that make the bigger picture. Mayo Clin Proc 2020 Sep;95(9):1819-1821. [doi: 10.1016/j.mayocp.2020.07.006] [Medline: 32861320]
23. Denfeld QE, Turrise S, MacLaughlin EJ, et al. Preventing and managing falls in adults with cardiovascular disease: a scientific statement from the American Heart Association. Circ Cardiovasc Qual Outcomes 2022 Jun;15(6):e000108. [doi: 10.1161/HCQ.0000000000000108] [Medline: 35587567]
24. Gesualdo M, Scicchitano P, Carbonara S, et al. The association between cardiac and gastrointestinal disorders: causal or casual link? J Cardiovasc Med (Hagerstown) 2016 May;17(5):330-338. [doi: 10.2459/JCM.0000000000000351] [Medline: 26702598]
25. Ariel H, Cooke JP. Cardiovascular risk of proton pump inhibitors. Methodist Debakey Cardiovasc J 2019;15(3):214-219. [doi: 10.14797/mdcj-15-3-214] [Medline: 31687101]
26. Sherwood MW, Melloni C, Jones WS, Washam JB, Hasselblad V, Dolor RJ. Individual proton pump inhibitors and outcomes in patients with coronary artery disease on dual antiplatelet therapy: a systematic review. J Am Heart Assoc 2015 Oct 29;4(11):e002245. [doi: 10.1161/JAHA.115.002245] [Medline: 26514161]
27. Ishiyama Y, Hoshide S, Mizuno H, Kario K. Constipation-induced pressor effects as triggers for cardiovascular events. J Clin Hypertens (Greenwich) 2019 Mar;21(3):421-425. [doi: 10.1111/jch.13489] [Medline: 30761728]
28. Majeed MH, Ali AA, Khalil HA, Bacon D, Imran HM. A review of the pharmacological management of chronic pain in patients with heart failure. Innov Clin Neurosci 2019 Nov 1;16(11-12):25-27. [Medline: 32082939]
29. Baoqi Y, Dan M, Xingxing Z, et al. Effect of anti-rheumatic drugs on cardiovascular disease events in rheumatoid arthritis. Front Cardiovasc Med 2021 Feb 3;8:812631. [doi: 10.3389/fcvm.2021.812631] [Medline: 35187113]
30. Sholter DE, Armstrong PW. Adverse effects of corticosteroids on the cardiovascular system. Can J Cardiol 2000 Apr;16(4):505-511. [Medline: 10787466]
31. Shulman M, Miller A, Misher J, Tentler A. Managing cardiovascular disease risk in patients treated with antipsychotics: a multidisciplinary approach. J Multidiscip Healthc 2014 Oct 31;7:489-501. [doi: 10.2147/JMDH.S49817] [Medline: 25382979]
32. Cazzola M, Matera MG, Donner CF. Inhaled beta2-adrenoceptor agonists: cardiovascular safety in patients with obstructive lung disease. Drugs 2005;65(12):1595-1610. [doi: 10.2165/00003495-200565120-00001] [Medline: 16060696]

33. Son YJ, Kwon BE. Overactive bladder is a distress symptom in heart failure. Int Neurourol J 2018 Jun;22(2):77-82. [doi: 10.5213/inj.1836120.060] [Medline: 29991228]
34. Jamian L, Wheless L, Crofford LJ, Barnado A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. Arthritis Res Ther 2019 Dec 30;21(1):305. [doi: 10.1186/s13075-019-2092-7] [Medline: 31888720]
35. Lucini FR, Stelfox HT, Lee J. Deep learning-based recurrent delirium prediction in critically ill patients. Crit Care Med 2023 Apr 1;51(4):492-502. [doi: 10.1097/CCM.0000000000005789] [Medline: 36790184]
36. Strypsteen T, Bertrand A. End-to-end learnable EEG channel selection for deep neural networks with Gumbel-softmax. J Neural Eng 2021 Jul 20;18(4). [doi: 10.1088/1741-2552/ac115d] [Medline: 34225257]

Abbreviations

AEFS: autoencoder-inspired unsupervised feature selection
APPROACH: Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease
ATC: Anatomical Therapeutic Chemical
AUC-ROC: area under the receiver operating characteristic curve
CAD: coronary artery disease
CAE: concrete autoencoder
CAENW: concrete autoencoder with no weight adjustment
CAEWW: concrete autoencoder with weight adjustment
DAD: Discharge Abstract Database
EHD: electronic health data
ICD: International Classification of Diseases
ICD-10: International Classification of Diseases, Tenth Revision
ICD-10-CA: International Classification of Diseases, Tenth Revision, Canada
LS: Laplacian score
MCFS: unsupervised feature selection for multicluster data
NACRS: National Ambulatory Care Reporting System
PCA: principal component analysis
PFA: principal feature analysis
PIN: Pharmaceutical Information Network
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 19.09.23; peer-reviewed by TA El-Hafeez, T Wang; revised version received 06.06.24; accepted 08.06.24; published 26.07.24.

Please cite as:
Ghasemi P, Lee J
Unsupervised Feature Selection to Identify Important ICD-10 and ATC Codes for Machine Learning on a Cohort of Patients With Coronary Heart Disease: Retrospective Study
JMIR Med Inform 2024;12:e52896
URL: https://medinform.jmir.org/2024/1/e52896
doi:10.2196/52896

Generalization of a Deep Learning Model for Continuous Glucose Monitoring–Based Hypoglycemia Prediction: Algorithm Development and Validation Study

Jian Shao[1], PhD; Ying Pan[2], PhD; Wei-Bin Kou[3], PhD; Huyi Feng[4], PhD; Yu Zhao[1], BBA; Kaixin Zhou[1], PhD; Shao Zhong[2], PhD

1

2

3

4

Corresponding Author:
Shao Zhong, PhD

Abstract

Background: Predicting hypoglycemia while maintaining a low false alarm rate is a challenge for the wide adoption of continuous glucose monitoring (CGM) devices in diabetes management. One small study suggested that a deep learning model based on the long short-term memory (LSTM) network had better performance in hypoglycemia prediction than traditional machine learning algorithms in European patients with type 1 diabetes. However, given that many well-recognized deep learning models perform poorly outside the training setting, it remains unclear whether the LSTM model could be generalized to different populations or patients with other diabetes subtypes.

Objective: The aim of this study was to validate LSTM hypoglycemia prediction models in more diverse populations and across a wide spectrum of patients with different subtypes of diabetes.

Methods: We assembled two large data sets of patients with type 1 and type 2 diabetes. The primary data set including CGM data from 192 Chinese patients with diabetes was used to develop the LSTM, support vector machine (SVM), and random forest (RF) models for hypoglycemia prediction with a prediction horizon of 30 minutes. Hypoglycemia was categorized into mild (glucose=54-70 mg/dL) and severe (glucose<54 mg/dL) levels. The validation data set of 427 patients of European-American ancestry in the United States was used to validate the models and examine their generalizations. The predictive performance of the models was evaluated according to the sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

Results: For the difficult-to-predict mild hypoglycemia events, the LSTM model consistently achieved AUC values greater than 97% in the primary data set, with a less than 3% AUC reduction in the validation data set, indicating that the model was robust and generalizable across populations. AUC values above 93% were also achieved when the LSTM model was applied to both type 1 and type 2 diabetes in the validation data set, further strengthening the generalizability of the model. Under different satisfactory levels of sensitivity for mild and severe hypoglycemia prediction, the LSTM model achieved higher specificity than the SVM and RF models, thereby reducing false alarms.

Conclusions: Our results demonstrate that the LSTM model is robust for hypoglycemia prediction and is generalizable across populations or diabetes subtypes. Given its additional advantage of false-alarm reduction, the LSTM model is a strong candidate to be widely implemented in future CGM devices for hypoglycemia prediction.

(JMIR Med Inform 2024;12:e56909) doi:10.2196/56909

KEYWORDS

hypoglycemia prediction; hypoglycemia; hypoglycemic; blood sugar; prediction; predictive; deep learning; generalization; machine learning; glucose; diabetes; continuous glucose monitoring; type 1 diabetes; type 2 diabetes; LSTM; long short-term memory

Introduction

Diabetes is a serious long-term disease with considerable influence on global health [1]. Type 1 diabetes mellitus (T1DM) is a disease in which the pancreas produces little or no insulin [2], whereas insulin resistance and insufficient insulin are the primary contributors to the development of type 2 diabetes mellitus (T2DM) [3]. Although the pathogenic mechanisms of

T1DM and T2DM are different, glucose-lowering treatments such as insulin administration are the common leading cause of hypoglycemia events in patients with both diabetes subtypes [4]. Severe hypoglycemia is a frequent phenomenon in patients with T1DM, with an annual prevalence of 30%-40% [5]. Although the risk of severe hypoglycemia in patients with T2DM is relatively lower, 46%-58% of these patients were reported to have experienced mild hypoglycemia symptoms over a 6-month period [6]. Patients experiencing frequent hypoglycemia events have 1.5-6.0 times increased risks of cardiovascular events and mortality than those without such events [7]. Patients with T2DM from Southeast Asia appear to have an elevated risk of hypoglycemia, as these patients are more often treated with a premixed insulin formulation, are younger, and have a lower BMI than those of their counterparts from Western countries [8-11]. Given that demographic and clinical factors such as ethnic group, diabetes subtype, and BMI are all important components of the complex risk profile of hypoglycemia, accurate risk prediction and prevention of hypoglycemia across populations and diabetes types remain significant challenges in diabetes management.

Recently, continuous glucose monitoring (CGM) has demonstrated good potential to predict hypoglycemia. For patients who wear insulin pumps or those who require multiple daily insulin injections, hypoglycemia prediction based on CGM data could provide a timely warning of impending hypoglycemia for the individual to take immediate action and increase their glucose levels. CGM devices are designed to produce time-series data by recording interstitial glucose concentrations within a relatively short interval of 5-15 minutes over a few days. Therefore, it is possible to leverage the early glucose readings to predict hypoglycemia events over the short-to-medium time horizon. Time-series forecast algorithms such as autoregressive and moving-average algorithms were first adopted to utilize the short-term temporal features of CGM data to predict hypoglycemia [12-15]. A small study including 17 patients with T1DM showed that these CGM-based algorithms achieved 86% sensitivity but only 58% specificity in hypoglycemia prediction [16]. Similar results from studies implementing these time-series forecast algorithms indicated that the low specificity might frequently generate false alarms, leading to discontinuation of CGM use in hypoglycemia prevention [17,18].

To improve the sensitivity and particularly the specificity of hypoglycemia prediction, both traditional machine learning algorithms such as support vector machine (SVM) and random forest (RF) models, along with deep learning models such as the convolutional neural network and long short-term memory network (LSTM) have been used to leverage more temporal features of CGM data [19-25]. When the features, including the mean of glucose and range of time in hyperglycemia, based on CGM data collected over the previous 6 hours were fed into the RF model, hypoglycemia prediction achieved a sensitivity of 93% and a specificity of 91% in a study of 112 patients with T1DM [26]. More recently, when an LSTM deep learning model was implemented on CGM data for hypoglycemia prediction, it achieved a sensitivity of 97% with remarkably few false alarms (0.9 false alarms per week) on a test data set including 10 patients with T1DM, thereby illuminating a path toward the widespread clinical adoption of CGM in hypoglycemia prediction [27].

However, a well-known challenge in implementing predictive models is their generalization [28]. The predictive performance of models could be substantially reduced when used in a setting that is not well-represented by the training data set [29,30]. This is particularly relevant in the case of hypoglycemia prediction, as the previously developed models for this purpose were mostly trained on a small data set of patients with T1DM from Western populations. In addition, the lack of a common test data set rendered the comparison of predictive performances between models unreliable. With recent improvements in measurement accuracy, CGM devices have also gained momentum and have begun to be adopted more widely for the management of T2DM, including in developing countries. Therefore, the established hypoglycemia prediction models should be validated in more diverse populations and over a wide spectrum of patients with different types of diabetes.

We hypothesized that the promising LSTM model for hypoglycemia prediction from CGM data could maintain good predictive performance in different settings for different populations. In this study, we assembled two large CGM data sets from China and the United States, both including patients with T1DM and patients with T2DM. We developed the LSTM model on the Chinese data set and then examined the model performance in the data set from European-Americans in the United States. Apart from exploring the model's generalization ability for T1DM and T2DM separately, we also compared the predictive performance of the LSTM model with that of SVM and RF models to further indicate its translational potential.

Methods

Ethical Considerations

The study protocol was approved by the ethics committees of Kunshan Hospital Affiliated to Jiangsu University (2023-03-014-H01-K01) and the study was performed in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from each participant before taking the measurements. The data analyzed were anonymized. All participants volunteered to participate in the project with no compensation provided.

Data Collection

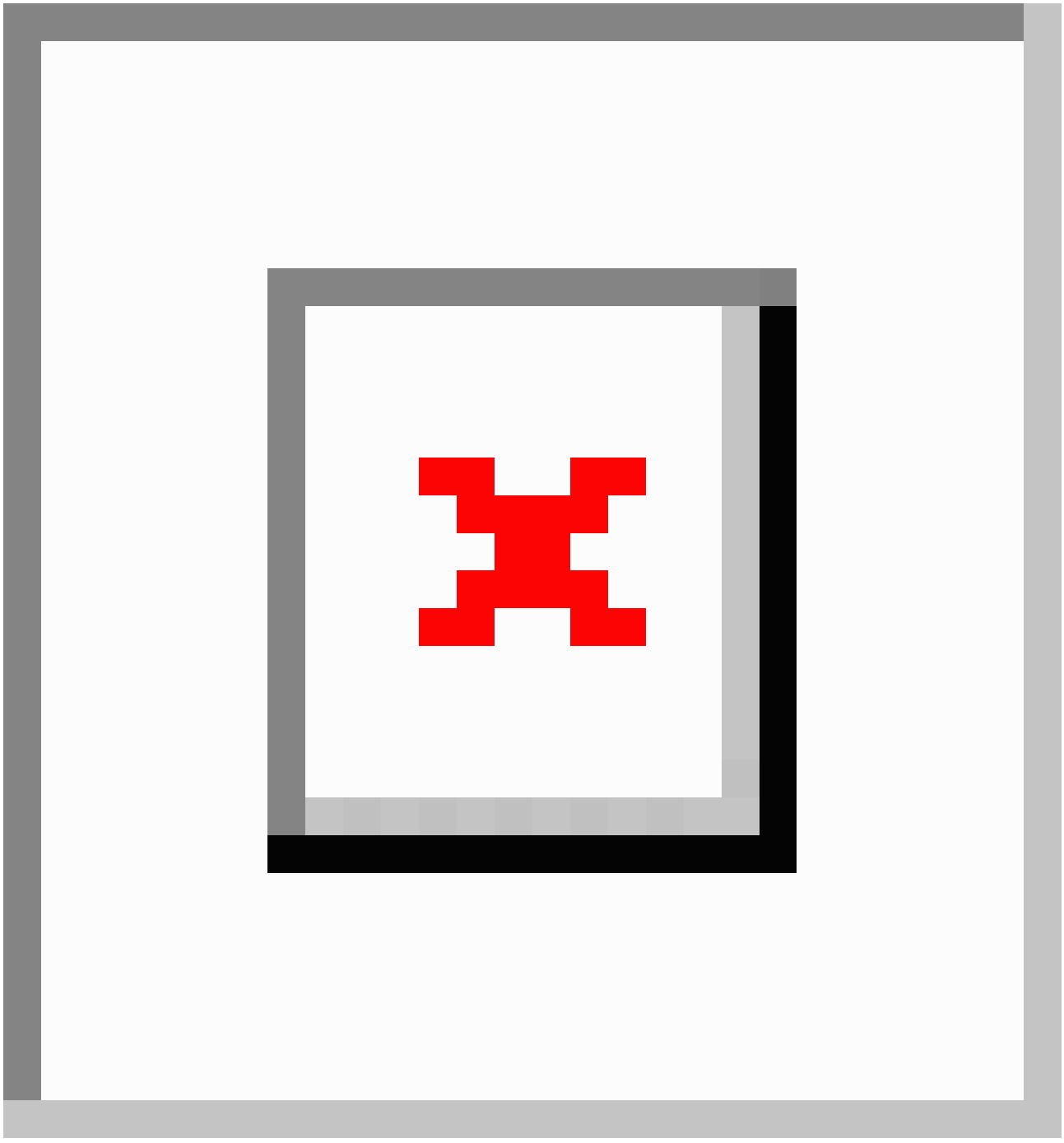
We collected a primary data set comprising 1578 days of CGM data collected from 264 Chinese people with diabetes to develop a deep learning model for hypoglycemia prediction. The individuals' glucose levels were monitored using the Medtronic MiniMed CGM device, which requires calibration according to self-monitored blood glucose levels. This CGM device can record glucose levels every 5 minutes over 3 days.

The mean absolute relative difference (MARD) was used to evaluate the quality of the CGM data. The MARD represents the average of the absolute error between all CGM values and matched reference values. A small MARD indicates that the CGM readings are close to the reference glucose value, whereas a larger MARD percentage indicates greater discrepancies between the CGM and reference glucose values. Each individual

had at least 5 self-monitoring of blood glucose (SMBG) measurements. As reference glucose values, the SMBG was used to calculate the MARD of CGM data. The data for 72 participants were filtered out because their MARD was higher than 15%, leaving data for 192 participants with 808 days of CGM data for analysis.

To examine whether the deep learning model trained and developed with data from the Chinese population could be generalized to a different population, we assembled a large validation data set that mainly comprised data from individuals of European-American ancestry. The validation data set shared by the A1c-Derived Average Glucose study group includes 507 participants and 7299 days of CGM data, also collected with Medtronic MiniMed devices [31]. After filtering out individuals without diabetes, 427 patients with either T1DM or T2DM were included to validate the model. This validation data set was split into two groups: the T1DM group of 268 participants with 3932 days of CGM data and the T2DM group of 159 participants with 2259 days of CGM data. Figure 1 provides the flowchart of exclusion criteria for the primary data set and validation data set.

Figure 1. Flowchart of exclusion criteria for the primary data set and validation data set. MARD: mean absolute relative difference.



Outcome

The glucose values reported by CGM devices were classified into three categories: nonhypoglycemic level (glucose>70 mg/dL), mild hypoglycemic level (glucose=54-70 mg/dL), and severe hypoglycemic level (glucose<54 mg/dL) according to the international consensus on CGM utility [32].

Data Preprocessing

The primary data set consisting of 192 patients was randomly split into three disjoint data sets, namely the training data set, development data set, and test data set, at a 7:1.5:1.5 ratio. The training data set was used to train the model, whereas the development data set was used to select the hyperparameters in the training process. The test data set was used to evaluate the performance of the developed model.

The CGM sensor may fail to detect a valid glucose level, resulting in the CGM device missing glucose values continuously. To preserve as much of the CGM data as possible, we divided an individual's CGM data into different segments at the time points of missing data rather than discarding all of the CGM data. A segment was removed if it was shorter than 6 hours (72 data points). We set each glucose value reported by the CGM device as a predictive target if there were sufficient data prior to the target time at which the predictive target was located. The data used to predict the hypoglycemic level of the predictive target were retrieved from a 6-hour time window spanning from −390 minutes to −30 minutes of the target time. After preprocessing the primary data set, the training, development, and test data sets included 100,879, 21,895, and 21,324 samples generated from 134, 29, and 29 participants, respectively. Similarly, the T1DM group and T2DM group from the validation data set contained 712,018 and 405,224 samples generated from 268 and 159 participants, respectively.

Model Development

We used the common bidirectional LSTM model containing both forward and backward layers to capture the long-range temporal features in the time-series CGM data and to combine these features with context factors [33]. Each LSTM layer consists of 128 memory cells [34]. We chose a set of context factors, including gender, age, diabetes type, and hemoglobin A_{1c} value, to capture the background risk of hypoglycemia and enhance the model's predictive performance [26]. Therefore, each input data sample included 72 points of CGM data collected during 6 hours and the context factors. The output was the probability of the target glucose value being at the nonhypoglycemic level, mild hypoglycemic level, and severe hypoglycemic level.

We trained the LSTM model to predict the categories of a CGM value within 30 minutes on the prediction horizon. The training process would be terminated if the accuracy failed to increase for 10 consecutive epochs. We used root mean square propagation [35] as the optimizer and set the mini batch size to 64. The LSTM model was developed using the Python package Keras [36]. We also developed models to implement the SVM and RF algorithms for comparison. The SVM model was developed using the radial basis function as the kernel function, which was also used in previous studies of hypoglycemia prediction [37]. The RF model included 100 trees and was developed with the Scikit-learn Python package [38] under default parameters. The input to the SVM and RF models was the same as that used for the LSTM model.

Model Evaluation

Sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) were used to evaluate model performance. The label for each sample was the category of a single CGM data point. Sensitivity and specificity indicate the proportion of the labels of CGM data points that were correctly predicted. The DeLong method was used to measure the 95% CIs for the AUC values [39]. All methods of evaluation were developed using Python and the pROC R package [40].

Results

Characteristics of the Data Set

Table 1 summarizes the characteristics of the primary data set and the validation data set. As expected, the average age of patients with T1DM was lower than that of the patients with T2DM in both data sets (Wilcoxon rank sum test, $P<.001$).

Table . Characteristics of the primary data set and validation data set.

Variables		Primary data set		Validation data set	
		Type 2 diabetes (n=175)	Type 1 diabetes (n=17)	Type 2 diabetes (n=159)	Type 1 diabetes (n=268)
Age (years), mean (SD)		53.30 (11.78)	40.59 (13.02)	55.64 (9.32)	43.06 (12.85)
Women, n (%)		51 (29.14)	11 (64.71)	81 (50.94)	140 (52.24)
Predictive targets, n					
	Nonhypoglycemia	129,609	12,029	396,415	660,111
	Mild hypoglycemia	1350	336	5985	28,287
	Severe hypoglycemia	608	166	2824	23,620
Hemoglobin A_{1c} (%), mean (SD)		7.69 (1.71)	8.46 (2.22)	7.01 (1.24)	7.51 (1.30)

Model Performance on the Primary Test Data Set

Using the primary data set from 192 individuals, the three models of LSTM, SVM, and RF were trained and we then evaluated their performance based on the AUC. At the mild hypoglycemic level, the LSTM model achieved an AUC of 97.22% (95% CI 96.78%-97.66%), which was significantly higher than the AUC of 94.33% (95% CI 93.13%-95.53%) and 94.81% (95% CI 93.72%-95.91%) achieved by the SVM and RF models, respectively (both $P<.001$). At the severe hypoglycemic level, the LSTM model achieved an AUC of 99.64% (95% CI 99.53%-99.76%), which was significantly higher than the AUC of 98.30% (95% CI 98.00%-98.60%) and 97.88% (95% CI 96.93%-98.83%) achieved by the SVM and RF models, respectively (both $P<.001$). These results demonstrated that the LSTM model could outperform the SVM and RF models in predicting hypoglycemia.

Model Generalization on the Validation Data Set

We then utilized the validation data set from 427 European-Americans to evaluate the generalization of the LSTM model developed from our primary data set of 192 Chinese individuals. The LSTM model achieved an AUC of 94.61% (95% CI 94.51%-94.71%) for mild hypoglycemia, which was significantly higher than the AUC of 92.59% (95% CI 92.48%-92.71%) and 91.43% (95% CI 91.28%-91.58%) achieved by the SVM and RF models, respectively (both $P<.001$). The LSTM model achieved an AUC of 96.40% (95% CI 96.25%-96.55%) for severe hypoglycemia, which was significantly higher than the AUC of 95.27% (95% CI 95.15%-95.39%) and 95.17% (95% CI 95.01%-95.32%) achieved by SVM and RF models, respectively (both $P<.001$). Although AUC values of the LSTM model decreased by approximately 3% in the validation data set compared to those from the primary test data set, the overall AUC was still higher than 94%, indicating that the LSTM model could accurately predict hypoglycemia in a different population.

Next, the generalizability of the LSTM model to various disease subtypes was evaluated in the subgroups of T1DM and T2DM from the validation data set. For T1DM, the LSTM model achieved an AUC of 93.49% (95% CI 93.38%-93.61%) at the mild hypoglycemia level, which was significantly higher than the AUC of 90.92% (95% CI 90.78%-91.06%) and 89.74% (95% CI 89.57%-89.92%) achieved by the SVM and RF models, respectively (both $P<.001$). In addition, the LSTM model achieved an AUC of 95.89% (95% CI 95.73%-96.05%) at the severe hypoglycemia level, which was significantly higher than the AUC of 94.06% (95% CI 93.91%-94.21%) and 94.53% (95% CI 94.37%-94.70%) achieved by the SVM and RF models, respectively (both $P<.001$).

For T2DM, the LSTM model achieved an AUC of 96.83% (95% CI 96.66%-97.01%) at the mild hypoglycemia level, which was significantly higher than the AUC of 95.72% (95% CI 95.51%-95.93%) and 94.08% (95% CI 93.73%-94.43%) achieved by the SVM and RF models, respectively (both $P<.001$). In addition, the LSTM model achieved an AUC of 97.65% (95% CI 97.27%-98.04%) at the severe hypoglycemia level, which was significantly higher than the AUC of 96.02% (95% CI 95.70%-96.34%) and 95.71% (95% CI 95.23%-96.19%) achieved by the SVM and RF models, respectively (both $P<.001$).

The AUCs of the LSTM model were consistently higher than those from the SVM and RF models in both the T1DM and T2DM data sets. Taken together, these results demonstrated that the LSTM model could be generalized to different diabetes subtypes without significant loss of predictive performance.

Comparison of the False Alarm Rate

Finally, we examined whether the LSTM model could achieve a low false alarm rate (ie, high specificity) under satisfactory sensitivity. According to previous studies of hypoglycemia prediction, we set the model parameters to fix the satisfactory sensitivity level at 90% and 95% for mild and severe hypoglycemia prediction, respectively [21,26,37]. As shown in Table 2, while maintaining a sensitivity of 90% for mild hypoglycemia, which is difficult to predict, the LSTM model could achieve a specificity of 88.43%, which was higher than the specificity obtained from the SVM and RF models. For severe hypoglycemia, when a higher satisfactory sensitivity rate of 95% was set, the LSTM model achieved a specificity of 87.34%, which was higher than that obtained from the SVM model. Moreover, the RF model could not achieve a sensitivity of 95% for the severe hypoglycemic level. Taken together, these results demonstrated that the LSTM model could maintain a lower false alarm rate than the SVM and RF models in clinically practical settings.

Table . Specificity and sensitivity of the three models on the validation data set.

	Mild hypoglycemic level		Severe hypoglycemic level	
	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)
LSTM[a]	88.43	90.00	87.34	95.00
SVM[b]	82.57	90.00	80.67	95.00
RF[c]	82.65	90.00	Not determined	Not achieved

[a]LSTM: long short-term memory.

[b]SVM: support vector machine.

[c]RF: random forest.

Discussion

Principal Findings

In this study, we assembled two large CGM data sets from China and the United States to develop and validate an LSTM deep learning model for hypoglycemia prediction. The LSTM model could maintain good predictive performance when applied to data sets from a different ethnic population or any common subtype of diabetes. The LSTM model could also predict both mild and severe hypoglycemia with higher accuracy than the traditional SVM and RF models. While targeting clinically meaningful high sensitivity, the LSTM model could achieve high specificity, thereby reducing the rate of false alarms.

Compared with the models tested without external validation in most previous studies of hypoglycemia prediction, we developed an LSTM model and validated the model in a data set from a different population to examine its generalizability [27]. There are considerable differences in dietary structure and clinical practice between China and the United States, which are among the many factors that might affect the risk of hypoglycemia. Previous studies demonstrated that clinical models trained in one population could result in an AUC reduction as great as 15% when applied to a distinct population [41-43]. However, the LSTM model derived from our Chinese training data set maintained high prediction performance (AUC>93%) with only a minor loss of 3% in the US data set, indicating good generalizability of the model. As CGM devices are becoming more widely adopted, the generalizability of the LSTM model could be further improved by training the model with data from multiple populations or can be fine-tuned for the target population using a transfer-learning approach [44].

We also examined the generalizability of the LSTM model on another dimension of diabetes pathogenicity. Given the different pathogenic mechanisms between T1DM and T2DM, hypoglycemia occurring in different diabetes subtypes would be expected to be preceded by various patterns of glucose fluctuation, which could be leveraged by the LSTM model for prediction. Therefore, the model was expected to lose predictive performance when the training and validation data sets had different proportions of diabetes subtypes. Indeed, we observed a higher AUC value for T2DM than for T1DM in the validation data set, which was likely due to the fact that our training data set primarily consisted of individuals with T2DM. However, for either subtype of diabetes, the LSTM model consistently maintained an AUC value above 93%, indicating the good generalizability of the model. With the increasing popularity of CGM usage in the management of all subtypes of diabetes, the LSTM model could be further improved by using larger training data sets with a wider representation of the various diabetes subtypes.

Achieving high sensitivity has been the main focus of previous models for hypoglycemia prediction, as severe hypoglycemia requires immediate external intervention [15,32]. With the sacrifice of high specificity, false alarms became an obstacle for the safe and widespread use of CGM devices [45-47]. False-alarm fatigue could lead to users ignoring the true alarms of hypoglycemia and contribute to the discontinuation of CGM use [45]. Moreover, glucose control could be compromised, as CGM users may frequently take action to elevate their glucose level when a false alarm is generated [46]. Therefore, it is imperative to balance the false alarm rate with sufficient sensitivity of the prediction. In this study, we demonstrated that the LSTM model would generate fewer false alarms than the traditional machine learning models under satisfactory sensitivity rates of 90% and 95% for mild and severe hypoglycemia, respectively. Therefore, the balanced hypoglycemia prediction performance from the LSTM model demonstrated that it has potential to promote the use of CGM in a variety of clinical settings.

One reason for the better predictive performance of the LSTM model than the SVM and RF models might be that the LSTM algorithm is more suitable for analyzing sequential data. CGM data are a type of sequential data that are generated in time order. The LSTM algorithm consists of memory cells that learn the sequential nature of observations within CGM data [48]. The input of one memory cell is the glucose value taken at one time point and then the LSTM takes all of the glucose values as inputs sequentially. Every memory cell retains the relevant information and discards irrelevant information for the predictive task, and then the relevant information in one cell is delivered to the next cell [49-53]. With this sequential structure, LSTM networks incorporate CGM data from the past to accurately make predictions of hypoglycemia risk in the near future.

Limitations

There are several limitations of this study. Although we tested the generalizability of the LSTM model using two data sets from China and the United States, further validation might still be required for application of the model in other countries. Similarly, as only T1DM and T2DM were included in our data sets, the model should be tested with wider and more representative training data sets to validate its utility on other minority subtypes of diabetes. Moreover, data from only one CGM device manufacturer were available for this study. Thus, it is unknown whether the model would perform equally well with data collected from other devices such as factory-calibrated CGM or noninvasive CGM devices. However, given that all of the devices were strictly calibrated by finger-stick glucose values, the fluctuation patterns and temporal dependence of CGM data, which are key factors for the LSTM prediction task, should be largely captured by any certified CGM device. Moreover, the performance of the LSTM model for hypoglycemia prediction will need to be further validated in a CGM data set without missing data.

Conclusions

We developed an accurate LSTM model for mild and severe hypoglycemia prediction using a large data set of 619 patients with diabetes from China and the United States. The model could be robustly generalized to different populations or any common subtype of diabetes. Moreover, while maintaining satisfactory levels of sensitivity, the model could also achieve high specificity, indicating its potential to mitigate the hypoglycemia false-alarm fatigue that is frequently observed in clinical practice. Taken together, we demonstrated that the

LSTM model is a strong candidate algorithm to be further tested and implemented for the wider clinical adoption of CGM.

Acknowledgments

We thank all of the involved clinicians and researchers for data collection and assistance. This study was funded by the National Key R&D Program of China (SQ2022YFB3200174) and Suzhou Science and Technology Project (SKY2022025).

Data Availability

Requests for access to the study data should be directed to the corresponding author.

Conflicts of Interest

None declared.

References

1. Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. Phys Ther 2008 Nov;88(11):1254-1264. [doi: 10.2522/ptj.20080020] [Medline: 18801858]
2. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. Lancet 2014 Jan;383(9911):69-82. [doi: 10.1016/S0140-6736(13)60591-7] [Medline: 25130995]
3. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet 2017 Jun 3;389(10085):2239-2251. [doi: 10.1016/S0140-6736(17)30058-2] [Medline: 28190580]
4. Cryer PE. The barrier of hypoglycemia in diabetes. Diabetes 2008 Dec;57(12):3169-3176. [doi: 10.2337/db08-1084] [Medline: 19033403]
5. Frier BM. The incidence and impact of hypoglycemia in type 1 and type 2 diabetes. Inter Diab Monitor 2009;21(6):210-218.
6. Silbert R, Salcido-Montenegro A, Rodriguez-Gutierrez R, Katabi A, McCoy RG. Hypoglycemia among patients with type 2 diabetes: epidemiology, risk factors, and prevention strategies. Curr Diab Rep 2018 Jun 21;18(8):53. [doi: 10.1007/s11892-018-1018-0] [Medline: 29931579]
7. International Hypoglycaemia Study Group. Hypoglycaemia, cardiovascular disease, and mortality in diabetes: epidemiology, pathogenesis, and management. Lancet Diabetes Endocrinol 2019 May;7(5):385-396. [doi: 10.1016/S2213-8587(18)30315-2] [Medline: 30926258]
8. Chan JCN, Malik V, Jia W, et al. Diabetes in Asia: epidemiology, risk factors, and pathophysiology. JAMA 2009 May 27;301(20):2129-2140. [doi: 10.1001/jama.2009.726] [Medline: 19470990]
9. Kalra S, Balhara YPS, Sahay BK, Ganapathy B, Das AK. Why is premixed insulin the preferred insulin? Novel answers to a decade-old question. J Assoc Physicians India 2013 Jan;61(1 Suppl):9-11. [Medline: 24482980]
10. Goh SY, Hussein Z, Rudijanto A. Review of insulin-associated hypoglycemia and its impact on the management of diabetes in Southeast Asian countries. J Diabetes Investig 2017 Sep;8(5):635-645. [doi: 10.1111/jdi.12647] [Medline: 28236664]
11. Aschner P, Sethi B, Gomez-Peralta F, et al. Insulin glargine compared with premixed insulin for management of insulin-naïve type 2 diabetes patients uncontrolled on oral antidiabetic drugs: the open-label, randomized GALAPAGOS study. J Diabetes Complications 2015 Aug;29(6):838-845. [doi: 10.1016/j.jdiacomp.2015.04.003] [Medline: 25981123]
12. Eren-Oruklu M, Cinar A, Quinn L, Smith D. Estimation of future glucose concentrations with subject-specific recursive linear models. Diabetes Technol Ther 2009 Apr;11(4):243-253. [doi: 10.1089/dia.2008.0065] [Medline: 19344199]
13. Yang J, Li L, Shi Y, Xie X. An ARIMA model with adaptive orders for predicting blood glucose concentrations and Hypoglycemia. IEEE J Biomed Health Inform 2019 May;23(3):1251-1260. [doi: 10.1109/JBHI.2018.2840690] [Medline: 29993728]
14. Eren-Oruklu M, Cinar A, Rollins DK, Quinn L. Adaptive system identification for estimating future glucose concentrations and hypoglycemia alarms. Automatica (Oxf) 2012 Aug;48(8):1892-1897. [doi: 10.1016/j.automatica.2012.05.076] [Medline: 22865931]
15. Dassau E, Cameron F, Lee H, et al. Real-time hypoglycemia prediction suite using continuous glucose monitoring: a safety net for the artificial pancreas. Diabetes Care 2010 Jun;33(6):1249-1254. [doi: 10.2337/dc09-1487] [Medline: 20508231]
16. Bayrak ES, Turksoy K, Cinar A, Quinn L, Littlejohn E, Rollins D. Hypoglycemia early alarm systems based on recursive autoregressive partial least squares models. J Diabetes Sci Technol 2013 Jan 1;7(1):206-214. [doi: 10.1177/193229681300700126] [Medline: 23439179]
17. Tansey M, Laffel L, Cheng J, et al. Satisfaction with continuous glucose monitoring in adults and youths with type 1 diabetes. Diabet Med 2011 Sep;28(9):1118-1122. [doi: 10.1111/j.1464-5491.2011.03368.x] [Medline: 21692844]
18. Ramchandani N, Arya S, Ten S, Bhandari S. Real-life utilization of real-time continuous glucose monitoring: the complete picture. J Diabetes Sci Technol 2011 Jul 1;5(4):860-870. [doi: 10.1177/193229681100500407] [Medline: 21880227]

19. Georga EI, Protopappas VC, Ardigò D, Polyzos D, Fotiadis DI. A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions. Diabetes Technol Ther 2013 Aug;15(8):634-643. [doi: 10.1089/dia.2012.0285] [Medline: 23848178]
20. Jensen MH, Christensen TF, Tarnow L, Seto E, Dencker Johansen M, Hejlesen OK. Real-time hypoglycemia detection from continuous glucose monitoring data of subjects with type 1 diabetes. Diabetes Technol Ther 2013 Jul;15(7):538-543. [doi: 10.1089/dia.2013.0069] [Medline: 23631608]
21. Mosquera-Lopez C, Dodier R, Tyler NS, et al. Predicting and preventing nocturnal hypoglycemia in type 1 diabetes using big data analytics and decision theoretic analysis. Diabetes Technol Ther 2020 Nov;22(11):801-811. [doi: 10.1089/dia.2019.0458] [Medline: 32297795]
22. Gu W, Zhou Z, Zhou Y, He M, Zou H, Zhang L. Predicting blood glucose dynamics with multi-time-series deep learning. Presented at: SenSys '17: 15th ACM Conference on Embedded Network Sensor Systems; Nov 5 to 8, 2017; Delft, The Netherlands. [doi: 10.1145/3131672.3136965]
23. Chen J, Li K, Herrero P, Zhu T, Georgiou P. Dilated recurrent neural network for short-time prediction of glucose concentration. Presented at: 3rd International Workshop on Knowledge Discovery in Healthcare Data co-located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018); Jul 13, 2018; Stockholm, Sweden. [doi: 10.1007/s41666-020-00068-2]
24. Doike T, Hayashi K, Arata S, Mohammad KN, Kobayashi A, Niitsu K. A blood glucose level prediction system using machine learning based on recurrent neural network for Hypoglycemia prevention. Presented at: 2018 16th IEEE International New Circuits and Systems Conference (NEWCAS); Jun 24 to 27, 2018; Montreal, QC. [doi: 10.1109/NEWCAS.2018.8585468]
25. Li J, Ma X, Tobore I, et al. A novel CGM metric-gradient and combining mean sensor glucose enable to improve the prediction of nocturnal hypoglycemic events in patients with diabetes. J Diabetes Res 2020 Nov;2020:8830774. [doi: 10.1155/2020/8830774] [Medline: 33204733]
26. Dave D, DeSalvo DJ, Haridas B, et al. Feature-based machine learning model for real-time hypoglycemia prediction. J Diabetes Sci Technol 2021 Jul;15(4):842-855. [doi: 10.1177/1932296820922622] [Medline: 32476492]
27. Mosquera-Lopez C, Dodier R, Tyler N, Resalat N, Jacobs P. Leveraging a big dataset to develop a recurrent neural network to predict adverse glycemic events in type 1 diabetes. IEEE J Biomed Health Inform 2019 Apr 17. [doi: 10.1109/JBHI.2019.2911701] [Medline: 30998484]
28. Zhang Y, Wu H, Liu H, Tong L, Wang MD. Improve model generalization and robustness to dataset bias with bias-regularized learning and domain-guided augmentation. arXiv. Preprint posted online on Oct 12, 2019. [doi: 10.48550/arXiv.1910.06745]
29. Kortylewski A, Egger B, Schneider A, Gerig T, Morel-Forster A, Vetter T. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Jun 16 to 17, 2019; Long Beach, CA, USA URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8972688 [accessed 2024-05-24] [doi: 10.1109/CVPRW.2019.00279]
30. Tian Y, Chen W, Zhou T, Li J, Ding K, Li J. Establishment and evaluation of a multicenter collaborative prediction model construction framework supporting model generalization and continuous improvement: a pilot study. Int J Med Inform 2020 Sep;141:104173. [doi: 10.1016/j.ijmedinf.2020.104173] [Medline: 32531725]
31. Nathan DM, Kuenen J, Borg R, et al. Translating the A1C assay into estimated average glucose values. Diabetes Care 2008 Aug;31(8):1473-1478. [doi: 10.2337/dc08-0545] [Medline: 18540046]
32. Danne T, Nimri R, Battelino T, et al. International consensus on use of continuous glucose monitoring. Diabetes Care 2017 Dec;40(12):1631-1640. [doi: 10.2337/dc17-1600] [Medline: 29162583]
33. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput 2000 Oct;12(10):2451-2471. [doi: 10.1162/089976600300015015] [Medline: 11032042]
34. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
35. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning. Lecture 6a. Overview of mini-batch gradient descent. Computer Science University of Toronto. URL: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf [accessed 2024-05-15]
36. Keras. URL: https://keras.io/ [accessed 2024-05-13]
37. Oviedo S, Contreras I, Quirós C, Giménez M, Conget I, Vehi J. Risk-based postprandial hypoglycemia forecasting using supervised learning. Int J Med Inform 2019 Jun;126:1-8. [doi: 10.1016/j.ijmedinf.2019.03.008] [Medline: 31029250]
38. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011 Nov 1;12:2825-2830. [doi: 10.5555/1953048.2078195]
39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [doi: 10.2307/2531595] [Medline: 3203132]
40. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12:77. [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]

41. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993 Nov 24;270(20):2478-2486. [doi: 10.1001/jama.1993.03510200084037] [Medline: 8230626]
42. Adrie C, Francais A, Alvarez-Gonzalez A, et al. Model for predicting short-term mortality of severe sepsis. Crit Care 2009 May;13(3):R72. [doi: 10.1186/cc7881] [Medline: 19454002]
43. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016 Jun 22;353:i3140. [doi: 10.1136/bmj.i3140] [Medline: 27334381]
44. Torrey L, Shavlik J. Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: IGI Global; 2010:242-264. [doi: 10.4018/978-1-60566-766-9]
45. Shivers JP, Mackowiak L, Anhalt H, Zisser H. “Turn it off!”: diabetes device alarm fatigue considerations for the present and the future. J Diabetes Sci Technol 2013 May 1;7(3):789-794. [doi: 10.1177/193229681300700324] [Medline: 23759412]
46. Cryer PE. Glycemic goals in diabetes: trade-off between glycemic control and iatrogenic hypoglycemia. Diabetes 2014 Jul;63(7):2188-2195. [doi: 10.2337/db14-0059] [Medline: 24962915]
47. Wong JC, Foster NC, Maahs DM, et al. Real-time continuous glucose monitoring among participants in the T1D Exchange clinic registry. Diabetes Care 2014 Oct;37(10):2702-2709. [doi: 10.2337/dc14-0303] [Medline: 25011947]
48. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid 2017 Sep 18;10(1):841-851. [doi: 10.1109/TSG.2017.2753802]
49. Xu Z, Li S, Deng W. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. Presented at: ACPR 2015: 3rd IAPR Asian Conference on Pattern Recognition; Nov 3, 2015; Kuala Lumpur, Malaysia. [doi: 10.1109/ACPR.2015.7486482]
50. Shi X, Jin Y, Dou Q, Heng PA. LRTD: long-range temporal dependency based active learning for surgical workflow recognition. Int J Comput Assist Radiol Surg 2020 Sep;15(9):1573-1584. [doi: 10.1007/s11548-020-02198-9] [Medline: 32588246]
51. Liao J, Liu L, Duan H, et al. Using a convolutional neural network and convolutional long short-term memory to automatically detect aneurysms on 2D digital subtraction angiography images: framework development and validation. JMIR Med Inform 2022 Mar 16;10(3):e28880. [doi: 10.2196/28880] [Medline: 35294371]
52. Athanasiou M, Fragkozidis G, Zarkogianni K, Nikita KS. Long short-term memory–based prediction of the spread of influenza-like illness leveraging surveillance, weather, and Twitter data: model development and validation. J Med Internet Res 2023 Feb 6;25:e42519. [doi: 10.2196/42519] [Medline: 36745490]
53. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of Google trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill 2020 Apr 14;6(2):e18828. [doi: 10.2196/18828] [Medline: 32234709]

Abbreviations

AUC: area under the receiver operating characteristic curve
CGM: continuous glucose monitoring
LSTM: long short-term memory
MARD: mean absolute relative difference
RF: random forest
SMBG: self-monitoring of blood glucose
SVM: support vector machine
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus

Edited by C Lovis; submitted 21.02.24; peer-reviewed by G Lim; revised version received 07.04.24; accepted 04.05.24; published 24.05.24.

Please cite as:
Shao J, Pan Y, Kou WB, Feng H, Zhao Y, Zhou K, Zhong S
Generalization of a Deep Learning Model for Continuous Glucose Monitoring–Based Hypoglycemia Prediction: Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e56909
URL: https://medinform.jmir.org/2024/1/e56909
doi:10.2196/56909

Time Series AI Model for Acute Kidney Injury Detection Based on a Multicenter Distributed Research Network: Development and Verification Study

Suncheol Heo[1,*], MS; Eun-Ae Kang[2,*], BA; Jae Yong Yu[1,*], PhD; Hae Reong Kim[1], PhD; Suehyun Lee[3], PhD; Kwangsoo Kim[4], PhD; Yul Hwangbo[5], MD, PhD; Rae Woong Park[6], MD, PhD; Hyunah Shin[7], BA; Kyeongmin Ryu[7], BA; Chungsoo Kim[8], PharmD; Hyojung Jung[5], BA; Yebin Chegal[9], BA; Jae-Hyun Lee[10,11], MD, PhD; Yu Rang Park[1], PhD

1

2

3

4

5

6

7

8

9

10

11

[*]these authors contributed equally

Corresponding Author:
Yu Rang Park, PhD

Abstract

Background: Acute kidney injury (AKI) is a marker of clinical deterioration and renal toxicity. While there are many studies offering prediction models for the early detection of AKI, those predicting AKI occurrence using distributed research network (DRN)–based time series data are rare.

Objective: In this study, we aimed to detect the early occurrence of AKI by applying an interpretable long short-term memory (LSTM)–based model to hospital electronic health record (EHR)–based time series data in patients who took nephrotoxic drugs using a DRN.

Methods: We conducted a multi-institutional retrospective cohort study of data from 6 hospitals using a DRN. For each institution, a patient-based data set was constructed using 5 drugs for AKI, and an interpretable multivariable LSTM (IMV-LSTM) model was used for training. This study used propensity score matching to mitigate differences in demographics and clinical characteristics. Additionally, the temporal attention values of the AKI prediction model's contribution variables were demonstrated for each institution and drug, with differences in highly important feature distributions between the case and control data confirmed using 1-way ANOVA.

Results: This study analyzed 8643 and 31,012 patients with and without AKI, respectively, across 6 hospitals. When analyzing the distribution of AKI onset, vancomycin showed an earlier onset (median 12, IQR 5-25 days), and acyclovir was the slowest compared to the other drugs (median 23, IQR 10-41 days). Our temporal deep learning model for AKI prediction performed well for most drugs. Acyclovir had the highest average area under the receiver operating characteristic curve score per drug (0.94), followed by acetaminophen (0.93), vancomycin (0.92), naproxen (0.90), and celecoxib (0.89). Based on the temporal attention values of the variables in the AKI prediction model, verified lymphocytes and calcvancomycin ium had the highest attention, whereas lymphocytes, albumin, and hemoglobin tended to decrease over time, and urine pH and prothrombin time tended to increase.

Conclusions: Early surveillance of AKI outbreaks can be achieved by applying an IMV-LSTM based on time series data through an EHR-based DRN. This approach can help identify risk factors and enable early detection of adverse drug reactions when prescribing drugs that cause renal toxicity before AKI occurs.

(JMIR Med Inform 2024;12:e47693) doi:10.2196/47693

KEYWORDS

adverse drug reaction; real world data; multicenter study; distributed research network; common data model; time series AI; time series; artificial intelligence; machine learning; adverse reaction; adverse reactions; detect; detection; toxic; toxicity; renal; kidney; nephrology; pharmaceutical; pharmacology; pharmacy; pharmaceutics

Introduction

Acute kidney injury (AKI) is associated with a mortality rate of 40% - 70% in hospitalized patients who develop AKI and causes significant kidney damage even after recovery, leading to dialysis, longer hospital stays, and increased costs of care [1-4]. Early detection of AKI increases the likelihood of AKI prevention, associated morbidity, and costs [5]. As no specific treatment can reverse AKI and the recognition of patients at risk of AKI before diagnosis contributes to better clinical outcomes than treatment after AKI occurs [6], early detection of AKI is essential for prompt therapeutic intervention.

Several studies have attempted to predict AKI occurrence. With the increasing availability of clinical databases, they developed models to predict the occurrence of AKI using electronic health records (EHRs) [7-18]. Although these studies used EHRs, the number of patients in the patient population was small because they focused on specific patients, such as surgical patients, patients with sepsis, and older adults. There have also been a number of studies using artificial intelligence (AI) models to predict AKI. Although attempts have been made to predict the occurrence of AKI early, few models have provided clear rationales and explanations [19-21]. Therefore, time series data analysis is required for AKI prediction models to reflect the temporal information between variables [22]. Time series analysis for AKI is necessary because the length of time each patient stays in a hospital or intensive care unit can differ from person to person, and the frequency of measurements can vary from values that are measured continuously (eg, blood pressure) to laboratory values that are measured on an as-needed basis. Recently, an interpretable multivariable long short-term memory (IMV-LSTM) method for considering time series data has been published [23]; however, little research has been conducted on this method.

To address these issues, this study aimed to apply and validate a multicenter-based explainable time series AI model for predicting the occurrence of AKI caused by specific nephrotoxic drugs in 6 hospitals in South Korea by using a large clinical database with a common data model (CDM) through a distributed research network (DRN).

Methods

Ethical Considerations

This study was approved by the institutional review committees of Severance Hospital (SH; 4-2021-1209), Gangnam Severance Hospital (GSH; 3-2021-0005), Konyang University Hospital (KYUH; 2021-10-003-001), Ajou University Hospital (AJUH; AJIRB-MED-MDB-21 - 676), Seoul National University Cancer Hospital (SNUH; E-2207-151-1342), and the National Cancer Center (NCC; 2022-0184). All retrospective data were anonymized and appropriate measures were taken to protect participant information.

Study Design

This retrospective observational cohort study analyzed the EHRs from 6 hospitals in South Korea between 1994 and 2021 to predict AKI. The EHRs were converted to OMOP-CDM (Observational Medical Outcomes Partnership Common Data Model) version 5.3.1. The 6 hospitals were SH, GSH, KYUH, AJUH, SNUH, and the NCC. An overall diagram of the cohort composition is provided in Figure 1.

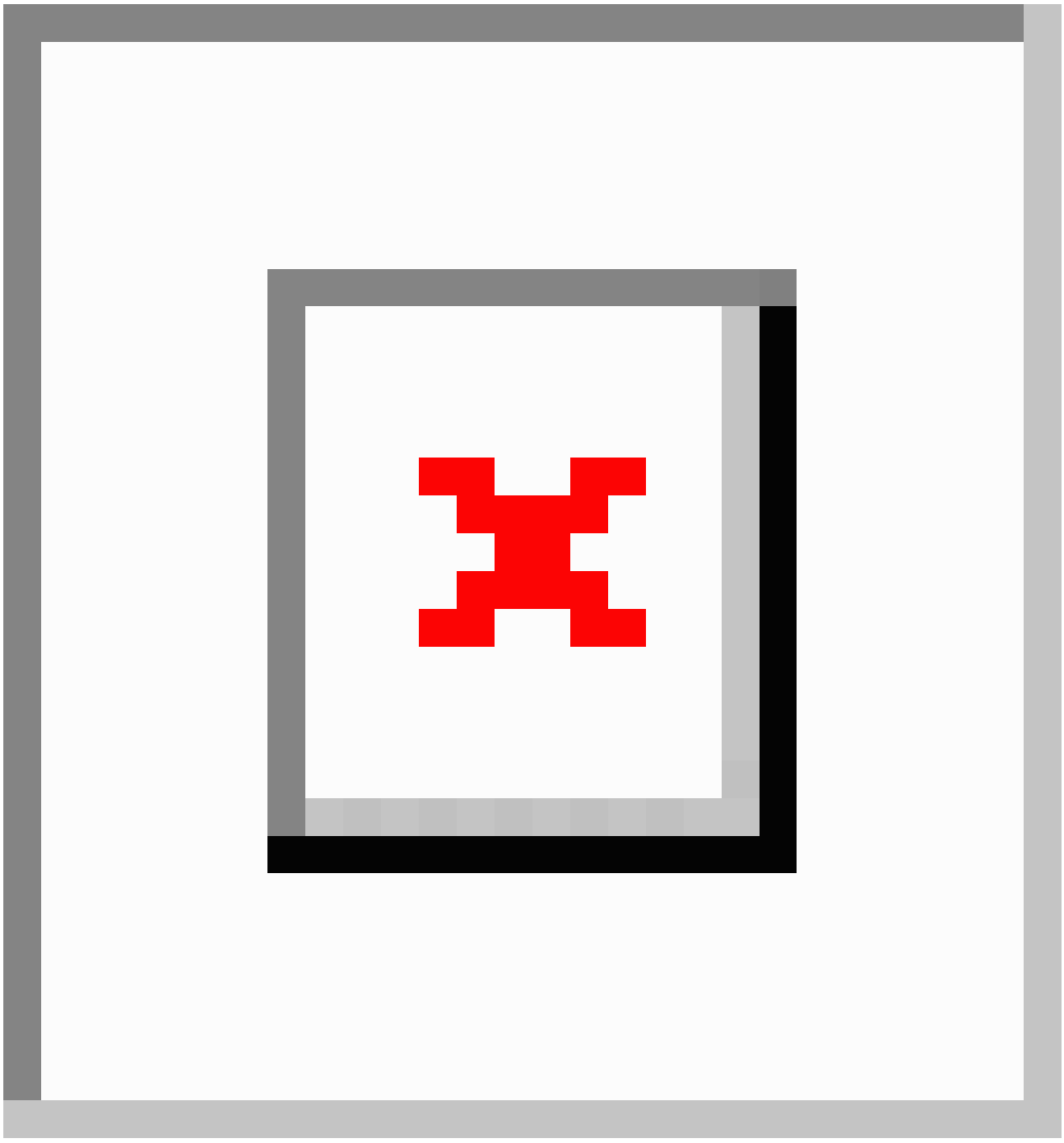
In our cohorts, we adopted criteria based on serum creatinine (SCr) to define AKI according to the Kidney Disease: Improving Global Outcomes (KDIGO) Clinical Practice Guidelines and the previously defined AKI classification stages mapped in the “injury” category [24-26]. The criterion is an increase in the SCr level to 2 times the baseline value. As an alternative to the baseline SCr levels, we defined the upper limit of normal (ULN) value of SCr as 1.2 mg/dl [27].

The targeted drugs were selected from 5 medications associated with a high risk of AKI according to the US Food and Drug Administration (FDA) and previous studies: acetaminophen; vancomycin; 2 nonsteroidal anti-inflammatory drugs (NSAIDs), naproxen and celecoxib; and 1 antiviral drug, acyclovir [21,22].

The inclusion criteria were as follows: (1) the target drugs were administered, (2) the patient had a visit record of at least 30 days prior to the observation period, and (3) the patient underwent at least 2 SCr tests during the preobservation period (0 - 60 days before the study). The exclusion criteria were as follows: (1) the patient had at least 1 SCr test outside the ULN value in the preobservation period. Participants were divided into case and control cohorts based on whether they met previously defined AKI criteria for 60 days after taking the first medicine. The observation period refers to the time range before and after the initial medication intake for each patient within the cohort. The cohort definitions were created using ATLAS, a web-based tool (Observational Health Data Sciences and Informatics), and are available as JSON files on GitHub [28].

To adjust for differences between cases and controls to reduce the effect of confounding variables, we used propensity score matching (PSM). Covariates included were age, sex, and SCr value at baseline. We normalized the covariates by applying standard scaling to ensure consistent dimensions across variables. A propensity score for each patient was generated by logistic regression. Patients were matched in a 1:3 ratio using a K-nearest neighbor (K-NN) algorithm using the Python *scikit-learn* library.

Figure 1. Overall flowchart for predicting acute kidney injury events. AJUH: Ajou University Hospital; GSH: Gangnam Severance Hospital; IMV-LSTM: interpretable multivariable long short-term memory; KYUH: Konyang University Hospital; NCC: National Cancer Center; OMOP: Observational Medical Outcomes Partnership; SH: Severance Hospital; SNUH: Seoul National University Cancer Hospital; ULN, upper limit of normal.



Candidate Predictors for Time Series

Candidate predictors were extracted from several key domains within the CDM containing per-patient observational data using structured query language tools in Python. Age and sex were used in the *person* domain, clinical laboratory tests in the *measurement* domain, medications in the *drug exposure* domain, diagnostic records in the *condition occurrence* domain, and surgical/procedure records in the *procedure occurrence* domain. Lab tests were treated as continuous variables, while other medications, conditions, and procedures were treated as binary variables. Statistical methods were used to select the variables. To identify predictors, we tested the statistical significance of the difference between the enrollment time of the cohort and the onset date of AKI using a 2-tailed paired *t* test and the McNemar test for continuous and dichotomous variables, respectively. To create a time series table, the candidate variables were pivoted into columns and dates were placed into rows. Missing values were handled in the following ways: forward fill for laboratory tests and diagnoses and zero fill for medications and treatments. The window size for predictions used 4-week sequence data and was processed by shifting the data of the prediction cycle by 2 weeks.

AKI Prediction Modeling

LSTM models based on recurrence have been designed to process time series data [29]. Attention-based LSTM models were initially proposed for learning words and the relationships between words in natural language processing [30,31] and later evolved into a key component of deep learning, becoming one of the methods used to provide interpretations, including importance scores for predicted outcomes.

As an advanced LSTM model, we used the IMV-LSTM module for the learning model, which is a multivariate LSTM neural network for the prediction and interpretation of multivariate time series [23]. This model improves on the LSTM-attention model, which can predict variable importance using multivariate inputs to configure variable-wise hidden states and mix both temporal and variable levels of attention for improved interpretability. The model was trained for 200 epochs with a batch size of 64 and a learning rate of 1e-3. An Adam optimizer was used with early stopping after 20 epochs. The data set was divided into training, test, and validation sets at a 6:2:2 ratio. Prediction performance was evaluated using the area under the receiver operating characteristic curve (AUROC) value. Additionally, we used the accuracy, precision, F_1-score, and area under the precision-recall curve (AUPRC) to ensure robustness for unbalanced data.

In this study, AKI prediction models were created for each hospital and drug. Each model had a different selection of candidate variables. To interpret the predictors in each model, variable- and temporal-wise attention scores were extracted from all trained models. These scores were then aggregated by calculating the overall temporal attention score, which was obtained by taking the weighted average of the temporal attention value over the attention value for each predictor variable. The resulting scores were plotted as heat maps for interpretation.

Statistical Analysis

We used statistical packages based on Python and R (R Project for Statistical Computing) for the statistical analysis. First, to compare the AKI and non-AKI groups, we calculated significance using the χ^2 test for categorical variables and an independent-sample 2-tailed *t* test for continuous variables. Second, to identify differences in the pattern of AKI occurrence between cohorts and drugs, a histogram was plotted for patients in each cohort from the date of cohort entry (the first day of administration of the target drug) to the date of AKI occurrence. Differences between drugs were analyzed using an independent-sample2-tailed *t* test. Third, we compared the distribution of the aggregated temporal attention scores with the actual trained data with box plots of the data for 4 weeks at 1-week intervals. A repeated ANOVA test was performed to identify temporal differences.

Results

Demographic and Clinical Characteristics

The demographics of the 31,012 patients without AKI and the 8643 patients with AKI across the 6 hospitals after PSM are shown in Table 1.

Table . Demonstration and clinical characteristics of patients with and without AKI across 6 hospitals after propensity score matching. The *P* values were obtained by conducting a 2-sample *t* test to compare the means for cases and controls.

	Case group							Control group							*P* value
	SH[a] (n=3028)	GSH[b] (n=491)	KYUH[c] (n=539)	AJUH[d] (n=1003)	SNUH[e] (n=2616)	NCC[f] (n=966)	Total (n=8643)	SH (n=11,018)	GSH (n=1699)	KYUH (n=1954)	AJOU (n=3328)	SNUH (n=9558)	NCC (n=3465)	Total (n=31,012)	
Age (years), mean (SD)	61.83 (15.23)	62.6 (15.04)	67.91 (13.64)	59.52 (15.79)	57.89 (16.49	60.55 (12.24)	60.65 (15.5)	61.14 (15.37)	62.26 (14.81)	67.71 (13.55)	59.67 (15.77)	57.46 (16.61)	60.14 (12.71)	60.21 (15.61)	.02
Gender, n (%)															
Male	1965 (64.89)	286 (58.25)	360 (66.79)	628 (62.61)	1652 (63.15)	569 (58.9)	5460 (63.17)	7092 (64.37)	985 (58.32)	950 (48.62)	1765 (53.03)	4602 (48.15)	2018 (58.24)	17,412 (56.15)	<.001
Female	1063 (35.11)	205 (41.75)	179 (33.21)	375 (37.39)	964 (36.85)	397 (41.1)	3183 (36.83)	3926 (35.63)	704 (41.68)	1004 (51.38)	1563 (46.97)	4956 (51.85)	1447 (41.76)	13,600 (43.85)	<.001
Sepsis, n (%)	325 (10.73)	50 (10.18)	7 (1.3)	139 (13.86)	40 (1.53)	1 (0.1)	562 (6.5)	373 (3.39)	37 (2.19)	21 (1.07)	128 (3.85)	41 (0.43)	1 (0.03)	601 (1.94)	<.001
Diabetes mellitus, n (%)	915 (30.22)	117 (23.83)	65 (12.06)	223 (22.23)	300 (11.47)	59 (6.11)	1679 (19.43)	2462 (22.35)	266 (15.75)	195 (9.98)	593 (17.82)	836 (8.75)	167 (4.82)	4519 (14.57)	<.001
Chronic kidney disease, n (%)	142 (4.69)	3 (0.61)	24 (4.45)	13 (1.3)	18 (0.69)	2 (0.21)	202 (2.34)	271 (2.46)	8 (0.47)	51 (2.61)	11 (0.33)	40 (0.42)	3 (0.09)	384 (1.24)	<.001
Chronic liver disease, n (%)	462 (15.26)	56 (11.41)	52 (9.65)	139 (13.86)	506 (19.34)	16 (1.66)	1231 (14.24)	766 (6.95)	82 (4.85)	71 (3.63)	126 (3.79)	759 (7.94)	42 (1.21)	1846 (5.95)	<.001
Hypoalbuminemia, n (%)	7 (0.23)	2 (0.41)	0	0	0	0	9 (0.1)	9 (0.08)	2 (0.12)	0	0	0	0	11 (0.04)	.01
Hypotension, n (%)	12 (0.4)	4 (0.81)	0	1 (0.1)	4 (0.15)	0	21 (0.24)	30 (0.27)	3 (0.18)	0	4 (0.12)	28 (0.29)	0	65 (0.21)	.56
Hypertension, n (%)	1462 (48.28)	179 (36.46)	77 (14.29)	351 (35)	264 (10.09)	71 (7.35)	2404 (27.81)	4178 (37.92)	455 (26.94)	360 (18.42)	1143 (34.34)	924 (9.67)	229 (6.61)	7289 (23.5)	<.001
Neoplasm (active cancers), n (%)	2112 (69.75)	320 (65.17)	255 (47.31)	652 (65)	2081 (79.55)	581 (60.14)	6001 (69.43)	5512 (50.03)	531 (31.44)	417 (21.34)	1282 (38.52)	4450 (46.56)	2470 (71.28)	14,662 (47.28)	<.001
Heart failure, n (%)	266 (8.78)	13 (2.65)	34 (6.31)	33 (3.29)	62 (2.37)	6 (0.62)	414 (4.79)	629 (5.71)	20 (1.18)	105 (5.37)	60 (1.8)	127 (1.33)	10 (0.29)	951 (3.07)	<.001
Obesity, n (%)	3 (0.1)	2 (0.41)	0	2 (0.2)	3 (0.11)	0	10 (0.12)	36 (0.33)	3 (0.18)	4 (0.2)	10 (0.3)	50 (0.52)	0	103 (0.33)	<.001
Peripheral vascular disease, n (%)	25 (0.83)	9 (1.83)	19 (3.53)	18 (1.79)	10 (0.38)	0	81 (0.94)	51 (0.46)	26 (1.54)	69 (3.53)	35 (1.05)	37 (0.39)	2 (0.06)	220 (0.71)	.03
Liver dysfunction, n (%)	62 (2.05)	2 (0.41)	2 (0.37)	40 (3.99)	31 (1.19)	2 (0.21)	139 (1.61)	71 (0.64)	4 (0.24)	25 (1.28)	46 (1.38)	59 (0.62)	3 (0.09)	208 (0.67)	<.001
Anemia, n (%)	541 (17.87)	25 (5.09)	35 (6.49)	89 (8.87)	154 (5.89)	4 (0.41)	848 (9.81)	903 (8.2)	53 (3.14)	68 (3.48)	176 (5.29)	282 (2.95)	7 (0.2)	1489 (4.8)	<.001
Prior kidney surgery, n (%)	0	0	1 (0.19)	3 (0.3)	14 (0.54)	0	18 (0.21)	0	0	0	0	11 (0.12)	1 (0.03)	12 (0.04)	<.001
Laboratory values (before medication), mean (SD)															
Serum creatinine (mg/dL)	0.46 (0.74)	0.79 (0.23)	0.99 (0.37)	0.87 (0.37)	0.87 (0.47)	1.01 (0.29)	0.75 (0.59)	0.44 (0.51)	0.76 (0.19)	0.98 (0.22)	0.84 (0.2)	0.83 (0.18)	0.98 (0.16)	0.72 (0.4)	<.001
Glucose (mg/dL)	126.89 (58.33)	132.1 (52.11)	150.75 (82.36)	187.0 (23.0)	121.48 (48.26)	125.85 (50.05)	126.93 (56.28)	121.72 (55.7)	124.3 (49.43)	140.11 (78.81)	151.5 (39.5)	115.65 (40.5)	120.88 (46.23)	120.95 (51.85)	<.001
Potassium (mmol/L)	3.92 (0.6)	4.21 (0.47)	4.14 (0.6)	4.17 (0.55)	4.22 (0.5)	4.29 (0.56)	4.11 (0.57)	3.93 (0.55)	4.22 (0.4)	4.09 (0.54)	4.15 (0.46)	4.22 (0.43)	4.28 (0.52)	4.11 (0.51)	.48

	Case group							Control group							*P* value
	SH[a] (n=3028)	GSH[b] (n=491)	KYUH[c] (n=539)	AJUH[d] (n=1003)	SNUH[e] (n=2616)	NCC[f] (n=966)	Total (n=8643)	SH (n=11018)	GSH (n=1699)	KYUH (n=1954)	AJOU (n=3328)	SNUH (n=9558)	NCC (n=3465)	Total (n=31012)	
Sodium (mmol/L)	138.6 (4.16)	137.27 (4.36)	136.49 (4.61)	138.48 (4.23)	138.46 (3.96)	138.31 (3.67)	138.3 (4.14)	139.44 (3.75)	138.31 (3.58)	137.74 (3.73)	139.27 (3.85)	139.71 (3.23)	138.96 (3.37)	139.28 (3.6)	<.001
BUN (blood urea nitrogen) (mg/dL)	17.83 (9.66)	15.44 (6.58)	17.67 (8.87)	14.68 (6.36)	14.72 (7.61)	15.55 (6.13)	16.13 (8.28)	16.82 (7.48)	15.78 (5.86)	17.19 (7.62)	14.9 (6.41)	14.67 (5.14)	14.63 (4.96)	15.67 (6.47)	<.001

[a]SH: Severance Hospital.

[b]GSH: Gangnam Severance Hospital.

[c]KYUH: Konyang University Hospital.

[d]AJUH: Ajou University Hospital.

[e]SNUH: Seoul National University Cancer Hospital.

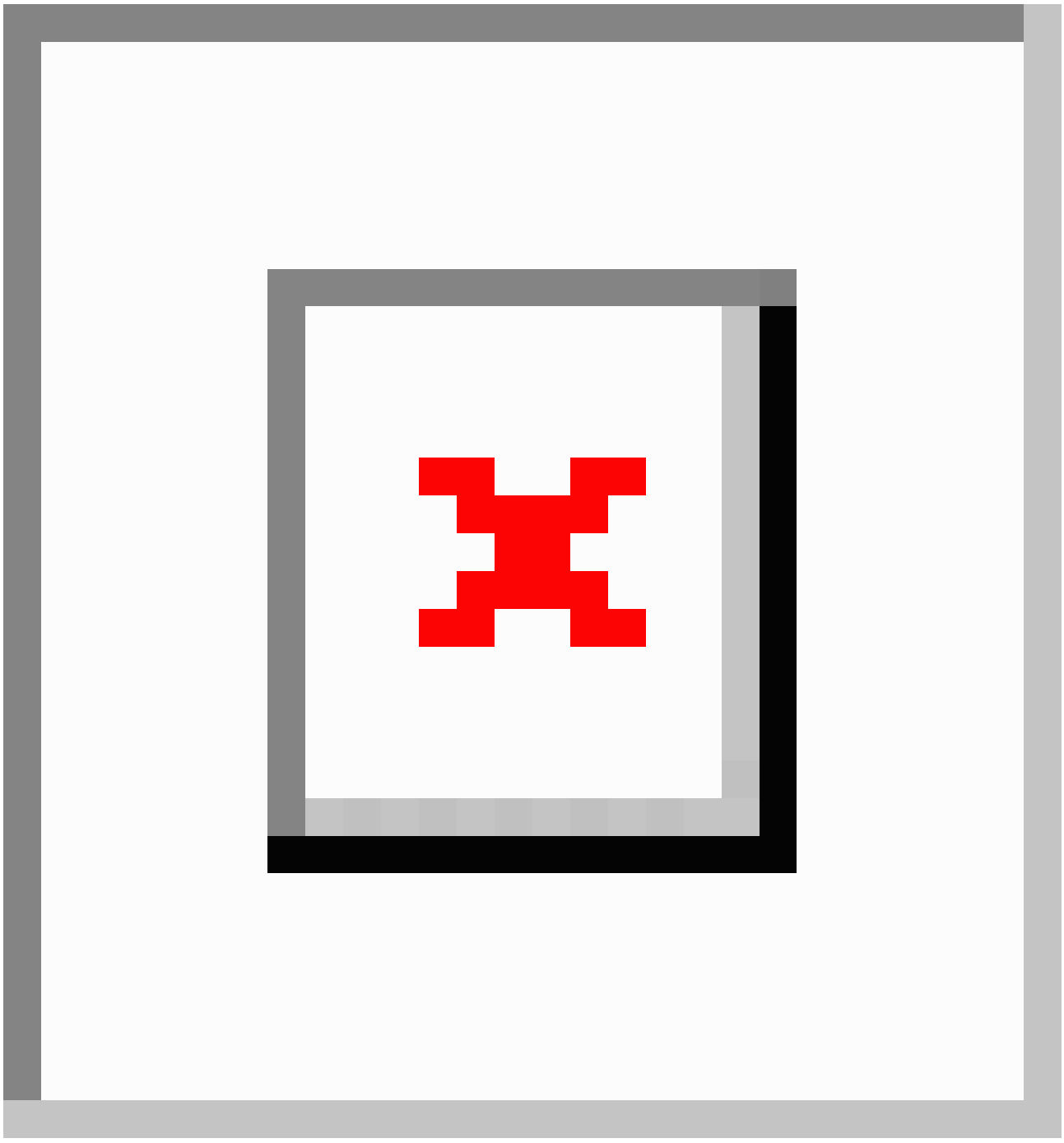
[f]NCC: National Cancer Center.

By hospital, the cohort consisted of 14,046 patients from SH, 2180 from GSH, 2493 from KYUH, 4331 from AJUH, 13,174 from SNUH, and 4431 from NCC after matching the propensity scores. Propensity matching was performed using age, sex, and SCr levels at baseline. As for the changes in covariates, the difference in mean age decreased from 6.39 (60.65 – 54.26) to 0.44 (60.65 – 60.21) years, the difference in male ratio decreased from 16.49% (63.17% – 46.68%) to 7.02% (63.17% – 56.15%), and the difference in SCr at baseline decreased from 0.14 mg/dL (0.71 – 0.58 mg/dL) to 0.03 mg/dL (0.71 – 0.67 mg/dL) (Table S2 in Multimedia Appendix 1). There were still statistically significant differences in PSM age (60.66, SD 15.86 vs 60.22, SD 15.94 years; *P*=.03), gender (63.17% vs 56.15% male; *P*<.001), and SCr at baseline (0.71, SD 0.61 vs 0.68, SD 0.41 mg/dL; *P*<.001). Patients who developed AKI had more severe neoplasms (ie, active cancers; 70.6% vs 44.26%; *P*<.001) and chronic liver disease (15.83% vs 6.55%; *P*<.001). Moreover, the analysis revealed the following differences: sepsis (6.5% vs 1.94%; *P*<.001), diabetes mellitus (19.43% vs 14.57%; *P*<.001), hypertension (27.81% vs 23.5%; *P*<.001), anemia (9.81% vs 4.8%; *P*<.001), and heart failure (4.79% vs 3.07%; *P*<.001). There was no significant difference between hypotension (0.24% vs 0.21%; *P*=.56), potassium (4.11, SD 0.57 vs 4.11, SD 0.51 mmol/L; *P*=.48), or renal artery stenosis (0.07% vs 0.03%; *P*=.13). Hypoalbuminemia, obesity, peripheral vascular disease, renal artery stenosis, liver dysfunction, and prior kidney surgery had low incidence rates (<2%).

Distribution of Adverse Drug Events

To analyze the differences in drug patterns at the time of AKI occurrence, we assessed the pattern for each drug. The median number of days for the occurrence of AKI among the drug and cohort patients in the entire hospital was 17 (IQR 7-33 days). Vancomycin appeared after a median period of 12 days, followed by naproxen (18 days), acetaminophen (19 days), celecoxib (22 days), and acyclovir (23 days). When comparing the IQR values, celecoxib (10 - 41 days) and Acyclovir (10 - 41 days) showed a relatively broad distribution, whereas acetaminophen (9 - 34 days) and naproxen (8 - 34 days) were distributed over 25 days. Vancomycin (5 - 25 days) exhibited the narrowest distribution. Celecoxib and acyclovir tended to be relatively distributed compared to acetaminophen, vancomycin, and naproxen. We compared the onset times of all drug pairs and hospital pairs to check the similarity in AKI occurrence (Figure 2). The patterns between specific drugs was similar for celecoxib and acyclovir (*P*=.88) and for acetaminophen and naproxen (*P*=.57). The patterns between hospitals were similar for SH and AJUH (*P*=.98), SH and GSH (*P*=.36), GSH and AJUH (*P*=.42), GSH and NCC (*P*=.24), and SNUH and NCC (*P*=.26).

Figure 2. (**A**) Comparison of acute kidney injury (AKI) onset time between drugs and (**B**) AKI onset time between hospitals. The *P* values were obtained by conducting independent 2-tailed *t* tests between each aggregated pair. AJUH: Ajou University Hospital; GSH: Gangnam Severance Hospital; KYUH: Konyang University Hospital; NCC: National Cancer Center; SH: Severance Hospital; SNUH: Seoul National University Cancer Hospital.

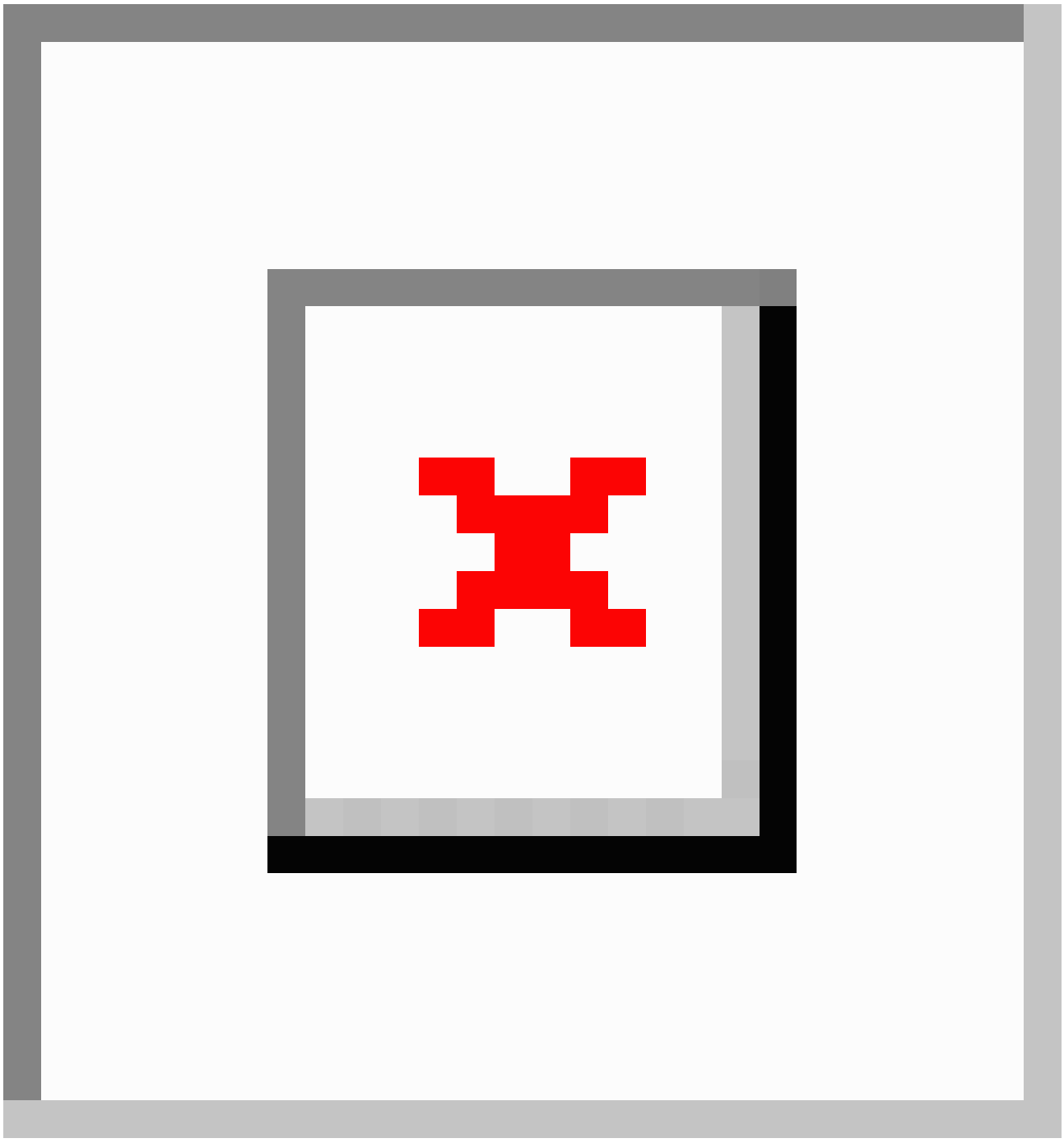


AKI Prediction Model Performance

The AUROC for each drug and hospital to evaluate the AKI predictive model, based on respective test sets (internal validation) is shown in Figure 3. A total of 26 trained models achieved a high AUROC value, of 0.92 on average, with each verification data set. In addition, among the averages of the drugs, acyclovir had the highest average AUROC score of 0.94, followed by acetaminophen (0.93), vancomycin (0.92), naproxen (0.90), and celecoxib (0.89). The highest AUROC value (0.97) was observed for the model of SH's celecoxib and acyclovir, SNUH's vancomycin, and KYUH's acyclovir prescription patients.

Multimedia Appendix 2 presents data on the precision, accuracy, F_1-score, and AUPRC of each predictive model. Overall, the average accuracy of the AKI prediction models was 0.88, whereas the average AUPRC and F_1-scores were both 0.78. The acyclovir prescription model achieved the highest accuracy score (0.91), followed by vancomycin (0.90), acetaminophen (0.89), naproxen (0.89), and celecoxib (0.86). Individually, the acyclovir SH model showed the best performance, with an AUPRC of 0.92 and an accuracy of 0.91.

Figure 3. Receiver operating characteristic (ROC) curves and areas under the curve (AUCs) of the acute kidney injury prediction model for each hospital and each drug. The square brackets indicate the 95% CI. AJUH: Ajou University Hospital; GSH: Gangnam Severance Hospital; KYUH: Konyang University Hospital; NCC: National Cancer Center; SH: Severance Hospital; SNUH: Seoul National University Cancer Hospital.



Temporal Feature Importance of the AKI Prediction Model

To interpret the AKI prediction model, we demonstrated the temporal attention values of each contributing variable in the 4 weeks prior to AKI onset, which were weighted aggregates from the model for each drug and hospital, as shown in Figure 4A. The temporal change pattern of the actual data corresponding to each variable in the 4 weeks prior to AKI onset is shown in Figure 4B. We also confirmed the difference in the distribution of highly important features between the case and control data using a 1-way ANOVA. The attention scores for all variables across all hospitals are detailed in Multimedia Appendix 3

The last week of lymphocytes (attention score at −1 week: 0.41) and the second week of calcium (attention score at −3 weeks: 0.41) showed the highest attention scores, followed by albumin (attention score at −1 week: 0.37; attention score at −4 weeks: 0.37), hemoglobin (attention score at −4 weeks: 0.37), and cholesterol (attention score at −4 weeks: 0.37). In Figure 5, the distribution of data by variable between the 2 groups was confirmed using actual data. There was a difference in data distribution between the case and control groups from the beginning. The values of lymphocytes, albumin, and hemoglobin in the case group decreased over time, while urine pH and prothrombin time in the case group tended to increase over time.

Figure 4. (**A**) Temporal attention score of important features of acute kidney injury (AKI) prediction model and (**B**) distribution of data over time (*P* values: repeated measures ANOVA). The figure shows the change over a 4-week period prior to the AKI event.

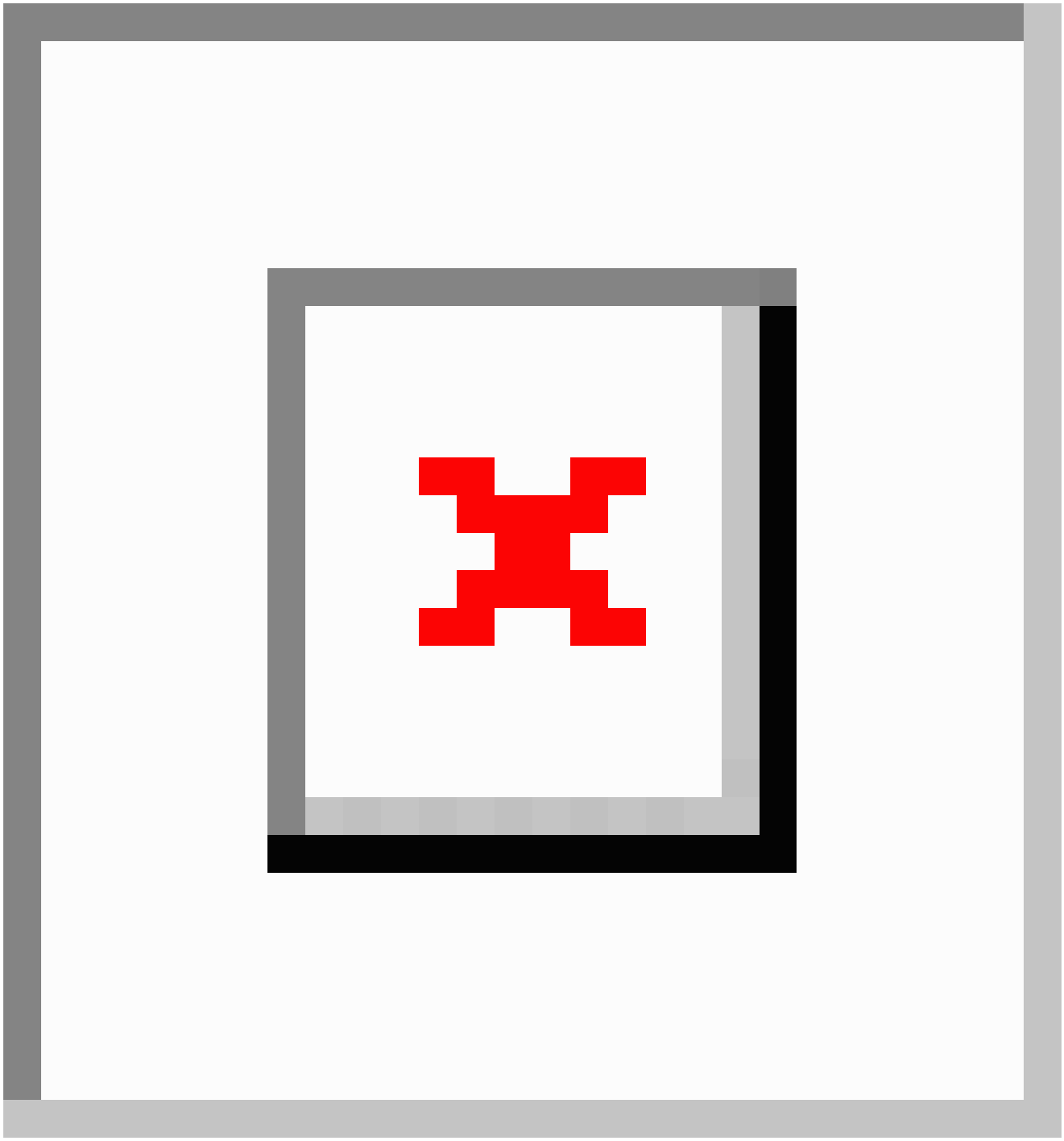
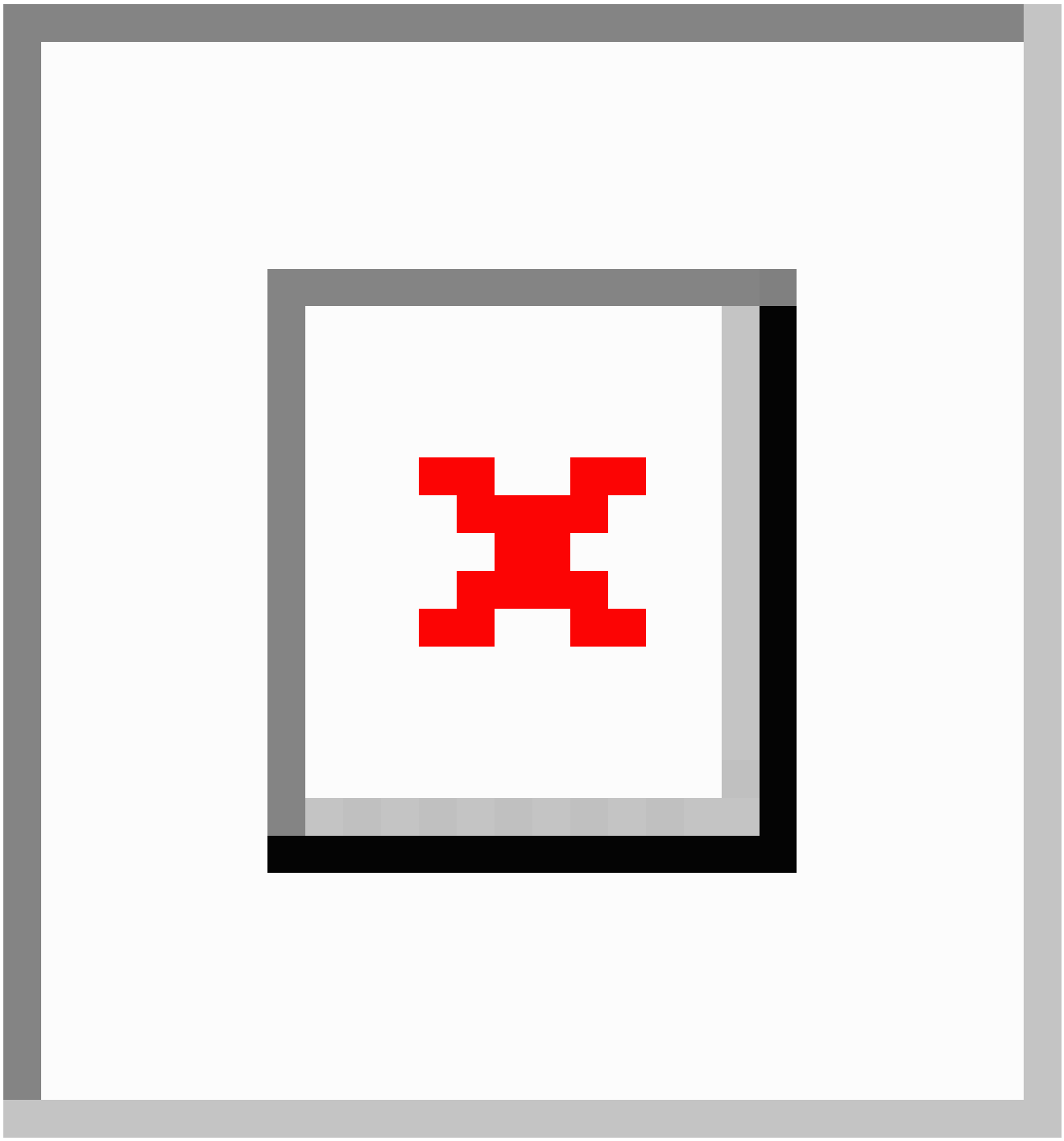


Figure 5. Acute kidney injury (AKI) onset time after drug administration at various medical centers. The red lines shows the median value. AJUH's naproxen and GSH's acyclovir cases were excluded as the number of AKI case groups was less than 20. AJUH: Ajou University Hospital; GSH: Gangnam Severance Hospital; KYUH: Konyang University Hospital; NCC: National Cancer Center; SH: Severance Hospital; SNUH: Seoul National University Cancer Hospital.



Discussion

Principal Findings

In this study, we developed time series–based IMV-LSTM models to predict AKI in patients taking specific nephrotoxic drugs using CDM-based DRNs in a 6-hospital EHR-based system. The principal findings are as follows: first, this study provides an interpretation of the temporal importance of variables for predicting AKI, and the models also achieved high performance, with an average AUC of 0.92%. Second, our study is a scalable multicenter study using a DRN, which can contribute to understanding drug-induced AKI. To the best of our knowledge, this is the first study to build an AKI prediction model by applying a time series–based IMV-LSTM model to a CDM using EHR data from 6 hospitals.

We established a retrospective cohort of patients who took nephrotoxicity-inducing drugs at 6 hospitals. With respect to demographic characteristics, we observed variations in the overall patient count and prevalence of comorbidities when comparing individuals with AKI and without AKI across different hospitals. Nevertheless, the majority of patients who developed AKI at most hospitals were older than 60 years and

had a high prevalence of comorbidities, including cancer (n=6001, 69.43%), hypertension (n=2404, 27.81%), diabetes (n=1679, 19.43%), and chronic liver disease (n=1231, 14.24%), which is consistent with findings reported in previous studies [32-35].

The pattern of each drug's association with AKI (Figure 5) showed that the median number of days for AKI onset when using nephrotoxic drugs was 17 (IQR 7-33) days. The onset occurred earliest with vancomycin (12, IQR 5-25 days) and latest with acyclovir (23, IQR 10-41 days). In previous studies [36,37], the time to onset of vancomycin-induced AKI showed a similar pattern to our results. We also found differences in the AKI onset between different classes within the same NSAID, and the multicenter AKI cohort showed similarities between hospitals. The finding of similar patterns in the AKI onset in the multicenter cohort supports the reliability of the AKI cohort and increases the explanatory power of AKI prediction models.

AKI is common among inpatients [38,39]. Previous models predicted AKI in the intensive care unit (ICU) and in surgical patients during hospital admission. For example, Zimmerman et al [40] predicted the occurrence of AKI in ICU inpatients (AUC 0.783), and Tseng et al [41] developed a predictive score for the development of AKI after cardiac surgery (AUC 0.839). Hsu et al [42] developed a risk score function for community-acquired AKI for inpatients (AUC 0.818). Koyner et al [10] developed a model to predict AKI in hospitalized patients (AUC 0.90). Despite this progress, few studies have applied time series deep learning to provide a temporal interpretation of drug-induced AKI, and our model stands out because it can predict nephrotoxic drug–induced AKI in a diverse hospital population.

This study achieved improved performance compared to previous AKI studies using recurrent neural networks (RNNs) [21,43-45]. Our model improves performance up to an AUC of 0.97 and an overall average of 0.92, which outperforms previous studies showing AKI prediction with RNN-based methods by Kim et al [43] in hospitalized patients (AUC 0.927), Rank et al [45] in cardiac surgery patients (AUC 0.893), and Xu et al [44] in inpatients (AUC 0.908). These results show promise for our model as a tool to predict AKI and facilitate early intervention and mitigation strategies for patients.

This study also provides additional interpretations regarding the temporal importance of features for AKI prediction. Some studies have reported information on interpretability or provided information about the interpretability of variables at the feature importance level [46]. However, our results show the importance of variables and the temporal importance of variables in the development of AKI. In this study, we highlight the vital role of temporal patterns of various indicators, such as lymphocytes, calcium, albumin, hemoglobin, and cholesterol, in predicting disease states, particularly the onset of AKI. The temporal pattern of lymphocytes increased gradually, peaking 1 week before AKI onset. The use of lymphocyte and neutrophil counts as predictive factors for AKI is consistent with other studies [47,48]. Calcium shows a pattern of peaking 3 weeks before AKI onset, and Prior studies showed an association between impacted calcium metabolism and AKI [49,50]. Albumin shows the highest pattern 1 and 4 weeks before onset, and low serum albumin levels (hypoalbuminemia) are a predictor of AKI [48,51,52]. Hemoglobin shows a pattern with a peak 4 weeks before onset, and previous studies have shown that the risk of AKI increases stepwise with a further decrease in hemoglobin concentration [53]. Temporal variations in variables based on reported laboratory data for the early detection of AKI emphasize the importance of monitoring and early intervention in populations.

In addition, this retrospective study can be followed by a subsequent study to validate the practicality of the AKI prediction model in clinical practice by applying it to a prediction system in a hospital EHR.

Limitations

This study has several limitations. First, because we used the CDM, it does not reflect the full range of clinical data. For example, we could not include admission records, which would have revealed a patient's condition. However, the use of CDM data allowed for a multicenter study that could be easily extended to other institutions that have converted to the CDM. Second, this was a retrospective study and could not address the underlying causes of AKI. Therefore, prospective studies are needed for validation with actual clinical data. Third, as with all retrospective studies, there may be unintentional patient selection bias and unaccounted-for confounders. However, to compensate for these limitations, we tried to equalize the distribution of patient characteristics through PSM. We also used a limited follow-up period to minimize the impact of these factors.

Conclusions

This study demonstrates the high performance of the IMV-LSTM method for AKI prediction using hospital EHR-based time series data. Our model can provide real-time assessment of AKI occurrence and individualized risk factors for AKI using time series data. We also demonstrated the robustness of our model through multicenter validation using a CDM through a DRN of 6 hospitals in South Korea, which also proves that scalability to other institutions that are converted to the CDM is possible. This may provide an objective quantitative tool for identifying patients at risk of developing AKI.

Acknowledgments

This study was supported by a grant from the Korea Institute of Drug Safety and Risk Management in 2021. This work was supported by the Bio-Industrial Technology Development Program (20014841) funded by the Ministry of Trade, Industry & Energy (Korea). We would like to thank the Medical Informatics Collaborative Unit members of Yonsei University College of Medicine for their assistance in data analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographics for each institution.
[XLSX File, 60 KB - medinform_v12i1e47693_app1.xlsx]

Multimedia Appendix 2
Model performance.
[XLSX File, 10 KB - medinform_v12i1e47693_app2.xlsx]

Multimedia Appendix 3
Feature importance.
[XLSX File, 67 KB - medinform_v12i1e47693_app3.xlsx]

References

1. Chertow GM, Burdick E, Honour M, Bonventre JV, Bates DW. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. J Am Soc Nephrol 2005 Nov;16(11):3365-3370. [doi: 10.1681/ASN.2004090740] [Medline: 16177006]
2. Waikar SS, Curhan GC, Ayanian JZ, Chertow GM. Race and mortality after acute renal failure. J Am Soc Nephrol 2007 Oct;18(10):2740-2748. [doi: 10.1681/ASN.2006091060] [Medline: 17855647]
3. Kee YK, Kim EJ, Park KS, et al. The effect of specialized continuous renal replacement therapy team in acute kidney injury patients treatment. Yonsei Med J 2015 May;56(3):658-665. [doi: 10.3349/ymj.2015.56.3.658] [Medline: 25837170]
4. Medina KRP, Jeong JC, Ryu JW, et al. Comparison of outcomes of mild and severe community- and hospital-acquired acute kidney injury. Yonsei Med J 2022 Oct;63(10):902-907. [doi: 10.3349/ymj.2021.0238] [Medline: 36168242]
5. Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V. Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak 2016 Mar 29;16:39. [doi: 10.1186/s12911-016-0277-4] [Medline: 27025458]
6. Khwaja A. KDIGO Clinical Practice Guidelines for acute kidney injury. Nephron Clin Pract 2012;120(4):c179-c184. [doi: 10.1159/000339789] [Medline: 22890468]
7. Ugwuowo U, Yamamoto Y, Arora T, et al. Real-time prediction of acute kidney injury in hospitalized adults: implementation and proof of concept. Am J Kidney Dis 2020 Dec;76(6):806-814. [doi: 10.1053/j.ajkd.2020.05.003] [Medline: 32505812]
8. Low S, Vathsala A, Murali TM, et al. Electronic health records accurately predict renal replacement therapy in acute kidney injury. BMC Nephrol 2019 Jan 31;20(1):32. [doi: 10.1186/s12882-019-1206-4] [Medline: 30704418]
9. Chua HR, Zheng K, Vathsala A, et al. Health care analytics with time-invariant and time-variant feature importance to predict hospital-acquired acute kidney injury: observational longitudinal study. J Med Internet Res 2021 Dec 24;23(12):e30805. [doi: 10.2196/30805] [Medline: 34951595]
10. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. Crit Care Med 2018 Jul;46(7):1070-1077. [doi: 10.1097/CCM.0000000000003123] [Medline: 29596073]
11. Wilson FP, Greenberg JH. Acute kidney injury in real time: prediction, alerts, and clinical decision support. Nephron 2018;140(2):116-119. [doi: 10.1159/000492064] [Medline: 30071528]
12. Huang CY, Güiza F, De Vlieger G, et al. Development and validation of clinical prediction models for acute kidney injury recovery at hospital discharge in critically ill adults. J Clin Monit Comput 2023 Feb;37(1):113-125. [doi: 10.1007/s10877-022-00865-7] [Medline: 35532860]
13. Bedford M, Stevens P, Coulton S, et al. Development of risk models for the prediction of new or worsening acute kidney injury on or during hospital admission: a cohort and nested study. Health Serv Deliv Res 2016;4(6):1-160. [doi: 10.3310/hsdr04060] [Medline: 26937542]
14. Kim WH, Lee SM, Choi JW, et al. Simplified clinical risk score to predict acute kidney injury after aortic surgery. J Cardiothorac Vasc Anesth 2013 Dec;27(6):1158-1166. [doi: 10.1053/j.jvca.2013.04.007] [Medline: 24050856]
15. Kiers HD, van den Boogaard M, Schoenmakers MCJ, et al. Comparison and clinical suitability of eight prediction models for cardiac surgery-related acute kidney injury. Nephrol Dial Transplant 2013 Feb;28(2):345-351. [doi: 10.1093/ndt/gfs518] [Medline: 23222415]
16. Zhou LZ, Yang XB, Guan Y, et al. Development and validation of a risk score for prediction of acute kidney injury in patients with acute decompensated heart failure: a prospective cohort study in China. J Am Heart Assoc 2016 Nov 16;5(11):e004035. [doi: 10.1161/JAHA.116.004035] [Medline: 27852590]
17. Wilson T, Quan S, Cheema K, et al. Risk prediction models for acute kidney injury following major noncardiac surgery: systematic review. Nephrol Dial Transplant 2016 Feb;31(2):231-240. [doi: 10.1093/ndt/gfv415] [Medline: 26705194]

18. Sanchez-Pinto LN, Khemani RG. Development of a prediction model of early acute kidney injury in critically ill children using electronic health record data. Pediatr Crit Care Med 2016 Jun;17(6):508-515. [doi: 10.1097/PCC.0000000000000750] [Medline: 27124567]
19. Sutherland SM, Chawla LS, Kane-Gill SL, et al. Utilizing electronic health records to predict acute kidney injury risk and outcomes: workgroup statements from the 15th ADQI Consensus Conference. Can J Kidney Health Dis 2016;3:11. [doi: 10.1186/s40697-016-0099-4] [Medline: 26925247]
20. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May;1(5):206-215. [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
21. Dong J, Feng T, Thapa-Chhetry B, et al. Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care. Crit Care 2021 Aug 10;25(1):288. [doi: 10.1186/s13054-021-03724-0] [Medline: 34376222]
22. Lee JM, Hauskrecht M. Modeling multivariate clinical event time-series with recurrent temporal mechanisms. Artif Intell Med 2021 Feb;112:102021. [doi: 10.1016/j.artmed.2021.102021] [Medline: 33581828]
23. Guo T, Lin T, Antulov-Fantulin N, editors. Exploring interpretable LSTM neural networks over multi-variable data. In: Proceedings of the 36th International Conference on Machine Learning: PMLR; 2019.
24. Kellum JA, Lameire N, Aspelin P, et al. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO Clinical Practice Guideline for acute kidney injury. Kidney Int Suppl 2012;2(1):1-138. [doi: 10.1038/kisup.2012.1]
25. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P, Acute Dialysis Quality Initiative workgroup. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care 2004 Aug;8(4):R204-R212. [doi: 10.1186/cc2872] [Medline: 15312219]
26. Mehta RL, Kellum JA, Shah SV, et al. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Crit Care 2007;11(2):R31. [doi: 10.1186/cc5713] [Medline: 17331245]
27. Hosten AO. BUN and creatinine. In: Clinical Methods: The History, Physical, and Laboratory Examinations, 3rd edition: Butterworths; 1990.
28. Development and verification of a time series AI model for acute kidney injury detection based on a multicenter distributed research network. GitHub. URL: https://github.com/DigitalHealthcareLab/22MOACDM [accessed 2023-07-08]
29. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
30. Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv. Preprint posted online on Sep 20, 2015. [doi: 10.48550/arXiv.1508.04025]
31. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv. Preprint posted online on Sep 1, 2014. [doi: 10.48550/arXiv.1409.0473]
32. Harirforoosh S, Asghar W, Jamali F. Adverse effects of nonsteroidal antiinflammatory drugs: an update of gastrointestinal, cardiovascular and renal complications. J Pharm Pharm Sci 2013;16(5):821-847. [doi: 10.18433/j3vw2f] [Medline: 24393558]
33. Perazella MA. Drug-induced nephropathy: an update. Expert Opin Drug Saf 2005 Jul;4(4):689-706. [doi: 10.1517/14740338.4.4.689] [Medline: 16011448]
34. Khan S, Loi V, Rosner MH. Drug-induced kidney injury in the elderly. Drugs Aging 2017 Oct;34(10):729-741. [doi: 10.1007/s40266-017-0484-4] [Medline: 28815461]
35. Finlay S, Bray B, Lewington AJ, et al. Identification of risk factors associated with acute kidney injury in patients admitted to acute medical units. Clin Med (Lond) 2013 Jun;13(3):233-238. [doi: 10.7861/clinmedicine.13-3-233] [Medline: 23760694]
36. van Hal SJ, Paterson DL, Lodise TP. Systematic review and meta-analysis of vancomycin-induced nephrotoxicity associated with dosing schedules that maintain troughs between 15 and 20 milligrams per liter. Antimicrob Agents Chemother 2013 Feb;57(2):734-744. [doi: 10.1128/AAC.01568-12] [Medline: 23165462]
37. Filippone EJ, Kraft WK, Farber JL. The nephrotoxicity of vancomycin. Clin Pharmacol Ther 2017 Sep;102(3):459-469. [doi: 10.1002/cpt.726] [Medline: 28474732]
38. Kate RJ, Pearce N, Mazumdar D, Nilakantan V. A continual prediction model for inpatient acute kidney injury. Comput Biol Med 2020 Jan;116:103580. [doi: 10.1016/j.compbiomed.2019.103580] [Medline: 32001013]
39. Kellum JA, Romagnani P, Ashuntantang G, Ronco C, Zarbock A, Anders HJ. Acute kidney injury. Nat Rev Dis Primers 2021 Jul 15;7(1):52. [doi: 10.1038/s41572-021-00284-z] [Medline: 34267223]
40. Zimmerman LP, Reyfman PA, Smith ADR, et al. Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements. BMC Med Inform Decis Mak 2019 Jan 31;19(Suppl 1):16. [doi: 10.1186/s12911-019-0733-z] [Medline: 30700291]
41. Tseng PY, Chen YT, Wang CH, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 2020 Jul 31;24(1):478. [doi: 10.1186/s13054-020-03179-9] [Medline: 32736589]
42. Hsu CN, Liu CL, Tain YL, Kuo CY, Lin YC. Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: development and validation study. J Med Internet Res 2020 Aug 4;22(8):e16903. [doi: 10.2196/16903] [Medline: 32749223]

43. Kim K, Yang H, Yi J, et al. Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: external validation and model interpretation. J Med Internet Res 2021 Apr 16;23(4):e24120. [doi: 10.2196/24120] [Medline: 33861200]
44. Xu J, Hu Y, Liu H, et al. A novel multivariable time series prediction model for acute kidney injury in general hospitalization. Int J Med Inform 2022 May;161:104729. [doi: 10.1016/j.ijmedinf.2022.104729] [Medline: 35279551]
45. Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. NPJ Digit Med 2020;3:139. [doi: 10.1038/s41746-020-00346-8] [Medline: 33134556]
46. Wang Y, Wei Y, Yang H, Li J, Zhou Y, Wu Q. Utilizing imbalanced electronic health records to predict acute kidney injury by ensemble learning and time series model. BMC Med Inform Decis Mak 2020 Sep 21;20(1):238. [doi: 10.1186/s12911-020-01245-4] [Medline: 32957977]
47. Weller S, Varrier M, Ostermann M. Lymphocyte function in human acute kidney injury. Nephron 2017;137(4):287-293. [doi: 10.1159/000478538] [Medline: 28662513]
48. Ishikawa M, Iwasaki M, Namizato D, et al. The neutrophil to lymphocyte ratio and serum albumin as predictors of acute kidney injury after coronary artery bypass grafting. Sci Rep 2022 Sep 14;12(1):15438. [doi: 10.1038/s41598-022-19772-7] [Medline: 36104386]
49. Wang B, Li D, Gong Y, Ying B, Cheng B. Association of serum total and ionized calcium with all-cause mortality in critically ill patients with acute kidney injury. Clinica Chimica Acta 2019 Jul;494:94-99. [doi: 10.1016/j.cca.2019.03.1616]
50. Thongprayoon C, Cheungpasitporn W, Mao MA, Sakhuja A, Erickson SB. Admission calcium levels and risk of acute kidney injury in hospitalised patients. Int J Clin Pract 2018 Apr;72(4):e13057. [doi: 10.1111/ijcp.13057] [Medline: 29314467]
51. Wiedermann CJ, Wiedermann W, Joannidis M. Hypoalbuminemia and acute kidney injury: a meta-analysis of observational clinical studies. Intensive Care Med 2010 Oct;36(10):1657-1665. [doi: 10.1007/s00134-010-1928-z] [Medline: 20517593]
52. Thongprayoon C, Cheungpasitporn W, Mao MA, Sakhuja A, Kashani K. U-shape association of serum albumin level and acute kidney injury risk in hospitalized patients. PLoS One 2018;13(6):e0199153. [doi: 10.1371/journal.pone.0199153] [Medline: 29927987]
53. Walsh M, Garg AX, Devereaux PJ, Argalious M, Honar H, Sessler DI. The association between perioperative hemoglobin and acute kidney injury in patients having noncardiac surgery. Anesth Analg 2013 Oct;117(4):924-931. [doi: 10.1213/ANE.0b013e3182a1ec84] [Medline: 24023017]

Abbreviations

ADR: adverse drug reaction
AI: artificial intelligence
AJUH: Ajou University Hospital
AKI: acute kidney injury
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CDM: common data model
DRN: distributed research network
EHR: electronic health record
FDA: Food and Drug Administration
GSH: Gangnam Severance Hospital
ICU: intensive care unit
IMV-LSTM: interpretable multivariable long short-term memory
KDIGO: Kidney Disease Improving Global Outcomes
KYUH: Konyang University Hospital
MIMIC: Medical Information Mart for Intensive Care
NCC: National Cancer Center
NSAID: nonsteroidal anti-inflammatory drug
OMOP-CDM: Observational Medical Outcomes Partnership Common Data Model
PSM: propensity score matching
RNN: recurrent neural network
SCr: serum creatinine
SH: Severance Hospital
SNUH: Seoul National University Cancer Hospital
ULN: upper limit of normal

Edited by C Lovis; submitted 30.03.23; peer-reviewed by M Oja, X Zhang; revised version received 08.07.23; accepted 19.05.24; published 05.07.24.

Please cite as:
Heo S, Kang EA, Yu JY, Kim HR, Lee S, Kim K, Hwangbo Y, Park RW, Shin H, Ryu K, Kim C, Jung H, Chegal Y, Lee JH, Park YR
Time Series AI Model for Acute Kidney Injury Detection Based on a Multicenter Distributed Research Network: Development and Verification Study
JMIR Med Inform 2024;12:e47693
URL: https://medinform.jmir.org/2024/1/e47693
doi:10.2196/47693

Prediction of In-Hospital Cardiac Arrest in the Intensive Care Unit: Machine Learning–Based Multimodal Approach

Hsin-Ying Lee[1,*], MD; Po-Chih Kuo[2,*], PhD; Frank Qian[3,4], MPH, MD; Chien-Hung Li[2], MS; Jiun-Ruey Hu[5], MPH, MD; Wan-Ting Hsu[6], MS; Hong-Jie Jhou[7], MD; Po-Huang Chen[8], MD; Cho-Hao Lee[9], MD; Chin-Hua Su[10], MSc; Po-Chun Liao[10], MSc; I-Ju Wu[10], MD; Chien-Chang Lee[10,11], MD, ScD

1
2
3
4
5
6
7
8
9
10
11

[*]these authors contributed equally

Corresponding Author:
Chien-Chang Lee, MD, ScD

Abstract

Background: Early identification of impending in-hospital cardiac arrest (IHCA) improves clinical outcomes but remains elusive for practicing clinicians.

Objective: We aimed to develop a multimodal machine learning algorithm based on ensemble techniques to predict the occurrence of IHCA.

Methods: Our model was developed by the Multiparameter Intelligent Monitoring of Intensive Care (MIMIC)–IV database and validated in the Electronic Intensive Care Unit Collaborative Research Database (eICU-CRD). Baseline features consisting of patient demographics, presenting illness, and comorbidities were collected to train a random forest model. Next, vital signs were extracted to train a long short-term memory model. A support vector machine algorithm then stacked the results to form the final prediction model.

Results: Of 23,909 patients in the MIMIC-IV database and 10,049 patients in the eICU-CRD database, 452 and 85 patients, respectively, had IHCA. At 13 hours in advance of an IHCA event, our algorithm had already demonstrated an area under the receiver operating characteristic curve of 0.85 (95% CI 0.815 - 0.885) in the MIMIC-IV database. External validation with the eICU-CRD and National Taiwan University Hospital databases also presented satisfactory results, showing area under the receiver operating characteristic curve values of 0.81 (95% CI 0.763-0.851) and 0.945 (95% CI 0.934-0.956), respectively.

Conclusions: Using only vital signs and information available in the electronic medical record, our model demonstrates it is possible to detect a trajectory of clinical deterioration up to 13 hours in advance. This predictive tool, which has undergone external validation, could forewarn and help clinicians identify patients in need of assessment to improve their overall prognosis.

(JMIR Med Inform 2024;12:e49142) doi:10.2196/49142

KEYWORDS

cardiac arrest; machine learning; intensive care; mortality; medical emergency team; early warning scores

Introduction

The prognosis of in-hospital cardiac arrest (IHCA) is poor as it represents the culmination of heterogeneous multi-organ dysfunction, with few treatments [1]. IHCA has an incidence of 9 to 10 per 1000 admissions and a mortality rate of 80% - 100% [2]. Therefore, clinical guidelines emphasize the urgent need for early identification of patients at risk for IHCA [3]. Early warning scores were developed to facilitate early identification of impending clinical deterioration and trigger

rapid interventions [4]. However, many traditional early warning scores are limited by considerable variation in discrimination in different populations and are often not sufficiently sensitive [5].

Recent research indicates that the implementation of the electronic Cardiac Arrest Risk Triage (eCART) score has significantly decreased the incidence of IHCA at UChicago Medicine [6]. However, the inclusion of laboratory data in eCART substantially diminishes the practicality and immediacy of this scoring system. Moreover, other studies have reported that calculating the Modified Early Warning Score (MEWS) 0.5 hours before a cardiac arrest can significantly increase the survival-to-discharge rate in patients experiencing IHCA [7]. Nonetheless, a 0.5-hour lead time is often insufficient for a prompt reaction during a patient's rapid deterioration. Given the continuously generated real-time information, such as vital signs, a time-varying model could be constructed for more timely and early identification of IHCA.

The aim of our study was to develop a recurrent neural network–based model using the electronic health records (EHRs) of a single tertiary medical center to predict incident IHCA. We hypothesized that variations in physiological parameters, evaluated in the context of known comorbidities, could help to predict incident cardiac arrest. We also aimed to validate the model in an independent cohort and compare it to a previous scoring system.

Methods

Ethics Approval

Given the retrospective study design, the Research Ethics Committee of the National Taiwan University Hospital (NTUH) approved this study (project approval 202206108RINB) and waived the requirement for obtaining informed consent.

Data Source

Predictive models were developed using the Multiparameter Intelligent Monitoring of Intensive Care (MIMIC)–IV v0.4 database and were externally validated using the Electronic Intensive Care Unit Collaborative Research Database (eICU-CRD) v2.0 [8,9]. Pre-existing institutional review board approval was waived given the deidentified nature of this public data set (Massachusetts Institute of Technology: 0403000206; Beth Israel Deaconess Medical Center: 2001-P-001699/14) [8]. One author who completed the Collaborative Institutional Training Initiative examination (certificate 57186438 for author HJJ) obtained access to the database and performed the data extraction. To assess the performance of our model in practical applications, we collected clinical data from the electronic medical records of the NTUH, spanning from 2008 to 2018. To decrease patient heterogeneity and feature variability, we applied the same inclusion criteria and data processing workflow to the 3 databases. We extracted data on patients older than 20 years who were hospitalized in intensive care units (ICUs) for at least 24 hours. Patients were excluded if they were encoded with a deceased status but without an IHCA labeling defined as below. We employed 5-fold cross-validation in our training cohort, randomly dividing the data set into 5 equally sized subsets. Four of these folds (80% of the MIMIC-IV cohort) were used for training, while the remaining fold (20% of the MIMIC-IV cohort) was reserved for internal validation. Performance metrics were recorded for each iteration, resulting in five distinct performance scores. These scores were then averaged to derive a singular more robust performance estimate for the model. Finally, external validation was performed on the entire eICU-CRD cohort.

Disease Outcome Ascertainment

In the MIMIC-IV cohort, patients were marked with IHCA if they were either labeled with a time-stamped database-specific procedure code (22,5466 cardiac arrest) or diagnosed with the *International Classification of Diseases, Ninth Revision* (*ICD-9*), Procedure Coding System (PCS) code 9960 (cardiopulmonary resuscitation, not otherwise specified). Although the MIMIC-IV database contained both *ICD-9* and *International Statistical Classification of Diseases, Tenth Revision* (*ICD-10*) codes, we did not convert *ICD-9*-PCS code 9960 to the *ICD-10*-PCS code, as the most approximately equivalent indicated code 5A1.2012 (performance of cardiac output, single, manual) represented variable definitions. For the eICU-CRD cohort, patients were classified with IHCA if they either presented with a time-stamped database-specific procedure note indicating cardiopulmonary resuscitation or were administered epinephrine, either as a bolus of 1 mg/10 ml or an infusion rate of 30 mg/250 ml at 100 ml/hr, with an associated administration time. In both the MIMIC-IV and eICU-CRD cohorts, the control group was defined as patients who were not labeled as having experienced an IHCA or being deceased, and the reference time was set as the ICU discharge time. For IHCA patients with multiple labelings, we only selected the time of the first label as the reference time. The data collection method in the NTUH database involves identifying patients with specific *ICD* codes (*ICD-9* 427.5; *ICD-10* T46.2, 145.8, 146.9). Patients who have been diagnosed with the aforementioned codes followed by the initiation of cardiopulmonary resuscitation or bolus epinephrine injection will be classified as patients who experienced IHCA.

Data Curation and Features Extraction

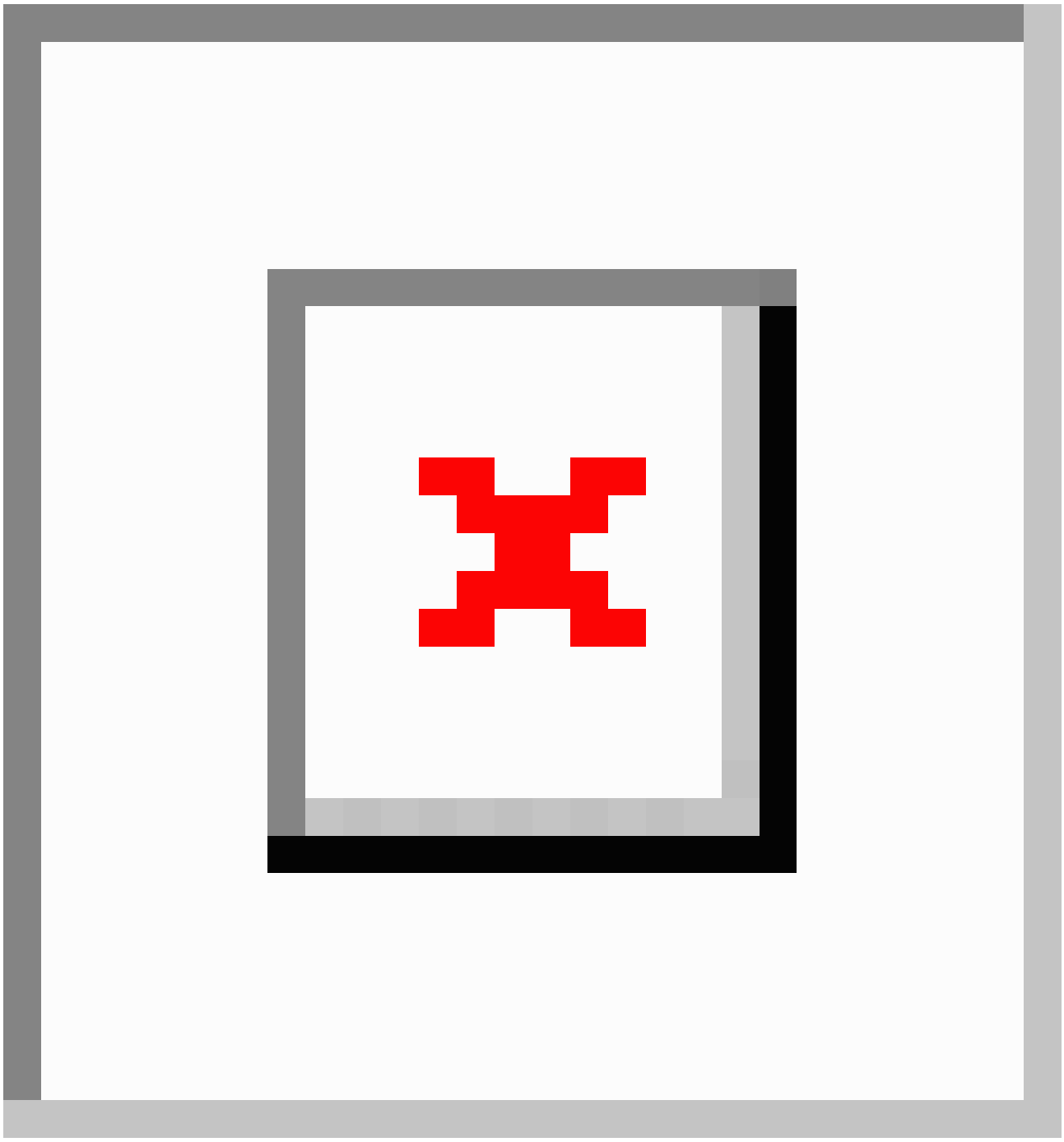
Two types of features were extracted: time-independent baseline features and time-varying physiologic readings from bedside monitors. Baseline features, which are variables registered at the time of admission, consisted of three types: (1) demographic information such as gender, age, ethnicity, type of ICU admission, and BMI; (2) chronic comorbidities, as identified by combined comorbidity score and Elixhauser Comorbidity Index [10,11]; (3) presenting illness, as identified by *ICD* codes for acute cardiac disease, respiratory insufficiency, sepsis, and potential reversible causes of cardiac arrest, popularly known as the *H*'s (hyperkalemia, hypokalemia, hypothermia, hypoxemia, hypovolemia, hydrogen ion, eg, acidosis) and *T*'s (spontaneous tension pneumothorax, thrombosis, cardiac tamponade) by resuscitation guidelines [12]. Physiologic readings, which consisted of 6 vital signs: heart rate (HR), respiratory rate, O_2 saturation (SpO_2), systolic blood pressure (sBP), diastolic blood pressure, and mean arterial pressure, were extracted on an hourly basis. For all patients, vital signs in the 24 hours prior to the reference time were recorded. To balance

model utility with adequate accuracy, we only investigated the risk of cardiac arrest starting from 13 hours prior to the event. To overcome the time series' irregularity, specific rules were applied to combine multiple vital signs in the same hour (Multimedia Appendix 1). The remaining missing values in vital signs were filled with the last observation carried forward method. To eliminate the misguidance of our imbalanced data set, we tested the two following remedies: synthetic minority oversampling technique (SMOTE) and near miss algorithm [13,14]. We employed SMOTE in the following training with a nearest neighbor interpolation of 1 as it yielded a better performance compared to the near miss algorithm (Figure S1 in Multimedia Appendix 1). After applying SMOTE, the numbers of IHCA patients and control patients were equal, signifying data balance.

Model Development

Our predictive model was encoded in three layers (Figure 1). First, random forest (RF) was responsible for classifying the baseline features [15]. For hyperparameter optimization, the number of estimators was set to 5, the maximum depth was set to 20, and Gini impurity was used to determine the split. Nodes are expanded until all leaves contain fewer than 2 samples [16]. Second, recurrent neural network with the long short-term memory (LSTM) architecture stored the vital signs trajectories in an hourly pattern [17]. There were 3 hidden layers and 8 cells each, with a tangent and a sigmoid activation function. The learning rate was set to 0.001, and a dropout rate of 0.4 was applied for regularization [18]. The Adam algorithm was adapted for optimizing network weights [19]. Last, the support vector machine (SVM) with a radial basis function kernel integrates the RF and LSTM models to generate the final prediction. The SVM predicts the identical target outcome by learning the relationship between the predictions from two base models (RF and LSTM) and the target outcomes in the training set [20]. All the models were implemented in Python 3.8.3 (Python Software Foundation) with TensorFlow 2.1.0, pandas 1.1.2, scikit-learn 0.24.2, and NumPy 1.19.1 libraries.

Figure 1. Illustration of the modeling framework. Each patient's data from the electronic health record were used as input for our model. Four preprocessing steps are carried out on the vital signs to obtain fixed-interval data. All features go through SMOTE to overcome data imbalance and are split into training and testing groups. Baseline features are inputted to random forest, and vital signs are inputted into LSTM for prediction. Support vector machine then integrates both models. AUROC: area under the receiver operating characteristic curve; LSTM: long short-term memory; SHAP: Shapley Additive Explanations.



Evaluation Strategy

To identify the perfect algorithm, the following machine learning (ML) techniques were evaluated in terms of prediction performance. First, based on the baseline data's time independency and binary structure, logistic regression (LR), *k*-nearest neighbor (KNN), extreme gradient boosting (XGBoost) tree, and SVM were compared with RF for model fitness. In the LR model, we applied an L2 penalty with a stopping tolerance set at 1e-4, and the model underwent a maximum of 100 iterations. For the KNN algorithm, we set the parameter K to 2, utilizing Euclidean distance as the chosen metric. In the XGBoost model, the number of estimators was configured to 5 with a maximum depth of 5 and a learning rate of 0.1. Hyperparameter optimization was carried out through a grid search. In the SVM, we used a radial basis function with an L2 penalty, setting the regularization parameter to 1. The SVM model was executed with a stopping tolerance of 1e-3, and no limit was imposed on the maximum number of iterations. For the time-dependent vital signs trajectories, the incorporation of memory gates in LSTM indicates its superiority in handling long sequence data. Thus, no other model comparison was made.

To compare different stacking techniques, LR was also implemented for comparison with SVM. Last, as we aim to use neural networks to accommodate our feature's complexity, we connected this 3-layer model by engaging a deep neural network in baseline data prediction and final stacking. The hyperparameters of the deep neural network were set at an epoch of 30, batch size of 24, and the Adam algorithm as optimizer. Model performance was assessed based on discrimination and calibration using the internal validation cohort, as quantified by the area under the receiver operating characteristic curve (AUROC) with mean values and 95% CIs [21]. Sensitivity and specificity metrics are presented by two binary classifications, including a predefined threshold of 0.5 and an optimal cutoff determined by the Youden index [22]. We used the Brier score to assess accuracy and visualized calibration curves across deciles based on observed and expected cardiac arrest numbers [23].

Model Interpretation

The importance of baseline features in the RF model was ranked based on "gain," the cumulative improvement in accuracy of the nodes attributed to a specific feature. To focus more on the local impact of each vital sign at the patient level, we employed the Shapley Additive Explanations (SHAP) method to explain how our LSTM model makes predictions during a specific timepoint [24].

Comparison With Previous Prediction Score

The Cardiac Arrest Risk Triage (CART), a commonly used cardiac arrest prediction model, was calculated to put the prediction results in perspective with prior studies [25]. A previously described "early warning score efficiency curve" was created to compare CART and our prediction model [26]. By plotting the percentage of detected events within 13 hours followed by the observations above the predefined threshold, a 0.5 probability in our model, and a score of 20 in the CART model, we could demonstrate the changes of cumulative incidence as the event time approached. Due to the large number of missing data for temperature and neurological status in our development cohort, we were unable to compare our risk prediction tool against the MEWS or Acute Physiology and Chronic Health Evaluation.

Results

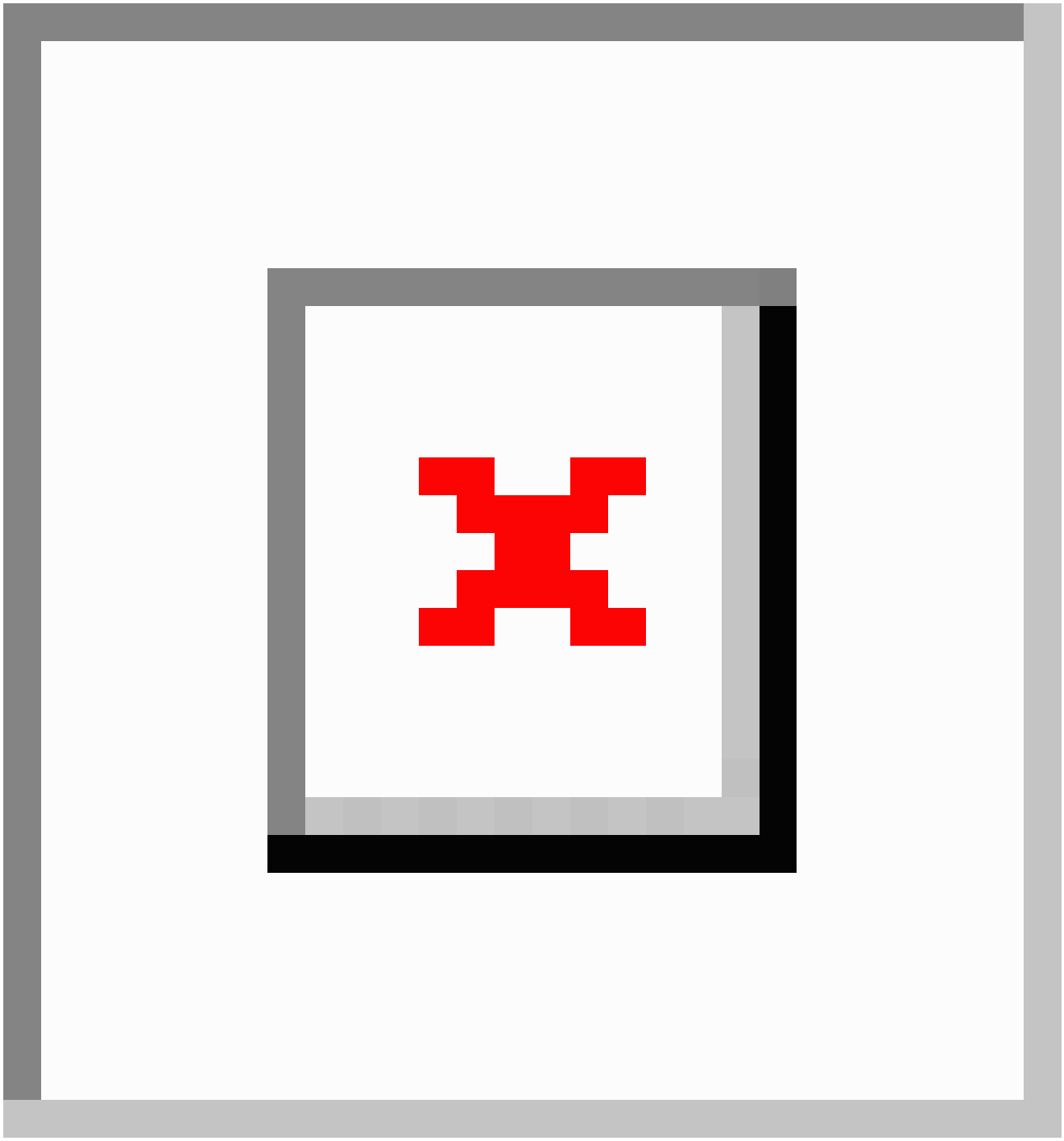
Patient Characteristics

A total of 34,633 patients in the MIMIC-IV database and 79,643 patients in the eICU-CRD database were included in our analysis. After processing the vital signs data, a total of 452 IHCA patients and 23,457 control patients from MIMIC-IV were used for model development, whereas 85 IHCA patients and 9964 control patients from eICU-CRD were used for external validation. Table S1 in Multimedia Appendix 1 shows the baseline characteristics of the IHCA group and the control group for the two cohorts. IHCA patients were significantly older (P<.001) and scored higher on combined comorbidity scores and the Elixhauser Comorbidity Index. In terms of presenting illness, myocardial infarction, pneumonia, respiratory failure, and the 5 *H*'s and 5 *T*'s were more prevalent in IHCA patients than among control patients.

Prediction From Time-Independent Data

Patient demographics, comorbidities, and presenting illness were first classified by RF. Figure 2 demonstrates the discrimination of the RF model (AUROC 0.80, 95% CI 0.779 - 0.844; sensitivity 0.71; specificity 0.78; F_1-score 0.79. The top five important features listed by RF include the presence of respiratory failure or acidosis, comorbid uncomplicated hypertension, comorbid fluid and electrolyte disorder, and initial ICU being the cardiac ICU.

Figure 2. Prediction from baseline features. (A) AUROC for evaluating the discriminatory ability of random forest on baseline features. (B) Feature importance derived from the random forest model. AUROC: area under the receiver operating characteristic curve.



Modeling of Time-Dependent Data

The trajectories of six vital signs were modeled with respect to time. Figure S2 A in Multimedia Appendix 1 illustrates that in the MIMIC-IV cohort, the control group exhibited a constant value of all six vital signs throughout the 24-hour collecting period. However, the vital signs of the IHCA patients were characterized by progressive deterioration in the last several hours. Of note, throughout the 24-hour monitoring period, patients who developed cardiac arrest exhibited, on average, a 12-mmHg lower sBP, 1.5% lower SpO_2, and a 9-bpm higher resting HR compared to the control group. However, the exact timing of the start of deterioration could not be clearly marked on the plot. A similar vital signs trajectory was seen in the eICU-CRD cohort (Figure S2 B in Multimedia Appendix 1).

Prediction From Time-Dependent Data

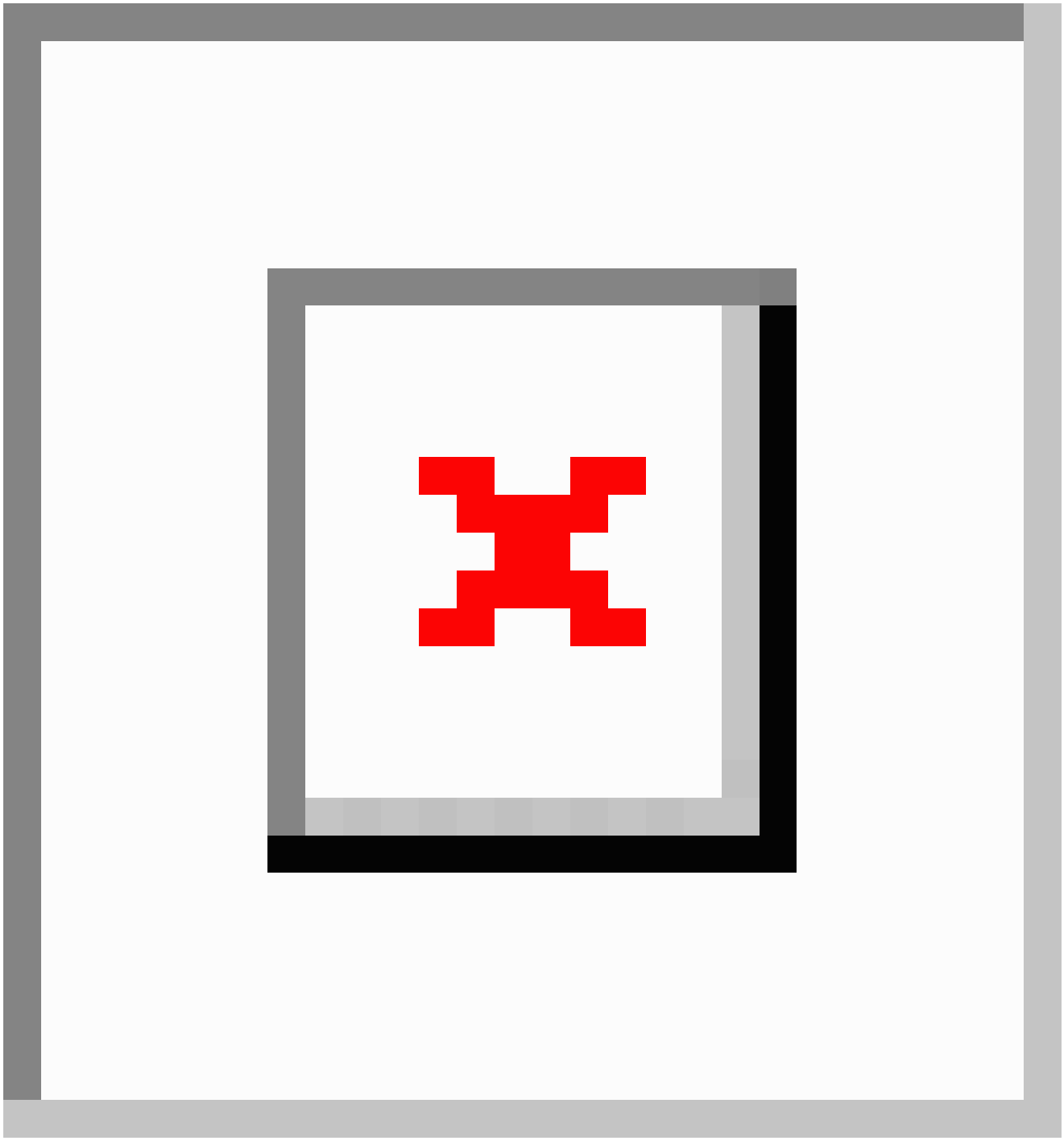
The hourly AUROC values for predicting cardiac arrest are presented in Figure S3 in Multimedia Appendix 1, which shows the results after SMOTE and cross-validation. A steady rise in AUROC was observed in the hours leading up to cardiac arrest with a sharp increase in the preceding 3 hours.

Performance of the SVM-Based Stacking Model

In the final step of model construction, we stacked the LSTM model with the RF model and combined both predictions from baseline features and vital signs. AUROCs of the stacked model

exhibited consistently better predictions compared with the baseline and vital signs–only model, with the highest AUROC of 0.91 (95% CI 0.874 - 0.935), sensitivity of 0.80, specificity of 0.86, and F_1-score of 0.85 1 hour prior to the event. Further evaluation of the stacked model presented an increase in sensitivity, specificity, negative predictive value, and F_1-score by the reduction of the time interval (Figure 3). However, the calibration plot showed a risk of overestimation and a steadily low Brier score throughout the 13 hours of prediction time (Figure S4 in Multimedia Appendix 1). Additionally, in Figure S5 in Multimedia Appendix 1, we compared the model performance using different cutoffs. We found that the optimal cutoff defined by the Youden index (at 13 hours: 0.29; at 12 hours: 0.25; at 11 hours: 0.38; at 10 hours: 0.25; at 9 hours: 0.28; at 8 hours: 0.26; at 7 hours: 0.28; at 6 hours: 0.30; at 5 hours: 0.26; at 4 hours: 0.38; at 3 hours: 0.30; at 2 hours: 0.34; at 1 hour: 0.35) presented with a better sensitivity compared with the predefined 0.5 cutoff; the largest difference was 14% at 12 hours prior to the event.

Figure 3. Performance of the stacked model in the Multiparameter Intelligent Monitoring of Intensive Care (MIMIC)–IV database. AUROCs (95% CIs) of the long short-term memory (LSTM) model with vital signs as input (orange plot), RF model with baseline features as input (gray plot), and stacked model after integration of RF and LSTM (blue plot) are shown. The three models' exact AUROCs, sensitivity, specificity, NPV, and F_1-score of the stacked model are listed in the table. AUROC: area under the receiver operating characteristic curve; NPV: negative predictive value; RF: random forest.



External Validation

We performed external validation of the stacked model in the eICU-CRD database. The results showed the best performance at 1 hour prior to IHCA with an AUROC of 0.89 (95% CI 0.849 - 0.920), sensitivity of 0.79, specificity of 0.83, and an F_1-score of 0.81. These findings align closely with the AUROC obtained from the MIMIC-IV data set (Figure 4). To further validate our model in an actual clinical scenario, we identified 1935 IHCA patients and 3692 control patients from the ICU of the NTUH. Additionally, our model demonstrated high prediction sensitivity and an AUROC of 0.945 when predicting IHCA 1 hour prior to its occurrence (Figure 5).

Figure 4. Performance of the stacked model in the Electronic Intensive Care Unit Collaborative Research Database (eICU-CRD). External validation of the stacked model is performed on the eICU-CRD. AUROC (95% CI) is plotted in a blue line; sensitivity is plotted in a gray line. AUROC: area under the receiver operating characteristic curve.

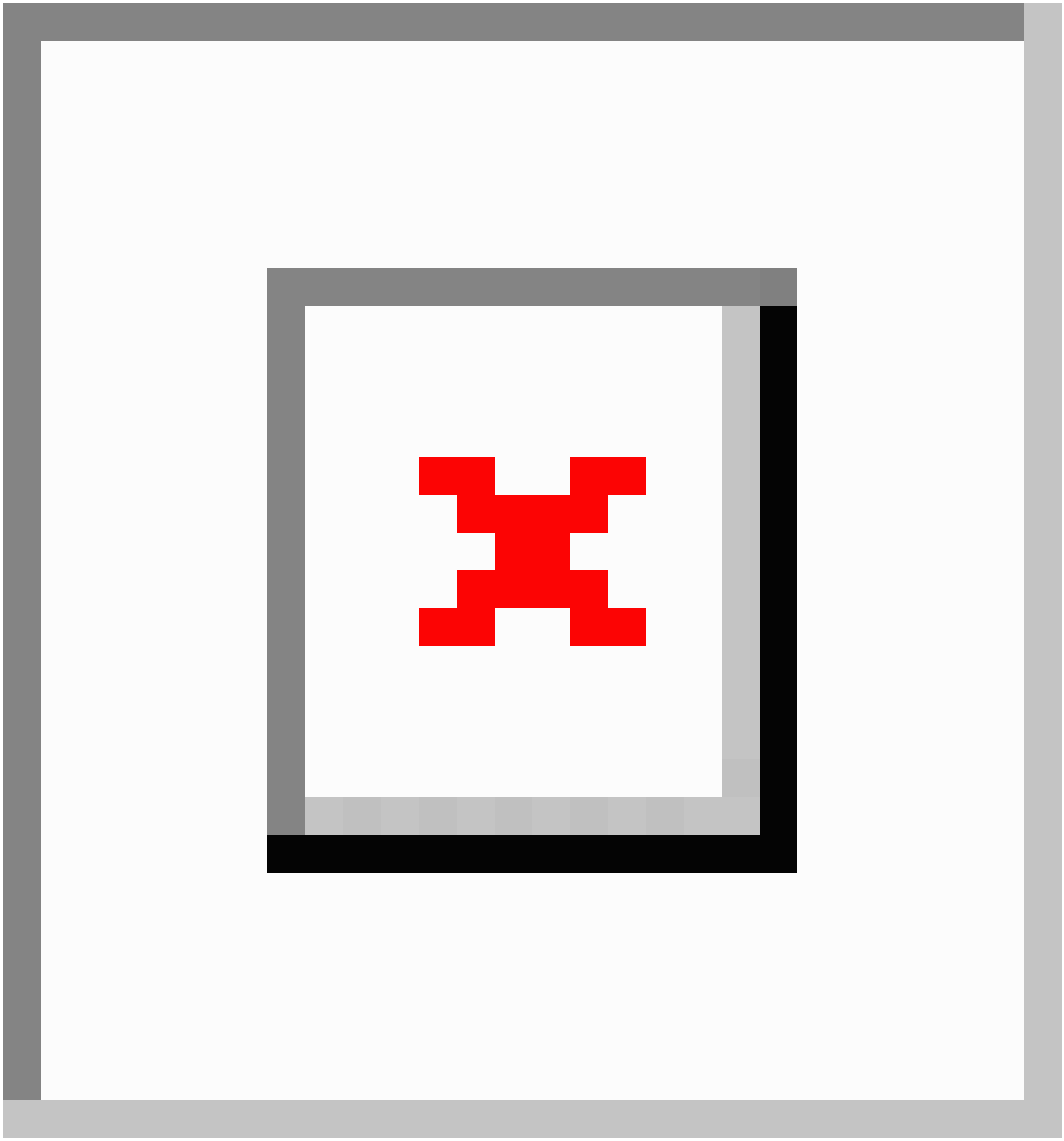
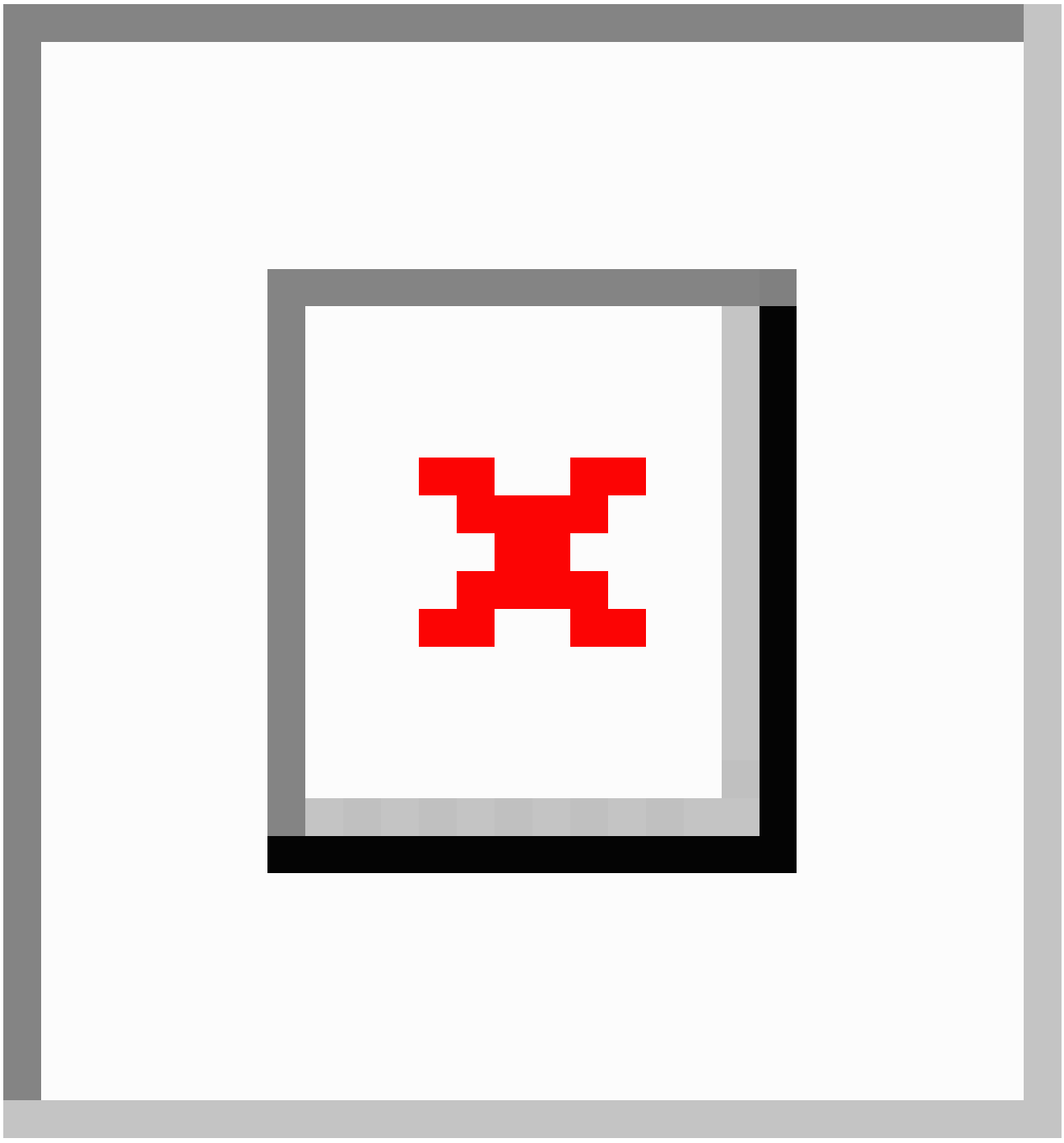


Figure 5. Performance of the stacked model in the clinical scenario. External validation of the stacked model is performed using data from 1935 in-hospital cardiac arrest patients and 3692 control patients collected from the National Taiwan University Hospital. AUROC is plotted in a blue line; sensitivity is plotted in a gray line. AUROC: area under the receiver operating characteristic curve.



Local Interpretation of the LSTM Model

We adopted the SHAP method to enable model explanation from an individual patient's perspective. In each box, SHAP values for specific vital signs are assigned, with positive SHAP values in red indicating a risk factor and negative SHAP values in blue indicating a protective factor. Figure S6 A in Multimedia Appendix 1 represents a patient from the MIMIC-IV database experiencing IHCA at 0 hours. As IHCA approaches, an increase in sBP from its average contributes to an elevated risk, with the most significant effect occurring 6 hours prior to IHCA. However, at 1 hour before IHCA, the most significant risk becomes a decrease in sBP from its average. Figure S6 B in Multimedia Appendix 1 illustrates another IHCA patient from the eICU-CRD database. In contrast to Figure S6 A in Multimedia Appendix 1, the most prominent feature at 1 hour prior to IHCA is a decrease in HR and SpO_2 from its baseline value. These figures showcase diverse presentations leading to IHCA in various patients, providing a valuable guideline for medical staff to identify the specific organ failure responsible for IHCA. The significance lies in enabling a swift response, incorporating timely interventions such as intubation for saturation drop and the administration of inotropic agents for decreased sBP. This approach ensures that medical staff will

not delay necessary treatments while determining the cause of IHCA.

Performance Compared With Different ML and Deep Learning Algorithms

Conventional statistics and supervised ML algorithms were compared to predict IHCA using only baseline features. RF demonstrated superior performance in terms of AUROC compared with XGBoost, LR, KNN, and SVM (Figure S7 in Multimedia Appendix 1). SVM also presented preferable results during the stacking operation compared with LR throughout the 13-hour prediction period. AUROCs at 1 hour prior to the incidence of IHCA were 0.91 versus 0.80 (Figure S8 in Multimedia Appendix 1). Finally, using a neural network to connect baseline, vital signs, and stacking predictions did not reveal an improving outcome (Figure S9 in Multimedia Appendix 1). After comparing several algorithms and combinations, RF, LSTM, and SVM predictions still yielded the most satisfactory results.

Detection Efficacy Compared to Previous Prediction Score

We compared the performance of our proposed model to that of the CART score. Overall, our model demonstrated better AUROC throughout the prediction period (Figure S10 in Multimedia Appendix 1). As illustrated in Figure S11 in Multimedia Appendix 1, it is evident that at 12 hours prior to cardiac arrest, our model was able to detect over 70% of patients at risk for IHCA, compared to the CART score that did not surpass a 65% detection rate even 1 hour prior to IHCA.

Discussion

Principal Findings

In this retrospective study of 34,633 patients in the MIMIC-IV database, we constructed a high-performance multimodal model (AUROC 0.91, 95% CI 0.874 - 0.935) that can predict IHCA up to 13 hours in advance using EHRs and high-resolution time series physiological readings. As the time of cardiac arrest approached, our model yielded a steady increase in the detection rate, finally reaching 89% 1 hour prior to the event. We also illustrated the impact of each vital sign on the prediction of cardiac arrest associated with individual patients through the use of SHAP values. Furthermore, we demonstrated the advantage of this ML algorithm over the CART score, which was derived using traditional regression models.

Comparison to Prior Work

As a ubiquitous activity in the hospital, several studies have demonstrated the importance of vital signs measurement in determining a patient's disease course [27]. Diastolic blood pressure, respiratory rate, and maximum HR have all been found to be significant and independent predictors of cardiac arrest [28]. However, maintaining a minimal model with only vital signs or adding lab data as predictors at the cost of decreasing model adaptability remains a dilemma [29,30]. The lactic acid level is the most representative laboratory biomarker in circulatory failure but had a high rate of missingness in the MIMIC-IV database (16,317/23,909, 68.2%). This motivated us to abandon utilizing lab results and assess if a nimbler model could be constructed with vital signs trends alone, overlaying the easily obtainable *ICD* codes and patient demographics as baseline features. Unsurprisingly, a significant increase in AUROC was discovered by adding demographics and comorbidities to the vital signs–only model. Furthermore, an SVM-based stacked model can address the predictive capabilities of underlying conditions and dynamic changes during disease deterioration. Stacking proves advantageous by compensating for the weaknesses of both models, with RF potentially struggling with highly correlated data and LSTM excelling in handling timely intricate information.

Distinct Advantages of Our Approach

The reason for not establishing an end-to-end neural network throughout the prediction stood out, as supervised ML algorithms retained the ability to determine the importance of each predictor and have better model explainability. Moreover, in the ensemble technique, stacking excels over both boosting and bagging due to its versatility in integrating diverse data domains and combining various types of models. Late fusion at the model level is also preferred over other fusion methods for mitigating feature discrepancies and enabling independent model training between the time-independent baseline and time-dependent vital signs. Additionally, the outperformance of SVM over LR in the stacking operation could be attributed to better data handling using the nonlinear kernel function. To evaluate the external validity of our model, we tested it on two distinct data sets—the eICU-CRD and NTUH databases—both representing patient groups with diverse ethnicities and disease backgrounds. Over a 10-year duration, we identified 1935 (34.3%) IHCA cases in NTUH. In contrast to prior IHCA prediction studies, such as Kwon et al's [31] 2.3% (n=1233) over 7 years, Chae et al's [32] 1.3% (n=1154) over 4 years, and Ding et al's [33] 23.09% over 5 years (n=1796), our clinical database demonstrated a higher IHCA incidence yet fewer cases [31-33]. This disparity is attributed to our ICU-focused validation database, in contrast to earlier studies that encompassed all patients who were hospitalized. Consequently, our approach ensures heightened data precision and a more nuanced understanding of patient dynamics through continuous monitoring within this critically ill cohort. Nevertheless, our high prediction quality in both independent databases ensures the credibility of our model across various demographic groups and subpopulations. The consistent performance across these data sets not only minimizes the possibility of overfitting but also validates the generalizability of our predictions.

Limitations of Our Methodology

Our study had limitations because we used data collected from one medical center. First, due to the nature of EHRs, we were unable to determine the reason for the multi-scale gaps and different frequencies of each input. Second, we did not include clinical interventions, body temperature, and mental status in our model. Clinical interventions may change the disease course or even terminate the deterioration process. Nevertheless, the complexity of the treatment record and the high frequency of missing values in temperature and mental status compelled us to omit these valuable predictors. Third, our identification of

IHCA relied on time-labeled database-specific procedure codes, *ICD* procedure codes, or administration of epinephrine in resuscitation dosages. In real-time clinical scenarios, delays in data entry may occur as documentation is considered secondary to patient care. Additionally, the accuracy of these codes is often operator dependent and may vary across different ICU policies. To minimize recording biases, we manually reviewed all IHCA vital signs data and only included reasonable measurements, ensuring that the identified IHCA timepoints correlated with the worst patient vital signs.

Conclusion

We built a multimodal ML model based on time serial vital signs and three types of baseline features, which were all easily accessible in the ICU. Our model showed high accuracy in detecting clinical deterioration leading to the development of IHCA up to 13 hours in advance in both the internal and external validation cohorts. A model like this could be integrated into a hospital's EHR system to identify high-risk patients and provide clinical decision support.

Acknowledgments

This study was funded by grant MOST-110-2622-8-002-017. No funding bodies had any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The data sets analyzed during this study are available in the Multiparameter Intelligent Monitoring of Intensive Care–IV repository [34] and Electronic Intensive Care Unit Collaborative Research Database repository [35].

Authors' Contributions

CCL has full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis, concept and design, critical revision of the manuscript for important intellectual content, obtaining funding, and supervision. HYL, PCK, FQ, HJJ, PHC, CH Lee, and IJW were responsible for drafting the manuscript, interpretation of the data, and critical revision of the manuscript for important intellectual content. CH Li was responsible for statistical analysis. JRH and WTH were responsible for the interpretation of the data and critical revision of the manuscript for important intellectual content. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables, figures, and material.
[DOCX File, 37539 KB - medinform_v12i1e49142_app1.docx]

References

1. Sinha SS, Sukul D, Lazarus JJ, et al. Identifying important gaps in randomized controlled trials of adult cardiac arrest treatments: a systematic review of the published literature. Circ Cardiovasc Qual Outcomes 2016 Nov;9(6):749-756. [doi: 10.1161/CIRCOUTCOMES.116.002916] [Medline: 27756794]
2. Holmberg MJ, Ross CE, Fitzmaurice GM, et al. Annual incidence of adult and pediatric in-hospital cardiac arrest in the United States. Circ Cardiovasc Quality Outcomes 2019 Jul 9;12(7):e005580. [doi: 10.1161/CIRCOUTCOMES.119.005580]
3. Morrison LJ, Neumar RW, Zimmerman JL, et al. Strategies for improving survival after in-hospital cardiac arrest in the United States: 2013 consensus recommendations: a consensus statement from the American Heart Association. Circulation 2013 Apr 9;127(14):1538-1563. [doi: 10.1161/CIR.0b013e31828b2770] [Medline: 23479672]
4. Spångfors M, Molt M, Samuelson K. In-hospital cardiac arrest and preceding National Early Warning Score (NEWS): a retrospective case-control study. Clin Med (Lond) 2020 Jan;20(1):55-60. [doi: 10.7861/clinmed.2019-0137] [Medline: 31941734]
5. Smith GB, Prytherch DR, Schmidt PE, Featherstone PI. Review and performance evaluation of aggregate weighted 'track and trigger' systems. Resuscitation 2008 May;77(2):170-179. [doi: 10.1016/j.resuscitation.2007.12.004] [Medline: 18249483]
6. Bartkowiak B, Snyder AM, Benjamin A, et al. Validating the electronic cardiac arrest risk triage (eCART) score for risk stratification of surgical inpatients in the postoperative setting: retrospective cohort study. Ann Surg 2019 Jun;269(6):1059-1063. [doi: 10.1097/SLA.0000000000002665] [Medline: 31082902]
7. Wang AY, Fang CC, Chen SC, Tsai SH, Kao WF. Periarrest Modified Early Warning Score (MEWS) predicts the outcome of in-hospital cardiac arrest. J Formos Med Assoc 2016 Feb;115(2):76-82. [doi: 10.1016/j.jfma.2015.10.016] [Medline: 26723861]

8. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
9. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5:180178. [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
10. Gagne JJ, Glynn RJ, Avorn J, Levin R, Schneeweiss S. A combined comorbidity score predicted mortality in elderly patients better than existing scores. J Clin Epidemiol 2011 Jul;64(7):749-759. [doi: 10.1016/j.jclinepi.2010.10.004] [Medline: 21208778]
11. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8-27. [doi: 10.1097/00005650-199801000-00004] [Medline: 9431328]
12. Soar J, Nolan JP, Böttiger BW, et al. European Resuscitation Council Guidelines for Resuscitation 2015: Section 1. Executive summary. Resuscitation 2015 Oct;95:1-80. [doi: 10.1016/j.resuscitation.2015.07.016] [Medline: 26477410]
13. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013 Mar 22;14:106. [doi: 10.1186/1471-2105-14-106] [Medline: 23522326]
14. Zhang JP, Mani I. KNN approach to unbalanced data distributions: a case study involving information extraction. Presented at: International Conference on Machine Learning (ICML 2003), Workshop on Learning from Imbalanced Data Sets; Aug 21, 2003; Washington, DC URL: https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1603053 [accessed 2024-07-12]
15. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002 Dec;2/3 [FREE Full text]
16. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Machine Learning Res 2012 Feb;13:281-305 [FREE Full text]
17. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 1;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9117894]
18. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Machine Learning Res 2014 Jun;15:1929-1958 [FREE Full text]
19. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Jan 30, 2017 URL: http://arxiv.org/abs/1412.6980 [accessed 2021-02-16]
20. Wang J, Feng K, Wu J. SVM-based deep stacking networks. Proc AAAI Conference Artif Intelligence 2019 Jul 17;33(1):5273-5280. [doi: 10.1609/aaai.v33i01.33015273]
21. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 6;162(1):W1-W73. [doi: 10.7326/M14-0698] [Medline: 25560730]
22. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. Biom J 2008 Jun;50(3):419-430. [doi: 10.1002/bimj.200710415] [Medline: 18435502]
23. Rolke W, Gongora CG. A chi-square goodness-of-fit test for continuous distributions against a known alternative. Comput Stat 2020 May 14;36:1885-1900. [doi: 10.1007/s00180-020-00997-x]
24. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on Nov 25, 2017 URL: https://arxiv.org/abs/1705.07874 [accessed 2024-07-02]
25. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs*. Crit Care Med 2012 Jul;40(7):2102-2108. [doi: 10.1097/CCM.0b013e318250aa5a] [Medline: 22584764]
26. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013 Apr;84(4):465-470. [doi: 10.1016/j.resuscitation.2012.12.016] [Medline: 23295778]
27. Smith GB. Vital signs: vital for surviving in-hospital cardiac arrest? Resuscitation 2016 Jan;98:A3-A4. [doi: 10.1016/j.resuscitation.2015.10.010] [Medline: 26597106]
28. Churpek MM, Yuen TC, Huber MT, Park SY, Hall JB, Edelson DP. Predicting cardiac arrest on the wards: a nested case-control study. Chest 2012 May;141(5):1170-1176. [doi: 10.1378/chest.11-1301] [Medline: 22052772]
29. Kennedy CE, Aoki N, Mariscalco M, Turley JP. Using time series analysis to predict cardiac arrest in a PICU. Pediatr Crit Care Med 2015 Nov;16(9):e332-e339. [doi: 10.1097/PCC.0000000000000560] [Medline: 26536566]
30. Ueno R, Xu L, Uegami W, et al. Value of laboratory results in addition to vital signs in a machine learning algorithm to predict in-hospital cardiac arrest: a single-center retrospective cohort study. PLoS One 2020 Jul 23;15(7):e0235835. [doi: 10.1371/journal.pone.0235835] [Medline: 32658901]
31. Kwon JM, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in-hospital cardiac arrest. J Am Heart Assoc 2018 Jun 26;7(13):e008678. [doi: 10.1161/JAHA.118.008678] [Medline: 29945914]
32. Chae M, Han S, Gil H, Cho N, Lee H. Prediction of in-hospital cardiac arrest using shallow and deep learning. Diagnostics (Basel) 2021 Jul 13;11(7):1255. [doi: 10.3390/diagnostics11071255] [Medline: 34359337]
33. Ding X, Wang Y, Ma W, et al. Development of early prediction model of in-hospital cardiac arrest based on laboratory parameters. Biomed Eng Online 2023 Dec 6;22(1):116. [doi: 10.1186/s12938-023-01178-9] [Medline: 38057823]

34. Medical Information Mart for Intensive Care. URL: https://mimic.mit.edu [accessed 2024-07-12]
35. eICU Collaborative Research Database. URL: https://eicu-crd.mit.edu/about/eicu/ [accessed 2024-07-12]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CART: Cardiac Arrest Risk Triage
eCART: electronic Cardiac Arrest Risk Triage
EHR: electronic health record
eICU-CRD: Electronic Intensive Care Unit Collaborative Research Database
HR: heart rate
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
ICU: intensive care unit
IHCA: in-hospital cardiac arrest
KNN: *k*-nearest neighbor
LR: logistic regression
LSTM: long short-term memory
MEWS: Modified Early Warning Score
MIMIC: Multiparameter Intelligent Monitoring of Intensive Care
ML: machine learning
NTUH: National Taiwan University Hospital
PCS: Procedure Coding System
RF: random forest
sBP: systolic blood pressure
SHAP: Shapley Additive Explanations
SMOTE: synthetic minority oversampling technique
SpO_2: O_2 saturation
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 19.05.23; peer-reviewed by R Sun, S Tiwari, T Hou; revised version received 11.02.24; accepted 23.04.24; published 23.07.24.

Please cite as:
Lee HY, Kuo PC, Qian F, Li CH, Hu JR, Hsu WT, Jhou HJ, Chen PH, Lee CH, Su CH, Liao PC, Wu IJ, Lee CC
Prediction of In-Hospital Cardiac Arrest in the Intensive Care Unit: Machine Learning–Based Multimodal Approach
JMIR Med Inform 2024;12:e49142
URL: https://medinform.jmir.org/2024/1/e49142
doi:10.2196/49142

Original Paper

Enhancing Health Equity by Predicting Missed Appointments in Health Care: Machine Learning Study

Yi Yang[1], MSc; Samaneh Madanian[1], PhD; David Parry[2], PhD

[1]Auckland University of Technology, Auckland, New Zealand

[2]Murdoch University, Perth, Australia

Corresponding Author:
Samaneh Madanian, PhD
Auckland University of Technology
6 St Paul Street, AUT WZ Building
Auckland, 1010
New Zealand
Phone: 64 99219999 ext 6539
Email: sam.madanian@aut.ac.nz

Abstract

Background: The phenomenon of patients missing booked appointments without canceling them—known as Did Not Show (DNS), Did Not Attend (DNA), or Failed To Attend (FTA)—has a detrimental effect on patients' health and results in massive health care resource wastage.

Objective: Our objective was to develop machine learning (ML) models and evaluate their performance in predicting the likelihood of DNS for hospital outpatient appointments at the MidCentral District Health Board (MDHB) in New Zealand.

Methods: We sourced 5 years of MDHB outpatient records (a total of 1,080,566 outpatient visits) to build the ML prediction models. We developed 3 ML models using logistic regression, random forest, and Extreme Gradient Boosting (XGBoost). Subsequently, 10-fold cross-validation and hyperparameter tuning were deployed to minimize model bias and boost the algorithms' prediction strength. All models were evaluated against accuracy, sensitivity, specificity, and area under the receiver operating characteristic (AUROC) curve metrics.

Results: Based on 5 years of MDHB data, the best prediction classifier was XGBoost, with an area under the curve (AUC) of 0.92, sensitivity of 0.83, and specificity of 0.85. The patients' DNS history, age, ethnicity, and appointment lead time significantly contributed to DNS prediction. An ML system trained on a large data set can produce useful levels of DNS prediction.

Conclusions: This research is one of the very first published studies that use ML technologies to assist with DNS management in New Zealand. It is a proof of concept and could be used to benchmark DNS predictions for the MDHB and other district health boards. We encourage conducting additional qualitative research to investigate the root cause of DNS issues and potential solutions. Addressing DNS using better strategies potentially can result in better utilization of health care resources and improve health equity.

(JMIR Med Inform 2024;12:e48273) doi:10.2196/48273

KEYWORDS

Did Not Show; Did Not Attend; machine learning; prediction; decision support system; health care operation; data analytics; patients no-show; predictive modeling; appointment nonadherence; health equity

Introduction

Adding to the existing pressures on the health care system [1,2], further substantial disruptions are caused when patients fail to attend their prescheduled appointments [3]. This is defined as Did Not Show (DNS), which is a scheduled but not utilized clinical appointment that patients failed to attend without canceling or rescheduling. This phenomenon is also known as Did Not Attend (DNA) or Failed To Attend (FTA). Causes include the patient forgetting about their appointment, miscommunication [4], logistical difficulties, appointment scheduling conflicts, and family/work commitments [3,5].

DNS can adversely affect patients' well-being, cause them and the system financial stress, and disturb health care operations and systems. Globally, DNS has an overall rate of 23%, with a wide geographical variation (13.2% in Oceania, 19.3% in

Europe, 23.5% in North America, 27.8% in Asia, and 43% in South America [6]). DNS is expensive for health systems; for example, estimated annual losses amounting to £790 million (over US $1 billion) were found in the United Kingdom [7] and $564 million in the United States [8]. It affects both primary and secondary health care [9], although secondary care losses are higher.

Patients mostly fail to comply with their clinical appointments when symptoms become less severe or unnoticeable [10,11], which might deteriorate underlying syndromes [12,13]. Patients are more likely to demand immediate medical attention when contracting serious health issues or require acute and emergency care if they miss scheduled health care appointments [12,14-16].

Eliminating DNS is hard to achieve, and its adverse effects necessitate methods and approaches for managing DNS such as sending digital reminders by text, phone, and email [17,18]. These approaches have not been very effective, as they are time-consuming and costly, and the health care system still faces DNS issues. Overbooking [3,19], open access [20], and DNS penalty approaches have also been used to enhance clinical slot utilization but can cause longer waiting times for patients and overtime for clinical staff [21].

Inspired by the success of artificial intelligence (AI) in different sectors, including health care [22,23], we considered the application of AI for DNS management via predicting the probability of DNS appointments [13,19,24,25]. AI and its subset techniques, such as machine learning (ML), are powerful for extracting cognitive insights from massive amounts of data [26,27].

The predicted DNS probabilities proved to be successful in providing the required information for DNS management [25] and supporting health care managers in making informed decisions for prioritizing patients and delivering clinical assistance. This enables health care providers to reschedule and reuse limited clinical resources for urgent cases while also expanding access to health care services for patients from diverse backgrounds, thereby promoting health care equity.

Therefore, clinical capabilities and medical resources can be used more effectively and efficiently, decreasing patients' wait times, increasing their satisfaction, and enhancing health productivity.

Most studies concerned with predicting DNS have mainly comprised small data sets or specific groups of people to develop models for DNS learning and prediction; however, DNS tends to be varied across populations. For example, longer distances to a medical facility increase DNS [8], but this finding was contradicted in another study [28]. Likewise, patients with chronic illnesses adhere to their scheduled appointments [13], while other studies [29] have shown that patients with more severe diseases have a higher DNS rate. Even within a single medical organization, DNS factors vary across different clinics [14]. These examples highlight the inconsistent nature of DNS predictors, showcasing the complexity of predicting tasks in this domain. Such variations pose challenges in creating a universal formula or model to effectively address DNS prediction issues on a global scale.

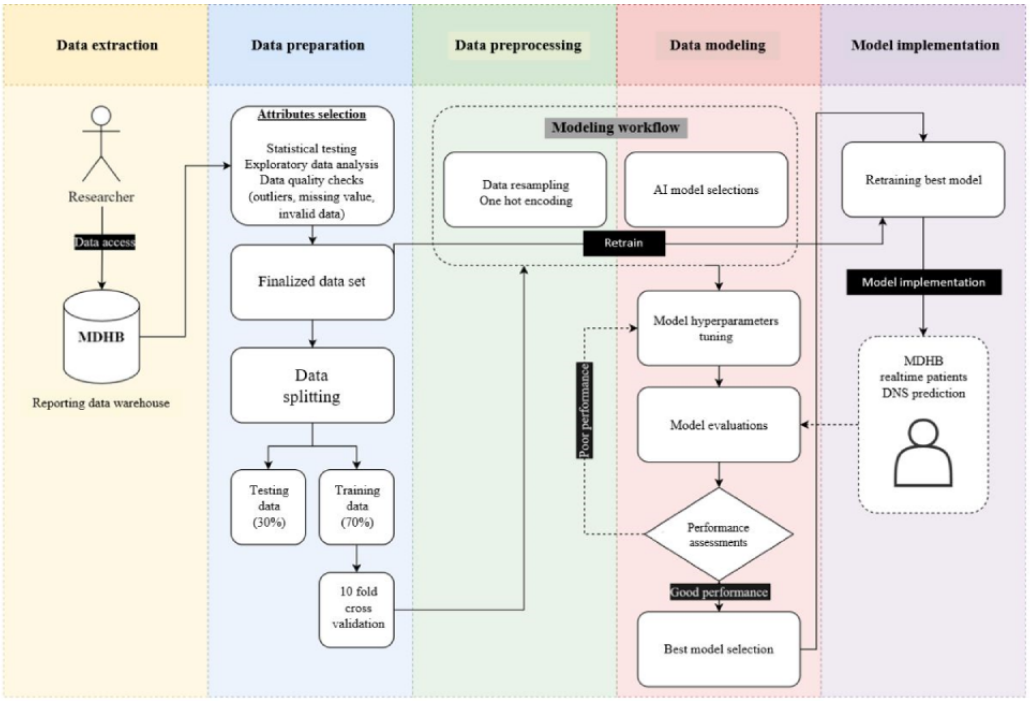
Considering the very limited DNS research in New Zealand and the complexity of developing a general DNS predictive model, we concentrated on the DNS issue in the MidCentral District Health Board (MDHB) hospital as a proof of concept. MDHB is located in the center of the North Island, New Zealand, covering a land area of over 8912 km^2 and with a population of over 191,100 people. In this region, about 18% of people are aged 65 years or older, with over 20% being Māori, and a higher proportion than the national average resides in more deprived areas [30]. These demographic factors could lead to inequity in access to health care services. To support MDHB in addressing health equity and providing additional support for patients, this study aimed to develop ML models and compare their performance in predicting the probabilities of future DNS appointments at MDHB. This study utilized a data set spanning 5 years of collected data.

Methods

Overview

Our research was organized into the following phases (Figure 1)**.** The initial phase involved *data extraction*, defining the data set to be used, and outlining the data extraction process. The *data preparation* phase involved conducting exploratory data analysis (EDA) to profile data and exclude irrelevant observations from the research. Subsequently, the data set was split into 2 parts—70% (454,831 records) for training and 30% (194,927 records) for testing. To avoid data linkage, the training and testing data sets were not mixed during the ML modeling phase. Moreover, the training set underwent a 10-fold cross-validation strategy to prevent bias as much as possible and fully utilize its limited training information. Next, the *data preprocessing* phase involved cleaning and transforming the cross-validation sets, ensuring that the training set was ready for the data modeling stage. A 10-fold cross-validation resampling strategy was applied to further optimize the utilization of the 70% training data. In the *data modeling* phase, we used 3 ML algorithms and tuned their hyperparameters to identify the best performance among the algorithms. Finally, in the *model evaluation* phase, various evaluation metrics were employed to determine the best-performing ML model for DNS prediction.

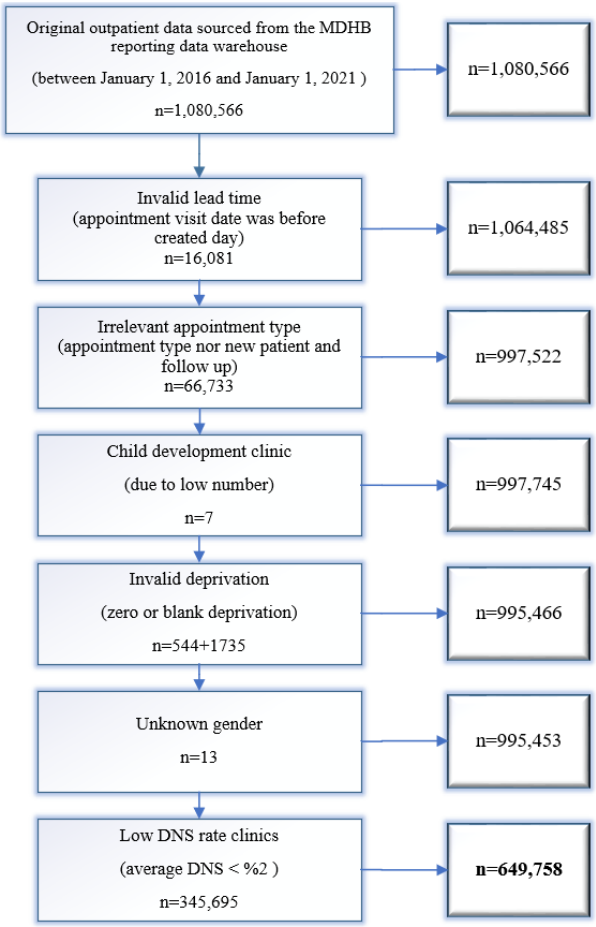
Figure 1. Research flow and procedure. AI: artificial intelligence; MDHB: MidCentral District Health Board.



Data Access and Extraction

Our data were sourced from MDHB reporting SQL farm and contained only outpatient visits with no link to other data sets. This significantly mitigated risks related to patient reidentification. Data deidentification and encryption were applied before data access, and New Zealand National Health Index numbers were encrypted to protect patients' privacy. We acquired 1,080,566 outpatient visit records from 38 clinics between January 1, 2016, and December 31, 2020, satisfying the research requirements with almost 57,000 DNS incidents (5% of the entire data set). The steps of data exclusion are presented in Figure 2. Because not many missing records were identified in the data sets, those with missed values were directly excluded.

Figure 2. Research data exclusion. DNS: Did Not Show; MDHB: MidCentral District Health Board.



Ethical Considerations

This study received ethics approval from the Auckland University of Technology (AUT; 20/303) and MDHB (2020.008.003), following which data access to the MDHB reporting data warehouse was granted.

Data Preparation

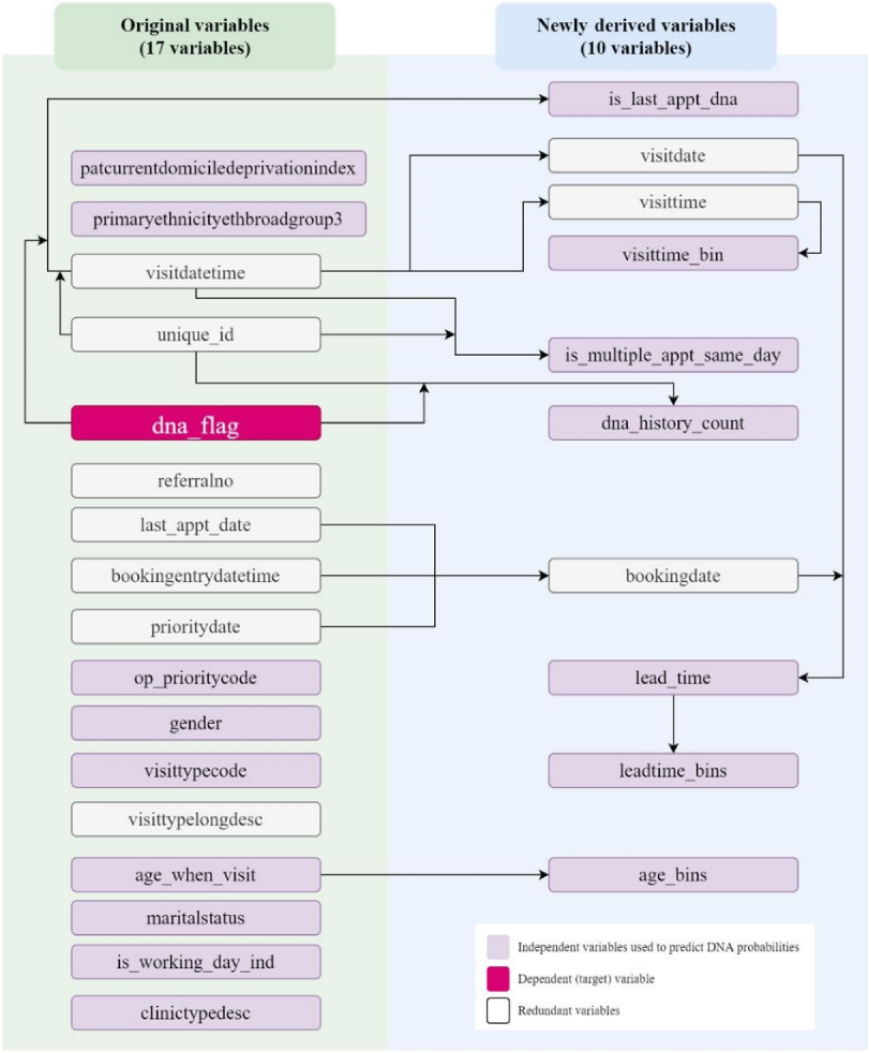
Phase Description

In this phase, understanding the data was important to adequately prepare them for the experiments. The data preparation process included data transformation and derivation (Figure 3). Following suggestions from the literature, new research variables were derived and introduced because some valuable DNS predictors were absent in the MDHB data set. For example, no direct information was available on the patients' DNS history [21,31], appointment lead time [31,32], or latest appointment DNS outcome [13]. The lead time was calculated by comparing the difference in days between the appointment creation date and the visit date. Appointments with longer lead times were expected to have greater DNA probability than those with shorter lead times [29].

Therefore, to better understand patient behavior and DNS patterns, we derived 10 new variables on top of the original variables (Figure 3). These attributes were introduced to support us in understanding when patients were more likely to miss their appointments in general and to identify regular nonadherent patients.

Initially, we extracted a data set with 17 columns and over 1 million records (Multimedia Appendix 1). Informed by the literature review [14,29,31-33], we derived and introduced another 10 variables on top of the original data and increased the data columns to 27. Among all the variables, 16 (59%) were used for ML modeling, and the redundant ones were excluded. The *dna_flag* attribute was the dependent (target) variable. Figure 3 demonstrates the original variables in addition to 10 newly derived ones.

Figure 3. Variable transformation and derivation visualization.



Cardinality Reduction

We conducted a cardinality reduction analysis to reduce variable categories with low frequency and small samples. The data set mostly included categorical variables, with numeric variables being rare. Each categorical level is called a cardinality, which means how many distinct values are in a column. In our data set, some categorical variables had fewer levels, such as patient gender—M (male), F (female), and U (unknown)—while others had hundreds of variations, such as suburbs or diagnosis codes.

Developing ML models often involves numerous categorical attributes, necessitating examination of the variables' cardinality, as most ML algorithms are distance-based and require converting categorical variables to numeric values. Categorical variables with high cardinality levels will derive massive new columns and expand the data set. This expansion increases model complexity, elevates computational costs, and decreases model generalization, which makes handling the data set challenging [34]. Therefore, we investigated the cardinality of our research variables and deployed a reduction strategy accordingly.

Cardinality reduction analysis was conducted to reduce the number of categories within variables with low frequency and small sample sizes. Following suggestions from the literature, new research variables were also derived and introduced, including patients' prior DNS history [14,16,21] and the appointment lead time [14,16,29,32].

Statistical Test

The chi-square test was used for analyzing homogeneity among different groups within variables [35] and for testing the independence between categorical variables [36]. The chi-square statistics (χ^2) and their *P* values were calculated to investigate whether different levels of a variable contributed differently to DNS events.

The confidence level (α=.05) was adopted as the *P* value threshold in the chi-square test. A *P* value less than .05 provided enough confidence to reject the null (H_0) hypothesis and accept the alternative hypothesis (H_A). The tested categorical variable was associated with DNS events [36]. Hence, we may consider using it for future prediction.

After the data preparation process, 16 variables were selected to predict the target *dna_flag*. Among them, 12 modeling predictors were nominal variables, including binary variables (Multimedia Appendix 2). We, therefore, conducted the chi-square (χ^2) statistical test to investigate the relationship

between those predictors and DNS events (Table 1). The chi-square was calculated using the following equation, where O and E are observed and expected values [36,37]:

After preparing the data set and before developing the ML models, an EDA was conducted to gain a deeper understanding of the research data landscape. EDA is a fundamental data analysis required before hypothesis and modeling formulation [38]. Its findings can be used to verify misleading models at a later stage [38] and reveal unexpected patterns [39]. The EDA helped uncover patients' DNS patterns through data aggregation and data visualization analysis. Finally, the EDA findings were validated against the ML model outcomes to verify their accuracy.

Table 1. Chi-square test on categorical variables.

Categorical variables	Chi-square statistic	Chi-square *P* value
dna_history_count	118,461	<.01
is_last_appt_dna	77,600	<.01
Clinictypedesc	35,201	<.01
age_bins	34,810	<.01
primaryethnicityethbroadgroup3	17,098	<.01
leadtime_bins	11,048	<.01
maritalstatus_group	10,527	<.01
visit_type_group	3525	<.01
visittime_bin	3447	<.01
patcurrentdomiciledeprivationindex	2655	<.01
is_multiple_appt_same_day	1913	<.01
op_prioritycode_group	1,496	<.01
is_working_day_ind	1,244	<.01
Gender	4	.06

Data Preprocessing

Due to the high number of categorical variables in our data set, the one-hot encoding technique was used in the preprocessing phase. Because distance-based algorithms can only deal with numerical values, in the cardinality reduction section, we used the one-hot encoding method to convert our categorical variables to numbers. After the conversion, different variables were introduced to our training data set, also known as indicator variables. For example, the variable gender derived 3 variables, *gender_male*, *gender_female*, and *gender_unknown*. Each of those variables can have a value of either 1 or 0.

As the predictive performance of classifiers is highly impacted by the selection of the hyperparameters [40], we conducted hyperparameter tuning to optimize our algorithms' learning process. We further optimized this process using the Grid Search method to boost the performance of our chosen models. Table 2 outlines specific details regarding the hyperparameters utilized.

Table 2. Hyperparameter tuning of the data modeling.

Models and hyperparameters	R package	Range	Purpose
Logistic regression			
penalty	Glmnet	1e-10- 1	Total amount of regularization used to prevent overfit and underfit
Random forest			
Trees	Ranger	300- 1000	Number of trees in the forest
Min_n	Ranger	3-10	Minimum amount of data to further split a node
Mtry	Ranger	3-5	Maximum number of features that will be randomly sampled to split a node
XGBoost[a]			
Trees	XGBoost	300-1000	Number of trees in the forest
Min_n	XGBoost	3-10	Minimum amount of data to further split a node
mtry	XGBoost	3-10	Maximum number of features that will be randomly sampled to split a node
tree_depth	XGBoost	3-12	Maximum depth of the tree

[a]XGBoost: Extreme Gradient Boosting.

Data Modeling

Addressing the imbalanced data set posed the main data modeling challenge. The annual DNS rate for MDHB was around 5%, which means 95% of the appointments were attended visits. This imbalance significantly affected the accuracy of our ML model in predicting attended cases. To tackle this issue, various internal and external strategies exist [41,42]. In this study, we employed an external approach that involved utilizing standard algorithms intended for a balanced data set but applying resampling techniques to the trained data set to reduce the negative impact caused by the unequal class. Our focus was on the resampling strategy, known for its effectiveness in handling imbalanced classification issues and its portability [42].

The resampling strategy involved 2 methods: (1) oversampling, where the size of the minority class is increased randomly to approach the majority class in a class-imbalance data set [43,44]; and (2) undersampling, where the size of the majority class decreases randomly to align with the minority class [43,44]. This strategy falls under both the oversampling and undersampling categories. Given the lack of definitive guidance on the effectiveness of these methods [42-44], we adopted both and compared their results.

Since we dealt with a binary classification prediction problem, supervised and classification algorithms were selected. Algorithms with good interpretability were also considered to explain which predictive variables influence DNS prediction more significantly. In a study concerning variable importance, tree-based models, such as random forest (RF) and gradient-boosted decision trees, were shown to inherently possess features that measure variable importance [45].

For the imbalanced data set, we used ensembling methods due to their proven advantages [46,47]. The following algorithms were chosen for developing DNS prediction models: logistic regression (LR), RF, and Extreme Gradient Boosting (XGBoost).

LR was chosen because it is a suitable analysis method across multiple fields for managing binary classification [48]. Our research concerned a supervised classification problem to predict whether a future outpatient appointment will become a DNS visit. With the response variable (*dna_flag*) offering dichotomous outcomes—either yes (1) or no (0)—LR stood as a fitting choice due to its proficiency in predicting binary outcomes and its established effectiveness in prior studies [7,13,33,49]. Tree-based ensembling algorithms were also chosen for their proven ability to deal with imbalanced data sets and model explainability [46,47]. RF can effectively handle combining random resampling strategies in imbalanced prediction. Tree-ensembling methods have more advanced prediction ability than a single model because they integrate prediction strength from several base learners [50].

Model Implementation and Evaluation

We used 10-fold cross-validation for model selection and bias reduction. The hyperparameters were tuned to boost each classifier's performance. We followed suggestions from the literature suggestions to use sensitivity, specificity, and the area under the receiver operating characteristic (AUROC) curve to quantify the models' prediction strength for the imbalance problem prediction.

During this phase, we used the testing data to validate the best predictive model chosen based on the model evaluation criteria. For this study, data before 2021 were used in the data modeling process. We coordinated with MDHB to access outpatient appointments from 2021 for model validation. Specifically, we used both weekly and monthly data for prediction, comparing these with actual appointment outcomes to validate the model. The benefit of using a new data set for validation was to assess model bias and goodness of fit outside the research environment. Positive performance and high prediction accuracy would

indicate potential real-life implementation of our research model after further investigation.

Results

Our study only included new patients and follow-up appointments. Therefore, we analyzed DNS costs limited to new patient and follow-up outpatient services over the last 5 years. The MDHB provided us with costing information for 34 different departments, and we calculated the DNS cost for each department (Table 3). In 2020, there were 2812 new patient DNS visits and 6240 follow-up DNA visits causing a loss of at least $2.9 million (US $1.8 million) at MDHB. More information regarding this calculation is provided in Multimedia Appendix 3 [51].

Each department was assigned a corresponding outpatient appointment price for a new patient and follow-up outpatient appointment services. We aggregated the total DNS occurrences of new patients and follow-up appointments, multiplying corresponding unit prices to quantify their financial impact. For instance, in 2020, there were 301 new patients and 745 follow-up patients who missed their scheduled bookings, which caused a revenue loss of $300,442 (US $190,000) in the orthopedics department.

Although the initial research expected to address the DNS issue for all outpatient clinics and patients at the MDHB, due to the broad scope of the DNS, we concentrated on clinics with a higher percentage of DNS and narrowed down the research scope to prioritize workloads. To successfully build a model for our focused patient groups, we eliminated as many irrelevant data points as possible. Then, data used for the model training were more fit for purpose for the high-needs population.

The modeling data set was created using 649,758 records and 17 columns (Figures 1 and 3). We developed ML models based on LR, RF, and XGBoost algorithms, with hundreds of hyperparameter combinations in our data modeling. To evaluate the models' prediction performance, accuracy, sensitivity, specificity, AUROC curves, and cost (computation time) were calculated (Table 4). The aim was to identify the best model and hyperparameters that resulted in optimal sensitivity and AUROC performances. Model prediction accuracy is critical; however, it was not a primary concern in this research as we dealt with an imbalanced data set [52].

Table 4 presents a summary comparison of the models' performance. As shown in the table, the LR-based model was the fastest and RF the slowest in terms of computation time. LR had the lowest AUROC scores (ie, the low DNS events prediction accuracy), while RF and XGBoost had a similar area under the curve (AUC) performance (around 0.92).

The undersampling strategy significantly improved our models' sensitivity. Sensitivity was chosen over accuracy because we were dealing with an imbalanced data set [52]. Sensitivity quantified the models' ability to correctly predict positive (DNS) cases that help detect high-risk DNS patients. RF and XGBoost had a very close sensitivity of 0.82. However, considering the computation cost factor, XGBoost had the lowest modeling time. XGBoost with undersampling was our best ML model for the DNS prediction. Its ROC curve is illustrated in Figure 4.

A further investigation was also performed to identify the top predicting factors for each model (Multimedia Appendix 4). The purpose of calculating variable significance scores was not to plug them into a calculation formula but to showcase which variables were more relatively critical in calculating the risk of DNS. Variable importance is critical to AI model development, as variables do not contribute evenly to the final prediction. Therefore, we focused on the most influential predictors and excluded irrelevant ones by scoring the variables' prediction contributions [53]. Variable importance is a measurement quantifying the relationship between an independent variable and the dependent [46].

The results shown in Multimedia Appendix 4 matched the chi-square statistical test results (Table 1). The leading factors were determined and selected using the variable (feature) importance. It was evident that the *dna_history_count* variable was the most influential predictor following *is_last_appt_dna*, *age_when_visit*, and *lead_time*. Additionally, *ethnicity* played an important role in constructing the XGBoost model for the DNS prediction.

We also aggregated outpatient appointment data and ranked the observed DNS rate of all outpatient clinics (Multimedia Appendix 5). We carried out this analysis to initiate an understanding of how disease type might influence the DNS rate.

Table 3. DNS[a] costs in 2020 at the MDHB[b] hospital[c].

Clinics	NP[d] DNS count	NP DNS price	Total NP DNS cost	FU[e] DNS count	FU DNS price	Total FU cost	Total DNS cost
Orthopedics	301	$346	$104,143	745	$263	$196,299	$300,442
Diabetes	90	$452	$40,658	576	$307	$176,643	$217,302
Ophthalmology	221	$239	$52,776	874	$174	$152,322	$205,099
Pediatric medicine	124	$600	$74,366	327	$395	$129,271	$203,637
Ear nose throat	253	$358	$90,571	367	$269	$98,744	$189,316
Gynecology	177	$403	$71,322	386	$280	$108,124	$179,446
Hematology	75	$632	$47,389	232	$348	$80,834	$128,223
Cardiology	109	$490	$53,397	245	$299	$73,259	$126,656
Radiation oncology	42	$505	$21,194	350	$293	$102,652	$123,846
General surgery	147	$387	$56,856	208	$309	$64,369	$121,225
Audiology	268	$214	$57,302	272	$214	$58,157	$115,459
Neurology	153	$617	$94,408	38	$400	$15,204	$109,612
Gastroenterology	68	$506	$34,393	186	$362	$67,401	$101,794
Medical oncology	18	$650	$11,703	229	$360	$82,327	$94,030
Dental	136	$244	$33,132	193	$244	$47,019	$80,151
Renal medicine	5	$559	$2,793	181	$344	$62,201	$64,995
Respiratory lab	38	$479	$18,192	121	$347	$42,021	$60,213
Obstetrics	101	$227	$22,906	143	$227	$32,431	$55,337
Respiratory sleep	20	$271	$5412	153	$271	$41,403	$46,815
Urology	65	$357	$23,178	85	$274	$23,253	$46,432
Dietetics	93	$175	$16,302	168	$175	$29,449	$45,751
General medicine	44	$517	$22,747	69	$322	$22,200	$44,948
Respiratory	39	$479	$18,671	70	$347	$24,309	$42,980
Dermatology	66	$316	$20,877	60	$236	$14,174	$35,051
Oral and maxillofacial	23	$296	$6799	124	$203	$25,185	$31,984
Endocrinology	25	$525	$13,127	34	$332	$11,284	$24,411
Rheumatology	18	$647	$11,643	31	$345	$10,693	$22,336
Plastic surgery (excluding burns)	18	$296	$5321	69	$203	$14,014	$19,335
GI[f] endoscopy	0	$506	$0	52	$362	$18,843	$18,843
Community pediatrics	20	$600	$11,994	10	$395	$3953	$15,948
Infectious diseases	7	$738	$5169	19	$534	$10,152	$15,321
Neurosurgery	1	$507	$507	29	$448	$12,990	$13,496
Podiatry	17	$207	$3522	47	$207	$9737	$13,259
Aged ATR[g] health	18	$244	$4394	35	$244	$8545	$12,939
Under 65 ATR	3	$244	$732	5	$244	$1221	$1953
Cardiothoracic	0	$573	$0	4	$425	$1698	$1698
Anesthetics	9	0	$0	3	$0	$0	$0

[a]DNS: Did Not Show.

[b]MDHB: MidCentral District Health Board.

[c]A currency exchange rate of NZD $1=US $0.61 is applicable for the listed costs.

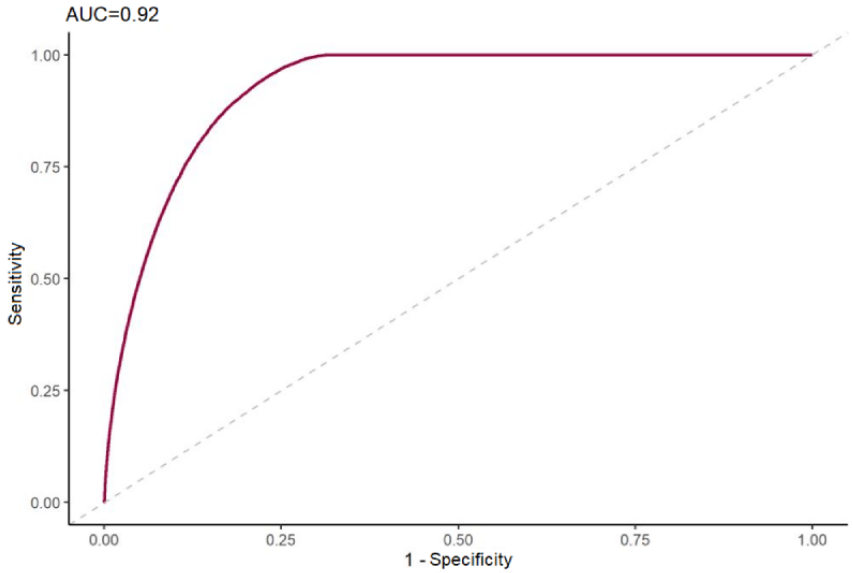
[d]NP: new patient.
[e]FU: follow-up.
[f]GI: gastrointestinal.
[g]ATR: assessment, treatment, and rehabilitation.

Table 4. Comparison of the ML[a] models' performance.

Classifier and resampling strategy	Sensitivity	Specificity	AUC[b]	Accuracy	Modeling cost
Logistic regression					
Undersampling (under_ratio=2)	0.5146	0.9227	0.8474	0.8897	Less than 1 hour (5 minutes)
Oversampling (over_ratio=0.5)	0.5091	0.9247	0.8592	0.8911	Less than 1 hour (14 minutes)
Random forest					
Undersampling (under_ratio=2)	0.8243	0.8524	0.9236	0.8501	Over 8 hours (8.4)
Oversampling (over_ratio=0.5)	0.5940	0.9260	0.9220	0.8990	Over 137 hours
XGBoost[c]					
Undersampling (under_ratio=2)	0.8278	0.8490	0.9239	0.9117	Over 4 hours (4.8)
Oversampling (over_ratio=0.5)	0.8297	0.8549	0.9267	0.8529	Over 51 hours (51.83)

[a]ML: machine learning.
[b]AUC: area under the curve.
[c]XGBoost: Extreme Gradient Boosting.

Figure 4. The receiver operating characteristic (ROC) of the best classifier, Extreme Gradient Boosting (XGBoost). AUC: area under curve.



Discussion

Principal Findings

Our results are comparable to similar previously published analyses [9], although the AUC for XGBoost was slightly higher in our case. This may be due to the data selection and local characteristics. We initially built a generic DNS prediction model for all outpatient clinics at MDHB. However, in light of the literature and DNS complexity, the project scope was narrowed down to clinics with higher DNS rates. As discussed previously in this paper, we excluded irrelevant and missed data, invalid lead time appointments, and clinics with very low DNS rates. This approach improved the ML models' performance and made sense from an operational perspective. The developed models provided insights useful for understanding the contributing factors for DNS. We found that patient DNS history, appointment characteristics, work commitments, and socioeconomic status substantially contributed to DNS events.

Patient DNS History

Understanding patients' DNS history was crucial for predicting future DNS patterns (Table 5) and developing the ML models. This also aligned with the chi-square test results (Table 1), which ranked the *dna_history_count* and *is_last_appt_dna* variables as the most important factors. Total DNS counts and the latest appointment's DNS outcome are pivotal for calculating

the probabilities of future DNS occurrences. These factors are consistent with the findings in the literature [14-16,21,32,54].

Managing DNS involves identifying patients with low adherence to scheduled visits for additional attention. Centralizing and managing DNS history can provide a comprehensive view, preventing data silos or gaps. Centralized monitoring can enhance the visibility of recurring DNS incidents and proactively alert clinicians of potential DNS cases. Our models account for changes in DNS behavior. To reduce the prediction bias, we screen for the most recent appointment DNS outcome (*is_last_appt_dna*).

Table 5. Top prediction variables in the developed ML[a] models.

Algorithm and variable importance ranking	Undersampling model	Oversampling model
Logistic regression		
1	dna_history_count	dna_history_count
2	is_working_day	is_working_day
3	is_last_appt_dna	is_multiple_appt_same_day
4	is_multiple_appt_same_day	is_last_appt_dna
5	lead_time	lead_time
Random forest		
1	dna_history_count	dna_history_count
2	is_last_appt_dna	age_when_visit
3	lead_time	lead_time
4	age_bins	is_last_appt_dna
5	clinic_type_desc	clinic_type_desc
XGBoost[a]		
1	dna_history_count	dna_history_count
2	is_last_appt_dna	age_when_visit
3	age_when_visit	is_last_appt_dna
4	lead_time	ethnicity
5	Ethnicity	lead_time

[a]XGBoost: Extreme Gradient Boosting.

Appointment Characteristics

Certain appointments expected more nonadherence, with distinct predictors related to appointment characteristics such as "working day" and "high lead time." Longer lead times correlated with increased DNS probability, while appointments on working days were more prone to DNS than nonworking days. These findings align with reports from [33,54,55] and emphasize the significant impact of appointment lead time on DNS prediction, as also indicated in [8,14,16,32,33,54]. This underscores how appointment characteristics directly affect DNS outcomes immediately after scheduling. Therefore, incorporating ML-predicted DNS risk estimations during appointment scheduling could automatically flag higher DNS probability for proactive management.

Furthermore, our analysis of the *op_priorityc*ode variable (Multimedia Appendix 1) indicated that, in general, patients with more serious health conditions were more likely to attend their appointments. This observation is reflected in Multimedia Appendix 5, which compares the DNS rates of different clinics with the overall average DNS rate of 0.053% (depicted red line). For example, patients visiting the audiology clinic had a potential DNS rate of 19.1% compared to a 0.9% DNS rate for the radiation oncology clinic. Our analysis of the *op_priorityc*ode variable was based on categorical data types reflecting appointment urgency and not based on a detailed analysis of each patient's diagnosis.

Work Commitments

Our findings suggest that patients struggled to adhere to appointments on working days or during working hours. Younger adults, particularly those between 20 and 30 years of age, had higher DNS rates due to work commitments, while older adults aged 65 years and above rarely missed their visits.

Furthermore, the XGBoost-based model highlighted that being single was an indicator of DNS visits (Figure 4). This could relate to time constraints among young professionals, a finding consistent with other studies [8,28,33,56]. For this group, a targeted reminder system could be developed to concentrate on appointments with higher DNA probability compared to the

DNS risk threshold. Consequently, the population-based reminding system could help optimize resource allocation, including staff efforts and costs.

Socioeconomic Status

We explored the deprivation index and clustered patient populations by using their ethnicity (Multimedia Appendix 6). Our findings indicated a strong association between European and Māori ethnicities and DNS outcomes, ranked among the top 5 predicting factors (Multimedia Appendix 4). Māori and Pacific populations had the highest DNS rates, in line with other research findings [56], while the European ethnicity had the lowest DNS rates. Māori and Pacific populations tended to reside in areas characterized by higher deprivation rates, whereas the percentage of other ethnicities living in higher deprivation regions decreased when the deprivation index increased.

In New Zealand, Māori and Pacific ethnical groups required increased health care attention [57] to ensure equity in the health care system. As indicated in Table 6, a larger proportion of these ethnic groups are situated in suburbs and areas with higher deprivation indexes (such as 8, 9, and 10) [58]. The higher deprivation index was also a strong indicator of socioeconomic deprivation geographically [58]. According to the New Zealand Index of Deprivation, neighborhoods with higher deprivation were more likely to experience adverse living conditions such as damp, cold, and crowded housing.

Moreover, regions with higher deprivation exhibit higher rates of unemployment, increased dependence on benefits, and more single-parent families [58]. Consequently, these living conditions and income disparities made patients living in these regions more susceptible to illness, while also encountering more barriers and obstacles in addressing their medical needs.

At MDHB, dedicated working groups were established to support Māori and Pacific patients in attending their scheduled hospital appointments. Our research reiterates the importance and necessity of those working groups, acknowledging the value of their work. Moreover, our model can support them further by providing tangible DNS probability scores to prioritize patients who require additional attention and support.

Table 6. Percentage of population residing at each deprivation level [58].

Deprivation level	Māori, n (%)	Pacific, n (%)	European, n (%)	Asian, n (%)	Other, n (%)
1	3113 (7)	293 (1)	37,314 (86)	2077 (5)	835 (2)
2	4951 (9)	429 (1)	46,405 (85)	1470 (3)	1071 (2)
3	6367 (13)	489 (1)	42,565 (84)	613 (1)	821 (2)
4	14,736 (14)	1747 (2)	84,728 (79)	4574 (4)	1593 (1)
5	14,400 (13)	3398 (3)	83,568 (77)	6015 (6)	1590 (1)
6	14,103 (15)	1759 (2)	74,351 (79)	2974 (3)	1248 (1)
7	13,442 (17)	3601 (5)	58,187 (75)	1858 (2)	870 (1)
8	36,843 (19)	5402 (3)	148,605 (75)	5434 (3)	1988 (1)
9	40,642 (24)	7324 (4)	111,319 (67)	5443 (3)	2442 (1)
10	31,998 (35)	6283 (7)	52,064 (56)	1610 (2)	521 (1)

Operational and Managerial Implications

The total DNS loss incurred by the MDHB hospital was around $2.9 million (US $1.8 million) in 2020. Notably, we observed that clinics with less life-threatening diseases (diabetes, audiology, and dental) had higher DNS rates. Considering our use of MDHB data, we expect to identify similar patterns in other district health boards for which the same DNS predicting factors can be applied for DNS management.

While the primary objective of our research was to calculate DNS risk for promoting health equity, we believe that leveraging DNS prediction can aid in managing limited health care resources more efficiently. By quantifying the DNS probability for future appointments on a scale from 0.00 to 1, clinicians or hospital operation managers can develop more personalized health care services for their patients. This leads to enhancing equity in accessing health care services for a wider population.

The predictions derived can support MDHB managers in designing, planning, and implementing more informed DNS management strategies. For example, a DNS appointments threshold (eg, 0.7) can be set, and all appointments with predicted odds greater than 0.7 can be selected, releasing 70% of resources and allocating some (or all) to the remaining 30% of patients with a higher DNS risk. Potentially, these released resources can subsidize interventions to support attendance. Without DNS prediction, the hospital cannot decide where to focus on solving the DNS problem and must invest money uniformly for every patient, leading to equality rather than equity in health care service access. Equality is not fit for purpose, especially considering the high attendance rate of 95% over the past 5 years, indicating that most patients attend appointments without additional support. However, for more optimum use of health care resources, other policies and guidance for appointment scheduling should be considered [59].

Potential Interventions to Reduce DNS

DNS Suggests Life Hardships

When patients miss medical appointments, it is a critical indicator suggesting they may be experiencing hardships in their lives [15,54,60]. Considering that a higher DNS rate correlates with a higher deprivation index, we can assume that

people residing in these areas may face greater transportation limitations. Moreover, people with severe mental health or addiction issues may not be able to independently visit their doctors [15]. These vulnerable groups require additional and ongoing appointment assistance. Unfortunately, they have been historically disadvantaged and marginalized by the current health care system [61].

The DNS prediction model we developed can help health care practitioners identify patients at higher risk of DNS. Targeted DNS improvement solutions can be designed based on predicted DNS probability, patient demography, and clinical history. This type of application can leverage the DNS prediction model to help identify and deliver patient-centric medical services to patients requiring additional help. Some examples are discussed in the subsequent sections.

Expanding Integrated Health Care Networks

For patients not facing life-threatening illnesses or requiring long-term health management (such as patients with diabetes), expanding services closer to patients might help meet their needs. MDHB could consider deploying clinicians to outsourced sites to supervise practitioners or attend to patients directly. Moreover, increasing collaborations with primary health care networks, promoting nurse-led services, and contracting private specialists can also be viable options for decreasing DNS rates. Developing a one-stop medical hub with multidisciplinary clinics for patients with lower clinical risk could encourage attendance and reduce DNS visits [19]. This is consistent with the New Zealand Ministry's latest health care system reform strategies, which aim to uplift health care equity [61]. The reform emphasizes the establishment of more locality networks in the community, resonating well with our research findings.

After-Hour Appointment Slots

To support young adults who are occupied by daily work, it might be favorable to increase more after-hour service slots in clinics when possible. If more appointment slots can be organized before or after working hours, working professionals may have more chances to adhere to their clinical appointments. Piloting more weekend clinics can also be a choice to meet younger generations' needs. In consonance with our suggestion, the recent New Zealand health care reform also promoted more affordable after-hours services [61]. Additionally, offering transportation assistance and improved wraparound well-being support for patients with a high-risk score could increase attendance. At-home patient visits could also be offered and delivered to patients facing severe transport limitations.

Limitations

Despite the success of our DNS prediction model, we need to acknowledge that it has some limitations. First, our model was trained on 5-year period data from MDHB. The single data source prevented us from exploring other critical dimensions such as household data or beneficiary data. We believe adding those data points would improve the prediction model and discover more patients' DNS patterns.

Furthermore, we pairwise compared the attribute *dna_flag* with other DNS predictor factors. However, future research should consider investigating and analyzing the association between variables and adding further variables to the conditioning set. This expanded analysis would offer deeper insights into patients' DNS behaviors.

Conclusions

To the best of our knowledge, this study represents one of the first attempts in New Zealand to develop ML prediction models supporting DNS management. We successfully developed and tested ML models to predict probabilities of outpatient appointments' DNS. Our selected model had an AUROC of 0.92 and a sensitivity performance of 0.82.

Acknowledgments

The authors would like to thank the New Zealand MidCentral District Health Board (MDHB) for their support of this study. We appreciate the advice, help, and support from the MDHB data analytics team, Dr Richard Fong, and Mr Rahul Alate. Without their contribution, this study would not have been possible.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of variables and their definitions.
[DOCX File , 16 KB - medinform_v12i1e48273_app1.docx]

Multimedia Appendix 2
Data type of original and newly derived variables.
[DOCX File , 15 KB - medinform_v12i1e48273_app2.docx]

Multimedia Appendix 3
Outpatient appointment prices.
[DOCX File , 19 KB - medinform_v12i1e48273_app3.docx]

Multimedia Appendix 4
Leading predicting factors of the best Extreme Gradient Boosting (XGBoost) model.
[DOCX File , 212 KB - medinform_v12i1e48273_app4.docx]

Multimedia Appendix 5
Did Not Show (DNS) rates of all outpatient clinics of the MidCentral District Health Board (MDHB) hospital.
[DOCX File , 329 KB - medinform_v12i1e48273_app5.docx]

Multimedia Appendix 6
Did Not Show (DNS) rates among different deprivation groups and ethnicities.
[DOCX File , 126 KB - medinform_v12i1e48273_app6.docx]

References

1. Tun SYY, Madanian S, Mirza F. Internet of things (IoT) applications for elderly care: a reflective review. Aging Clin Exp Res 2021 Apr;33(4):855-867. [doi: 10.1007/s40520-020-01545-9] [Medline: 32277435]
2. Madanian S. The use of e-health technology in healthcare environment: The role of RFID technology. 10th International Conference on e-Commerce in Developing Countries: with focus on e-Tourism (ECDC); 15-16 April 2016; Isfahan, Iran: IEEE; 2016 Presented at: 10th International Conference on e-Commerce in Developing Countries (ECDC); April 15-16; Isfahan, Iran p. 1-5 URL: https://ieeexplore.ieee.org/document/7492974 [doi: 10.1109/ECDC.2016.7492974]
3. Alaeddini A, Yang K, Reddy C, Yu S. A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Manag Sci 2011 Jun;14(2):146-157 [FREE Full text] [doi: 10.1007/s10729-011-9148-9] [Medline: 21286819]
4. Kaplan-Lewis E, Percac-Lima S. No-show to primary care appointments: why patients do not come. J Prim Care Community Health 2013 Oct;4(4):251-255. [doi: 10.1177/2150131913498513] [Medline: 24327664]
5. DeFife JA, Conklin CZ, Smith JM, Poole J. Psychotherapy appointment no-shows: rates and reasons. Psychotherapy (Chic) 2010 Sep;47(3):413-417. [doi: 10.1037/a0021168] [Medline: 22402096]
6. Dantas L, Fleck J, Cyrino Oliveira FL, Hamacher S. No-shows in appointment scheduling - a systematic literature review. Health Policy 2018 Apr;122(4):412-421 [FREE Full text] [doi: 10.1016/j.healthpol.2018.02.002] [Medline: 29482948]
7. Blæhr E, Søgaard R, Kristensen T, Væggemose U. Observational study identifies non-attendance characteristics in two hospital outpatient clinics. Dan Med J 2016 Oct;63(10) [FREE Full text] [Medline: 27697132]
8. Davies ML, Goffman RM, May JH, Monte RJ, Rodriguez KL, Tjader YC, et al. Large-scale no-show patterns and distributions for clinic operational research. Healthcare (Basel) 2016 Mar 16;4(1) [FREE Full text] [doi: 10.3390/healthcare4010015] [Medline: 27417603]
9. Nelson A, Herron D, Rees G, Nachev P. Predicting scheduled hospital attendance with artificial intelligence. NPJ Digit Med 2019 Apr 12;2(1):26 [FREE Full text] [doi: 10.1038/s41746-019-0103-3] [Medline: 31304373]
10. Samuels RC, Ward VL, Melvin P, Macht-Greenberg M, Wenren LM, Yi J, et al. Missed appointments: factors contributing to high no-show rates in an urban pediatrics primary care clinic. Clin Pediatr (Phila) 2015 Sep 12;54(10):976-982. [doi: 10.1177/0009922815570613] [Medline: 25676833]
11. Hayton C, Clark A, Olive S, Browne P, Galey P, Knights E, et al. Barriers to pulmonary rehabilitation: characteristics that predict patient attendance and adherence. Respir Med 2013 Mar;107(3):401-407 [FREE Full text] [doi: 10.1016/j.rmed.2012.11.016] [Medline: 23261311]
12. French LR, Turner KM, Morley H, Goldsworthy L, Sharp DJ, Hamilton-Shield J. Characteristics of children who do not attend their hospital appointments, and GPs' response: a mixed methods study in primary and secondary care. Br J Gen Pract 2017 Jul;67(660):e483-e489 [FREE Full text] [doi: 10.3399/bjgp17X691373] [Medline: 28630057]
13. Goffman RM, Harris SL, May JH, Milicevic AS, Monte RJ, Myaskovsky L, et al. Modeling patient no-show history and predicting future outpatient appointment behavior in the Veterans Health Administration. Mil Med 2017 May;182(5):e1708-e1714. [doi: 10.7205/MILMED-D-16-00345] [Medline: 29087915]
14. Mohammadi I, Wu H, Turkcan A, Toscos T, Doebbeling BN. Data analytics and modeling for appointment no-show in community health centers. J Prim Care Community Health 2018;9:2150132718811692 [FREE Full text] [doi: 10.1177/2150132718811692] [Medline: 30451063]
15. Williamson AE, Ellis DA, Wilson P, McQueenie R, McConnachie A. Understanding repeated non-attendance in health services: a pilot analysis of administrative data and full study protocol for a national retrospective cohort. BMJ Open 2017 Feb 14;7(2):e014120 [FREE Full text] [doi: 10.1136/bmjopen-2016-014120] [Medline: 28196951]
16. Lee G, Wang S, Dipuro F, Hou J, Grover P, Low L. Leveraging on predictive analytics to manage clinic no show and improve accessibility of care. : IEEE; 2017 Presented at: IEEE International Conference on Data Science and Advanced Analytics (DSAA); October 19-21; Tokyo, Japan p. 19-21 URL: https://ieeexplore.ieee.org/document/8259804 [doi: 10.1109/DSAA.2017.25]

17. Orskov ER, Fraser C. The effects of processing of barley-based supplements on rumen pH, rate of digestion of voluntary intake of dried grass in sheep. Br J Nutr 1975 Nov;34(3):493-500. [doi: 10.1017/s0007114575000530] [Medline: 36]
18. Prasad S, Anand R. Use of mobile telephone short message service as a reminder: the effect on patient attendance. Int Dent J 2012 Feb 18;62(1):21-26 [FREE Full text] [doi: 10.1111/j.1875-595X.2011.00081.x] [Medline: 22251033]
19. AlMuhaideb S, Alswailem O, Alsubaie N, Ferwana I, Alnajem A. Prediction of hospital no-show appointments through artificial intelligence algorithms. Ann Saudi Med 2019;39(6):373-381 [FREE Full text] [doi: 10.5144/0256-4947.2019.373] [Medline: 31804138]
20. Kunjan K, Wu H, Toscos TR, Doebbeling BN. Large-scale data mining to optimize patient-centered scheduling at health centers. J Healthc Inform Res 2019 Mar 4;3(1):1-18 [FREE Full text] [doi: 10.1007/s41666-018-0030-0] [Medline: 35415421]
21. Lenzi H, Ben AJ, Stein AT. Development and validation of a patient no-show predictive model at a primary care setting in Southern Brazil. PLoS One 2019 Apr 4;14(4):e0214869 [FREE Full text] [doi: 10.1371/journal.pone.0214869] [Medline: 30947294]
22. Chen T, Madanian S, Airehrour D, Cherrington M. Machine learning methods for hospital readmission prediction: systematic analysis of literature. J Reliable Intell Environ 2022 Jan 30;8(1):49-66. [doi: 10.1007/s40860-021-00165-y]
23. Madanian S, Parry D, Adeleye O, Poellabauer C, Mirza F, Mathew S, et al. Automatic speech emotion recognition using machine learning: digital transformation of mental health. 2022 Presented at: Pacific Asia Conference on Information Systems; July 5-9; Taipei, Taiwan and Sydney, Australia URL: https://aisel.aisnet.org/pacis2022/45/
24. Kurasawa H, Hayashi K, Fujino A, Takasugi K, Haga T, Waki K, et al. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. J Diabetes Sci Technol 2016 May;10(3):730-736 [FREE Full text] [doi: 10.1177/1932296815614866] [Medline: 26555782]
25. Barrera Ferro D, Brailsford S, Bravo C, Smith H. Improving healthcare access management by predicting patient no-show behaviour. Decision Support Systems 2020 Nov;138:113398 [FREE Full text] [doi: 10.1016/j.dss.2020.113398]
26. Madanian S, Rasoulipanah HR, Yu J. Stress detection on social network: public mental health surveillance. 2023 Presented at: The 2023 Australasian Computer Science Week; January 31-February 3; Melbourne, Australia p. 170-175 URL: https://dl.acm.org/doi/10.1145/3579375.3579397 [doi: 10.1145/3579375.3579397]
27. Madanian S, Chen T, Adeleye O, Templeton J, Poellabauer C, Parry D, et al. Speech emotion recognition using machine learning — a systematic review. Intell Syst Appl 2023 Nov;20:200266 [FREE Full text] [doi: 10.1016/j.iswa.2023.200266]
28. Hamilton W, Round A, Sharp D. Patient, hospital, and general practitioner characteristics associated with non-attendance: a cohort study. Br J Gen Pract 2002 Apr;52(477):317-319 [FREE Full text] [Medline: 11942451]
29. Eid WE, Shehata SF, Cole DA, Doerman KL. Predictors of nonattendance at an endocrinology outpatient clinic. Endocr Pract 2016 Aug;22(8):983-989. [doi: 10.4158/EP161198.OR] [Medline: 27124692]
30. Living in MidCentral: geographic area and population. Te Whatu Ora Health New Zealand. Wellington, New Zealand URL: https://www.careers.mdhb.health.nz/living-in-midcentral [accessed 2023-10-16]
31. Topuz K, Uner H, Oztekin A, Yildirim M. Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network. Ann Oper Res 2017 Apr 4;263(1-2):479-499. [doi: 10.1007/s10479-017-2489-0] [Medline: 44]
32. Lin Q, Betancourt B, Goldstein BA, Steorts RC. Prediction of appointment no-shows using electronic health records. J Appl Stat 2020 Jul;47(7):1220-1234 [FREE Full text] [doi: 10.1080/02664763.2019.1672631] [Medline: 35707022]
33. Fiorillo CE, Hughes AL, I-Chen C, Westgate PM, Gal TJ, Bush ML, et al. Factors associated with patient no-show rates in an academic otolaryngology practice. Laryngoscope 2018 Mar 16;128(3):626-631 [FREE Full text] [doi: 10.1002/lary.26816] [Medline: 28815608]
34. García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. Cham, Switzerland: Springer; 2015.
35. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb) 2013;23(2):143-149 [FREE Full text] [doi: 10.11613/bm.2013.018] [Medline: 23894860]
36. Franke TM, Ho T, Christie CA. The chi-square test. Am J Eval 2011 Nov 08;33(3):448-458. [doi: 10.1177/1098214011426594]
37. The R Project for Statistical Computing.: The R Foundation URL: https://www.r-project.org/ [accessed 2023-10-15]
38. Behrens J, DiCerbo K, Yel N, Levy R. Exploratory data analysis. In: Handbook of Psychology: Research Methods in Psychology. New York, NY: John Wiley & Sons; 2012:2012.
39. Behrens JT. Principles and procedures of exploratory data analysis. Psychol Methods 1997 Jun;2(2):131-160. [doi: 10.1037/1082-989x.2.2.131]
40. Mantovani R, Horváth T, Cerri R, Vanschoren J. Hyper-parameter tuning of a decision tree induction algorithm. : IEEE; 2016 Presented at: 5th Brazilian Conference on Intelligent Systems (BRACIS); October 9-12; Recife, Brazil p. 37-42. [doi: 10.1109/BRACIS.2016.018]
41. Doucette J, Heywood M. GP classification under imbalanced data sets: active sub-sampling and AUC approximation. In: O'Neill M, Vanneschi L, Gustafson S, Alcázar A, Falco I, Cioppa A, et al, editors. Genetic Programming. Berlin, Germany: Springer; 2008:9-23.
42. Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Comput Intell 2004 Jan 28;20(1):18-36 [FREE Full text] [doi: 10.1111/j.0824-7935.2004.t01-1-00228.x]

43. Batuwita R, Palade V. Efficient resampling methods for training support vector machines with imbalanced datasets. : IEEE; 2010 Presented at: 2010 International Joint Conference on Neural Networks (IJCNN); July 18-23; Barcelona, Spain p. 1-8. [doi: 10.1109/IJCNN.2010.5596787]
44. Xu-Ying Liu, Jianxin Wu, Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst, Man, Cybern B 2009 Apr;39(2):539-550. [doi: 10.1109/tsmcb.2008.2007853]
45. Greenwell B, Boehmke B. Variable importance plots—An introduction to the vip package. R J 2020;12(1):343. [doi: 10.32614/rj-2020-013]
46. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 2011 Jul 29;11:51 [FREE Full text] [doi: 10.1186/1472-6947-11-51] [Medline: 21801360]
47. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. University of California Berkeley. 2004. URL: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf [accessed 2023-12-28]
48. Hosmer D, Lemeshow S, Sturdivant R. Applied Logistic Regression. Hoboken, NJ: John Wiley & Sons; 2013.
49. Devasahay SR, Karpagam S, Ma NL. Predicting appointment misses in hospitals using data analytics. Mhealth 2017;3:12 [FREE Full text] [doi: 10.21037/mhealth.2017.03.03] [Medline: 28567409]
50. Sahin EK. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Appl Sci 2020 Jun 30;2(7). [doi: 10.1007/s42452-020-3060-1]
51. Nationwide service framework library online. Te Whatu Ora Health New Zealand. 2023. URL: https://www.tewhatuora.govt.nz/our-health-system/nationwide-service-framework-library/ [accessed 2023-10-14]
52. Nejatian S, Parvin H, Faraji E. Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification. Neurocomputing 2018 Feb;276:55-66. [doi: 10.1016/j.neucom.2017.06.082]
53. Hastie T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer; 2017.
54. Harvey HB, Liu C, Ai J, Jaworsky C, Guerrier CE, Flores E, et al. Predicting no-shows in radiology using regression modeling of data available in the electronic medical record. J Am Coll Radiol 2017 Oct;14(10):1303-1309. [doi: 10.1016/j.jacr.2017.05.007] [Medline: 28673777]
55. Ma N, Khataniar S, Wu D, Ng S. Predictive analytics for outpatient appointments. : IEEE; 2014 Presented at: 2014 International Conference on Information Science & Applications (ICISA); May 6-9; Seoul, South Korea p. 6-9. [doi: 10.1109/ICISA.2014.6847449]
56. Lamba M, Alamri Y, Garg P, Frampton C, Rowbotham D, Gearry R. Predictors of non-attendance at outpatient endoscopy: a five-year multi-centre observational study from New Zealand. N Z Med J 2019 Jun 07;132(1496):31-38. [Medline: 31170131]
57. Maori health. New Zealand Ministry of Health. 2023. URL: https://www.health.govt.nz/our-work/populations/maori-health [accessed 2023-11-05]
58. Atkinson J. Socioeconomic deprivation indexes: NZDep and NZiDe. University of Otago. 2019. URL: https://www.otago.ac.nz/wellington/departments/publichealth/research/hirp/otago020194.html [accessed 2023-11-15]
59. Samorani M, LaGanga L. Outpatient appointment scheduling given individual day-dependent no-show predictions. Eur J Oper Res 2015 Jan;240(1):245-257 [FREE Full text] [doi: 10.1016/j.ejor.2014.06.034]
60. Wilcox A, Levi EE, Garrett JM. Predictors of non-attendance to the postpartum follow-up visit. Matern Child Health J 2016 Nov 25;20(Suppl 1):22-27. [doi: 10.1007/s10995-016-2184-9] [Medline: 27562797]
61. The new health system. New Zealand Department of the Prime Minister and Cabinet. 2021. URL: https://dpmc.govt.nz/our-business-units/transition-unit/response-health-and-disability-system-review/information [accessed 2023-11-15]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUROC: area under the receiver operating characteristic
AUT: Auckland University of Technology
DNA: Did Not Attend
DNS: Did Not Show
EDA: exploratory data analysis
FTA: Failed To Attend
LR: logistic regression
MDHB: MidCentral District Health Board
ML: machine learning
RF: random forest
ROC: receiver operating characteristic
XGBoost: Extreme Gradient Boosting

Edited by C Lovis; submitted 17.04.23; peer-reviewed by A Blasiak, D Gartner; comments to author 02.10.23; revised version received 07.11.23; accepted 04.12.23; published 12.01.24.

Please cite as:
Yang Y, Madanian S, Parry D
Enhancing Health Equity by Predicting Missed Appointments in Health Care: Machine Learning Study
JMIR Med Inform 2024;12:e48273
URL: https://medinform.jmir.org/2024/1/e48273
doi:10.2196/48273
PMID:38214974

Original Paper

Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study

Anne de Hond[1,2,3], MSc, PhD; Marieke van Buchem[1,2,3], MSc; Claudio Fanconi[3,4], MSc; Mohana Roy[5], MD; Douglas Blayney[5], MD; Ilse Kant[1,6], MSc, PhD; Ewout Steyerberg[1,2], MSc, PhD; Tina Hernandez-Boussard[3,7,8], MSc, PhD

[1]Clinical AI Implementation and Research Lab, Leiden University Medical Centre, Leiden, Netherlands
[2]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands
[3]Department of Medicine (Biomedical Informatics), Stanford Medicine, Stanford University, Stanford, CA, United States
[4]Department of Electrical Engineering and Information Technology, ETH Zürich, Zürich, Switzerland
[5]Department of Medical Oncology, Stanford Medicine, Stanford University, Stanford, CA, United States
[6]Department of Digital Health, University Medical Centre Utrecht, Utrecht, Netherlands
[7]Department of Biomedical Data Science, Stanford University, Stanford, CA, United States
[8]Department of Epidemiology & Population Health (by courtesy), Stanford University, Stanford, CA, United States

Corresponding Author:
Tina Hernandez-Boussard, MSc, PhD
Department of Medicine (Biomedical Informatics)
Stanford Medicine
Stanford University
1265 Welch Road
Stanford, CA, 94305
United States
Phone: 1 650 725 5507
Email: boussard@stanford.edu

Abstract

Background: Patients with cancer starting systemic treatment programs, such as chemotherapy, often develop depression. A prediction model may assist physicians and health care workers in the early identification of these vulnerable patients.

Objective: This study aimed to develop a prediction model for depression risk within the first month of cancer treatment.

Methods: We included 16,159 patients diagnosed with cancer starting chemo- or radiotherapy treatment between 2008 and 2021. Machine learning models (eg, least absolute shrinkage and selection operator [LASSO] logistic regression) and natural language processing models (Bidirectional Encoder Representations from Transformers [BERT]) were used to develop multimodal prediction models using both electronic health record data and unstructured text (patient emails and clinician notes). Model performance was assessed in an independent test set (n=5387, 33%) using area under the receiver operating characteristic curve (AUROC), calibration curves, and decision curve analysis to assess initial clinical impact use.

Results: Among 16,159 patients, 437 (2.7%) received a depression diagnosis within the first month of treatment. The LASSO logistic regression models based on the structured data (AUROC 0.74, 95% CI 0.71-0.78) and structured data with email classification scores (AUROC 0.74, 95% CI 0.71-0.78) had the best discriminative performance. The BERT models based on clinician notes and structured data with email classification scores had AUROCs around 0.71. The logistic regression model based on email classification scores alone performed poorly (AUROC 0.54, 95% CI 0.52-0.56), and the model based solely on clinician notes had the worst performance (AUROC 0.50, 95% CI 0.49-0.52). Calibration was good for the logistic regression models, whereas the BERT models produced overly extreme risk estimates even after recalibration. There was a small range of decision thresholds for which the best-performing model showed promising clinical effectiveness use. The risks were underestimated for female and Black patients.

Conclusions: The results demonstrated the potential and limitations of machine learning and multimodal models for predicting depression risk in patients with cancer. Future research is needed to further validate these models, refine the outcome label and predictors related to mental health, and address biases across subgroups.

(JMIR Med Inform 2024;12:e51925) doi:10.2196/51925

KEYWORDS

natural language processing; machine learning; artificial intelligence; oncology; depression; clinical decision support; decision support; cancer; patients with cancer; chemotherapy; mental health; prediction model; depression risk; cancer treatment; radiotherapy; diagnosis; validation; cancer care; care

Introduction

Background

Depression in patients with cancer occurs frequently around diagnosis and treatment and has been negatively associated with a patient's prognosis, quality of life, and treatment adherence [1-5]. Despite affecting up to 20% of patients with cancer and far exceeding the prevalence in the general population (8.4% in the United States [6]), depression is underdiagnosed and often untreated [1,3,7-9]. Constrained clinician time and a strong focus on anticancer treatment may contribute to the insufficient identification of patients at risk for depression [10-13]. Early detection of depression in patients with cancer may enable timely mental health support to augment the anticancer treatment.

Clinical decision support tools with artificial intelligence (AI) technologies could synthesize the abundance of data collected during treatment to help clinicians identify which patients may need specific attention and steer additional mental health resources to those at high risk. A recent review [14] of AI models developed for depression risk in primary care [15], elderly care [16,17], and social media posts [18-20] highlights how AI tools have the potential for early identification of mental health issues. However, oncology-specific applications are rare, and those that do exist are developed on selected small samples that may not generalize to clinical care settings [21,22]. This leaves a gap in oncological care for mental health.

Objective

We aimed to develop a prediction model for early identification of patients at risk for depression within the first month of chemo- or radiotherapy treatment. We assessed the relevance of different data modalities for predictive performance in a retrospective cohort study.

Methods

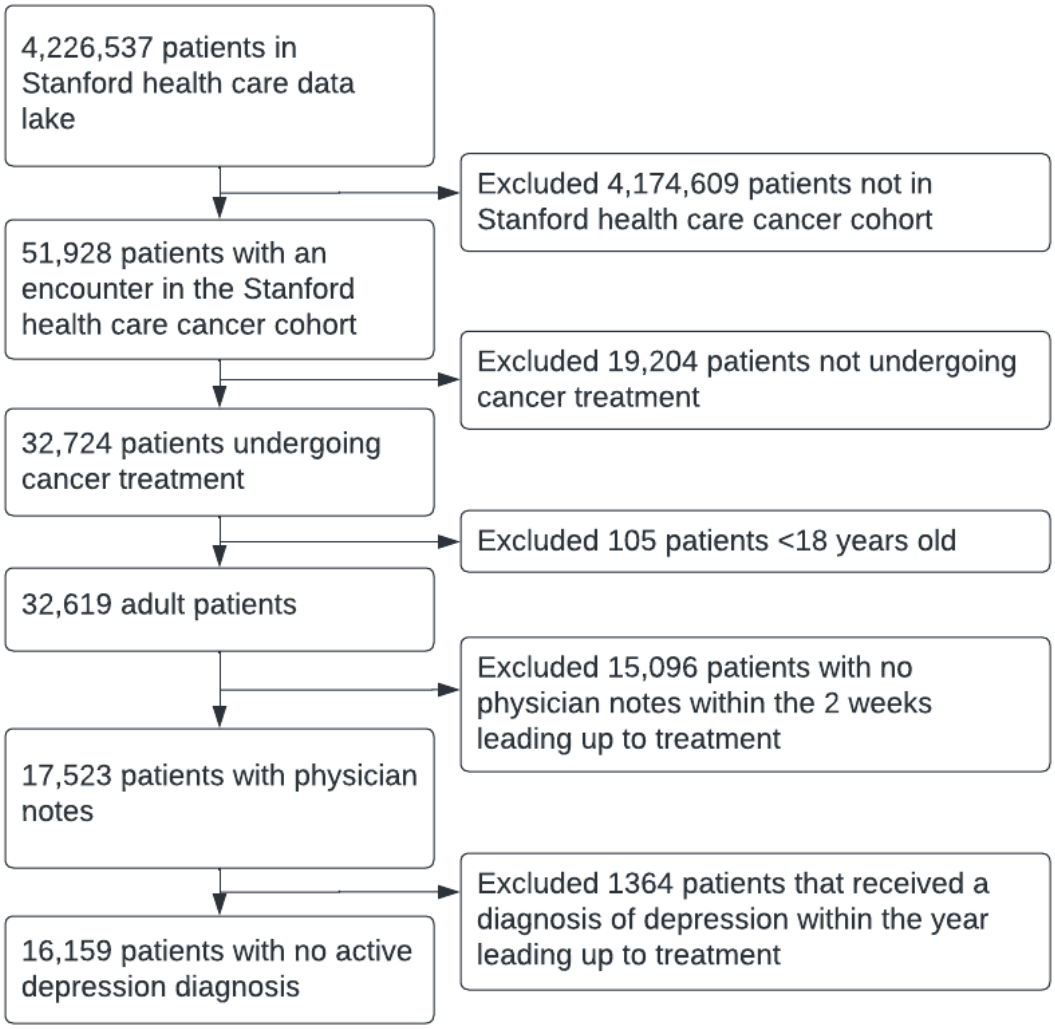
Data Source and Patient Population

This retrospective observational study used data from the integration of 3 health care organizations: an academic medical center (AMC), a primary and specialty care alliance (PSC), and a community medical center (CMC). These organizations offer a wide spectrum of specialized and advanced health care services for complex medical conditions, operating in over 600 clinics. The PSC, established in 2011, comprises more than 70 primary and specialty clinics throughout the California Bay Area. The CMC provides a range of inpatient and outpatient services in the Tri-Valley region of East Bay and was acquired by the AMC in 2015. Following the merger and acquisition, all health care settings adopted the same Epic-based electronic health record (EHR; Epic Systems Corporation) system. Patients for the study were identified from a clinical data warehouse that consolidated patient data from the AMC, PSC, and CMC from 2008 to 2021 [23]. The EHR system was initiated in 2005, and by 2008, the data had reached a state of robustness and high quality. The study concluded in 2021 to ensure that all patients who visited the clinic during the extended period were comprehensively captured.

As an integral component of the EHR system, the MyHealth portal and web interface are seamlessly incorporated into the EHR. This integration includes a patient portal, enabling patients to engage with their health care teams through secure email communication. Patient-generated emails were systematically gathered from the MyHealth patient portal. These email exchanges feature structured subject lines, with patients selecting from a predefined set of categories such as "Non-Urgent Medical Question," "Prescription Question," "Visit Follow-Up Question," "Test Results Question," "Update My Health Information," "Scheduling Question," and "Ordered Test Question." The email body allows for free-text input but is limited to 1000 characters. Importantly, all incoming emails are meticulously triaged to the appropriate members of the patient's health care team, including clerical, scheduling, clinical, or other team members, who take the necessary actions or provide responses as needed [24].

Adult patients receiving chemo- or radiotherapy treatment were included in the cohort. Given the data-intensive nature of the techniques used [25], our objective was to encompass all eligible patients throughout the entire available period at the time of our analysis. The start of cancer treatment was defined as the first patient encounter that registered chemotherapy (including targeted and immunotherapy) or radiotherapy ("chemotherapy" and "clinical procedure codes" in Multimedia Appendix 1). We excluded patients who did not receive cancer treatment (eg, patients seen for a second opinion only), were younger than 18 years, and had no clinician notes within the 2 weeks leading up to the treatment (Figure 1). We also excluded patients with a depression diagnosis within the year leading up to treatment as we aimed to focus on individuals who are at risk of developing depression during or after their treatment (Figure 1). It was assumed that these patients were already receiving treatment for their depression or at least had additional support offered to them.

Figure 1. Flowchart of cohort selection.



Ethical Considerations

This study was approved by the Stanford institutional review board (#47644). Informed consent was waived for this retrospective study for access to personally identifiable health information as it would not be reasonable, feasible, or practical. The data are housed in the Stanford Nero Computing Platform, which is a highly secure, fully integrated internal research data platform meeting all security standards for high risk and protected health information data. The security is managed and monitored, and the platform is updated and adapted to meet regulatory changes.

Predictive Outcome

Depression was defined in consultation with oncologist coauthors (DWB and MR) as a depression diagnosis via the *International Classification of Diseases* (ICD)-9 and ICD-10 codes obtained from EHR data ("ICD depression codes" in Multimedia Appendix 1). This end point was chosen as it was the most conservative and has been shown to correlate reasonably well with clinical opinion [26]. Depression risk was predicted within 1 month of cancer treatment. This time window was chosen as depression prevalence is highest during diagnosis and the acute phase of cancer treatment [27].

Structured Data Predictors

The following variables were obtained from structured EHR fields: sex (male and female), age, insurance status (private, Medicare, Medicaid, and other or not identified), cancer stage (I, II, III, IV, and missing), hospitalized in the previous month (yes or no), 1 or more emergency department visits in the previous month (yes or no), the Charlson comorbidity score [28], and the number of emails sent in the month prior to treatment (none, 1-3, 4, or more) based on a previous study [24]. Insurance status was recoded into 4 comprehensive categories (private, Medicare, Medicaid, and other or not identified). Cancer stage was also recoded to contain the 4 main stages (I, II, III, IV, and missing). Whether or not patients sent emails at night in the previous month was also included as insomnia and depression are intimately related [29]. Binary variables were added indicating whether a patient had previously received a depression diagnosis; depressant medication; or a referral to a psychiatrist, psychologist, or social worker. Finally, race and ethnicity (Hispanic, non-Hispanic Asian, non-Hispanic Black, and non-Hispanic White) was included in one of the sensitivity analyses (see below). The ethnicities "Latino" and "Hispanic" were merged into 1 category (Hispanic). The categorical predictors were converted into dummy variables.

Descriptive statistics were reported in terms of percentages for categorical variables and the mean and SD for continuous variables. We analyzed the cancer and insurance information that was closest to, but preceding, the patient's start of treatment. We stratified descriptive statistics according to outcome (depression diagnosis or not) and messaging behavior (active email communicator in the past month or not).

Unstructured Text Predictors

Unstructured text included patient emails with the subject "Non-Urgent Medical Question" sent through a secure patient portal and clinician notes [24].

A Bidirectional Representations from Transformers (BERT) model was trained on a subset of manually labeled emails to classify each email as being "concerning for depression" or not (see the Multimedia Appendix 2 [30-33] for further details on the annotation strategy and model development). Automatically sent emails; copies of previously sent emails; and emails containing questionnaires, appointment requests, and medication refill requests were removed from the set of patient emails. Emails with less than 30 words were removed from the data set. Each email in the final data set was truncated to a maximum token length of 512. This BERT model assigned each patient email a classification score ranging from 0 (not concerning for depression at all) to 1 (most concerning for depression). These email classification scores were summarized at the patient level by calculating the minimum email classification score in the previous month, the maximum score in the previous month, and the mean score in the previous month. These email classification features were then included as structured data in the subsequent model developments.

Clinician notes that were shorter than 100 words or longer than 5000 were removed as these contained erroneous entries or long copies of previous notes, respectively. Notes with mentions of clinical trials, duplicates, and empty notes were also removed. We merged the most recent clinical notes (at most 3) created within the 2 weeks before the start of treatment. The merged notes were decomposed into chunks of at most 25 sequences (to avoid computational issues), each sequence consisting of 256 tokens.

Model Development

For all models, data were randomly split into the same two-thirds for the train set and one-third for the test set. A total of 6 models were trained to assess the value of multimodal data for this use case.

First, a machine learning (ML) model was developed based on the structured EHR data (model 1), email classification scores (model 2), and the combination of the 2 (model 3). The following ML algorithms were compared for these models: least absolute shrinkage and selection operator (LASSO) logistic regression, a decision tree, random forest, gradient boosting decision trees, *k*-nearest neighbor, and naive Bayes.

LASSO logistic regression is a regularized regression approach, providing both variable selection and shrinkage of regression coefficients. A decision tree is a nonparametric algorithm consisting of a hierarchical tree structure. A random forest combines the predictions of many independently built decision trees into 1 prediction. Gradient boosting decision trees essentially optimize random forest estimation by gradient boosting. The *k*-nearest neighbor algorithm is also nonparametric and uses proximity to previously seen data points to make predictions. Finally, naive Bayes is a generative algorithm that models the distribution of its predictors to make predictions.

The hyperparameters of these models (see Tables S1-S3 in Multimedia Appendix 2) were optimized using Bayesian optimization and 5×10-fold cross-validation. The final ML models were trained on all training data with optimized hyperparameters. The best-performing ML algorithms were the basis for extension with unstructured data.

We trained BERT models based on the clinician notes (model 4), the structured EHR data in combination with the clinician notes (model 5), and the structured EHR data in combination with the email classification scores and the clinician notes (model 6). BERT models are deep learning language models that learn contextual relations between words in a text. Models 5 and 6 made use of a modality-specific deep learning architecture to combine the different data modalities in the modeling process (see Multimedia Appendix 2 for more details) [34]. We used a pretrained DistilBERT model [32] as it required less computation than BERT or ClinicalBERT models [33]. The hyperparameters were tuned on 80% and validated on 20% of the training data. The model parameters of the best-performing epoch on the validation data were chosen for further analyses. Probability estimates were recalibrated via isotonic regression for all models [35].

Statistical Analysis

Model discrimination was quantified by the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) on the test data. Calibration was assessed through calibration plots, with a calibration intercept and slope as summary performance measures [36]. CIs were obtained via bootstrapping (based on 1000 iterations).

As an initial assessment of clinical usefulness, we performed a decision curve analysis for all 6 models plotting net benefit (NB) across a range of decision probability thresholds [37,38]. NB is defined as the number of true-positive classifications penalized for false-positive classifications [39]. The models have the potential to improve clinical decision-making when they have higher NB than 2 baseline strategies: label all as high risk for developing depression and label none as high risk for developing depression.

Sensitivity Analysis

Sensitivity analyses were performed on the best-performing model to evaluate the impact of modeling choices on model outcomes. Additional models considered different prediction windows (45 days, 2 months, 3 months, and 6 months after the start of cancer treatment). Moreover, patients dying within these prediction windows are a potential competing risk for patients at risk for depression. We therefore removed these patients from the train and test data and repeated the analyses. Variants of outcome definitions such as predicting a prescription of

antidepressant medication and a referral to a psychiatrist, psychologist, or social worker within 1 month of cancer treatment ("antidepressant medication" and "mental health referral" in Multimedia Appendix 1) were considered. These definitions were chosen as they might indicate a patient experiencing depression without being officially diagnosed. We also trained a model on the combined outcome of either receiving a depression diagnosis, antidepressant medication prescription, or referral to a psychiatrist, psychologist, or social worker.

Fairness Analysis and Including Race and Ethnicity

To identify potential fairness issues for specific demographic groups, AUROC, calibration slope, and intercept were compared across sex and race and ethnicity groups [40]. In addition, race and ethnicity was added as a confounder to assess its effect on subgroup model performance.

Software, Data, and Reporting

All analyses were performed in Python 3.9.7 (Python Software Foundation). Code is available in a git repository [41]. We followed the MINIMAR reporting guidelines (see Multimedia Appendix 2) [42].

Results

Descriptive Statistics

A total of 16,159 patients starting cancer treatment between 2008 and 2021 were included in the analyses, of whom 437 (2.7%) received a diagnosis of depression within 1 month of cancer treatment (Table 1 and Figure 1). The 437 patients receiving a depression diagnosis within 1 month of treatment were, on average, younger, more likely to be female, more likely to be non-Hispanic White, and less likely to be non-Hispanic Asian (Table 1). Moreover, patients with a depression diagnosis made more emergency department visits (Table 1). They were also more likely to have received a previous depression diagnosis more than a year before the start of treatment, a prescription for antidepressant medication, and a mental health referral.

Patients who sent emails (4816/16,159, 29.8%) were more likely to be non-Hispanic White or Asian and be privately insured (Table S4 in Multimedia Appendix 2). On average, they were less likely to be hospitalized but made more emergency department visits 1 month prior to treatment and had a higher Charlson comorbidity score; they were also more likely to have previously received a depression diagnosis, antidepressant medication, and a mental health referral.

Table 1. Descriptive statistics of the cancer cohort.

Descriptive statistics	All (N=16,159)	No depression diagnosis within 1 month after onset of treatment (n=15,722, 97.3%)	Depression diagnosis within 1 month after onset of treatment (n=437, 2.7%)
Demographics			
Sex (female), n (%)	8568 (53)	8296 (52.8)[a]	272 (62.2)[a]
Age (years), mean (SD)	62 (15)	62 (15)[a]	60 (14)[a]
Race and ethnicity, n (%)			
Hispanic	1870 (11.6)	1812 (11.5)[a]	58 (13.3)[a]
Non-Hispanic Asian	3582 (22.2)	3525 (22.4)[a]	57 (13)[a]
Non-Hispanic Black	422 (2.6)	410 (2.6)[a]	<20 (<5)[a]
Non-Hispanic White	8864 (54.9)	8583 (54.6)[a]	281 (64.3)[a]
Other	1421 (8.8)	1392 (8.9)[a]	29 (6.6)[a]
Insurance characteristics, n (%)			
Private	8745 (54.1)	8496 (54)	249 (57)
Medicare	2590 (16)	2514 (16)	76 (17.4)
Medicaid	1917 (11.9)	1860 (11.8)	57 (13)
Other or not identified	2907 (18)	2852 (18.1)	55 (12.6)
Treatment characteristics, mean (SD)			
Number of hospitalizations one month prior to treatment	2083 (13)	2016 (13)	67 (15)
Number of emergency department visits 1 month prior to treatment	945 (6)	895 (6)[a]	50 (11)[a]
Charlson comorbidity score	6.9 (3.8)	6.9 (3.8)	6.9 (3.9)
Tumor type, n (%)			
Breast	1772 (11)	1739 (11.1)	33 (7.6)
Lung	1001 (6.2)	973 (6.2)	28 (6.4)
Prostate	777 (4.8)	764 (4.9)	<20 (<5)
Colon and rectum	543 (3.4)	525 (3.3)	<20 (<5)
Non-Hodgkin lymphoma	535 (3.3)	527 (3.4)	<20 (<5)
Other	3459 (21.4)	3364 (21.4)	95 (21.7)
Missing	8072 (50)	7830 (49.8)	242 (55.4)
Cancer stage, n (%)			
Stage I	1492 (9.2)	1466 (9.3)	26 (5.9)
Stage II	1499 (9.3)	1468 (9.3)	31 (7.1)
Stage III	1329 (8.2)	1294 (8.2)	35 (8)
Stage IV	1758 (10.9)	1699 (10.8)	59 (13.5)
Missing	10081 (62.4)	9795 (62.3)	286 (65.4)
Patient email information (1 month prior to treatment)			
Sent 1 or more emails, n (%)	4070 (25.2)	3943 (25.1)	127 (29.1)
Email length in words, mean (SD)	49 (35)	49 (35)	49 (35)
Sent emails at night, n (%)	308 (1.9)	296 (1.9)	<20 (<5)
Mental health history, n (%)			
History of depression diagnosis	400 (2.5)	343 (2.2)[a]	57 (13)[a]

Descriptive statistics	All (N=16,159)	No depression diagnosis within 1 month after onset of treatment (n=15,722, 97.3%)	Depression diagnosis within 1 month after onset of treatment (n=437, 2.7%)
History of antidepressant medication	2219 (13.7)	2030 (12.9)[a]	189 (43.2)[a]
History of mental health referral	2707 (16.8)	2563 (16.3)[a]	144 (33)[a]

[a]This was tested at the 5% significance level.

Performance Statistics

The best-performing ML models were based on LASSO logistic regression (Table 2; Tables S1 and S3 in Multimedia Appendix 2). The model based on structured data alone had an AUROC of 0.74 (95% CI 0.71-0.78). The combination of structured data with email classification scores also had an AUROC of 0.74 (95% CI 0.71-0.78), while a model based solely on email classification scores had an AUROC of 0.54 (95% CI 0.52-0.56). At a high level of sensitivity (0.9 at a decision threshold of 1%; Table 3), the PPV of the best-performing model based on structured data was low (0.04; Table 3). At higher decision thresholds (3% and 10%; Table 3), the PPV was increased to 0.07 and 0.17, respectively, but this came at a cost of sensitivity (0.63 and 0.19).

The BERT model based on the clinician notes performed worst and had an AUROC of 0.50 (95% CI 0.49-0.52; Table 2). Combining structured EHR data with clinician notes did improve AUROC performance (0.71, 95% CI 0.68-0.75; Table 2) and so did adding email classification scores (0.70, 95% CI 0.67-0.73; Table 2).

Calibration was acceptable for all ML models. The BERT-based models tended to produce overly extreme risk estimates even after recalibration.

The decision curve analysis showed a small range of decision thresholds for which the best-performing model (LASSO logistic regression based on structured data) had higher NB than the treat all or treat no one strategies (Figure 2). At a decision threshold of 3%, the model with structured EHR data had a NB of 0.01. This represents a net increase of 1 true positive patient at risk for depression per 100 patients without increasing any false positives (at the start of treatment). At a threshold of 10%, the model had a NB of only 0.002, so 2 net true positives per 1000 patients.

Table 2. Discrimination and calibration for predicting depression risk within 1 month after the onset of treatment (test data).

Type of data	AUROC[a] (95% CI)	Calibration intercept (95% CI)	Calibration slope (95% CI)
Structured EHR[b] data	0.74 (0.71 to 0.78)	0.07 (–0.09 to 0.24)	0.93 (0.77 to 1.09)
Patient emails	0.54 (0.52 to 0.56)	–0.02 (–0.18 to 0.14)	1.0 (0.52 to 1.48)
Structured EHR data and patient emails	0.74 (0.71 to 0.78)	0.07 (–0.09 to 0.24)	0.91 (0.76 to 1.07)
Clinician notes	0.5 (0.49 to 0.52)	–0.05 (–0.21 to 0.11)	0.94 (–1.32 to 3.2)
Structured EHR data and clinician notes	0.71 (0.68 to 0.75)	–0.09 (–0.25 to 0.07)	1.92 (1.57 to 2.28)
Structured EHR data, clinician notes, and patient emails	0.7 (0.67 to 0.73)	–0.16 (–0.32 to –0.0)	2.46 (1.98 to 2.93)

[a]AUROC: area under the receiver operating characteristics curve.

[b]EHR: electronic health record.

Table 3. Sensitivity, specificity, PPV[a], and NPV[b] at different decision thresholds for predicting depression risk within 1 month after the onset of treatment (test data).

Threshold and analysis	Structured EHR[c] data	Patient emails	Structured EHR data and patient emails	Clinician notes	Structured EHR data and clinician notes	Structured EHR data, clinician notes, and patient emails
1%						
Sensitivity (n/N)	0.9 (140/156)	1.0 (156/156)	0.87 (136/156)	1.0 (156/156)	1.0 (156/156)	1.0 (156/156)
Specificity (n/N)	0.35 (1847/5231)	0.0 (0/5231)	0.37 (1915/5231)	0.0 (0/5231)	0.0 (0/5231)	0.0 (0/5231)
PPV (n/N)	0.04 (140/3524)	0.03 (156/5387)	0.04 (136/3452)	0.03 (156/5387)	0.03 (156/5387)	0.03 (156/5387)
NPV (n/N)	0.99 (1847/1863)	N/A[d]	0.99 (1915/1935)	N/A	N/A	N/A
3%						
Sensitivity (n/N)	0.63 (98/156)	0.13 (20/156)	0.58 (90/156)	1.0 (156/156)	0.55 (86/156)	0.67 (104/156)
Specificity (n/N)	0.75 (3912/5231)	0.95 (4962/5231)	0.77 (4032/5231)	0.0 (0/5231)	0.82 (4293/5231)	0.71 (3735/5231)
PPV (n/N)	0.07 (98/1417)	0.07 (20/289)	0.07 (90/1289)	0.03 (156/5387)	0.08 (86/1024)	0.06 (104/1600)
NPV (n/N)	0.99 (3912/3970)	0.97 (4962/5098)	0.98 (4032/4098)	N/A	0.98 (4293/4363)	0.99 (3735/3787)
10%						
Sensitivity (n/N)	0.19 (29/156)	0.0 (0/156)	0.19 (30/156)	0.0 (0/156)	0.0 (0/156)	0.0 (0/156)
Specificity (n/N)	0.97 (5086/5231)	1.0 (5231/5231)	0.97 (5071/5231)	1.0 (5231/5231)	1.0 (5231/5231)	1.0 (5231/5231)
PPV (n/N)	0.17 (29/174)	N/A	0.16 (30/190)	N/A	N/A	N/A
NPV (n/N)	0.98 (5086/5213)	0.97 (5231/5387)	0.98 (5071/5197)	0.97 (5231/5387)	0.97 (5231/5387)	0.97 (5231/5387)

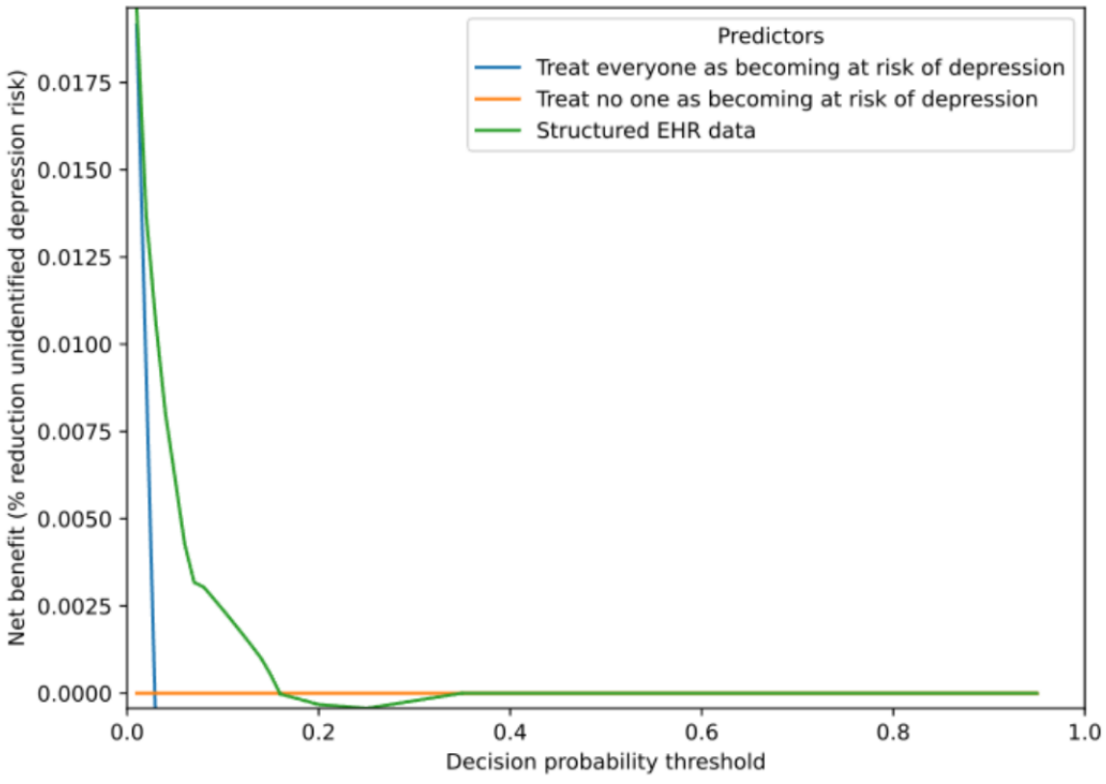
[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]EHR: electronic health record.

[d]N/A: not available.

Figure 2. This decision curve analysis (DCA) plots net benefit for the baseline treat all and treat none strategies and the best-performing prediction model (LASSO logistic regression on structured data). EHR: electronic health record; LASSO: least absolute shrinkage and selection operator.



Sensitivity Analysis

The models predicting depression risk within 45 days (AUROC 0.73, 95% CI 0.69-0.76), 2 months (AUROC 0.73, 95% CI 0.70-0.76), 3 months (AUROC 0.73, 95% CI 0.71-0.76), and 6 months (AUROC 0.72, 95% CI 0.70-0.74) of cancer treatment obtained similar discrimination and calibration compared to the base model predicting depression risk within 1 month (LASSO logistic regression; Table S5 in Multimedia Appendix 2). In the test data, a total of 24 (0.4%) patients died within 1 month after starting treatment. Omitting patients dying within the time frames of interest (1 month-6 months) had no impact on model performance (Table S6 in Multimedia Appendix 2). The model trained to predict depression medication (LASSO logistic regression) also obtained similar discrimination (0.75, 95% CI 0.73-0.78; Table S7 in Multimedia Appendix 2) and calibration compared to the base model predicting depression risk via depression diagnosis. The model trained to predict a referral to a psychiatrist, psychologist, or social worker obtained a lower AUROC of 0.62 (95% CI 0.60-0.64; Table S7 in Multimedia Appendix 2) and comparable calibration.

Fairness Analysis and Including Race and Ethnicity

The fairness analysis showed that model discrimination was similar for male patients (AUROC 0.73, 95% CI 0.67-0.80; Table S8 in Multimedia Appendix 2) and female patients (AUROC 0.74, 95% CI 0.70-0.78; Table S8 in Multimedia Appendix 2). The calibration plot showed that depression risk was underestimated for female patients and overestimated for male patients (Figure S1 in Multimedia Appendix 2). Discrimination was best for the non-Hispanic Black patients (AUROC 0.92, 95% CI 0.84-0.99; Table S8 in Multimedia Appendix 2), with respect to the non-Hispanic White patients (AUROC 0.74, 95% CI 0.69-0.78; Table S8 in Multimedia Appendix 2) and the non-Hispanic Asian patients (AUROC 0.75, 95% CI 0.63-0.87; Table S8 in Multimedia Appendix 2), and it was worst for Hispanic patients (AUROC 0.71, 95% CI 0.62-0.80; Table S8 in Multimedia Appendix 2). Predictions were underestimated for the non-Hispanic Black patients and overestimated for the non-Hispanic Asian patients (Figure S1 in Multimedia Appendix 2). Adding race and ethnicity as a feature to the best-performing model did not improve model discrimination or calibration (AUROC 0.74, 95% CI 0.71-0.78 vs 0.74, 95% CI 0.70-0.77; Table S9 in Multimedia Appendix 2).

Discussion

Principal Findings

This study developed a prediction model to identify patients with cancer at risk for depression within 1 month of chemo- or radiotherapy treatment. We used data from a large comprehensive cancer center with over 16,000 patients. The best-performing models (LASSO logistic regression with structured data with or without patient email classification scores) had reasonable AUROC and calibration. The LASSO logistic regression model with structured data demonstrates a small improvement in NB over the baseline strategy of labeling no one as at risk for depression. Multimodal BERT models (trained on structured data and unstructured text) did not perform better than the best-performing ML model trained solely on structured data.

To date, depression in patients with cancer is underdiagnosed, and studies show that patients with depression are up to 3 times more likely to be noncompliant with medical treatment recommendations [3,43,44]. Treatment adherence is a high priority, given the evidence demonstrating statistically significant associations between treatment nonadherence and patient outcomes, including cancer progression, low-value health care use, and worse survival [45-48]. Therefore, an AI model—which flags patients at risk for depression with minimal clinical input and workflow disruption—is needed at the point of care to prompt clinicians to intervene early and improve patient well-being and anticancer outcomes.

This model may be used in preparation for clinical consultations to more efficiently use the limited time allotted to oncologist-patient interaction to facilitate any needed additional mental health support. By harnessing a combination of structured EHR data and unstructured text data from patient emails and clinician notes, the tool can offer a comprehensive assessment of a patient's depression risk and help synthesize this information at point of care for the provider. With the ability to establish personalized risk assessments, determine clinical use thresholds, and address potential biases in risk assessment, a clinical decision support tool developed from this work has the potential to significantly enhance the quality of care and mental health outcomes for these vulnerable patients. As the study recognizes the need for ongoing validation, refinement, and bias mitigation, it underscores the dynamic and adaptable nature of this tool in improving cancer care and treatment adherence. This tool can be a valuable addition to the health care system, ultimately improving mental health outcomes and treatment adherence for these vulnerable patients.

The created model has good performance, although our label (receiving a depression diagnosis) depends heavily upon the accurate recognition of depression by the care team. The model's clinical usefulness depends on the acceptability of the test trade-off. The best-performing model had a high false-positive rate at high levels of sensitivity, and the decision curve analysis showed a test trade-off of 100 assessments for 1 additional true positive patient at a decision threshold of 3%. If these assessments can be done nearly for free (eg, a quick check during a patient visit) and if we already miss all future depressions, then this small improvement may be welcome, although this warrants further validation and testing in the clinical environment. The high false-positive rate and small NB of the best-performing model are likely affected by the moderate discrimination and low event rate [49]. In future developments, the NB may be increased by focusing on improving the labeling of the outcome variable. In addition, richer input data not available to us at the time of analysis could improve model discrimination, like information on lifestyle habits, self-reported mental health assessments, and clinical and pathological factors.

As depression presents differently across sex, race, and ethnicity [50-52], algorithmic fairness forms an important concern when predicting depression risk. We found discrepant model calibration across race, ethnicity, and sex even when controlling

for race, ethnicity, and sex in the model. These results align with previous findings that showed poor calibration for minority groups [53,54] and stress the importance of algorithmic fairness assessment in the depression domain. The differences in calibration may be caused by different (recorded) depression rates among groups. This could result in a disproportionate number of missed patients in need of additional mental health resources in specific groups. For example, female and non-Hispanic Black patients might consistently receive a lower predicted risk score than their actual risk. A next step could be to apply bias mitigation techniques for in- or postprocessing during model development, like threshold selection and recalibration within specific groups [55]. Moreover, more diverse data may be collected to adequately capture the differences in symptomatology between different groups. For example, we may include appetite disturbances that are reported more by women and comorbid alcohol and substance abuse that are reported more by men [50].

We also found discrepant model discrimination across race and ethnicity, with the highest AUROC for the non-Hispanic Black group. These findings diverge from the literature, where the AUROC of the minority groups is usually lower compared to the majority group [56]. However, caution is needed when interpreting this finding, due to the very low number of positive cases in this group (less than 20). More data should be collected to better investigate these differences.

The models based solely on text information (patient emails and clinician notes) performed on par with a random coin toss. This implies that the signal-to-noise ratio in this type of data may be too low to be of prognostic value for this specific use case. This might be particularly true for patient emails, where the frequency of the emails varied widely between patients. However, it is important to note that unstructured text, such as patient emails and clinician notes, can potentially provide valuable information that is not captured in structured data. Therefore, multimodal models that incorporate both structured and unstructured data have the potential to improve clinical predictions. Increasing and regularizing the frequency of digital contact between patient and clinician may aid future research on multimodal models in this field, for example, through digital systems for monitoring patient-reported outcomes [57,58]. Digital communication with the aid of chat robots such as ChatGPT [59] provides further direction to better capture patients' mental health status. This finding also implies that structured data contains strong predictors for depression risk, for example, a history of depression or mental illness, which is well established in the literature and should be considered for future model developments [60-62].

Limitations

This study had limitations. First, we used the ICD codes for depression diagnosis as indicators of depression risk. This provided a clear and detectable label for our outcome event in the EHR. However, not all patients experiencing depression will receive a coded depression diagnosis with a related ICD code as underdiagnosis is a common problem [3,9]. It is possible that depression may have been diagnosed elsewhere and not recorded in our EHR, that depressive symptoms may have existed and not been recorded or ignored by the oncology-focused clinicians, or that the patient did not express their depressive symptoms to their oncology-focused clinician. In addition, some inconsistencies persisted between the ICD-9 and ICD-10 codes, with the ICD-10 codes including depression associated with bipolar disorder. This may have compromised the accuracy of our predictive models in this exploratory study and should be considered for future research.

Moreover, changing the outcome of interest to either antidepressant medication or a referral to a psychiatrist, psychologist, or social worker did not change the accuracy of the predictive models. An explanation might be that patients with depression are often treated with antidepressants by primary care doctors. For antidepressant medication, it is important to note that there may have been overascertainment as this medication is also used to treat more severe and chronic forms of anxiety. This should be considered when interpreting our results and warrants further study.

Second, the modeling approach was focused on a point-of-care solution, meaning we used clinically meaningful end points (eg, 1 month after starting cancer treatment) and used a diverse patient population. Although this provides the potential for broad application across multiple cancer types, the diversity in cancer types and cancer stages might have introduced noise and impacted model performance.

Third, we used cut-off values for clinician notes that were too short or too long to keep the modeling computationally feasible. This may have led to information loss. Future research may investigate ways of retaining this information when preprocessing texts. Finally, we used data from a single integrated health system for model development, albeit comprised of 3 sites (academic hospital, community hospital, and community practice network). As the cultural background of patients and some data are specific to this health system, our results may not generalize to other populations. Further validation on data sets with different demographics and examination of the mechanisms driving potential biases are needed.

Conclusions

This study demonstrated the potential and limitations of using structured and unstructured text data for predicting depression risk in patients with cancer using a variety of ML and multimodal models. After further validation and mitigating biases across subgroups, these models have the potential to improve patient outcomes by alerting clinicians of the possible need to escalate support among this vulnerable patient population. Future studies might improve the prediction of depression risk in patients with cancer by refining the outcome label, expanding the predictors related to mental health, and devoting part of the digital patient communication to mental health aspects.

Acknowledgments

We like to thank Max Schuessler, Vaibhavi Shah, and Angelo Capodici for their help with the annotations of patient emails. Our special thanks go to Dr David Spiegel for reviewing this article for psychiatric relevance and accuracy. This work was funded by the Leids Universiteits Fonds/Slingelands Fonds, the Prins Bernhard Cultuurfonds or Crone-Haver Droeze Fonds, and Fonds Dr Catharine van Tussenbroek. These funders played no role in study design, data collection, analysis, and interpretation of data, or writing the manuscript.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to the protected nature of the patient data. Requests to access these data sets should be directed to boussard@stanford.edu.

Authors' Contributions

AdH, MvB, and THB were responsible for the conceptualization and design of the study. AdH performed the data extraction. AdH and CF performed the data analysis. MR and DB provided clinical advice and recommendations on usability and clinical relevance. AdH drafted the original manuscript. All authors had full access to all the data, critically analyzed, reviewed, contributed, and approved the final manuscript.

Conflicts of Interest

DB reports institutional research funding from BeyondSpring, leadership roles or stock ownership in Artelo and Madora, and personal fees from G1 Therapeutics, Bristol Myers Squibb, Merck & Co Inc, and Eli Lilly and Company all outside the submitted work. THB is a board member and stockholder of Athelo Health, a stockholder at Verantos, Inc, and a consultant for Grai-Matter outside the submitted work. The other authors declare no competing interests.

Multimedia Appendix 1
Tabular metadata appendix.
[XLS File (Microsoft Excel File), 47 KB - medinform_v12i1e51925_app1.xls]

Multimedia Appendix 2
Supplemental methods and results.
[DOCX File , 117 KB - medinform_v12i1e51925_app2.docx]

References

1. Linden W, Vodermaier A, Mackenzie R, Greig D. Anxiety and depression after cancer diagnosis: prevalence rates by cancer type, gender, and age. J Affect Disord 2012;141(2-3):343-351. [doi: 10.1016/j.jad.2012.03.025] [Medline: 22727334]
2. Smith HR. Depression in cancer patients: pathogenesis, implications and treatment (review). Oncol Lett 2015;9(4):1509-1514 [FREE Full text] [doi: 10.3892/ol.2015.2944] [Medline: 25788991]
3. Pitman A, Suleman S, Hyde N, Hodgkiss A. Depression and anxiety in patients with cancer. BMJ 2018;361:k1415 [FREE Full text] [doi: 10.1136/bmj.k1415] [Medline: 29695476]
4. Colleoni M, Mandala M, Peruzzotti G, Robertson C, Bredart A, Goldhirsch A. Depression and degree of acceptance of adjuvant cytotoxic drugs. Lancet 2000;356(9238):1326-1327. [doi: 10.1016/S0140-6736(00)02821-X] [Medline: 11073026]
5. Grassi L, Indelli M, Marzola M, Maestri A, Santini A, Piva E, et al. Depressive symptoms and quality of life in home-care-assisted cancer patients. J Pain Symptom Manage 1996;12(5):300-307 [FREE Full text] [doi: 10.1016/s0885-3924(96)00181-9] [Medline: 8942125]
6. National Survey on Drug Use and Health (NSDUH). Substance Abuse and Mental Health Services Administration. 2020. URL: https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health [accessed 2023-12-15]
7. Walker J, Hansen CH, Martin P, Symeonides S, Ramessur R, Murray G, et al. Prevalence, associations, and adequacy of treatment of major depression in patients with cancer: a cross-sectional analysis of routinely collected clinical data. Lancet Psychiatry 2014;1(5):343-350 [FREE Full text] [doi: 10.1016/S2215-0366(14)70313-X] [Medline: 26360998]
8. Caruso R, Breitbart W. Mental health care in oncology. Contemporary perspective on the psychosocial burden of cancer and evidence-based interventions. Epidemiol Psychiatr Sci 2020;29:e86 [FREE Full text] [doi: 10.1017/S2045796019000866] [Medline: 31915100]
9. Mitchell AJ, Chan M, Bhatti H, Halton M, Grassi L, Johansen C, et al. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. Lancet Oncol 2011;12(2):160-174. [doi: 10.1016/S1470-2045(11)70002-X] [Medline: 21251875]
10. Vinckx MA, Bossuyt I, de Casterlé BD. Understanding the complexity of working under time pressure in oncology nursing: a grounded theory study. Int J Nurs Stud 2018;87:60-68. [doi: 10.1016/j.ijnurstu.2018.07.010] [Medline: 30055374]

11. Dreismann L, Goretzki A, Ginger V, Zimmermann T. What if… I asked cancer patients about psychological distress? barriers in psycho-oncological screening from the perspective of nurses-a qualitative analysis. Front Psychiatry 2021;12:786691 [FREE Full text] [doi: 10.3389/fpsyt.2021.786691] [Medline: 35153856]
12. Söllner W, DeVries A, Steixner E, Lukas P, Sprinzl G, Rumpold G, et al. How successful are oncologists in identifying patient distress, perceived social support, and need for psychosocial counselling? Br J Cancer 2001;84(2):179-185 [FREE Full text] [doi: 10.1054/bjoc.2000.1545] [Medline: 11161373]
13. Steven B, Lange L, Schulz H, Bleich C. Views of psycho-oncologists, physicians, and nurses on cancer care-a qualitative study. PLoS One 2019;14(1):e0210325 [FREE Full text] [doi: 10.1371/journal.pone.0210325] [Medline: 30650112]
14. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019;21(11):116. [doi: 10.1007/s11920-019-1094-0]
15. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep 2021;11(1):1980 [FREE Full text] [doi: 10.1038/s41598-021-81368-4] [Medline: 33479383]
16. Arun V, Prajwal V, Krishna M, Arunkumar BV, Padma SK, Shyam V. A boosted machine learning approach for detection of depression. 2018 Presented at: 2018 IEEE Symposium Series on Computational Intelligence (SSCI); November 18-21, 2018; Bangalore, India. [doi: 10.1109/ssci.2018.8628945]
17. Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF, Aizenstein HJ. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int J Geriatr Psychiatry 2015;30(10):1056-1067 [FREE Full text] [doi: 10.1002/gps.4262] [Medline: 25689482]
18. Aldarwish MM, Ahmad HF. Predicting depression levels using social media posts. 2017 Presented at: 2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS); March 22-24, 2017; Bangkok, Thailand. [doi: 10.1109/isads.2017.41]
19. Deshpande M, Rao V. Depression detection using emotion artificial intelligence. 2017 Presented at: 2017 International Conference on Intelligent Sustainable Systems (ICISS); December 07-08, 2017; Palladam, India. [doi: 10.1109/iss1.2017.8389299]
20. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. J Med Internet Res 2018;20(12):e11817 [FREE Full text] [doi: 10.2196/11817] [Medline: 30522991]
21. Papachristou N, Puschmann D, Barnaghi P, Cooper B, Hu X, Maguire R, et al. Learning from data to predict future symptoms of oncology patients. PLoS One 2018;13(12):e0208808 [FREE Full text] [doi: 10.1371/journal.pone.0208808] [Medline: 30596658]
22. Chen L, Ma X, Zhu N, Xue H, Zeng H, Chen H, et al. Facial expression recognition with machine learning and assessment of distress in patients with cancer. Oncol Nurs Forum 2021;48(1):81-93 [FREE Full text] [doi: 10.1188/21.ONF.81-93] [Medline: 33337433]
23. Sun R, Bozkurt S, Winget M, Cullen MR, Seto T, Hernandez-Boussard T. Characterizing patient flow after an academic hospital merger and acquisition. Am J Manag Care 2021;27(10):e343-e348 [FREE Full text] [doi: 10.37765/ajmc.2021.88764] [Medline: 34668676]
24. Coquet J, Blayney DW, Brooks JD, Hernandez-Boussard T. Association between patient-initiated emails and overall 2-year survival in cancer patients undergoing chemotherapy: evidence from the real-world setting. Cancer Med 2020;9(22):8552-8561 [FREE Full text] [doi: 10.1002/cam4.3483] [Medline: 32986931]
25. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14(1):137 [FREE Full text] [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]
26. Trinh NHT, Youn SJ, Sousa J, Regan S, Bedoya CA, Chang TE, et al. Using electronic medical records to determine the diagnosis of clinical depression. Int J Med Inform 2011;80(7):533-540 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.03.014] [Medline: 21514880]
27. Krebber AMH, Buffart LM, Kleijn G, Riepma IC, de Bree R, Leemans CR, et al. Prevalence of depression in cancer patients: a meta-analysis of diagnostic interviews and self-report instruments. Psychooncology 2014;23(2):121-130 [FREE Full text] [doi: 10.1002/pon.3409] [Medline: 24105788]
28. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383 [FREE Full text] [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
29. Benca RM, Peterson MJ. Insomnia and depression. Sleep Med 2008;9(Suppl 1):S3-S9 [FREE Full text] [doi: 10.1016/S1389-9457(08)70010-8] [Medline: 18929317]
30. Zhang. Improved Adam optimizer for deep neural networks. 2018 Presented at: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS); June 4-6, 2018; Banff, AB, Canada.
31. Lamproudis, Henriksson, Dalianis. Developing a clinical language model for Swedish: continued pretraining of generic BERT with in-domain data. In: Recent Advances in Natural Language Processing. Kerrville, TX: Association for

Computational Linguistics; 2021 Presented at: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021); September 1-3, 2021; Held Online.
32. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019 Presented at: 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019; December 13, 2019; Vancouver, BC, Canada. [doi: 10.1090/mbk/121/79]
33. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Presented at: 2nd Clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL 2019; June 7, 2019; Minneapolis, USA p. 72-78. [doi: 10.18653/v1/w19-1909]
34. Fanconi C, van Buchem M, Hernandez-Boussard T. Natural language processing methods to identify oncology patients at high risk for acute care with clinical notes. AMIA Jt Summits Transl Sci Proc 2023;2023:138-147 [FREE Full text] [Medline: 37350895]
35. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. ICML 2001;1:609-616 [FREE Full text]
36. van Calster B, Nieboer D, Vergouwe Y, de Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167-176 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.12.005] [Medline: 26772608]
37. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health 2022;4(12):e853-e855 [FREE Full text] [doi: 10.1016/S2589-7500(22)00188-1] [Medline: 36270955]
38. Vickers AJ, van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6 [FREE Full text] [doi: 10.1136/bmj.i6] [Medline: 26810254]
39. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
40. Röösli E, Bozkurt S, Hernandez-Boussard T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. Sci Data 2022;9(1):24 [FREE Full text] [doi: 10.1038/s41597-021-01110-7] [Medline: 35075160]
41. Predicting depression for cancer patients. gitlab. URL: https://gitlab.com/a.a.h.de_hond/predicting-depression-for-cancer-patients [accessed 2024-01-04]
42. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (Minimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27(12):2011-2015 [FREE Full text] [doi: 10.1093/jamia/ocaa088] [Medline: 32594179]
43. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med 2000;160(14):2101-2107 [FREE Full text] [doi: 10.1001/archinte.160.14.2101] [Medline: 10904452]
44. Gold SM, Köhler-Forsberg O, Moss-Morris R, Mehnert A, Miranda JJ, Bullinger M, et al. Comorbid depression in medical diseases. Nat Rev Dis Primers 2020;6(1):69 [FREE Full text] [doi: 10.1038/s41572-020-0200-2] [Medline: 32820163]
45. Makubate B, Donnan PT, Dewar JA, Thompson AM, McCowan C. Cohort study of adherence to adjuvant endocrine therapy, breast cancer recurrence and mortality. Br J Cancer 2013;108(7):1515-1524 [FREE Full text] [doi: 10.1038/bjc.2013.116] [Medline: 23519057]
46. Wu EQ, Johnson S, Beaulieu N, Arana M, Bollu V, Guo A, et al. Healthcare resource utilization and costs associated with non-adherence to imatinib treatment in chronic myeloid leukemia patients. Curr Med Res Opin 2010;26(1):61-69 [FREE Full text] [doi: 10.1185/03007990903396469] [Medline: 19905880]
47. Hershman DL, Shao T, Kushi LH, Buono D, Tsai WY, Fehrenbacher L, et al. Early discontinuation and non-adherence to adjuvant hormonal therapy are associated with increased mortality in women with breast cancer. Breast Cancer Res Treat 2011;126(2):529-537 [FREE Full text] [doi: 10.1007/s10549-010-1132-4] [Medline: 20803066]
48. Giese-Davis J, Collie K, Rancourt KMS, Neri E, Kraemer HC, Spiegel D. Decrease in depression symptoms is associated with longer survival in patients with metastatic breast cancer: a secondary analysis. J Clin Oncol 2011;29(4):413-420 [FREE Full text] [doi: 10.1200/JCO.2010.28.4455] [Medline: 21149651]
49. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health 2022;4(12):e853-e855 [FREE Full text] [doi: 10.1016/S2589-7500(22)00188-1] [Medline: 36270955]
50. Altemus M, Sarvaiya N, Epperson CN. Sex differences in anxiety and depression clinical perspectives. Front Neuroendocrinol 2014;35(3):320-330 [FREE Full text] [doi: 10.1016/j.yfrne.2014.05.004] [Medline: 24887405]
51. Barnes DM, Keyes KM, Bates LM. Racial differences in depression in the United States: how do subgroup analyses inform a paradox? Soc Psychiatry Psychiatr Epidemiol 2013;48(12):1941-1949 [FREE Full text] [doi: 10.1007/s00127-013-0718-7] [Medline: 23732705]
52. Hooker K, Phibbs S, Irvin VL, Mendez-Luck CA, Doan LN, Li T, et al. Depression among older adults in the United States by disaggregated race and ethnicity. Gerontologist 2019;59(5):886-891 [FREE Full text] [doi: 10.1093/geront/gny159] [Medline: 30561600]

53. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
54. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. Lancet Digit Health 2021;3(4):e241-e249 [FREE Full text] [doi: 10.1016/S2589-7500(21)00022-4] [Medline: 33766288]
55. Pfohl S, Xu Y, Foryciarz A, Ignatiadis N, Genkins J, Shah N. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. 2022 Presented at: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency; June 21-24, 2022; Seoul Republic of Korea. [doi: 10.1145/3531146.3533166]
56. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
57. Denis F, Basch E, Septans AL, Bennouna J, Urban T, Dueck AC, et al. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. JAMA 2019;321(3):306-307 [FREE Full text] [doi: 10.1001/jama.2018.18085] [Medline: 30667494]
58. Basch E, Stover AM, Schrag D, Chung A, Jansen J, Henson S, et al. Clinical utility and user perceptions of a digital system for electronic patient-reported symptom monitoring during routine cancer care: findings from the PRO-TECT trial. JCO Clin Cancer Inform 2020;4:947-957 [FREE Full text] [doi: 10.1200/CCI.20.00081] [Medline: 33112661]
59. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv Preprint posted online on March 4, 2022. [doi: 10.48550/arXiv.2203.02155]
60. Stafford L, Komiti A, Bousman C, Judd F, Gibson P, Mann GB, et al. Predictors of depression and anxiety symptom trajectories in the 24 months following diagnosis of breast or gynaecologic cancer. Breast 2016;26:100-105 [FREE Full text] [doi: 10.1016/j.breast.2016.01.008] [Medline: 27017248]
61. Fervaha G, Izard JP, Tripp DA, Aghel N, Shayegan B, Klotz L, et al. Psychological morbidity associated with prostate cancer: rates and predictors of depression in the RADICAL PC study. Can Urol Assoc J 2021;15(6):181-186 [FREE Full text] [doi: 10.5489/cuaj.6912] [Medline: 33212008]
62. Wojnarowski C, Firth N, Finegan M, Delgadillo J. Predictors of depression relapse and recurrence after cognitive behavioural therapy: a systematic review and meta-analysis. Behav Cogn Psychother 2019;47(5):514-529 [FREE Full text] [doi: 10.1017/S1352465819000080] [Medline: 30894231]

Abbreviations

AI: artificial intelligence
AMC: academic medical center
AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
CMC: community medical center
EHR: electronic health record
ICD: International Classification of Diseases
LASSO: least absolute shrinkage and selection operator
ML: machine learning
NB: net benefit
NPV: negative predictive value
PPV: positive predictive value
PSC: primary and specialty care alliance

Edited by C Lovis; submitted 17.08.23; peer-reviewed by L Liu, Y Chu; comments to author 03.10.23; revised version received 11.11.23; accepted 08.12.23; published 18.01.24.

Please cite as:
de Hond A, van Buchem M, Fanconi C, Roy M, Blayney D, Kant I, Steyerberg E, Hernandez-Boussard T
Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study
JMIR Med Inform 2024;12:e51925
URL: https://medinform.jmir.org/2024/1/e51925
doi:10.2196/51925
PMID:38236635

Original Paper

Improving Prediction of Survival for Extremely Premature Infants Born at 23 to 29 Weeks Gestational Age in the Neonatal Intensive Care Unit: Development and Evaluation of Machine Learning Models

Angie Li[1], MD; Sarah Mullin[1], PhD; Peter L Elkin[1], MD

Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, United States

Corresponding Author:
Angie Li, MD
Department of Biomedical Informatics
Jacobs School of Medicine and Biomedical Sciences
University at Buffalo
77 Goodell Street
Suite 540
Buffalo, NY, 14203
United States
Phone: 1 716 888 4858
Email: ali83@buffalo.edu

Abstract

Background: Infants born at extremely preterm gestational ages are typically admitted to the neonatal intensive care unit (NICU) after initial resuscitation. The subsequent hospital course can be highly variable, and despite counseling aided by available risk calculators, there are significant challenges with shared decision-making regarding life support and transition to end-of-life care. Improving predictive models can help providers and families navigate these unique challenges.

Objective: Machine learning methods have previously demonstrated added predictive value for determining intensive care unit outcomes, and their use allows consideration of a greater number of factors that potentially influence newborn outcomes, such as maternal characteristics. Machine learning–based models were analyzed for their ability to predict the survival of extremely preterm neonates at initial admission.

Methods: Maternal and newborn information was extracted from the health records of infants born between 23 and 29 weeks of gestation in the Medical Information Mart for Intensive Care III (MIMIC-III) critical care database. Applicable machine learning models predicting survival during the initial NICU admission were developed and compared. The same type of model was also examined using only features that would be available prepartum for the purpose of survival prediction prior to an anticipated preterm birth. Features most correlated with the predicted outcome were determined when possible for each model.

Results: Of included patients, 37 of 459 (8.1%) expired. The resulting random forest model showed higher predictive performance than the frequently used Score for Neonatal Acute Physiology With Perinatal Extension II (SNAPPE-II) NICU model when considering extremely preterm infants of very low birth weight. Several other machine learning models were found to have good performance but did not show a statistically significant difference from previously available models in this study. Feature importance varied by model, and those of greater importance included gestational age; birth weight; initial oxygenation level; elements of the APGAR (appearance, pulse, grimace, activity, and respiration) score; and amount of blood pressure support. Important prepartum features also included maternal age, steroid administration, and the presence of pregnancy complications.

Conclusions: Machine learning methods have the potential to provide robust prediction of survival in the context of extremely preterm births and allow for consideration of additional factors such as maternal clinical and socioeconomic information. Evaluation of larger, more diverse data sets may provide additional clarity on comparative performance.

(JMIR Med Inform 2024;12:e42271) doi:10.2196/42271

KEYWORDS

reproductive informatics; pregnancy complications; premature birth; neonatal mortality; machine learning; clinical decision support; preterm; pediatrics; intensive care unit outcome; health care outcome; survival prediction; maternal health; decision tree model; socioeconomic

Introduction

Preterm birth has long been a leading cause of infant mortality, with the lowest gestational age births associated with the highest rates of mortality [1]. In 2019, 59,506 infants were born at 31 weeks or less in the United States, and the infant mortality rate in this cohort was 18% [2]. When a patient is expected to deliver an extremely preterm infant, counseling on possible outcomes, methods of resuscitation, and anticipated course in the neonatal intensive care unit (NICU) ideally begins prior to birth. Many providers have used the National Institute of Child Health and Human Development (NICHD) risk calculator to initiate this discussion on the chances of infant mortality and severe morbidity after birth. The calculator is based on a logistic regression model using 5 prepartum factors (gestational age, estimated weight, sex, antenatal steroids, and multiple birth), derived from the preterm birth data of a network of US hospitals. With advances in NICU care and more knowledge about long-term outcomes, the calculator was updated in 2020 and maintains a similar performance (mean 0.744, SD 0.005) [3,4]. After initial resuscitation, several scoring systems are also available to predict mortality after a neonate arrives in the NICU [5-7]. However, they are less predictive with extremely low birth weight infants, as evidenced by the Score for Neonatal Acute Physiology With Perinatal Extension II (SNAPPE-II) survival model having a mean performance of 0.78 (SD 0.01) for infants weighing less than 1500 g at birth versus 0.91 (SD 0.01) overall. On review of several models, Clinical Risk Index for Babies (CRIB) had the highest performance in predicting very low birth weight neonate survival, with a mean of 0.88 (SD 0.02), although the CRIB and SNAPPE models were developed with data from geographically separate populations (Europe vs North America) [8].

Despite counseling supported by available risk calculators, decisions surrounding the continuation of life support and redirection to end-of-life care remain extremely difficult in the context of birth at the periviable preterm gestational ages because the postnatal course can be highly variable [9-11]. In addition, perceptions regarding the clinical situation can differ among providers and family members, and consideration of clinical and social context may be helpful [12,13].

Numerous machine learning models have been tested to improve the prediction of adult intensive care unit outcomes. The Medical Information Mart for Intensive Care III (MIMIC-III) database, which contains electronic health record (EHR) information of critical care patients at the Beth Israel Deaconess Medical Center from 2001 to 2012, has often been a source of data used in their development and testing [14-17]. Using the NICU data from MIMIC-III, this study builds and compares different types of machine learning algorithms that predict neonatal mortality and examines the value of incorporating features representing both structured and unstructured clinical elements for extremely preterm infants.

Methods

Ethical Considerations

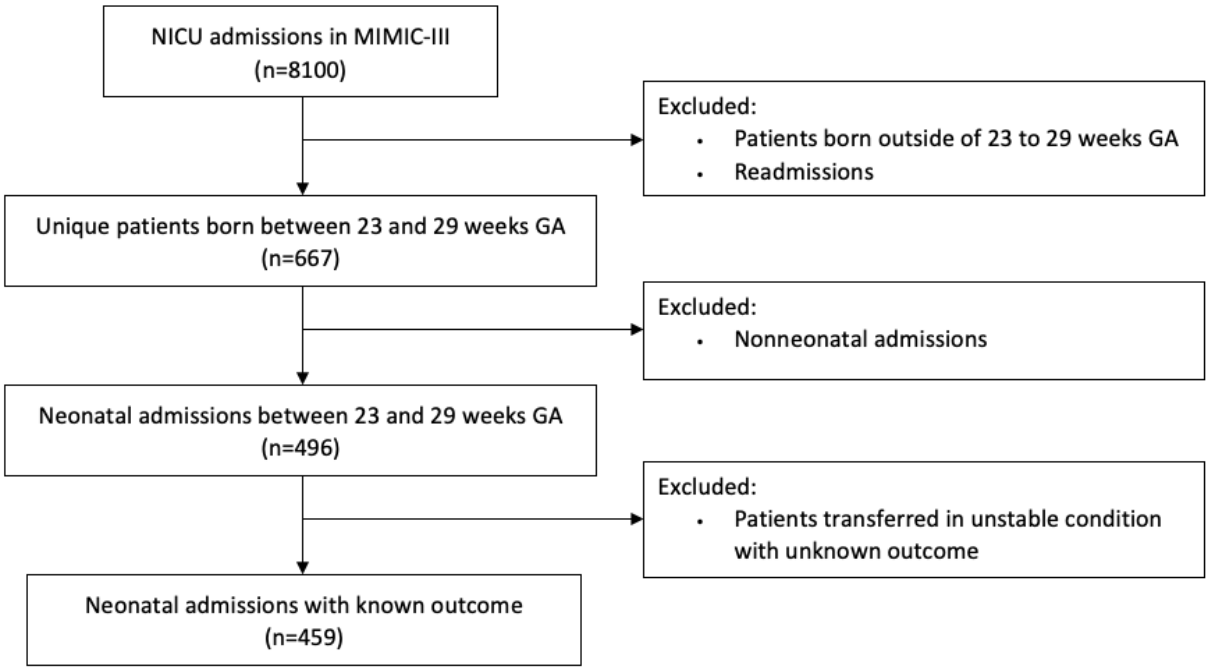
The institutional review board of the University at Buffalo determined the study (ID STUDY00003721) to be exempt as a secondary analysis of a publicly available data set. A data use agreement was obtained for the MIMIC-III database, which contains deidentified protected health information freely available for secondary analysis. The primary data collection for MIMIC-III was originally approved by the institutional review boards of Beth Israel Deaconess Medical Center and Massachusetts Institute of Technology with a waiver of individual patient consent, and no compensation was provided at that time.

Data Selection

Records of extremely preterm neonates admitted to the NICU in the MIMIC-III database were extracted using PostgreSQL (The PostgreSQL Global Development Group). A query was performed for admissions with *ICD-9* (*International Classification of Diseases, Ninth Revision*) codes corresponding to extremely preterm delivery less than 30 weeks as well as very low birth weight. From the resulting records, those of neonates born outside of 23 to 29 weeks were excluded, as well as duplicate records and readmissions. Some records corresponded to nonneonatal admissions, for example, where an infant had a prior history of preterm birth, and they were excluded. When the remaining records were reviewed, it was found that some neonates were transferred outside of the hospital for surgery and had an unknown outcome. These records were also excluded (Figure 1).

From the 459 neonatal admission records that were selected, the patients' demographics, vital signs, laboratory results, medications, procedures, and clinical text were queried from the database and reviewed. Of the available information, relevant elements were extracted based on factors found to be pertinent in previous scoring systems and expert knowledge. By manually curating the clinical text, including completed admission and discharge notes, we were able to incorporate features found only in unstructured form, including maternal clinical comorbidities and pregnancy complications. For this study, consideration of neonatal assessment and treatment was limited to data found initially at the time of NICU admission. The nonnumerical elements were encoded. Data that varied by clinical severity were encoded in that order, and the remaining categorical data underwent binary encoding. Median imputation was used to complete missing data.

Figure 1. Flowchart of selection criteria. GA: gestational age; MIMIC-III: Medical Information Mart for Intensive Care III; NICU: neonatal intensive care unit.



Ultimately, 83 features that could be used in machine learning algorithms were generated, of which approximately half represented maternal clinical and demographic information, with the remaining features representing infant findings at the time of admission (Multimedia Appendix 1).

Model Analysis

Several machine learning classification algorithms were implemented using Python 3.8 scikit-learn 1.2, and the resulting models were tested for their efficacy in predicting mortality. The same algorithms were also examined considering only prepartum features, assuming birth weight would be an estimated weight, to produce models that could be of assistance for clinicians counseling patients prior to an extremely preterm birth.

The performance of each model was endeavored to be optimized. To ensure that feature value range did not drive performance, standard scaling as well as min-max scaling were applied to quantitative features and used for models that were dependent upon distance calculations (eg, logistic regression, neural network, and support vector machine [SVM]). The final reported models used standard scaling due to improved performance over min-max scaling. Scaling was not performed for models invariant to monotonic transformations, such as random forest [18]. For the decision tree–based models, the hyperparameters of number of trees and maximum depth were adjusted. Number of trees began at 50 estimators and was increased by 50 until performance plateaued, which was at 250 trees with a maximum depth of 6 for the random forest method and 350 trees with a maximum depth of 5 for AdaBoost. The k value in the k-nearest neighbor algorithm was adjusted from the default value of 3 up to 20 (approximating the square root of the number of samples), and performance peaked at 4 in the final model. Because of the expected relatively small and imbalanced class sizes (8.1% in the minority class), a held-out test set was not used, and 10-fold stratified cross-validation with an 80:20 training and testing ratio was performed to ensure similar ratios across folds [19]. Mean performance metrics for F_1-score, area under the receiver operating characteristic (AUROC), and average precision are reported, as well as log loss and Brier score, where a smaller value is ideal when considering imbalanced classification.

Features most correlated with the predicted outcome were determined for the higher-performing methods. For the logistic regression model, coefficients most positively and negatively associated with mortality could be determined. For the remaining machine learning models, the most influential features were either directly queried using an available scikit-learn method or through the calculation of feature permutation importance.

Results

Of the included neonatal patients, 37 of 459 (8.1%) expired during the admission period after birth. The average length of stay for infants who survived after initial admission was 62.5 (SD 37.3) days. The average gestational age of the neonates at birth was 27 (SD 1.67) weeks, and 236 (51.4%) were male versus 223 (48.6%) female. Birth weights ranged from 365 to 2165 g, with the average birth weight being 1016 (SD 278) g, and 441 neonates were considered to have a very low birth weight (<1500 g). The average maternal age was 31.4 (SD 6.02) years. In terms of race and ethnicity, the majority of the included infants were in a category considered to be White (n=278, 60.1%), followed by Black (n=69, 15%), unknown (n=42, 9.2%), other (n=25, 5.4%), Hispanic (n=25, 5.4%), Asian (n=16, 3.5%), and Native American (n=4, 0.9%; Table 1).

Table 1. Demographics of patients whose records were included in the study.

	Total (N=459), n (%)	Survived (n=422, 91.9%), n (%)	Expired (n=37, 8.1%), n (%)
Gestational age (weeks)			
23	7 (1.5)	2 (28.6)	5 (71.4)
24	40 (8.7)	28 (70)	12 (30)
25	41 (8.9)	36 (87.8)	5 (12.2)
26	52 (11.3)	49 (94.2)	3 (5.8)
27	87 (19)	84 (96.6)	3 (3.4)
28	106 (23.1)	98 (92.5)	8 (7.5)
29	126 (27.5)	125 (99.2)	1 (0.8)
Sex			
Male	236 (51.4)	214 (90.7)	22 (9.3)
Female	223 (48.6)	208 (93.3)	15 (6.7)
Race			
Asian	16 (3.5)	15 (93.7)	1 (6.3)
Black	69 (15)	62 (89.9)	7 (10.1)
Hispanic	25 (5.4)	23 (92)	2 (8)
Native American	4 (0.9)	3 (75)	1 (25)
White	278 (60.1)	255 (91.7)	23 (8.3)
Other	25 (5.4)	23 (92)	2 (8)
Unknown	42 (9.2)	41 (97.6)	1 (2.4)
Insurance			
Private	343 (74.7)	311 (90.7)	32 (9.3)
Government	116 (25.3)	113 (97.4)	3 (2.6)
Uninsured	2 (0.4)	0 (0)	2 (100)
Family religion			
Catholic	100 (21.8)	91 (91)	9 (9)
Protestant	24 (5.2)	22 (91.7)	2 (8.3)
Jewish	16 (3.5)	15 (93.7)	1 (6.3)
Other	30 (6.5)	25 (83.3)	5 (16.7)
Unknown	289 (63)	269 (93.1)	20 (6.9)
Type of delivery			
Cesarean section	356 (77.6)	331 (93)	25 (7)
Vaginal delivery	103 (22.4)	91 (88.3)	12 (11.7)
Pregnancy type			
Singleton	247 (53.8)	230 (93.1)	17 (6.9)
Multiple	212 (46.2)	192 (90.6)	20 (9.4)
Antenatal steroids			
Received	369 (80.4)	347 (94)	22 (6)
Partially received	71 (15.5)	65 (91.5)	6 (8.5)
Not received	19 (4.1)	14 (73.7)	5 (26.3)

Logistic regression, Naïve Bayes, k-nearest neighbor, SVM, random forest, AdaBoost, and neural network classifiers were compared for efficacy in predicting mortality (Figure 2 and Table 2). Standard scaling transformation improved performance only for the logistic regression, SVM, and neural network methods. The random forest model had the highest predictive

performance when considering overall AUROC (mean 0.91, SD 0.07), F_1-score (0.67), and Brier score (0.06). The AdaBoost model had the next highest AUROC (mean 0.88, SD 0.10); however, the F_1-score (0.45) was low due to poor precision. On the other hand, the neural network model yielded the top F_1-score (0.67) and Brier score (0.05) despite having a lower AUROC (mean 0.84, SD 0.16). SVM was overall next best performing model (mean 0.86, SD 0.13; F_1-score 0.62; Brier score 0.06), followed by logistic regression (mean 0.82, SD 0.16; F_1-score 0.61; Brier score 0.08). The Naïve Bayes (mean 0.74, SD 0.22; F_1-score 0.40; Brier score 0.25) and k-nearest neighbor (mean 0.64, SD 0.13; F_1-score 0.34; Brier score 0.07) methods were the worst performing.

Figure 2. Receiver operating characteristic curves for the highest-performing models in Table 2. A: Logistic regression; B: SVM (support vector machine); C: Random forest; D: AdaBoost; E: Neural networks, F: Naïve Bayes; AUROC: area under the receiver operating characteristic; FP: false positive; TP: true positive.

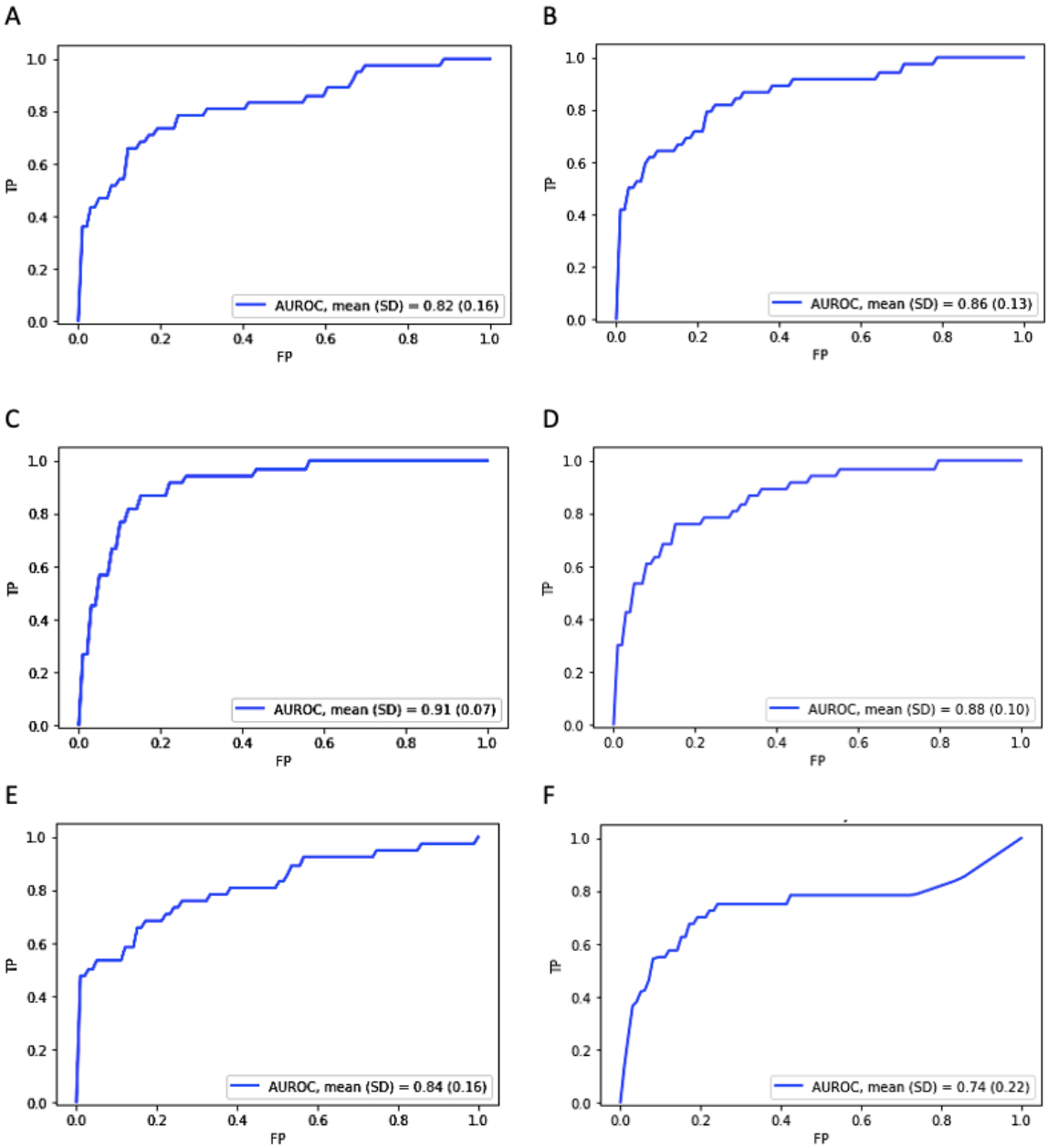


Table 2. AUROC[a], average precision, F1-score, log loss, and Brier scores for 10-fold stratified cross-validation predicting mortality using initial neonatal intensive care unit admission features (lower log loss and Brier scores are ideal when considering imbalanced classification).

Method	AUROC, mean (SD)	Precision, mean (SD)	F_1-score	Log loss score	Brier score
Logistic regression	0.82 (0.16)	0.55 (0.25)	0.61	0.35	0.08
SVM[b]	0.86 (0.13)	0.61 (0.24)	0.62	0.20	0.06
Random forest	0.91 (0.07)	0.61 (0.22)	0.67	0.19	0.06
AdaBoost	0.88 (0.10)	0.55 (0.25)	0.45	0.80	0.07
Neural network	0.84 (0.16)	0.65 (0.24)	0.67	0.30	0.05
Naïve Bayes	0.74 (0.22)	0.39 (0.17)	0.40	3.90	0.25
K-nearest neighbor	0.64 (0.13)	0.24 (0.16)	0.34	1.74	0.07

[a]AUROC: area under the receiver operating characteristic.

[b]SVM: support vector machine.

On post hoc chi-square analysis of the categorical variables, the factors that most influenced the outcome were insurance status, initial breathing assessment of the infant, and presence of a serious fetal anomaly (Table 3). When examining Pearson correlation of continuous variables, higher levels of ventilation and blood pressure support as well as higher arterial blood gas base deficit were properties mildly to moderately correlated with mortality. Larger gestational age, birth weight, and higher APGAR (appearance, pulse, grimace, activity, and respiration) scores at birth negatively correlated with mortality to a similar degree (Table 4).

Similar features were most strongly associated with outcome in the machine learning–based models, although they varied in importance (Table 5). For example, in the random forest model, gestational age, birth weight, and initial oxygen level were of higher importance, whereas in the neural network model, initial blood pressure support and activity level were the most influential features.

Evaluation of classifiers using only prepartum features, assuming birth weight as the estimated weight, also yielded the highest performance measures with the random forest method (Table 6). The random forest features that were consistently of highest importance included gestational age, weight, and maternal age (Table 7).

Table 3. Chi-square: categorical features significantly associated with outcome.

Feature	Description	Chi-square (*df*)
un_ins	Uninsured	22.8 (1)
breathing1	Initial breathing assessment	21.4 (2)
anomaly	Serious fetal anomaly	20.8 (1)
airway1	Initial type of airway or ventilation	17.1 (4)
religion_jehovahs	Religion Jehovah's Witness	11.4 (1)
twintwin	Twin-twin transfusion syndrome	9.4 (1)
uncertain	Uncertain pregnancy dating	7.9 (1)
religion_other	Religion other	4.9 (1)
gov_ins	Medicaid or Medicare insurance	4.7 (1)
muscle1	Muscle tone	4.6 (4)

Table 4. Pearson correlation: correlation of continuous features with mortality.

Feature	Description	Correlation
FiO2_1	Initial amount of oxygen ventilation	0.28
BD1	Initial arterial blood gas base deficit	0.23
dopa1	Initial IV[a] dopamine rate	0.20
temp1	Initial temperature	0.14
pCO2_1	Initial arterial blood gas pCO_2[b]	0.13
G	Maternal gravidity	0.07
P	Maternal parity	0.06
PRBC1	Initial IV blood transfusion amount	0.06
maternal_age	Maternal age	0.05
gluc1	Initial glucose	0.03
bands1	Initial bands	0.03
multiple	Number of fetuses at delivery	0.02
pO2_1	Initial arterial blood gas pO_2[c]	–0.01[d]
wbc1	Initial white blood cells	–0.02
BPmean1	Initial mean blood pressure	–0.04
monos1	Initial monocytes	–0.05
HR1	Initial heart rate	–0.05
hct1	Initial hematocrit	–0.07
neuts1	Initial neutrophil count	–0.07
SaO2_1	Initial oxygen saturation	–0.20
birth_wt	Birth weight	–0.22
GA	Gestational age at birth	–0.32
apgar1	One-minute APGAR[e] score	–0.32
apgar5	Five-minute APGAR score	–0.35

[a]IV: intravenous.

[b]pCO_2: partial pressure of carbon dioxide

[c]pO_2: partial pressure of oxygen.

[d]Negative correlations with mortality imply a correlation with survival.

[e]APGAR: appearance, pulse, grimace, activity, and respiration.

Table 5. Features of highest importance in various models, listed in order of importance. Positive and negative associations with mortality can be calculated only in logistic regression models. For the tree-based random forest and AdaBoost algorithms, an impurity-based method was used to determine overall feature importance. For the remaining algorithms, importance was found via feature permutation[a].

Logistic regression: positively associated with mortality	Logistic regression: negatively associated with mortality	Random forest	AdaBoost	SVM[b]	Neural network
race_hispanic	GA	GA	neuts1	activity1	dopa1
color1	race_unk	birth_wt	hct1	GA	activity1
anomaly	apgar1	SaO2_1	SaO2_1	HTN	multiple
race_asian	gov_ins	BD1	wbc1	anomaly	uncertain
un_ins	activity1	apgar1	apgar1	breathL1	race_unk
dopa1	monos1	gluc1	monos1	breathR1	twintwin
abdomen1	breathL1	dopa1	temp1	twintwin	anomaly
pvt_ins	PRBC1	apgar5	HR1	birth_wt	muscle1
multiple	infert	FiO2_1	FiO2_1	antfont1	wbc1
FiO2_1	dm	neuts	bands1	caprefill1	abdomen1

[a]The descriptions of variable names are present in Multimedia Appendix 1.

[b]SVM: support vector machine.

Table 6. AUROC[a], average precision, F1-score, log loss, and Brier scores for 10-fold stratified cross-validation predicting mortality when only prepartum features are available (lower log loss and Brier scores are ideal when considering imbalanced classification).

Method	AUROC, mean (SD)	Precision, mean (SD)	F_1-score	Log loss score	Brier score
Logistic regression	0.77 (0.14)	0.41 (0.18)	0.51	0.29	0.07
SVM[b]	0.76 (0.10)	0.37 (0.15)	0.46	0.25	0.07
Random forest	0.80 (0.14)	0.54 (0.27)	0.59	0.22	0.06
AdaBoost	0.75 (0.17)	0.44 (0.29)	0.54	0.27	0.07
Neural network	0.76 (0.11)	0.44 (0.18)	0.53	0.31	0.07
Naïve Bayes	0.68 (0.21)	0.30 (0.11)	0.19	6.09	0.59
K-nearest neighbor	0.62 (0.12)	0.20 (0.12)	0.30	1.77	0.09

[a]AUROC: area under the receiver operating characteristic.

[b]SVM: support vector machine.

Table 7. Prepartum features of highest importance in various models, listed in order of importance[a].

Logistic regression: positively associated with mortality	Logistic regression: negatively associated with mortality	Random forest	AdaBoost	SVM[b]	Neural network
maternal_age	GA	GA	birth_wt	GA	un_ins
anomaly	race_unk	birth_wt	maternal_age	steroids	steroids
un_ins	dm	maternal_age	GA	P	HTN
asthma	depression	anomaly	G	infert	GA
religion_jehovahs	PTL	G	multiple	G	anomaly
pvt_ins	steroids	P	religion_unk	uncertain	twintwin
race_hispanic	gov_ins	un_ins	steroids	birth_wt	race_unk
twintwin	HTN	steroids	sex	sex	sex
uncertain	infert	uncertain	anomaly	multiple	P
multiple	P	twintwin	P	anomaly	SVD

[a]The descriptions of variable names are present in Multimedia Appendix 1.

[b]SVM: support vector machine.

Several of the important features found in the top-performing models were among those manually curated in unstructured form, including the presence of maternal hypertensive disease and diabetes, uncertain pregnancy dating (uncertain), fetal anomaly (anomaly), and twin-twin transfusion syndrome.

Discussion

Principal Findings

There is a potential for existing risk calculators to be outperformed by tree-based machine learning algorithms, as indicated by the higher performance of our random forest model versus SNAPPE-II in the context of extremely premature or very low birth weight infants (in fact AUROC increased to mean 0.92, SD 0.05 when only the neonates <1500 g were considered in the random forest model to directly compare to SNAPPE-II). Performance difference compared with CRIB is inconclusive, however. In terms of estimating neonatal mortality prior to preterm birth, although the point estimates of several of the machine learning algorithms using additional features extracted from the EHR were higher than that of the NICHD calculator, overlapping CIs preclude any conclusion about significant differences in performance.

Comparison to Prior Work

Examination of prior work further points to the importance of using data available from the EHR, including unstructured health data. For example, the relatively high-performing CRIB score includes the presence of fetal malformation as a variable. Saria et al [20] incorporated signal processing of short-term time series data from neonatal vital sign sensors to produce a model classifying infants at high risk for severe morbidity or mortality. To maintain accuracy over time, Meadow et al [11] proposed a longitudinal NICU survival model combining adverse events, imaging report information, and caretaker intuition. Hamilton et al [21] more recently applied tree-based machine learning in the context of preterm birth to determine clusters of pregnancy characteristics that were at the highest risk for severe neonatal morbidity or mortality.

Strengths and Limitations

This study is limited by a small data set with data from a single institution, which in turn limits the ability to establish statistical significance in performance differences and the variety of machine learning methods that can be examined. Because of the retrospective nature of the study, there is less control over the format of the data and the amount of missing data. Although a single-institution data set is usually considered a limitation, Rysavy et al [4] emphasized that extremely preterm neonatal outcomes are significantly influenced by the hospital of birth and suggested maintaining ongoing and updated prediction models from outcomes within hospital systems. Using machine learning would be ideal for this task, allowing for consideration of a number of features retrievable from the EHR with a high tolerance for missing or outlier data as the volume of data increases. Tree-based machine learning algorithms may be additionally advantageous due to their ability to iteratively combine numerous weakly predictive features into stronger predictors.

Knowledge of the most influential features, which was possible to visualize in the majority of the presented models, provides transparency. Understanding which factors contribute most to the prediction of outcomes in a model can help clinical providers derive greater intuition regarding how applicable the model is to a particular patient.

The inclusion of maternal information and pregnancy characteristics found in unstructured form in the MIMIC-III database allowed for consideration of factors beyond the numerical neonatal data. Some of these additional variables, such as the presence of fetal anomalies or twin-twin transfusion syndrome, were found to be of high importance in several top-performing models, especially in those used in the prepartum period prior to an anticipated extremely preterm delivery. This illustrates that machine learning–based models could potentially be helpful for continuity of care, starting in the prepartum timeframe with ongoing predictive ability after birth. Maternal demographic information had an influence on mortality prediction in some of the higher-performing models but not others. Although demographic data can provide additional knowledge of social context, unintended bias can also be introduced into the resulting model [22].

Future Directions

Future work anticipates further evaluation of these methods on larger, more diverse data sets to determine if there is a significant and reproducible performance advantage. Expanding the study to include additional data would also allow the evaluation of more powerful machine learning methods such as deep learning methods. Eventually, the maintenance of a more representative and up-to-date cohort for training could potentially be accomplished via collaborative or federated learning techniques across institutions [22,23]. To address the possibility of algorithmic bias, further work could include a comparison of prediction results using models with and without protected demographic features and a calculation of the level of discrimination that could result. Assessment of more data from underrepresented groups may also aid in producing increasingly accurate and less discriminatory models [24,25].

In this study, unstructured information was manually extracted from admission and discharge notes in the MIMIC-III database and allowed for consideration of additional relevant features in our models. This suggests that the use of natural language processing to better understand clinical context may further improve the prediction of outcomes of extremely preterm births. As automated natural language processing of clinical notes becomes more mature and prevalent, the use of these features gleaned from unstructured EHR data will be increasingly applicable [26].

Additional potential future directions include integrating with or adding functionalities found in other intensive care unit models, such as time series modeling, and predicting outcomes other than mortality, such as the development of comorbidities, discharge location, length of stay, and likelihood of readmission.

Conclusions

This study examined machine learning models produced from the MIMIC-III NICU data set and their predictive ability in the

clinically challenging situation of extremely preterm birth. The tree-based random forest model was found to have higher performance than the SNAPPE-II model when predicting the survival of extremely preterm infants of very low birth weight. Several other models, including those using only features that would be known prepartum, also appeared to have good predictive performance but failed to show a statistically significant difference from prior models. Features of highest importance in these models were explored and included traditional variables, such as gestational age and birth weight, but also information that may be found in unstructured form in the EHR. Evaluation of these and even more advanced machine learning methods on larger data sets may offer further clarity about performance differences, and natural language processing techniques would allow for greater use of unstructured clinical information.

Acknowledgments

This work was supported by a National Institutes of Health National Library of Medicine training grant (T15 LM012495-02). Generative artificial intelligence was not used in any portion of the paper writing.

Data Availability

The data sets analyzed in this study are available in the Medical Information Mart for Intensive Care III (MIMIC-III) Clinical Database [14].

Authors' Contributions

AL and PLE conceptualized study methodology. AL and SM participated in data curation, statistical analysis, and writing. All authors reviewed the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Categorical and continuous features.
[DOCX File , 21 KB - medinform_v12i1e42271_app1.docx]

Multimedia Appendix 2
TRIPOD checklist for model development.
[PDF File (Adobe PDF File), 825 KB - medinform_v12i1e42271_app2.pdf]

References

1. Stoll BJ, Hansen NI, Bell EF, Walsh MC, Carlo WA, Shankaran S, et al. Trends in care practices, morbidity, and mortality of extremely preterm neonates, 1993-2012. JAMA 2015;314(10):1039-1051 [FREE Full text] [doi: 10.1001/jama.2015.10244] [Medline: 26348753]
2. Ely D, Driscoll A. Infant mortality in the United States, 2019: data from the period linked birth/infant death file. Natl Vital Stat Rep 2021;70(14):1-17 [FREE Full text] [doi: 10.15620/cdc:111053]
3. Dance A. Survival of the littlest: the long-term impacts of being born extremely early. Nature 2020;582(7810):20-23. [doi: 10.1038/d41586-020-01517-z] [Medline: 32488165]
4. Rysavy MA, Horbar JD, Bell EF, Li L, Greenberg LT, Tyson JE, et al. Assessment of an updated neonatal research network extremely preterm birth outcome model in the Vermont Oxford Network. JAMA Pediatr 2020;174(5):e196294 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.6294] [Medline: 32119065]
5. Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: simplified newborn illness severity and mortality risk scores. J Pediatr 2001;138(1):92-100. [doi: 10.1067/mpd.2001.109608] [Medline: 11148519]
6. Parry G, Tucker J, Tarnow-Mordi W, UK Neonatal Staffing Study Collaborative Group. CRIB II: an update of the clinical risk index for babies score. Lancet 2003;361(9371):1789-1791. [doi: 10.1016/S0140-6736(03)13397-1] [Medline: 12781540]
7. Groenendaal F, de Vos MC, Derks JB, Mulder EJH. Improved SNAPPE-II and CRIB II scores over a 15-year period. J Perinatol 2017;37(5):547-551. [doi: 10.1038/jp.2016.276] [Medline: 28125092]
8. McLeod JS, Menon A, Matusko N, Weiner GM, Gadepalli SK, Barks J, et al. Comparing mortality risk models in VLBW and preterm infants: systematic review and meta-analysis. J Perinatol 2020;40(5):695-703. [doi: 10.1038/s41372-020-0650-0] [Medline: 32203174]
9. Andrews B, Myers P, Lagatta J, Meadow W. A comparison of prenatal and postnatal models to predict outcomes at the border of viability. J Pediatr 2016;173:96-100. [doi: 10.1016/j.jpeds.2016.02.042] [Medline: 26995702]

10. Dupont-Thibodeau A, Barrington KJ, Farlow B, Janvier A. End-of-life decisions for extremely low-gestational-age infants: why simple rules for complicated decisions should be avoided. Semin Perinatol 2014;38(1):31-37. [doi: 10.1053/j.semperi.2013.07.006] [Medline: 24468567]
11. Meadow W, Lagatta J, Andrews B, Lantos J. The mathematics of morality for neonatal resuscitation. Clin Perinatol 2012;39(4):941-956 [FREE Full text] [doi: 10.1016/j.clp.2012.09.013] [Medline: 23164189]
12. Hellmann J, Knighton R, Lee SK, Shah PS, Canadian Neonatal Network End of Life Study Group. Neonatal deaths: prospective exploration of the causes and process of end-of-life decisions. Arch Dis Child Fetal Neonatal Ed 2016;101(2):F102-F107. [doi: 10.1136/archdischild-2015-308425] [Medline: 26253166]
13. Steurer MA, Anderson J, Baer RJ, Oltman S, Franck LS, Kuppermann M, et al. Dynamic outcome prediction in a socio-demographically diverse population-based cohort of extremely preterm neonates. J Perinatol 2017;37(6):709-715. [doi: 10.1038/jp.2017.9] [Medline: 28206998]
14. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
15. Tang F, Xiao C, Wang F, Zhou J. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open 2018;1(1):87-98 [FREE Full text] [doi: 10.1093/jamiaopen/ooy011] [Medline: 31984321]
16. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018;83:112-134 [FREE Full text] [doi: 10.1016/j.jbi.2018.04.007] [Medline: 29879470]
17. Caicedo-Torres W, Gutierrez J. ISeeU: visually interpretable deep learning for mortality prediction inside the ICU. J Biomed Inform 2019;98:103269 [FREE Full text] [doi: 10.1016/j.jbi.2019.103269] [Medline: 31430550]
18. Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 2021;9(3):52 [FREE Full text] [doi: 10.3390/technologies9030052]
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830 [FREE Full text]
20. Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. Sci Transl Med 2010;2(48):48ra65 [FREE Full text] [doi: 10.1126/scitranslmed.3001304] [Medline: 20826840]
21. Hamilton EF, Dyachenko A, Ciampi A, Maurel K, Warrick PA, Garite TJ. Estimating risk of severe neonatal morbidity in preterm births under 32 weeks of gestation. J Matern Fetal Neonatal Med 2020;33(1):73-80. [doi: 10.1080/14767058.2018.1487395] [Medline: 29886760]
22. Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health 2022;1(5):e0000033 [FREE Full text] [doi: 10.1371/journal.pdig.0000033] [Medline: 36812504]
23. Nguyen TV, Dakka MA, Diakiw SM, VerMilyea MD, Perugini M, Hall JMM, et al. A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data. Sci Rep 2022;12(1):8888 [FREE Full text] [doi: 10.1038/s41598-022-12833-x] [Medline: 35614106]
24. Chen IY, Johansson FD, Sontag D. Why is my classifier discriminatory? 2018 Presented at: Proceedings of the 32nd International Conference on Neural Information Processing Systems; December 3-8, 2018; Montréal, Canada p. 3543-3554. [doi: 10.5555/3327144.3327272]
25. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021;4:123-144 [FREE Full text] [doi: 10.1146/annurev-biodatasci-092820-114757] [Medline: 34396058]
26. Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assoc 2022;29(7):1292-1302 [FREE Full text] [doi: 10.1093/jamia/ocac058] [Medline: 35475536]

Abbreviations

APGAR: appearance, pulse, grimace, activity, and respiration
AUROC: area under the receiver operating characteristic
CRIB: Clinical Risk Index for Babies
EHR: electronic health record
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
MIMIC-III: Medical Information Mart for Intensive Care III
NICHD: National Institute of Child Health and Human Development
NICU: neonatal intensive care unit
SNAPPE-II: Score for Neonatal Acute Physiology With Perinatal Extension II
SVM: support vector machine

Edited by C Lovis; submitted 30.08.22; peer-reviewed by M Casal-Guisande, F Meza; comments to author 17.11.22; revised version received 02.02.23; accepted 28.12.23; published 14.02.24.

Please cite as:
Li A, Mullin S, Elkin PL
Improving Prediction of Survival for Extremely Premature Infants Born at 23 to 29 Weeks Gestational Age in the Neonatal Intensive Care Unit: Development and Evaluation of Machine Learning Models
JMIR Med Inform 2024;12:e42271
URL: https://medinform.jmir.org/2024/1/e42271
doi:10.2196/42271
PMID:38354033

Original Paper

Advancing Accuracy in Multimodal Medical Tasks Through Bootstrapped Language-Image Pretraining (BioMedBLIP): Performance Evaluation Study

Usman Naseem[1], PhD; Surendrabikram Thapa[2], MS; Anum Masood[3,4,5], PhD

[1]School of Computing, Macquarie University, Sydney, Australia

[2]Department of Computer Science, Virginia Tech, Blacksburg, VA, United States

[3]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway

[4]Harvard Medical School, Harvard University, Boston, MA, United States

[5]Department of Radiology, Boston Children's Hospital, Boston, MA, United States

Corresponding Author:
Anum Masood, PhD
Department of Circulation and Medical Imaging
Norwegian University of Science and Technology
B4-135, Realfagbygget Building.
Gloshaugen Campus
Trondheim, 7491
Norway
Phone: 47 92093743
Email: anum.msd@gmail.com

Abstract

Background: Medical image analysis, particularly in the context of visual question answering (VQA) and image captioning, is crucial for accurate diagnosis and educational purposes.

Objective: Our study aims to introduce BioMedBLIP models, fine-tuned for VQA tasks using specialized medical data sets such as Radiology Objects in Context and Medical Information Mart for Intensive Care-Chest X-ray, and evaluate their performance in comparison to the state of the art (SOTA) original Bootstrapping Language-Image Pretraining (BLIP) model.

Methods: We present 9 versions of BioMedBLIP across 3 downstream tasks in various data sets. The models are trained on a varying number of epochs. The findings indicate the strong overall performance of our models. We proposed BioMedBLIP for the VQA generation model, VQA classification model, and BioMedBLIP image caption model. We conducted pretraining in BLIP using medical data sets, producing an adapted BLIP model tailored for medical applications.

Results: In VQA generation tasks, BioMedBLIP models outperformed the SOTA on the Semantically-Labeled Knowledge-Enhanced (SLAKE) data set, VQA in Radiology (VQA-RAD), and Image Cross-Language Evaluation Forum data sets. In VQA classification, our models consistently surpassed the SOTA on the SLAKE data set. Our models also showed competitive performance on the VQA-RAD and PathVQA data sets. Similarly, in image captioning tasks, our model beat the SOTA, suggesting the importance of pretraining with medical data sets. Overall, in 20 different data sets and task combinations, our BioMedBLIP excelled in 15 (75%) out of 20 tasks. BioMedBLIP represents a new SOTA in 15 (75%) out of 20 tasks, and our responses were rated higher in all 20 tasks (P<.005) in comparison to SOTA models.

Conclusions: Our BioMedBLIP models show promising performance and suggest that incorporating medical knowledge through pretraining with domain-specific medical data sets helps models achieve higher performance. Our models thus demonstrate their potential to advance medical image analysis, impacting diagnosis, medical education, and research. However, data quality, task-specific variability, computational resources, and ethical considerations should be carefully addressed. In conclusion, our models represent a contribution toward the synergy of artificial intelligence and medicine. We have made BioMedBLIP freely available, which will help in further advancing research in multimodal medical tasks.

(JMIR Med Inform 2024;12:e56627) doi:10.2196/56627

KEYWORDS

biomedical text mining; BioNLP; vision-language pretraining; multimodal models; medical image analysis

Introduction

Background

In recent decades, the fields of data analysis, machine learning, and deep learning have undergone remarkable advancements, with profound implications for various professional domains [1,2]. One of the most promising frontiers for these advancements is medical science, where data-driven models have the potential to bring about significant breakthroughs [3,4]. Medical data predominantly exist in the form of images and textual reports, encompassing x-ray images, medical records, and more. To harness the full potential of these data sources, a visual language model capable of extracting insights from both images and text becomes paramount. Visual language models, which are at the core of this research, represent a fusion of computer vision and natural language processing (NLP). These models possess the capability to understand and generate text-based descriptions for visual content, making them invaluable in contexts where both images and text are essential for comprehensive analysis.

This study explored and adapted visual language models specifically for medical data sets, building upon the foundation laid by existing models. The primary objective was to enhance the performance of these models when confronted with medical data, such as the Radiology Objects in Context (ROCO) [5] and Medical Information Mart for Intensive Care-Chest X-ray (MIMIC-CXR) [6] data sets. This was achieved through a comprehensive process of pretraining on medical data sets and rigorous fine-tuning, with the ultimate goal of determining the optimal model configurations and parameters. This study thus facilitates advancements in health care, contributing to more accurate diagnoses; streamlined medical reporting; and, ultimately, improved patient care. We have made BioMedBLIP models freely available, facilitating the progress of research in diverse medical applications involving multiple modalities [7].

Related Work

In the domain of visual language models and their applications within the medical field, several notable studies and advancements have paved the way for this research project. These works solve different problems within health care analytics and have played critical roles in shaping this study's foundation.

Within the medical domain, image captioning has emerged as a valuable tool that enables health care professionals and researchers to enhance their diagnostic and reporting processes. Image captioning technology allows for the automatic generation of textual descriptions for medical images, such as x-rays, magnetic resonance imaging (MRI) scans, and computed tomography (CT) scans. This capability brings about several significant benefits. First, it aids clinicians in the diagnostic process by providing detailed descriptions of medical images, helping medical professionals to quickly and accurately identify abnormalities or pathologies in the images, thus improving the efficiency and accuracy of diagnoses. Second, image captions serve as a means of clear and standardized communication among health care professionals, reducing the potential for misinterpretation when multiple experts are involved in the diagnostic process. Third, image captions make medical images more accessible to a broader audience, including patients, promoting health literacy and patient engagement. Moreover, in a clinical setting, image captions expedite the process of creating medical reports, improving the overall quality of patient records. They also play a valuable role in medical education and training, aiding in the learning and teaching of medical imaging and diagnostics. Pavlopoulos et al [8] proposed that biomedical image captioning can significantly expedite clinicians' diagnostic processes and presented a comprehensive survey covering various aspects of medical image captioning, including data sets and evaluation measures. Furthermore, the task of the automatic generation of medical image reports, introduced by Jing et al [9], aimed to streamline the reporting process for physicians, enhancing both efficiency and accuracy. To address this, Jing et al [9] used a hierarchical Long Short-Term Memory model, which was tested on 2 publicly available data sets, Indiana University X-Ray [10] and Pathology Education Informational Resource (PEIR)-Gross [9].

This connection between image captioning and medical report generation underscores the practical utility of visual language models in improving health care processes. In addition to these advancements, the field of medical visual question answering (VQA) has gained increasing relevance. Medical VQA tasks involve developing models capable of answering questions related to medical images and bridging the gap between textual queries and visual data. Lin et al [11] introduced various medical data sets and proposed methods to enhance model performance in medical VQA tasks. We used various data sets presented by Lin et al [11] in our experiments. Furthermore, Li et al [12] emphasized the significance of pretraining models on general images to capture meaningful representations of medical data, thus laying the groundwork for our approach. This insight served as our motivation to explore an approach of pretraining models on domain-specific medical data sets, with the aim of achieving enhanced performance for medical VQA (MedVQA) tasks. Notably, Li et al [12] encountered a limitation in their work, as they did not pretrain models using medical data sets. Their decision was influenced by computational resource constraints, and they believed that domain-specific pretraining would be the key to improving model performance in MedVQA tasks. To address this gap in the research landscape, we took the initiative to pretrain our model using medical data sets, thereby bridging the gap between general and medical image understanding.

Transformer models have become instrumental in a diverse array of applications in various vision and language (V+L) tasks, including medical VQA. The Transformer, proposed by Vaswani et al [13], represents a departure from traditional recurrent or convolutional neural networks. Its architecture replaces recurrent layers with a multihead self-attention encoder and decoder

structure. Compared to traditional recurrent neural network models, the Transformer significantly reduces training time, making it a scalable solution capable of handling a wide range of inputs and applicable to diverse V+L tasks, including the analysis of medical images.

Several prominent transformer-based models have had a significant impact on the landscape of NLP and multimodal tasks. One of the most influential models is Bidirectional Encoder Representations From Transformers (BERT). Proposed by Devlin et al [14], BERT has demonstrated its efficacy in a wide variety of NLP tasks. This is achieved through a pretraining phase where 15% of input sequences are masked. These masked tokens can be replaced with random words, original words, or MASK tokens. Subsequently, the transformer auto-encodes these tokens, and fine-tuning is applied to adapt the pretrained model to downstream tasks. The generalization capabilities of BERT are remarkable, making it adept at handling a wide array of semantic tasks. This is due in part to its bidirectional training, which allows the model to learn contextual information from both the left and right sides of a given word.

BERT's versatility allows it to be tailored for various applications, and one domain where it has shown great promise is the biomedical field. In the biomedical domain, text data often exhibit complex language patterns and domain-specific terminology. Lee et al [15] recognized the need for a model that could adapt to these linguistic intricacies and introduced BioBERT, a BERT-based model fine-tuned on biomedical text. BioBERT effectively addresses the word distribution shift from general data to biomedical data, making it a valuable tool for tasks such as biomedical text mining. The BioBERT model's workflow involves transferring weights from BERT, which is pretrained on general domain data, to BioBERT. Subsequently, BioBERT is pretrained on biomedical domain data, followed by fine-tuning and evaluation of various downstream tasks. This adaptation enhances BioBERT's performance in domain-specific tasks, such as biomedical text classification and named entity recognition.

The success of BERT and its adaptations has paved the way for exploring their application in multimodal tasks, where both text and image data are involved. For instance, VisualBERT, proposed by Li et al [16], was inspired by BERT and designed to capture rich semantics in V+L tasks. It uses a stack of Transformer layers and integrates pretrained object proposal systems for image feature extraction. In the training process, VisualBERT uses self-supervised learning with masked word tokens and performs image caption classification tasks with true and false captions. This approach enables the model to capture intricate relationships between text and image content, making it highly suitable for multimodal tasks where textual descriptions are needed for visual content.

Learning Cross-Modality Encoder Representations from Transformers (LXMERT), another notable model, builds upon the success of BERT and its variants [17]. Tan and Bansal [17] recognized the importance of interpreting the semantic meaning of both images and text while exploring the relationships between V+L. LXMERT's encoders, based on the Transformer architecture, are pretrained on large volumes of image-text pairs. The pretraining process, inspired by BERT, includes techniques such as adding random masks. Interestingly, LXMERT's pretraining approach has been found to outperform data augmentation, a common practice used to increase the amount of training data. Consequently, LXMERT is well suited for tasks that involve understanding and generating textual descriptions for visual content, such as image captioning and VQA.

As the field of V+L tasks continues to evolve, so do the transformer-based models designed to tackle them. The Vision Transformer (ViT), introduced by Dosovitskiy et al [18], represents an innovation to address the challenges of applying the Transformer architecture directly to computer vision tasks. ViT operates by dividing an image into 16×16 patches and processing them with position embeddings using a standard Transformer encoder. This approach has shown promise, but it demands substantial computational resources and extensive training data. Notably, ViT32 ViT was used by Eslami et al [19] as part of fine-tuned versions of Contrastive Language-Image Pretraining (CLIP), comparing the performance of different models in the medical domain.

Similarly, Universal Image-Text Representation (UNITER), proposed by Chen et al [20], takes inspiration from the BERT model. The UNITER model has demonstrated strong performance in various V+L tasks. The architecture of UNITER uses the Transformer as its core, with the image and text embedder working in tandem to encode image and text features into a common embedding space. This approach enables the generation of contextual embeddings, facilitating a better understanding of the relationships of V+L. These transformer-based models collectively represent a spectrum of approaches and adaptations within the broader field of V+L tasks.

The landscape of transformer-based models, ranging from BERT to ViT, demonstrates their adaptability and effectiveness in various domains, including NLP and multimodal tasks. CLIP, introduced by Radford et al [21], represents a significant step forward in the multimodal domain. CLIP was designed to connect images and text through a shared embedding space, enabling it to understand the relationship between the 2 modalities. By pretraining on a massive data set containing images and their associated textual descriptions, CLIP can align images with natural language descriptions, making it a versatile tool for a wide range of tasks. This novel approach has significant implications for the medical field, where visual data, such as medical images, often require textual descriptions for comprehensive analysis and interpretation.

Building upon the success of CLIP, PubMedCLIP emerged as a tailored solution for MedVQA. Eslami et al [19] recognized the need for a model specifically adapted to the medical domain and developed PubMedCLIP, a fine-tuned version of CLIP trained on a data set of medical image-text pairs from PubMed articles. This adaptation enables PubMedCLIP to better understand the nuances of medical images and text, resulting in improved performance on MedVQA tasks.

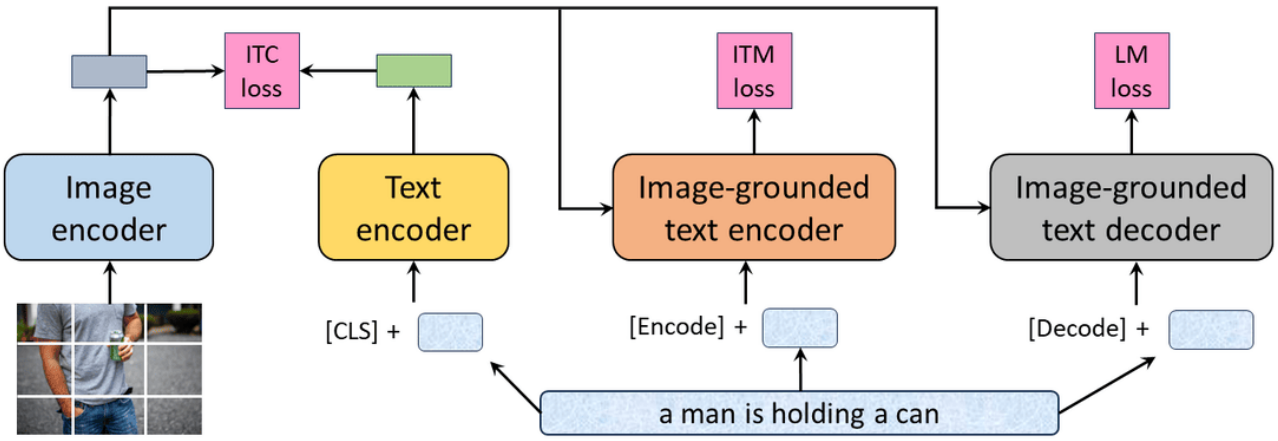
One of the recent works fusing medical imaging and text data is MedBLIP [22]. MedBLIP uses a trainable 3D vision encoder,

MedQFormer, which connects medical images with language models. However, MedBLIP could not significantly improve VQA performance and classification accuracy. For the experimental evaluation of MedBLIP, authors only used MRI scans and text, which will limit the use of MedBLIP in the case of other modalities, such as positron emission tomography, CT, and x-ray images.

Despite performing well in vision-language tasks, CLIP suffers from a number of limitations. Firstly, it is primarily focused on vision-language understanding tasks, such as image retrieval and VQA. This means that it is not well suited for generation tasks, such as image captioning. Secondly, CLIP is trained on a large data set of image-text pairs collected from the web. However, these data are often noisy and contain incorrect or misleading captions. This can lead to CLIP making mistakes when performing tasks that require an accurate understanding of the relationship between images and text.

To address these limitations, Li et al [23] proposed Bootstrapping Language-Image Pretraining (BLIP), a vision-language pretraining framework. BLIP, as shown in Figure 1, is a unified model that can be used for both understanding and generation tasks. This is achieved by incorporating a captioning module into the model, which allows BLIP to generate captions for images. In addition, BLIP addresses the issue of noisy web data by bootstrapping the captions. This means that a captioner generates synthetic captions and a filter removes the noisy ones. This results in a cleaner data set that can be used to train a more robust model. As a result of these improvements, BLIP has been shown to achieve state-of-the-art (SOTA) results on a wide range of vision-language tasks, including image retrieval, VQA, image captioning, and visual grounding. In addition, BLIP is more efficient to train and can be fine-tuned for downstream tasks with lesser data. Finally, BLIP is more interpretable than CLIP, as the captioning module allows users to understand how the model is reasoning about images.

Figure 1. Pretraining architecture of Bootstrapping Language-Image Pretraining (BLIP). CLS: classification; ITC: image-text contrastive; ITM: image-text matching; LM: language modeling.



Capitalizing on the strengths of BLIP, we propose BioMedBLIP by pretraining and fine-tuning the model with a medical data set to achieve SOTA results on medical vision-language tasks. Specifically, BLIP's unified approach to vision-language understanding and generation makes it well suited for tasks such as medical image classification, medical image retrieval, and medical image captioning. In addition, BLIP's ability to handle noisy data makes it well suited for training on medical data sets, which can often be noisy and contain incomplete or inaccurate information. We evaluated our pretrained model using various standard task-specific performance metrics.

Methods

In this section, we describe our pretraining process along with training strategies and resources used.

BioMedBLIP

Overview

BLIP, initially pretrained on general image data sets, possesses knowledge rooted in general image understanding. However, medical images exhibit distinct characteristics that differentiate them from general images. Many medical images are gray scale, such as x-rays and MRI scans, which results in a significant divergence between the general image domain and the medical image domain. To bridge this gap, we conducted the pretraining of BLIP using medical data sets, producing an adapted BLIP model tailored for medical applications.

As shown in Figure 1, BLIP is organized into 4 key modules.

- Visual transformer block (image encoder): the first module serves as an image encoder, using a visual transformer to extract features from medical images.
- BERT-based text encoder: the second module is a text encoder based on BERT. It processes textual data, ensuring a comprehensive understanding of medical texts.
- Cross-attention and binary classification: the third module shares parameters with the text encoder, facilitating joint image-text embeddings through cross-attention. It uses a binary classifier to confirm the pairing of images and text.
- Text decoder: the final module is a text decoder, which shares some components with the preceding encoders, such as feed-forward and cross-attention layers. However, it maintains its own causal self-attention layers. The text

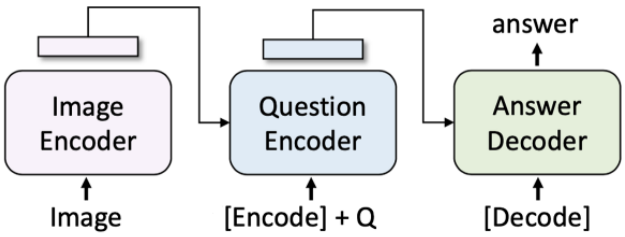
decoder generates text auto-regressively, and cross-entropy loss is applied during this process.

For BLIP, we explored various pretraining approaches. Initially, we attempted to pretrain BLIP from scratch. Subsequently, we pretrained BLIP from a provided checkpoint using the ROCO and MIMIC data sets. Further experimentation involved extending the checkpoint with the inclusion of the ROCO data set onto the MIMIC data set. To apply BLIP to downstream tasks, we followed the BLIP framework's process to refactor the model's modules and assembled an adapted model tailored for specific tasks.

BioMedBLIP VQA Generation Model

For VQA tasks, we adopted the structure provided by BLIP, as depicted in Figure 2. VQA tasks require the model to generate textual answers based on given images and question pairs. The process involves encoding the image to create image embeddings, producing image-question joint embeddings with the help of the question encoder, and using the answer decoder to generate the final answer.

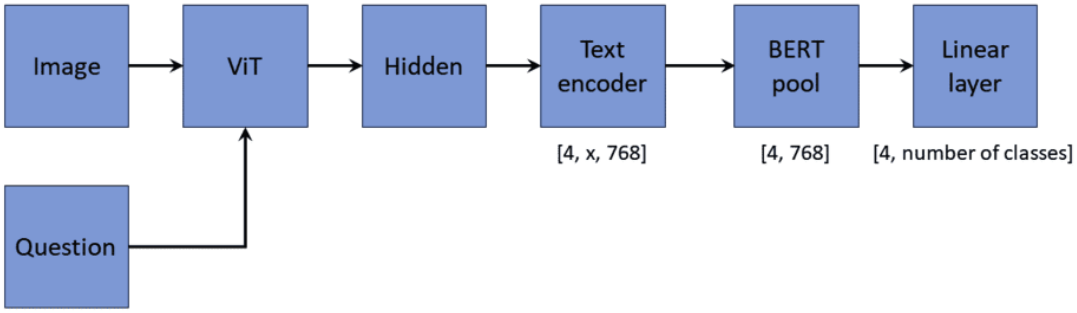
Figure 2. BioMedBLIP visual question answering generation model.



Modified BLIP Classification Model

The modified BLIP classification model, illustrated in Figure 3, shares similarities with the generation model. It generates joint image-text embeddings using the image encoder and text encoder. However, instead of using the answer decoder, a pooling layer is introduced to reduce the vector dimension. Subsequently, a linear classification layer is applied to produce multiple classification results.

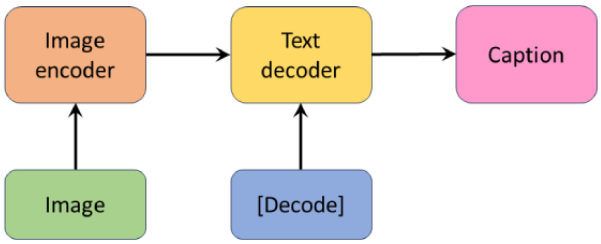
Figure 3. BioMedBLIP classification model. ViT: Vision Transformer.



BioMedBLIP Image Caption Model

The image caption model, presented in Figure 4, is composed of the image encoder and text decoder, following BLIP's implementation. Unlike VQA tasks, the image captioning task involves generating text based solely on images. Therefore, the text encoder is omitted, and the text decoder takes image embeddings provided by the image encoder and the *[Decode]* token as input to produce image captions.

Figure 4. BioMedBLIP image caption model.



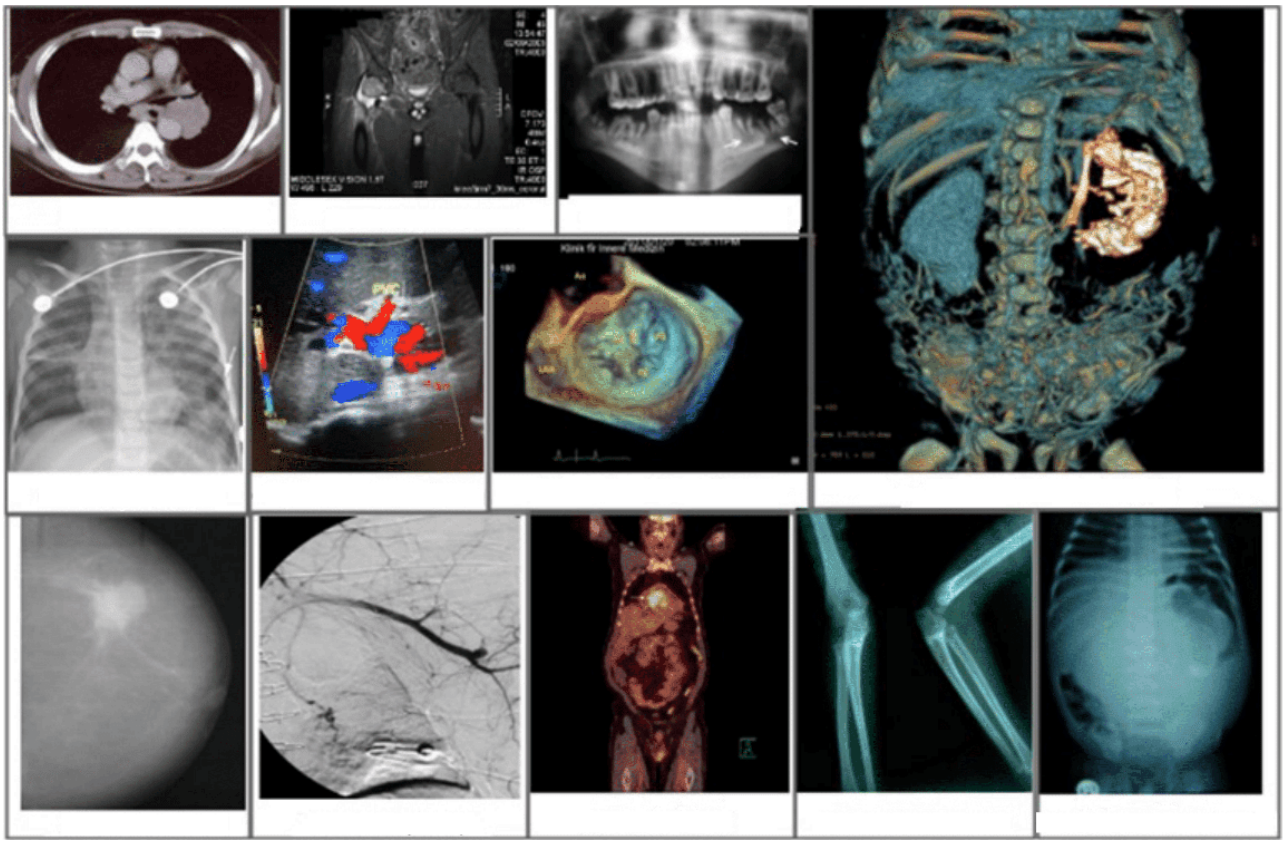
Data Sets

Our research leverages a diverse range of medical data sets, encompassing a variety of visual and textual medical data sources. These data sets serve as the foundation for pretraining and fine-tuning our visual language model.

ROCO Data Set

The ROCO data set [5] plays a pivotal role in pretraining BioMedBLIP. It encompasses over 81,000 radiology images representing multiple medical imaging modalities, including CT, ultrasound, x-ray, fluoroscopy, positron emission tomography, mammography, MRI, and angiography [5]. Our approach involved consolidating the training, validation, and test data into a single comprehensive JSON file, enabling the pretraining of BioMedBLIP. Notably, the captions in the ROCO data set were sourced from peer-reviewed scientific biomedical literature and downloaded from a GitHub link [24]. Some examples are shown in Figure 5.

Figure 5. Some images in the Radiology Objects in Context data set.



MIMIC-CXR Data Set

The MIMIC-CXR data set is a large data set that consists of 377,110 chest x-rays corresponding to 227,827 imaging studies [6]. Some examples of chest x-rays from this data set are shown in Figure 6. In our context, we used this data set for BioMedBLIP's pretraining. It is worth noting that each medical study extracted from the hospital's electronic health record system can be related to multiple chest x-rays. Our efforts focused on filtering the chest x-rays, retaining those with anteroposterior and posteroanterior positions, and ensuring that each medical report had a single associated chest x-ray. After preprocessing, we obtained 218,139 image-caption pairs, which were instrumental in pretraining BioMedBLIP. The medical studies are XML files, and we extracted the "Findings" and "Impressions" sections as the caption for medical images.

Figure 6. Chest x-rays in the Medical Information Mart for Intensive Care-Chest X-ray data set.

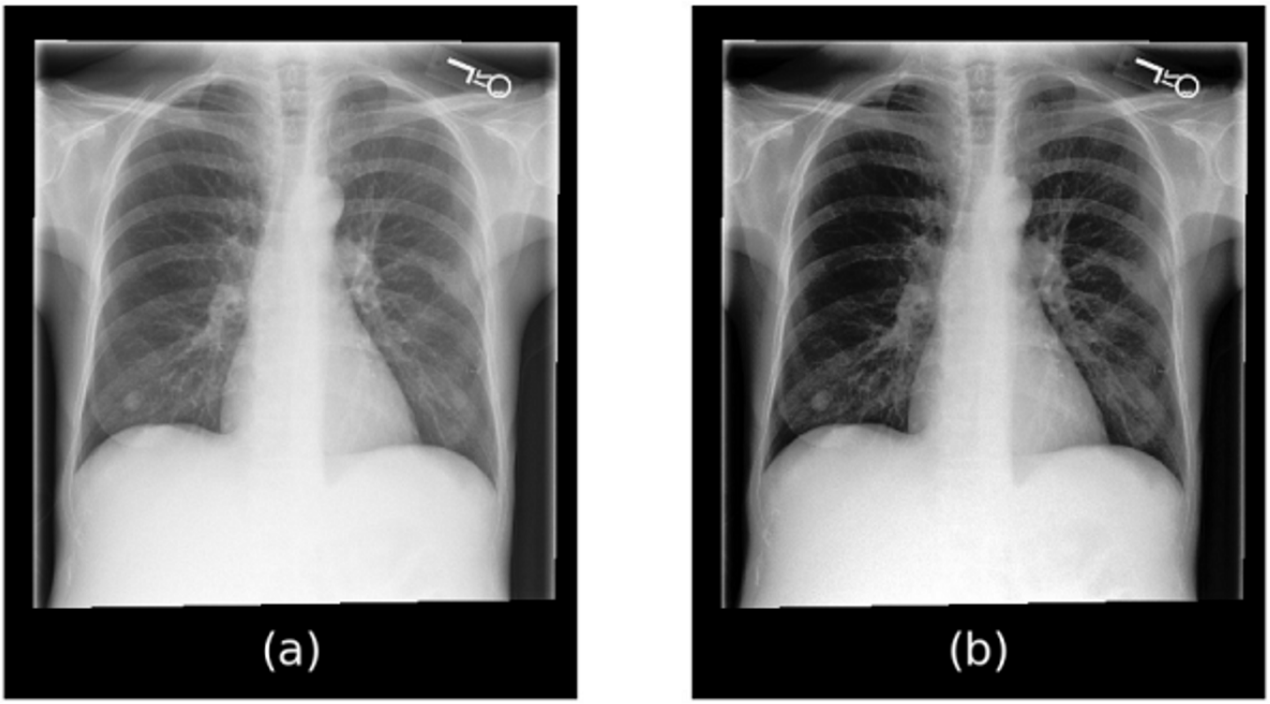
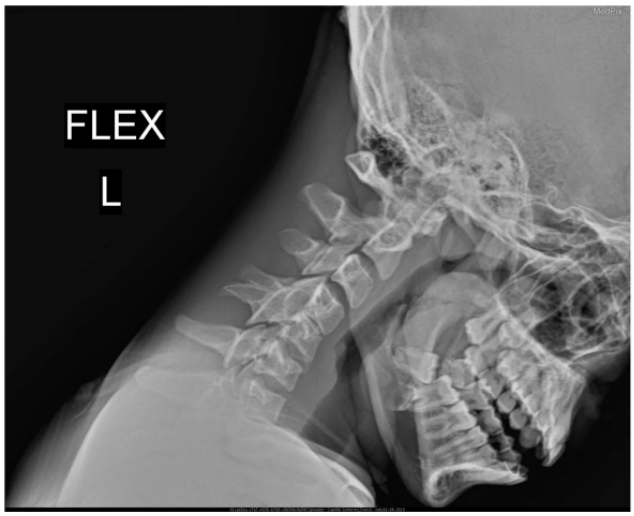


Image Cross-Language Evaluation Forum 2019 Data Set

The Image Cross-Language Evaluation Forum (ImageCLEF) 2019 data set [25,26], provided by the ImageCLEF organization for evaluation, served as a critical component in our work. This data set comes in 3 parts: training, validation, and testing sets. No preprocessing was performed on the data set, and it was leveraged for our VQA generation task. It contains 12,792; 2000; and 500 image-caption pairs for the training, validation, and testing sets, respectively. An example of a radiology image in the ImageCLEF data set is shown in Figure 7.

Figure 7. A radiology image from the Image Cross-Language Evaluation Forum 2019 data set.



Semantically-Labeled Knowledge-Enhanced Data Set

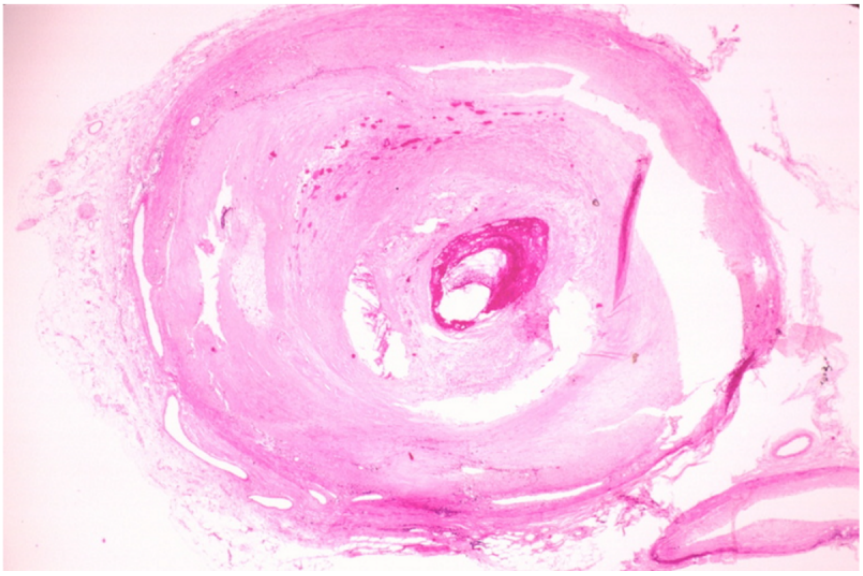
The Semantically-Labeled Knowledge-Enhanced (SLAKE) data set [27] was designed for medical VQA tasks [28]. We used this data set for VQA generation and VQA classification tasks. The SLAKE data set has both Chinese question-answer pairs and English question-answer pairs. Our data set preparation included filtering to retain only English question-answer pairs. After filtering, the SLAKE data set consisted of 4919, 1053, and 1061 image-caption pairs for the training, validation, and testing sets, respectively. Notably, the SLAKE data set features 2 different answer types, open and close, allowing us to assess model performance for open-ended and close-ended questions. An example of a radiology image from the SLAKE data set is shown in Figure 8.

Figure 8. A radiology image from the Semantically-Labeled Knowledge-Enhanced data set.

PathVQA Data Set

The PathVQA data set is a VQA data set for "AI pathologist" development [29]. It contains numerous pathology images together with questions and corresponding answers. All image-question-answer triplets are manually checked to ensure correctness. In our setting, we obtained a data set with 32,795 image-question-answer pairs after initial preprocessing. We further categorized questions into open-ended and close-ended questions using our own splitting, yielding 20,968; 5241; and 6552 image-question-answer pairs for the training, validation, and testing sets, respectively. All these images in the validation and testing sets were picked randomly. An example of a pathology image from the PathVQA data set is shown in Figure 9. We used this data set to perform the VQA generation and VQA classification tasks in our project.

Figure 9. A pathology image from the PathVQA data set.



VQA-RAD Data Set

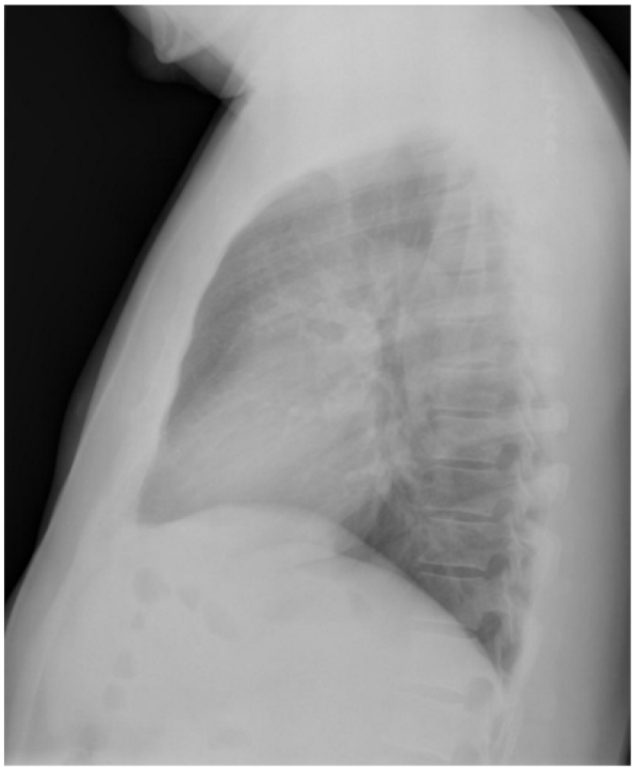
VQA-RAD is the first data set that was manually constructed. During the data collection process, clinicians asked natural questions about radiology images. Meanwhile, their reference answers would be provided [30]. It has radiology images together with question-answer pairs. We used the original data [31] splitting and did not perform any data preprocessing on this data set. It contains 2452, 614, and 452 image-question-answer pairs for the training, validation, and testing sets, respectively. An example of a radiology image from the VQA-RAD data set is shown in Figure 10. We used the VQA-RAD data set to implement the VQA generation and VQA classification tasks.

Figure 10. A radiology image from the Visual Question Answering in Radiology data set.

Open-I Data Set

The Open-I data set [32] is a compilation of chest x-ray images collected from open-source literature and biomedical image collections. Our focus was specifically on the chest x-ray images within this data set. We downloaded the data set from the official Open-I website, which comprises 2 parts: images and medical reports. The medical reports were stored as XML files, with the "Finding" and "Impression" sections extracted as captions for the images. Our downloaded version contained 2452, 614, and 452 image-caption pairs for the training, validation, and testing sets, respectively. An example of a radiology image from the Open-I data set is shown in Figure 11.

Figure 11. A radiology image from the Open-I data set.

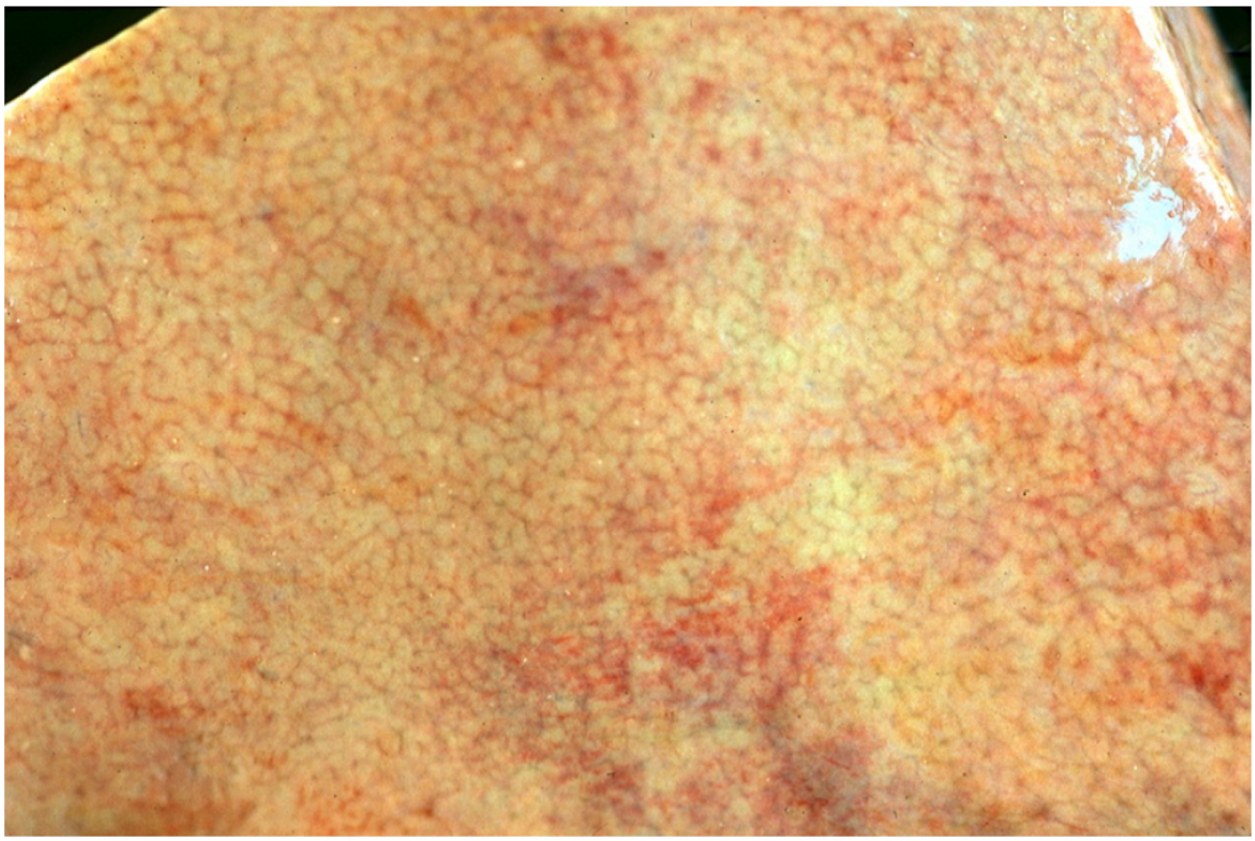


PEIR-Gross Data Set

The PEIR-Gross data set originated from the PEIR and contains 7442 image-caption pairs across 21 subcategories [9]. Our data set preparation involved splitting it into training and testing sets. We also generated a validation set by randomly selecting 10% of the training data. After preprocessing, we had 6029, 669, and 745 image-caption pairs for the training, validation, and testing sets, respectively. An example of a medical image from the

PEIR-Gross data set is shown in Figure 12. This data set was used for image captioning tasks.

Figure 12. A medical image from Pathology Education Informational Resource-Gross data set.



Implementation

Overview

In addition to the original BLIP base model checkpoint obtained from the original BLIP GitHub repository [33], we also meticulously pretrained a series of checkpoints on various medical data sets. These checkpoints serve as critical resources that we are making available for the broader research community and our clients. Here, we provide a comprehensive list of these checkpoints:

- Original BLIP base model checkpoint (ViT-Base): this checkpoint represents the foundational BLIP model pretrained on the LAION115M data set.
- Pretraining on combined-MED checkpoints: these checkpoints were derived from pretraining on a comprehensive data set combining SLAKE, Open-I, ImageCLEF, and PathVQA data sets from scratch, with versions available for both 20 and 50 epochs.
- Pretraining on SLAKE checkpoint: this checkpoint is the result of pretraining on the SLAKE data set from scratch, spanning 20 epochs.
- Pretraining on ROCO checkpoint: pretrained on the ROCO data set from scratch, this checkpoint encapsulates knowledge gained over 10 epochs.
- Pretraining on MIMIC-CXR checkpoint: for up to 10 epochs, this checkpoint embodies the insights obtained from pretraining on the MIMIC-CXR data set from scratch.
- Pretraining on ROCO from the BLIP original checkpoint: this checkpoint extends the pretraining on the ROCO data set from the existing BLIP checkpoint up to 50 epochs.
- Pretraining on MIMIC-CXR from the BLIP original checkpoint: similar to the ROCO extension, this checkpoint involves the pretraining on the MIMIC-CXR data set from the original BLIP checkpoint and spans up to 50 epochs.
- Pretraining on ROCO and MIMIC-CXR from the BLIP original checkpoint: this checkpoint represents an amalgamation of knowledge acquired from both ROCO and MIMIC-CXR data sets, building upon the original BLIP checkpoint and extending to 50 epochs.

Pretraining Details

To undertake the pretraining of the BLIP model, our approach involved several key steps. Initially, we used the ROCO data set for the pretraining of the BLIP original checkpoint, resulting in the creation of the BioMedBLIP-ROCO models. These models were uniquely identified by the number of epochs they were pretrained for, with “BioMedBLIP-ROCO-10,” for instance, signifying the original BLIP checkpoint pretrained on the ROCO data set for 10 epochs. This choice was informed by the fact that the original BLIP checkpoint had been trained on millions of standard images and, therefore, already possessed the fundamental knowledge required for a visual language model. Our attempt to pretrain BLIP from scratch yielded unsatisfactory performance. Subsequently, we embarked on pretraining the original BLIP checkpoint with the MIMIC-CXR data set to produce the BLIP-MIMIC models. Finally, we took

the BLIP-MIMIC checkpoints and further pretrained them using the ROCO data set, resulting in the creation of the BioMedBLIP-ROCO and MIMIC models. In this study, we retained checkpoints representing 10, 20, and 50 epochs for each of these models.

To optimize the pretraining process, we conducted a series of experiments aimed at identifying the most suitable hyperparameters. The selected hyperparameters for the pretraining process were as follows: initial learning rate: $3e^{-5}$; warmup learning rate: $1e^{-6}$; warmup steps: 3000; and optimizer: AdamW.

Fine-Tuning Details

The fine-tuning phase involved an extensive process encompassing various BLIP checkpoints and data sets, which included the BLIP original base model checkpoint, BioMedBLIP-ROCO-10, BioMedBLIP-ROCO-20, BioMedBLIP-ROCO-50, BioMedBLIP-MIMIC-10, BioMedBLIP-MIMIC-20, BioMedBLIP-MIMIC-50, BioMedBLIP-ROCO&MIMIC-10, BioMedBLIP-ROCO&MIMIC-20, and BioMedBLIP-ROCO&MIMIC-50. These checkpoints were fine-tuned on a selection of data sets, namely, ImageCLEF, SLAKE, VQA-RAD, PathVQA, Open-I (Indiana University X-RAY), and PEIR-Gross.

For the fine-tuning process, we adhered to the common practice of splitting the data sets into training, validation, and test sets. Typically, the training set constitutes 80% of the total data set, while the test set encompasses the remaining 20%. In certain cases, the data set authors had already performed the necessary data set splits, and we made no further modifications to these sets. In addition, we created an answer list to facilitate the evaluation of predicted VQA generation sentences, VQA classification labels, and image captions.

As part of our methodology, we used a YAML configuration file, which proved essential for adapting to different running environments. Through a series of meticulously designed experiments, we optimized the hyperparameters for each checkpoint's fine-tuning process. These hyperparameters included train batch size, test batch size, learning rates, and the number of epochs. Importantly, the optimal hyperparameters varied for each checkpoint when applied to different data sets, ensuring the fine-tuning process was meticulously tailored for each specific scenario.

Resources

This study relied on a combination of hardware and software resources to execute efficiently. We used 3 primary platforms for code execution: Google Colab (Google Inc), Google Cloud Platform (Google Inc), and the University of Sydney's Artemis HPC supercomputer. Google Colab we used had an Intel(R) Xeon(R) central processing unit (CPU) running at 2.30 GHz, an Nvidia Tesla P100-PCIE-16GB (Nvidia Corp) graphics processing unit (GPU), and 12.8 GB of RAM. The version of the Google Cloud platform we used had 4 CPU cores, a Tesla A100-PCIE-40GB GPU, and 26 GB of RAM. On the Artemis HPC platform [34], we had access to an impressive array of resources, including 7636 CPU cores, 45 TB of RAM, 108 NVIDIA V100 GPUs, 378 TB of storage, and 56 Gbps FDR InfiniBand networking. For our pretraining tasks, we specifically used 4 CPU cores, Tesla V100-PCIE GPUs, and 48 GB of RAM. Our code is available on HuggingFace [7].

Ethical Considerations

In our study, ethical considerations were meticulously observed to ensure compliance with relevant regulations and standards. We did not require ethics approval for this research because the data sets used, including ROCO, MIMIC-CXR, SLAKE, PathVQA, VQA-RAD, ImageCLEF, Open-I, and PEIR-Gross, are publicly available. These data sets have been previously collected, anonymized, and made accessible for research purposes by their respective organizations and custodians. No user data or human data were collected directly by us for this study. We strictly adhered to data use policies specified by the data providers, ensuring that all data handling was performed in accordance with ethical guidelines. In addition, the data sets used did not contain personally identifiable information, and appropriate measures were taken to maintain data privacy and security throughout the research process. Using publicly available data, we ensured that our research complied with institutional and local policies regarding the use of medical data for research purposes.

Results

Overview

In this section, we describe our evaluation metrics along with the results and findings. The purpose of our experimentation was to evaluate how adding domain-specific information helps in improving the performance of downstream tasks in the biomedical domain. For this reason, we have compared the results of our models with the original BLIP model.

Evaluation Metrics

In this subsection, we provide a detailed overview of the specific evaluation metrics used to assess the performance of our model across various downstream tasks.

- VQA generation: for VQA generation, we used exact match (EM) as the metric. EM assesses the model's performance by treating predictions that precisely match the ground truth as correct answers. It is particularly relevant for evaluating generative tasks.
- VQA classification metrics: for VQA classification, we used accuracy. Accuracy is a fundamental metric for classification tasks, quantifying the proportion of correctly classified instances.
- Report generation or image captioning metrics: Bilingual Evaluation Understudy (BLEU) is a metric that assesses the similarity between the generated answer and the reference answer, considering n-grams. BLEU-1, in particular, focuses on 1-grams.

VQA Generation Results

Overview

For the VQA eneration task, we used the SLAKE, VQA-RAD, PathVQA, and ImageCLEF data sets to check the performance of our models BLIP-Original, BioMedBLIP-ROCO, and BioMedBLIP-MIMIC&ROCO. These models were designed to generate answers for visual questions and underwent different pretraining strategies. Tables 1 and 2 present a comparison of BioMedBLIP models with the SOTA model.

Table 1. Comparison of BioMedBLIP models versus the SOTA[a] on VQA[b] generation tasks (part 1).

Data set	Original BLIP[c] (SOTA; accuracy)	BioMedBLIP models (accuracy)			
		ROCO[d]-10	ROCO-20	ROCO-30	MIMIC[e]-10
SLAKE[f]-Overall	77.95	80.87	80.11	80.21	78.51
SLAKE-Open	73.80	75.81	74.57	75.04	73.80
SLAKE-Close	87.26	88.70	88.70	89.42	85.82
VQA-RAD[g]-Overall	34.37	35.70	*37.03* [h]	35.03	26.16
VQA-RAD-Open	39.66	43.02	*46.37*	43.02	26.82
VQA-RAD-Close	30.88	30.51	30.88	29.78	25.74
PathVQA-Overall	*66.64*	63.00	64.65	55.46	51.45
PathVQA-Open	*43.78*	38.45	40.79	23.84	18.26
PathVQA-Close	88.35	87.62	*88.57*	87.16	84.75
ImageCLEF[i]	48.20	*58.27*	56.81	57.63	56.41

[a]SOTA: state of the art.

[b]VQA: visual question answering.

[c]BLIP: Bootstrapping Language-Image Pretraining.

[d]ROCO: Radiology Objects in Context.

[e]MIMIC: Medical Information Mart for Intensive Care.

[f]SLAKE: Semantically-Labeled Knowledge-Enhanced.

[g]VQA-RAD: Visual Question Answering in Radiology.

[h]Best performing models are italicized.

[i]ImageCLEF: Image Cross-Language Evaluation Forum.

Table 2. Comparison of BioMedBLIP models versus the SOTA[a] on VQA[b] generation tasks (part 2).

Data set	BioMedBLIP models (accuracy)				
	MIMIC[c]-20	MIMIC-50	MIMIC&ROCO-10	MIMIC&ROCO-20	MIMIC&ROCO-50
SLAKE[d]-Overall	79.92	70.50	*82.00* [e]	81.53	81.34
SLAKE-Open	75.50	65.12	76.28	76.28	*76.59*
SLAKE-Close	86.78	78.85	*90.87*	76.28	88.70
VQA-RAD[f]-Overall	30.38	25.50	32.59	29.93	35.70
VQA-RAD-Open	36.87	22.35	32.96	29.05	40.78
VQA-RAD-Close	26.10	27.57	*32.35*	30.51	*32.35*
PathVQA-Overall	53.31	50.89	61.74	54.09	60.27
PathVQA-Open	20.79	17.71	36.22	22.62	33.69
PathVQA-Close	85.91	84.14	87.32	85.64	86.92
ImageCLEF[g]	52.27	53.63	19.89	54.80	56.42

[a]SOTA: state of the art.

[b]VQA: visual question answering.

[c]MIMIC: Medical Information Mart for Intensive Care.

[d]SLAKE: Semantically-Labeled Knowledge-Enhanced.

[e]Best performing models are italicized.

[f]VQA-RAD: Visual Question Answering in Radiology.

[g]ImageCLEF: Image Cross-Language Evaluation Forum.

Results on the SLAKE Data Set

For the overall SLAKE data set, the BioMedBLIP MIMIC&ROCO-10 model exhibited the highest EM accuracy, reaching an impressive 82%, outperforming the BLIP original model and other variants. The results in SLAKE-Open highlighted the superiority of the BioMedBLIP MIMIC&ROCO-50 model, which achieved the best performance with an EM accuracy of 76.59%. Finally, in the SLAKE-Close category, the BioMedBLIP MIMIC&ROCO-10 model stood out with an impressive accuracy of 90.87%, demonstrating its strong performance in generating answers that match ground truth answers exactly.

Results on the VQA-RAD Data Set

The original BLIP (SOTA) model, serving as the baseline, achieved an EM score of 34.37 in the "VQA-RAD-Overall" category. However, it was surpassed by our BioMedBLIP model, specifically "ROCO-20," which demonstrated exceptional performance with an EM score of 37.03, indicating its effectiveness in generating accurate answers to visual questions. This trend continued in the "VQA-RAD-Open" data set, where BioMedBLIP-ROCO-20 outperformed the baseline with an EM score of 46.37, highlighting its strong performance in open-ended VQA tasks. Notably, the close-ended category also saw success for the BioMedBLIP models, with "MIMIC&ROCO-10" and "MIMIC&ROCO-50" achieving the top EM score of 32.35.

Results on the PathVQA Data Set

In the evaluation of VQA generation tasks on the PathVQA data set, the original BLIP (SOTA) model exhibited an EM score of 66.64, setting a high standard. Among the BioMedBLIP models, "ROCO-20" emerged as the top performer in the PathVQA-Overall and PathVQA-Open categories, achieving scores of 64.65 and 40.79, respectively. Particularly noteworthy is "MIMIC&ROCO-10," which achieved a competitive EM score of 61.74 in the PathVQA-Overall task. In the PathVQA-Close category, "BioMedBLIP-ROCO-20" stood out with an EM score of 88.57, surpassing the original BLIP (SOTA) model. The BioMedBLIP models "MIMIC&ROCO-10" and "MIMIC&ROCO-50" also displayed a strong performance. These results emphasize the effectiveness of different pretraining strategies and the potential for improved performance in VQA generation tasks using the PathVQA data set with BioMedBLIP.

Results on the ImageCLEF Data Set

The SOTA original BLIP model achieved an EM score of 48.20. In contrast, the BioMedBLIP models, which were pretrained with different data sets and epochs, demonstrated notable improvements. Notably, the BioMedBLIP model pretrained with ROCO (10 epochs) emerged as the top performer with an impressive EM score of 58.27, surpassing the SOTA model. This suggests that the use of the ROCO data set for pretraining significantly enhances the ability of the model to generate precise answers to visual questions on the ImageCLEF data set. Other variants of the BioMedBLIP model, pretrained with different data sets and epochs, also exhibited varying degrees of success in this task.

VQA Classification Results

Overview

For the VQA classification task, we used the SLAKE, VQA-RAD, and PathVQA data sets to check the performance

of our models BLIP-Original, BioMedBLIP-ROCO, and BioMedBLIP-MIMIC-CXR. These models were designed to generate answers for visual questions and underwent different pretraining strategies. Tables 3 and 4 present a comparison of the BioMedBLIP models with the SOTA models.

Table 3. Comparison of BioMedBLIP models versus the SOTA[a] on VQA[b] classification tasks (part 1).

Data set	Original BLIP[c] (SOTA; accuracy)	BioMedBLIP models (accuracy)			
		ROCO[d]-10	ROCO-20	ROCO-30	MIMIC[e]-10
SLAKE[f]-Overall	77.85	*81.06* [g]	80.21	80.04	78.70
SLAKE-Open	75.50	75.66	77.05	77.52	75.66
SLAKE-Close	81.49	83.31	*85.10*	84.86	83.41
VQA-RAD[h]-Overall	*40.35*	33.70	19.96	23.95	34.36
VQA-RAD-Open	20.67	27.37	25.10	26.33	*28.49*
VQA-RAD-Close	*51.84*	39.71	33.09	39.71	38.23
PathVQA-Overall	60.09	59.25	57.77	58.65	58.85
PathVQA-Open	*37.21*	33.60	30.06	33.17	34.05
PathVQA-Close	85.15	84.96	85.54	84.20	83.71

[a]SOTA: state of the art.
[b]VQA: visual question answering.
[c]BLIP: Bootstrapping Language-Image Pretraining.
[d]ROCO: Radiology Objects in Context.
[e]MIMIC: Medical Information Mart for Intensive Care.
[f]SLAKE: Semantically-Labeled Knowledge-Enhanced.
[g]Best performing models are italicized.
[h]VQA-RAD: Visual Question Answering in Radiology.

Table 4. Comparison of BioMedBLIP models versus the SOTA[a] on VQA[b] classification tasks (part 2).

Data set	BioMedBLIP models (accuracy)				
	MIMIC[c]-20	MIMIC-50	MIMIC&ROCO[d]-10	MIMIC&ROCO-20	MIMIC&ROCO-50
SLAKE[e]-Overall	77.57	74.18	73.90	80.89	69.10
SLAKE-Open	74.88	72.25	71.62	*77.90* [f]	71.32
SLAKE-Close	81.73	77.16	76.69	84.77	68.04
VQA-RAD[g]-Overall	33.70	34.15	34.59	29.49	31.49
VQA-RAD-Open	*28.49*	*28.49*	*28.49*	26.26	27.93
VQA-RAD-Close	37.13	37.87	38.69	31.62	33.82
PathVQA-Overall	58.04	36.86	60.41	60.24	*61.13*
PathVQA-Open	31.98	18.96	32.61	33.32	34.09
PathVQA-Close	84.17	54.80	59.70	85.38	*86.29*

[a]SOTA: state of the art.
[b]VQA: visual question answering.
[c]MIMIC: Medical Information Mart for Intensive Care.
[d]ROCO: Radiology Objects in Context.
[e]SLAKE: Semantically-Labeled Knowledge-Enhanced.
[f]Best performing models are italicized.
[g]VQA-RAD: Visual Question Answering in Radiology.

Results on the SLAKE Data Set

In the context of VQA classification tasks using the SLAKE data set, the evaluation results are presented in terms of accuracy, allowing for a comprehensive comparison between the original BLIP model and several BioMedBLIP variants, each pretrained with specific data sets and epochs. The BioMedBLIP models demonstrated their potential for significant improvements over the original BLIP model. In the "SLAKE-Overall" data set category, BioMedBLIP models pretrained with the ROCO data set consistently outperformed the original BLIP model, with ROCO-10 achieving an accuracy of 81.06. Furthermore, BioMedBLIP models pretrained with the MIMIC-CXR data set, particularly MIMIC-20, showcased strong performance. For the SLAKE open-ended data set subtask, the model pretrained with both MIMIC-CXR and ROCO for 20 epochs achieved an accuracy of 77.90, surpassing the original BLIP model. For the SLAKE close-ended data set, the ROCO-20 variant of BioMedBLIP stood out with an accuracy of 85.10, clearly surpassing the original BLIP model's performance. These results emphasize the significance of data set choice and pretraining duration, with BioMedBLIP models showcasing their potential for improved accuracy in classifying visual questions within the SLAKE data set.

Results on the VQA-RAD Data Set

In the VQA-RAD-Overall data set, encompassing both open-ended and closed-ended questions, none of the BioMedBLIP models outperformed the original BLIP model, highlighting the challenges in achieving superior accuracy in a mixed question type. However, in the VQA-RAD open-ended data set, all the BioMedBLIP models excelled, surpassing the original BLIP model's accuracy and showcasing their effectiveness in open-ended question answering. Four variants of BioMedBLIP models, namely, MIMIC-10, MIMIC-20, MIMIC-50, and MIMIC&ROCO-10, had the highest score of 28.49. For the VQA-RAD close-ended data set, the BioMedBLIP models did not surpass the original BLIP model, with the best performers being ROCO-10 and ROCO-30. This shows that further research is warranted on strategies to improve the performance on the close-ended VQA-RAD data set.

Results on the PathVQA Data Set

On the PathVQA-Overall data set, the original BLIP model achieved an accuracy of 60.09, while the BioMedBLIP models demonstrated varied performance. Notably, the model pretrained with both MIMIC-CXR and ROCO data sets for 50 epochs emerged as the top performer, achieving an accuracy of 61.13. This dual pretraining approach showed significant promise in enhancing classification accuracy. For the PathVQA-Open data set, the original BLIP model achieved the highest accuracy of 37.21, with BioMedBLIP models pretrained on ROCO and MIMIC-CXR data yielding slightly lower results. The highest performance among BioMedBLIP models came from the model pretrained with both data sets for 50 epochs, reaching an accuracy of 34.09, which is very close to the original BLIP model's accuracy. In contrast, the original BLIP model excelled on the PathVQA-Close data set, achieving an impressive accuracy of 85.15. Nevertheless, the BioMedBLIP model pretrained with both MIMIC-CXR and ROCO for 50 epochs outperformed the original BLIP model, achieving an accuracy of 86.29.

These results collectively indicate that the choice of pretraining strategy and the duration of training significantly influence the classification accuracy of BioMedBLIP models on the various VQA classification data sets, with the combined data set pretraining demonstrating notable advantages in certain contexts.

Image Captioning Task Results

We used the PEIR-Gross data set for the image captioning task to check the performance of our BioMedBLIP models. For the image captioning task, BLIP-original had a BLEU-1 score of 24.8. Similarly, when using the ROCO data set for training, the BioMedBLIP models had scores of 24.4, 24.6, and 25.1 for 10, 20, and 30 epochs respectively. Furthermore, when the MIMIC data set was used, the BLEU-1 score of 23.1, 23.9, and 24.1 was achieved by BioMedBLIP models when trained for 10, 20, and 50 epochs respectively. Additionally, when the training data combined MIMIC and ROCO, the BLEU-1 score of 23.9, 24.3, and 24.2 was achieved for 10, 20, and 30 epochs respectively by BioMedBLIP models. In terms of BLEU-1 scores, higher values are indicative of better performance. The results show that the BioMedBLIP-ROCO-50 model surpassed the original BLIP model in the image captioning task with a BLEU-1 score of 25.1 (over 1.2% improvement from the score of the original BLIP model), demonstrating that our approach has the potential to enhance the model's capabilities for generating captions. In contrast, all other models, including various pretraining strategies and epochs, exhibited slightly lower BLEU-1 scores. While some models may have exhibited minor variations in performance, it is essential to emphasize that, overall, our results were quite consistent. This consistency indicates that our pretraining strategies and fine-tuning approaches are robust and capable of producing reliable outcomes.

Discussion

Principal Findings

In the presented VQA generation, VQA classification, and report generation tasks, we aimed to assess the performance of our BioMedBLIP models against the SOTA original BLIP model using diverse medical image data sets and pretraining strategies. The findings indicate that our BioMedBLIP models, pretrained with specialized medical data sets, exhibit substantial improvements in generating answers for visual questions, as well as in classifying images and questions, depending on the specific data set and pretraining strategy used.

In this section, it is important to emphasize the general trends and insights that can be drawn from these results:

- Data set specificity: the choice of pretraining data sets, such as ROCO and MIMIC-CXR, significantly impacts model performance in both VQA generation and VQA classification tasks. Specialized medical data sets have proven to be valuable for enhancing model capabilities in medical image analysis and question answering.
- Pretraining duration: longer pretraining durations, as evidenced by models such as MIMIC&ROCO-50, have

shown their potential to improve classification accuracy in specific categories of questions, demonstrating the importance of considering pretraining strategies tailored to the task.
- Diverse performance: our models exhibit varying levels of success across different data sets and task categories. This underscores the need for flexibility in selecting pretraining strategies, depending on the specific goals and data sets of a given application.
- Image captioning: our approach, particularly the one followed for BioMedBLIP-ROCO-50, demonstrates improvements in generating captions for medical images, showcasing the model's capacity to excel in both VQA and image captioning tasks.
- Selection of epoch numbers: in our experiments, we observed that the BioMedBLIP models converge at varying numbers of epochs, depending on the data set and the specific downstream tasks involved. This variability is typical in deep learning, where a model's loss decreases up to a certain point and then may increase, indicating that longer training periods do not necessarily yield better performance.

These results underscore the potential of our BioMedBLIP models to excel in a wide range of medical image analysis and question answering tasks, with their performance varying depending on the specific data set and task at hand. BioMedBLIP was tested using 20 different data sets and task combinations. Our method excelled in 15 (75%) out of 20 tasks. BioMedBLIP represents a new SOTA in 15 (75%) out of 20 tasks, and our responses were rated higher in all 20 tasks (P<.005) in comparison to SOTA models. Regression analyses showed that our model's VQA generation has a statistically significant predictor (P<.002) on the SLAKE, PathVQA, and ImageCLEF data sets. In contrast, our model's VQA classification has a relatively lower predictor (P<.003) on the SLAKE, PathVQA, and VQA-RAD data sets in accordance with the regression analyses.

VQA Generation

For the SLAKE data set, our model, pretrained on a combination of general domain data sets and medical domain data sets (MIMIC-CXR and ROCO), consistently outperformed other models across various SLAKE data sets, including open-ended, close-ended, and aggregated types. In contrast to the study by Li et al [12], our results affirm the benefits of pretraining on both general and medical data sets, addressing limitations in their work. This underscores the advantage of a domain-specific model for specialized downstream tasks. Notably, our observations align with the findings of Eslami et al [35], indicating that a pretrained ViT, such as our BLIP model, possesses a comprehensive understanding of image content and long-range dependencies, essential for interpreting the SLAKE data set. Surprisingly, in the VQA-RAD data set, the model pretrained solely on the ROCO data set (for 20 epochs) excelled in open-ended and aggregated tasks, while the MIMIC-CXR and ROCO model performed better in the close-ended task. Contrary to expectations, the model pretrained on general domain data sets and ROCO outperformed the model pretrained on larger domain-specific data sets for PathVQA and ImageCLEF. We attribute this to the superior preprocessing of the ROCO data set, incorporating red bounding boxes that aid in learning crucial image regions. Moreover, our 50-epoch pretraining might not suffice for larger data sets, suggesting the need for further exploration with extended training. Overall, our models demonstrated superior performance on the medical data sets compared to the original BLIP model, emphasizing the efficacy of our approach in medical image analysis.

VQA Classification

In the exploration of VQA classification tasks, diverse models underwent experimentation and fine-tuning across data sets such as SLAKE, PathVQA, and VQA-RAD, encompassing open-ended, close-ended, and aggregated question types. Notably, the SLAKE data set consistently emerged as the data set where BioMedBLIP models consistently exhibited superior performance. Tables 3 and 4 highlight that, in the majority of cases, the BioMedBLIP models performed better than the original BLIP model.

Furthermore, our experiments delved into the impact of different epoch settings on model performance. An intriguing observation emerged, indicating that the relationship between a model's performance and the number of epochs is not consistently positive. This suggests the critical importance of judiciously selecting epoch configurations during the construction of visual language models for medical data sets, challenging the notion that more epochs always lead to improved accuracy.

Image Captioning

In the context of the image captioning task, our findings, while not entirely satisfactory, reveal promising aspects, particularly with the BLIP-ROCO 50 model. This variant surpasses the original BLIP model in BLEU measurements, hinting at potential improvements in using the BLIP model for image captioning. However, the overall performance of the modified models hovers around 23% to 25%, suggesting that the BLIP model may not be inherently well suited for image captioning tasks.

Further Insights

The development and application of our BioMedBLIP models have far-reaching implications across the health care and educational sectors. First and foremost, our models can significantly contribute to improving medical diagnosis and decision support systems. By enhancing the capacity to analyze medical images and answer visual questions, they have the potential to facilitate more accurate and timely health care interventions, ultimately benefiting patient outcomes. In medical education and training, our models can serve as valuable tools for students and professionals alike. Automated question answering capabilities can bridge knowledge gaps and improve learning outcomes in a field that demands continuous learning. Moreover, the automation of image analysis and question answering tasks has the potential to reduce the workload on medical professionals, allowing them to allocate more time to complex aspects of patient care. In terms of research, our models can expedite medical investigations by streamlining the analysis of extensive data sets, potentially leading to groundbreaking discoveries and advancements in the field. Finally, on a global

scale, the availability of advanced artificial intelligence models such as ours can improve medical services in underserved regions where access to specialized medical expertise is limited, thereby contributing to more equitable health care delivery. However, these opportunities are accompanied by ethical and societal responsibilities. Ensuring patient privacy, addressing biases in the data, and maintaining transparency in the development and deployment of artificial intelligence models are pivotal steps to maximize their positive impact while mitigating potential risks and pitfalls in the medical field and beyond.

Our BioMedBLIP models, while showing promise in medical image analysis, come with inherent limitations. First, their performance is heavily contingent on the quality and diversity of the training data. Limitations in data availability, such as smaller or less representative medical data sets, can hinder the model's ability to generalize to real-world scenarios and may introduce biases. Second, the variability in model performance across different data sets and task categories poses a challenge. Achieving optimal results often demands the fine-tuning of pretraining strategies for specific tasks, which may not always be straightforward in practical applications. While choosing the models appropriate for real-world applications, there are various considerations to be made. In our work, we have presented widely used metrics to evaluate model's performance. Different downstream tasks might have different metrics that are used for evaluation. In this case, comparison on the grounds of the most relevant metrics should be made. Moreover, the computational demands for pretraining models, especially in the context of medical tasks, can be substantial, potentially limiting the accessibility of our approach to settings with limited computational resources. We recommend training models for different epochs before selecting them for real-world applications. This is because for different data sets and tasks, the models tend to show convergence at different numbers of epochs. Finally, ethical and privacy concerns are paramount in the use of medical image data. Striving to ensure strict compliance with data protection regulations and maintaining patient privacy and data security are imperative in any real-world implementation.

Conclusions

In conclusion, our development and evaluation of BioMedBLIP models for medical image analysis tasks reveal both promise and practical considerations. These models have shown substantial potential in enhancing the interpretation of medical images and responding to visual questions in a health care context. The choice of pretraining data sets, including ROCO and MIMIC-CXR, plays a pivotal role in model performance, underscoring the importance of specialized medical data for training. Furthermore, a longer duration of pretraining, exemplified by the MIMIC&ROCO-50 model, has demonstrated the potential to elevate classification accuracy in specific question categories. However, our findings highlight the variability in performance across different data sets and tasks, necessitating a flexible approach to pretraining strategies. Moreover, our models have promising implications across health care and education. They can bolster medical diagnosis, decision support systems, and research efforts while also streamlining medical education and reducing the workload on health care professionals. The global accessibility of these models can bring specialized medical expertise to underserved regions.

Authors' Contributions

UN designed and conducted the experiments. UN, ST, and AM provided technical insights and analyzed the results. UN, ST, and AM jointly wrote the first draft and revised it. AM supervised the project. All authors approved the final version of this manuscript.

Conflicts of Interest

None declared.

References

1. Naseem U, Thapa S, Zhang Q, Hu L, Rashid J, Nasim M. Incorporating historical information by disentangling hidden representations for mental health surveillance on social media. Soc Netw Anal Min 2023 Dec 10;14:9. [doi: 10.1007/S13278-023-01167-9]
2. Naseem U, Thapa S, Zhang Q, Hu L, Masood A, Nasim M. Reducing knowledge noise for improved semantic analysis in biomedical natural language processing applications. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. 2023 Presented at: ClinicalNLP@ACL 2023; July 14, 2023; Toronto, ON. [doi: 10.18653/v1/2023.clinicalnlp-1.32]
3. Li F, Thapa S, Bhat S, Sarkar A, Abbott AL. A temporal encoder-decoder approach to extracting blood volume pulse signal morphology from face videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023 Presented at: CVPRW 2023; June 17-24, 2023; Vancouver, BC URL: https://doi.org/10.1109/CVPRW59228.2023.00635 [doi: 10.1109/cvprw59228.2023.00635]
4. Thapa S, Adhikari S. ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. Ann Biomed Eng 2023 Dec 16;51(12):2647-2651. [doi: 10.1007/s10439-023-03284-0] [Medline: 37328703]
5. Pelka O, Koitka S, Rückert J, Nensa F, Friedrich CM. Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Proceedings of the Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. 2018 Presented at: LABELS CVII STENT 2018; September 16, 2018; Granada, Spain. [doi: 10.1007/978-3-030-01364-6_20]

6. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019 Dec 12;6(1):317 [FREE Full text] [doi: 10.1038/s41597-019-0322-0] [Medline: 31831740]
7. biomedblip / biomedblip. Hugging Face. URL: https://huggingface.co/biomedblip/biomedblip/tree/main [accessed 2024-07-09]
8. Pavlopoulos J, Kougia V, Androutsopoulos I. A survey on biomedical image captioning. arXiv. Preprint posted online on May 26, 2019 [FREE Full text] [doi: 10.18653/v1/w19-1803]
9. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 Presented at: ACL 2018; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1240]
10. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, et al. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016 Mar;23(2):304-310 [FREE Full text] [doi: 10.1093/jamia/ocv080] [Medline: 26133894]
11. Lin Z, Zhang D, Tao Q, Shi D, Haffari G, Wu Q, et al. Medical visual question answering: a survey. Artif Intell Med 2023 Sep;143:102611. [doi: 10.1016/j.artmed.2023.102611] [Medline: 37673579]
12. Li Y, Wang H, Luo Y. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2020 Presented at: BIBM 2020; December 16-19, 2020; Seoul, South Korea. [doi: 10.1109/bibm49941.2020.9313289]
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online on June 12, 2017 [FREE Full text]
14. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on October 11, 2018 [FREE Full text] [doi: 10.18653/v1/n18-2]
15. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
16. Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: a simple and performant baseline for vision and language. arXiv. Preprint posted online on August 9, 2019 [FREE Full text]
17. Tan H, Bansal M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. arXiv. Preprint posted online on August 20, 2019 [FREE Full text] [doi: 10.18653/v1/d19-1514]
18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. Preprint posted online on October 22, 2020 [FREE Full text]
19. Eslami S, de Melo G, Meinel C. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? arXiv. Preprint posted online on December 27, 2021 [FREE Full text] [doi: 10.18653/v1/2023.findings-eacl.88]
20. Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, et al. UNITER: UNiversal Image-TExt Representation learning. arXiv. Preprint posted online on September 25, 2019 [FREE Full text] [doi: 10.1007/978-3-030-58577-8_7]
21. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. arXiv. Preprint posted online on February 26, 2021 [FREE Full text]
22. Chen Q, Hu X, Wang Z, Hong Y. MedBLIP: bootstrapping language-image pre-training from 3D medical images and texts. arXiv. Preprint posted online on May 18, 2023 [FREE Full text]
23. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv. Preprint posted online on January 28, 2022 [FREE Full text]
24. Radiology Objects in COntext (ROCO): a multimodal image dataset. GitHub. URL: https://github.com/razorx89/roco-dataset [accessed 2023-11-22]
25. Ionescu B, Müller H, Péteri R, Cid YD, Liauchuk V, Kovalev V, et al. ImageCLEF 2019: multimedia retrieval in medicine, lifelogging, security and nature. In: Proceedings of the Experimental IR Meets Multilinguality, Multimodality, and Interaction - 10th International Conference of the CLEF Association. 2019 Presented at: CLEF 2019; September 9-12, 2019; Lugano, Switzerland. [doi: 10.1007/978-3-030-28577-7_28]
26. Abacha AB, Hasan SA, Datla VV, Liu J, Demner-Fushman D, Muller H. VQA-Med: overview of the medical visual question answering task at ImageCLEF 2019. In: Proceedings of the Experimental IR Meets Multilinguality, Multimodality, and Interaction - 10th International Conference of the CLEF Association. 2019 Presented at: CLEF 2019; September 9-12, 2019; Lugano, Switzerland. [doi: 10.18653/v1/w19-5039]
27. SLAKE: a Semantically-Labeled Knowledge-Enhanced dataset for medical visual question answering. MedVQA. URL: https://www.med-vqa.com/slake//#gt-Download [accessed 2024-06-21]
28. Liu B, Zhan LM, Xu L, Ma L, Yang Y, Wu XM. SLAKE: a Semantically-Labeled Knowledge-Enhanced dataset for medical visual question answering. arXiv. Preprint posted online on February 18, 2021 [FREE Full text] [doi: 10.1109/isbi48211.2021.9434010]
29. He X, Cai Z, Wei W, Zhang Y, Mou L, Xing E, et al. Towards visual question answering on pathology images. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference

on Natural Language Processing (Volume 2: Short Papers). 2021 Presented at: ACL/IJCNLP 2021; August 1-6, 2021; Virtual Event. [doi: 10.18653/v1/2021.acl-short.90]
30. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. Sci Data 2018 Nov 20;5(1):180251 [FREE Full text] [doi: 10.1038/sdata.2018.251] [Medline: 30457565]
31. AIOZ AI - overcoming data limitation in medical visual question answering (MICCAI 2019). GitHub. URL: https://github.com/aioz-ai/MICCAI19-MedVQA [accessed 2023-11-22]
32. Open access biomedical image search engine. National Institutes of Health National Library of Medicine. URL: https://openi.nlm.nih.gov/ [accessed 2023-11-23]
33. salesforce/BLIP. GitHub. URL: https://github.com/salesforce/BLIP [accessed 2023-11-23]
34. Introduction to high performance computing. Sydney Informatics Hub, University of Sydney. URL: https://sydney-informatics-hub.github.io/training.artemis.introhpc/ [accessed 2023-11-24]
35. Eslami S, Meinel C, de Melo G. PubMedCLIP: how much does CLIP benefit visual question answering in the medical domain? In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023 Presented at: EACL 2023; May 2-6, 2023; Dubrovnik, Croatia. [doi: 10.18653/v1/2023.findings-eacl.88]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
BLEU: Bilingual Evaluation Understudy
BLIP: Bootstrapping Language-Image Pretraining
CLIP: Contrastive Language-Image Pretraining
CPU: central processing unit
CT: computed tomography
EM: exact match
GPU: graphics processing unit
ImageCLEF: Image Cross-Language Evaluation Forum
LXMERT: Learning Cross-Modality Encoder Representations From Transformers
MedVQA: medical visual question answering
MIMIC-CXR: Medical Information Mart for Intensive Care-Chest X-ray
MRI: magnetic resonance imaging
NLP: natural language processing
PEIR: Pathology Education Informational Resource
ROCO: Radiology Objects in Context
SLAKE: Semantically-Labeled Knowledge-Enhanced
SOTA: state of the art
V+L: vision and language
ViT: Vision Transformer
VQA: visual question answering

Edited by C Lovis; submitted 22.01.24; peer-reviewed by J Chen, T Ma; comments to author 24.02.24; revised version received 20.04.24; accepted 04.05.24; published 05.08.24.

Please cite as:
Naseem U, Thapa S, Masood A
Advancing Accuracy in Multimodal Medical Tasks Through Bootstrapped Language-Image Pretraining (BioMedBLIP): Performance Evaluation Study
JMIR Med Inform 2024;12:e56627
URL: https://medinform.jmir.org/2024/1/e56627
doi:10.2196/56627
PMID:39102281

Original Paper

Early Diagnosis of Hereditary Angioedema in Japan Based on a US Medical Dataset: Algorithm Development and Validation

Kouhei Yamashita[1], MD, PhD; Yuji Nomoto[2], MD; Tomoya Hirose[3], MD, PhD; Akira Yutani[4], PhD; Akira Okada[5], ME; Nayu Watanabe[5], MMG; Ken Suzuki[5], ME; Munenori Senzaki[5], MS; Tomohiro Kuroda[4], PhD

[1]Department of Hematology and Oncology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

[2]Department of Palliative Care Medicine, Niigata City General Hospital, Niigata, Japan

[3]Department of Traumatology and Acute Critical Medicine, Graduate School of Medicine, Osaka University, Osaka, Japan

[4]Division of Medical Information Technology and Administration Planning, Kyoto University Hospital, Kyoto, Japan

[5]Healthcare and Life Science, IBM Consulting, IBM Japan, Ltd, Tokyo, Japan

Corresponding Author:
Kouhei Yamashita, MD, PhD
Department of Hematology and Oncology
Graduate School of Medicine
Kyoto University
54 Shogoin-kawahara-cho, Sakyo-ku
Kyoto, 606-8507
Japan
Phone: 81 75 751 4964
Fax: 81 75 751 4963
Email: kouhei@kuhp.kyoto-u.ac.jp

Abstract

Background: Hereditary angioedema (HAE), a rare genetic disease, induces acute attacks of swelling in various regions of the body. Its prevalence is estimated to be 1 in 50,000 people, with no reported bias among different ethnic groups. However, considering the estimated prevalence, the number of patients in Japan diagnosed with HAE remains approximately 1 in 250,000, which means that only 20% of potential HAE cases are identified.

Objective: This study aimed to develop an artificial intelligence (AI) model that can detect patients with suspected HAE using medical history data (medical claims, prescriptions, and electronic medical records [EMRs]) in the United States. We also aimed to validate the detection performance of the model for HAE cases using the Japanese dataset.

Methods: The HAE patient and control groups were identified using the US claims and EMR datasets. We analyzed the characteristics of the diagnostic history of patients with HAE and developed an AI model to predict the probability of HAE based on a generalized linear model and bootstrap method. The model was then applied to the EMR data of the Kyoto University Hospital to verify its applicability to the Japanese dataset.

Results: Precision and sensitivity were measured to validate the model performance. Using the comprehensive US dataset, the precision score was 2% in the initial model development step. Our model can screen out suspected patients, where 1 in 50 of these patients have HAE. In addition, in the validation step with Japanese EMR data, the precision score was 23.6%, which exceeded our expectations. We achieved a sensitivity score of 61.5% for the US dataset and 37.6% for the validation exercise using data from a single Japanese hospital. Overall, our model could predict patients with typical HAE symptoms.

Conclusions: This study indicates that our AI model can detect HAE in patients with typical symptoms and is effective in Japanese data. However, further prospective clinical studies are required to investigate whether this model can be used to diagnose HAE.

(JMIR Med Inform 2024;12:e59858) doi:10.2196/59858

KEYWORDS

machine learning; screening; AI; prediction; rare diseases; HAE; electronic medical record; real world data; big data; angioedema; edema; ML; artificial intelligence; algorithm; algorithms; predictive model; predictive models; predictive analytics; predictive system; practical model; practical models; early warning; early detection; real world data; RWD; Electronic health record; EHR;

electronic health records; EHRs; EMR; electronic medical records; EMRs; patient record; patient record; health record; health records; personal health record; PHR

Introduction

The rare genetic disease hereditary angioedema (HAE) induces acute attacks of swelling in various regions of the body, including the face, hands, arms, legs, abdomen, genitals, buttocks, and throat. Gastrointestinal disturbances such as abdominal pain, nausea, and vomiting are frequently associated with edema. Laryngeal edema is rare, even though more than half of the patients with HAE encounter this life-threatening condition [1]. Its global prevalence is estimated to be 1 in 50,000 people, with no reported bias among different ethnic groups [2]. In Japan, about 1 in 250,000 people are diagnosed with HAE, which suggests that only 20% of potential HAE cases are identified [3], suggesting that many patients with HAE remain undiagnosed in Japan. Furthermore, in Japan, the mean duration from the first symptoms to diagnosis is 15.6 years [4], which is longer than that in Europe and the United States [5,6]. Early detection of undiagnosed patients is critical for effective treatment of HAE.

To overcome this situation in Japan, the Diagnostic Consortium to Advance the Ecosystem for Hereditary Angioedema (DISCOVERY) was established in 2021 [7]; it aimed to identify patients with undiagnosed HAE and provide them with appropriate treatment as early as possible.

In this study, we aimed to develop an artificial intelligence (AI) model that can detect suspected patients with HAE using medical history data (claims and electronic medical records [EMRs]) in the United States. We then sought to validate the model's performance in detecting HAE cases. In addition, we conducted a pilot study at Kyoto University Hospital (KUHP) using the EMR data to verify the model's applicability to medical data obtained from the Japanese population. The main objective of this study was to verify whether this model could identify patients with a history of HAE or related diseases.

Methods

Overview

First, we developed an AI model using medical history data from the United States as a reference. Thereafter, we applied the model to medical history data from Japan and verified its efficacy using a Japanese dataset. Note that we used a large dataset of patients from the United States as input for the model, considering that HAE is a rare disease.

Initial Model Development with US Dataset

Data Selection

The Merative MarketScan Explorys Claims-EMR Data Set (formerly IBM Watson Health) [8] was used to obtain patient-level linked claims and EMR data for US patients. The diagnoses and prescription histories of patients with edema or digestive symptoms from January 2012 to January 2021 were identified from the dataset and were used to build our model. Data from a total of 4,283,815 patients were used in the study.

To identify the diagnosis history of patients, the *International Classification of Diseases* (*ICD*) [9] code (ninth and 10th edition) available in this dataset was used. However, the *ICD* code for HAE (D84.1) represents "defects in the complement system," which is also applicable to other similar diseases. Therefore, we used the prescription history of drugs administered only for HAE (Table 1) to distinguish patients with HAE. We categorized the patients with a prescription history of these drugs as the "HAE group," representing patients presumed to have HAE.

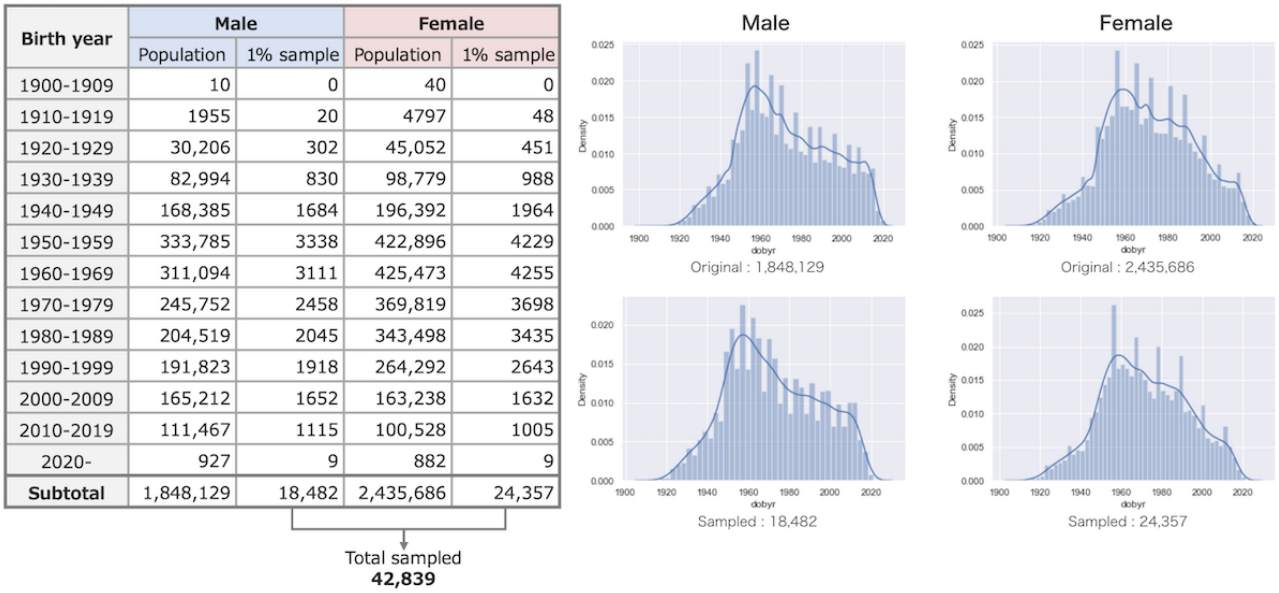
Table 1. US Food and Drug Administration–approved medications used only for hereditary angioedema (as of January 2022).

Proprietary name	Nonproprietary name	Product NDC[a]
BERINERT	Human C1-esterase inhibitor	63833-825
CINRYZE	Human C1-esterase inhibitor	42227-081
		42227-083
FIRAZYR	Icatibant acetate	54092-702
HAEGARDA	Human C1-esterase inhibitor	63833-828
		63833-829
KALBITOR	Ecallantide	47783-101
ORLADEYO	Berotralstat hydrochloride	72769-101
		72769-102
RUCONEST	C1 esterase inhibitor recombinant	70383-350
		69913-350
		71274-350
TAKHZYRO	lanadelumab-flyo	47783-644
Icatibant (Generic)	Icatibant acetate or Icatibant	0093-3066
		24201-207
		60505-6214
		63323-574
		68462-828
		69097-664
		71225-114
SAJAZIR	Icatibant	70709-013

[a]NDC: National Drug Code.

To maintain the demographic characteristics of the original data, the control group was randomly sampled from 1% of the remaining patients, with a fixed ratio of age groups and male-to-female ratio (Figure 1). Note that this was crucial to reduce the data volume to operate the model using limited computation resources (2 central processing units and 16 GB of memory). This was done considering the potential use of the model in various medical institutions in the future.

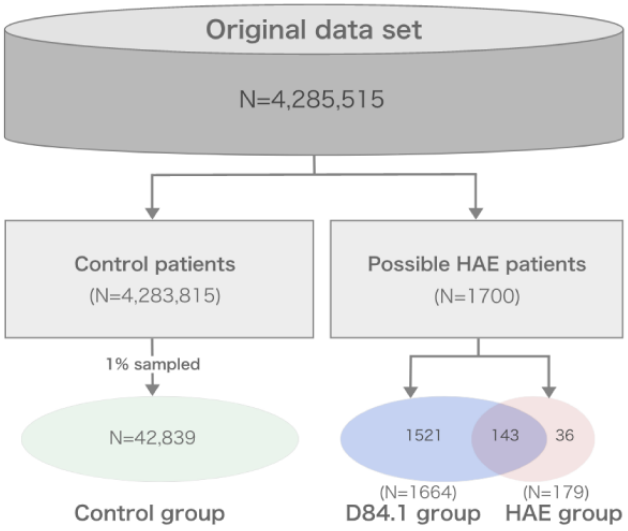
Figure 1. Comparison of the distribution of the 1% sampled data set with that of the population. dobyr: date of birth year.



Birth year	Male		Female	
	Population	1% sample	Population	1% sample
1900-1909	10	0	40	0
1910-1919	1955	20	4797	48
1920-1929	30,206	302	45,052	451
1930-1939	82,994	830	98,779	988
1940-1949	168,385	1684	196,392	1964
1950-1959	333,785	3338	422,896	4229
1960-1969	311,094	3111	425,473	4255
1970-1979	245,752	2458	369,819	3698
1980-1989	204,519	2045	343,498	3435
1990-1999	191,823	1918	264,292	2643
2000-2009	165,212	1652	163,238	1632
2010-2019	111,467	1115	100,528	1005
2020-	927	9	882	9
Subtotal	1,848,129	18,482	2,435,686	24,357

Finally, 3 groups were included for model development and validation (Figure 2): the HAE group with 179 patients, D84.1 (including individuals that likely have HAE but do not have a prescription history of HAE-specific treatments) with 1521 patients, and the control group with 42,839 patients.

Figure 2. Flowchart depicting the different patient groups created using the US data set; HAE: hereditary angioedema.



To develop the model, the *ICD* code was used to create features that described the diagnostic history of patients. As this dataset contained both *ICD-9* and *ICD-10* codes throughout the data period, we standardized the 2 *ICD* types. We assigned codes representing the same disease items from both *ICD-9* and *ICD-10* codes under a single ID.

Model Development

Feature Selection

We counted the number of types of *ICD* codes diagnosed in both the HAE and D84.1 groups, as these 2 groups should have similar features. Furthermore, the differences in *ICD* code types between the groups were required to create a model that can identify patients with HAE. We examined rank correlations between the 2 groups and found it to be approximately 0.08, which suggested that the 2 groups had different characteristics. We then examined specific *ICD* codes that were significantly ranked differently between the two groups and identified 25 such *ICD* codes, which were then used as the primary features in developing the model.

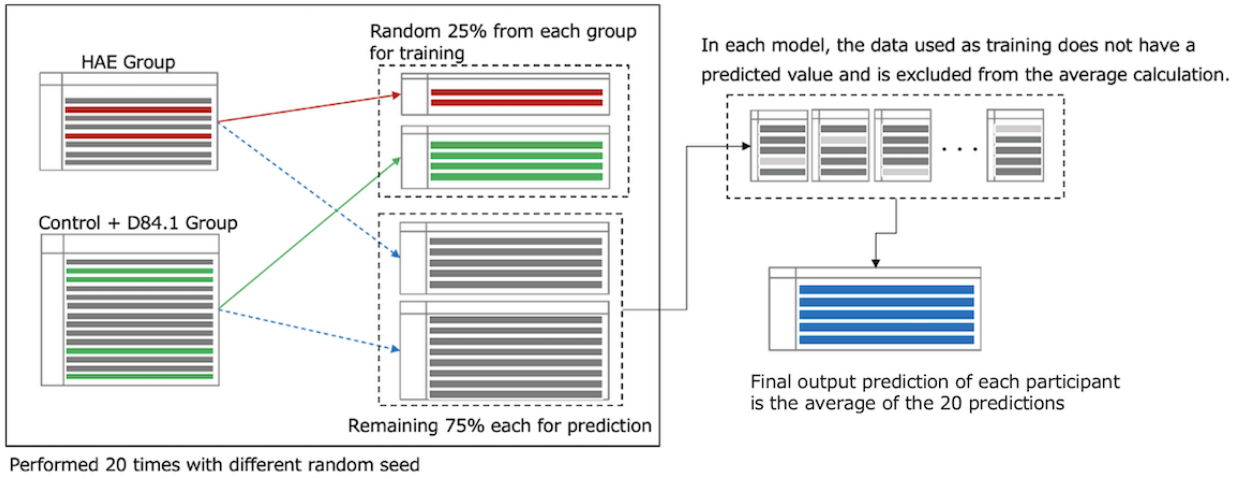
We also examined *ICD* codes that were diagnosed several times over a period of 1 year. This is important as patients with HAE tend to have repeated occurrences of swelling in various regions of the body [1], which can lead to the diagnosis of stomachaches and edemas. We counted the number of patients who had been diagnosed with stomachaches or edemas between 2 and 4 times per year and found a substantial difference between both groups. Considering that the medical record entry may overlap multiple times when changing the record types, we conducted the removal of duplicates based on the date and *ICD* code for each patient. Thereafter, we labeled a group of *ICD* codes related to abdominal pain or edemas and counted the number of times they were assigned in a 1-year span window for each patient based on this dataset. From this exploratory analysis, we included instances where individuals experienced four or more incidences of stomachaches and 3 or more incidences of edema per year as part of the main features of our model. The table of the explanatory variables is provided in Multimedia Appendix 1.

Model Building

The number of patients in the HAE group was extremely small compared with that in the control + D84.1 groups; thus, to avoid overfitting, we used bootstrap sampling [10,11] to create the model. A generalized linear model [12] with regularization terms [13,14] was adopted. We used $\log\left(\frac{p}{1-p}\right)$ for the link function to create a logistic model that would indicate the likelihood of the patient belonging to the HAE group. We chose logistic regression for evaluation as it allows for regression with regularization and is relatively easy to use for evaluating and interpreting feature importance by checking the coefficients. The estimation of the partial regression coefficients was calculated by the maximum likelihood method, which estimates parameters (known as maximum likelihood estimates) that maximize the likelihood of the given observed values. The regularization parameter λ was set to 1 to ensure that it was Lasso regularization.

We used 25% of the data from the HAE group and another 25% from the control + D84.1 groups to train the model, which was then used to predict the remaining 75% of each group. This modeling process was performed 20 times with different random seeds. The average predicted value was calculated as the final output for all the patients. In each trial, the sample used as training data did not have a predicted value and was excluded from the average value calculation (Figure 3). Upon applying the regularization using Lasso regression, the number of substantial features was sorted out during each calculation by mathematically adjusting the coefficients of some variables to 0. The number of sorted features varied with an average of 10; notably, different features were selected every time.

Figure 3. Training data extraction and prediction calculation of the constructed model. HAE: hereditary angioedema.



Evaluation Method and Threshold Setting

After obtaining the final value for each participant, we performed Welch *t* test on the 2 distributions to confirm that the 2 groups had different means. Subsequently, we defined the threshold value that yielded the most balanced classification accuracy using the receiver operating characteristic (ROC) curve. ROC curves help visualize the entire scenario of trade-offs between sensitivity and precision across a set of cutoff points. The volumes of the HAE and control + D84.1 groups were not equal; therefore, it was important to check the balance between sensitivity and precision rather than the accuracy itself.

Model Application to Japanese EMR Data

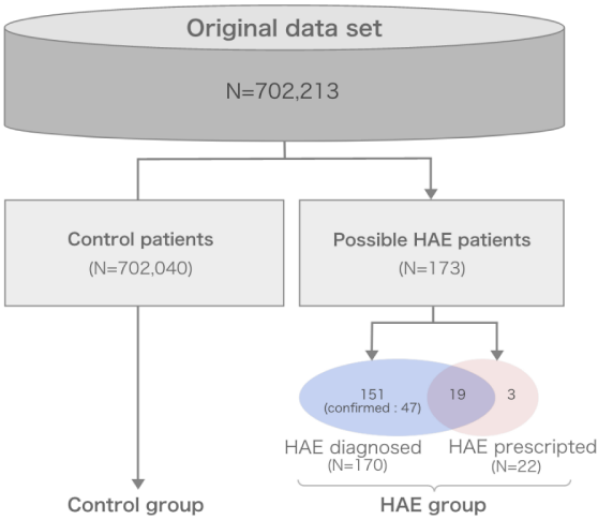
Data Extraction and Model Application

For the validation step using Japanese data, data were extracted from a data warehouse (DWH), which collects medical data from the EMR of the KUHP. Patient IDs in the DWH are pseudonymized. The medical data were obtained for a total of 702,213 patients, among which 22 had a history of HAE, 47 had a confirmed diagnosis of HAE, and 123 had a suspected diagnosis of HAE. The data for model validation included those associated with patients from all these groups (patients using drugs for HAE, patients with confirmed HAE, and patients suspected to have HAE). This was done because physicians may have suspected HAE for some patients if their symptoms were similar to those of patients with the condition. Therefore, these 3 types of patients were considered as the patients with HAE in the study (HAE group; Figure 4).

To adapt the model to Japanese data, we used the standard disease name codes widely used in Japan, as defined by the Medical Information System Development Center (MEDIS-DC) [15], instead of the *ICD* code. Although the *ICD* code is the basic classification code for diagnosis, the standard disease name codes have more subdivisions compared with the *ICD* code, and hence, they can provide a more precise clinical diagnosis. We converted the *ICD* codes using the standard disease name code master for *ICD-10* [16].

Patient data extracted from the DWH were transferred to the Google Cloud Platform server (a virtual private cloud environment) hosted at KUHP. The AI model and statistical programs were stored in a container and sent to the server. We then accessed the server through a virtual private network, which could only be accessed by the authors of this study. The model was applied to all patient data on this server.

Figure 4. Flowchart depicting the different patient groups obtained from the KUHP data set. HAE: hereditary angioedema.



Ethical Considerations

This study was approved by the Ethics Committee of the Kyoto University Hospital (approval number R3750). In this study, we used pseudonymized information that had already been processed, thus individual informed consent was not required. The pseudonymized medical data is made available for academic research in accordance with KUHP's privacy policy. Information regarding each study is publicly disclosed on the institution's website, where patients are informed of their right to opt out along with the opt-out procedure.

Results

Evaluation of the Initial Model

Welch *t* test indicated that the 2 patient groups did not have the same mean values, as suggested by the *P* value of 2.2e-16. Furthermore, the area under the ROC curve was 86.4%, which was obtained when only the HAE group was set as true and all the other groups as false. The best accuracy threshold of this ROC curve was calculated as 39%, with an accuracy of 99.6%. This is because the volume of the control + D84.1 group was larger than that of the HAE group. The true-positive (sensitivity) of this threshold was only 10.6%, with a precision of 54.3%.

As we aimed to identify patients likely to have HAE, we searched for a different threshold that could improve the sensitivity while keeping the precision at an acceptable level. Considering the fact that the prediction of the HAE group had  and , 0.075-0.125 could be a good threshold candidate. We confirmed the sensitivity and precision for the thresholds of 0.075, 0.1, and 0.125 to determine the most balanced threshold, as shown in Table 2.

Table 2. Cross-tabulation was calculated at 3 different threshold values using all data groups for a detailed evaluation of different scaled precisions and group sensitivities.

	Control group		D84.1 group	D84.1 and HAE[a]	HAE group (not D84.1)	Score (%)			
	100% scale converged	1% scale				1% scale precision	100% scale precision	Sensitivity 1[b]	Sensitivity 2[c]
Suspection statistic (threshold=0.1)									
Not suspected, n	4,279,500	42,795	1312	55	30	27.1	2.0	61.5	52.5
Suspected, n	4400	44	209	88	6	—[d]	—	—	—
Suspection statistic (threshold=0.075)									
Not suspected, n	4,277,900	42,779	1266	52	28	23.9	1.6	63.6	55.3
Suspected, n	6000	60	255	91	8	—	—	—	—
Suspection statistic (threshold=0.125)									
Not suspected, n	4,280,600	42,806	1343	62	33	28.5	2.4	56.6	46.9
Suspected, n	3300	33	178	81	3	—	—	—	—

[a]HAE: hereditary angioedema.

[b]Excluding "not D84.1" patients.

[c]Including "not D84.1" patients.

[d]Not applicable.

The threshold value of 0.1 had a sensitivity of 52.5% and precision of 27.1%, indicating that 1 out of 2 known HAE group participants can be correctly detected, and 1 out of 4 detected participants should correctly belong to the HAE group. If we exclude the HAE group participants who were not diagnosed with D84.1, the sensitivity was 61.5%. This result was calculated based on 1% of the sample size of the original control group; thus, by multiplying the number of all participants from the control group by 100, we obtained a 100% scale precision of 2%. This was 2 times better than the 1% precision goal set at the beginning of the study. This means that based on this model, 1 out of 50 suspected patients is highly likely to have HAE. Considering that HAE prevalence is estimated to be 1 in 50,000 people, we can expect to find undiagnosed patients with HAE quickly and efficiently using this model output.

From a conservative standpoint, the threshold value of 0.1 seems optimal. However, to identify more potential patients with HAE, it might be better to apply the 0.075 threshold, which has a sensitivity of 55.3% and a precision of 23.9%. If we recalculate the 100% scale precision in the same manner as described above, we obtain 1.6%. This means we can still achieve our goal of 1% precision while improving the sensitivity.

In addition, we need to consider the fact that the ratio of suspected patients in the US dataset can be calculated to be approximately 0.09% with a 0.1 threshold and 0.15% with a threshold of 0.075. If this model is to be used on a much smaller volume dataset compared with the US dataset, there is an approximately 2 times higher risk of obtaining zero suspected patients with a 0.1 threshold than with the 0.075 threshold.

Application of the Model to Japanese EMR Data

To verify the performance of this model using Japanese data, it was applied to patient data obtained from KUHP, and the output of potential patients with HAE was obtained based on the selected threshold. The diagnostic histories of these patients were stored at a single university hospital. Compared with the dataset used to build the original model, the variation and coverage of the entire diagnostic history were assumed to be relatively low. Therefore, we adopted a threshold value of 0.075 in this validation study to aggressively identify patients with HAE. We considered the HAE group (Figure 4) as the correct data for this validation.

As shown in Table 3, 65 of 173 patients with HAE were detected using this model, indicating a sensitivity of 37.6%.

Table 3. Cross-tabulation with precision and sensitivity scores of Kyoto University Hospital results.

	Control group	HAE[a] group			Score (threshold=0.075)	
		Prescripted	Prescripted and diagnosed	Diagnosed	Precision (%)	Sensitivity (%)
Suspection statistic (threshold=0.075)					3.2	31.8
Not suspected	701,829	2	13	93		
Suspected	211	1	6	58		

[a]HAE: hereditary angioedema.

Some patients in the HAE group did not have a diagnostic history specific to HAE (eg, abdominal pain, swelling, or edema) within the KUHP data. Their common symptoms might have been treated by their primary doctors or clinics and not at this university hospital. Furthermore, because HAE is a hereditary disorder, some patients may have been diagnosed through family testing. These factors appear to lead to a lower sensitivity score for the Japanese dataset than that for the US data.

The precision score was 23.6%, which is more than 14 times higher than that of the initial model. As mentioned in the Introduction section, only 20% of patients in Japan are diagnosed with HAE, which means that 80% of patients with HAE are undiagnosed. Therefore, the 211 patients from the control group who were suspected to have HAE in our model may be undiagnosed patients with HAE.

Discussion

Principal Findings

In this study, we developed an AI model for screening patients with HAE and validated its performance using 2 methods.

First, a large patient dataset was selected to build a model containing patient-level linked claims and EMR data from the United States. The advantage of this dataset is that it contains a long-term prescription and diagnostic history across multiple medical institutes. The diagnostic characteristics of patients with HAE were determined by analyzing the dataset. Based on these characteristics, we constructed a generalized linear model with regularization terms. At a threshold of 0.1, the sensitivity score was 52.5% and the precision score was 27.1% if patients with possible HAE were included in the correct answer group. When these were excluded from the correct answers, the sensitivity score was 61.5%.

We then applied this model to Japanese EMR data. This validation was conducted at a single university hospital using DWH data. Generally, patients often visit local hospitals and rarely visit university hospitals if they present with common symptoms. Considering this situation, data obtained from a single university may have some difficulty with model performance. Although the sensitivity score was lower than that of the US dataset (37.6%), the precision score reached 23.6% with a threshold value of 0.075. This implies that our model has a high possibility of identifying patients with undiagnosed HAE in Japan.

Limitations

Our study had several limitations. Generally, because HAE is a rare disease, patient group data (correct answer data in machine learning) are quite small. In addition, the variance in each patient's features was larger than that in common diseases. We also suggest possible limitations and countermeasures.

Family History

In our basic analysis of the HAE group, we found that some patients in the HAE group had a lower diagnostic history than others. We suspected that these patients had been diagnosed with HAE based on their family histories. Because our model uses the diagnostic history to calculate the probability, these cases are potentially difficult to detect.

US Patient Data Consists of Data From Multiple Hospitals

Our model may rely on the fact that US patient data consists of data from multiple hospitals. Collecting data from multiple hospitals will allow tracking of the records of a single patient across these hospitals and provide a more detailed medical history. For validation in the Japanese dataset, we could only use data from a single university hospital, which may be one of the reasons for the low sensitivity.

Potential Patients With HAE Might Be Included in the Control + D84.1 Groups

Since the HAE diagnosis rate was low, it is likely that there were more patients with HAE in the control + D84.1 groups. In our approach, we assigned the HAE group a prescription history of HAE drugs to keep the model conservative.

Possible Difference in Diagnostic Tendency Between the United States and Japan

If there are differences in how doctors make diagnostics between countries, we may need to customize the model or threshold to adapt it to Japan and other countries.

Comparison With Previous Work

Few previous studies have focused on screening patients for rare diseases based on diagnostic histories such as medical claims. Nonetheless, some studies have focused on a few rare diseases. For example, a previous study used AI models based on diagnostic history to identify patients with Pompe disease [17]. In this study, 104 patients were flagged by the model to have the disease, but only 19 were determined by specialists to have a high likelihood of having Pompe disease, rendering a precision score of 18.27% [17]. In comparison, our model recorded a precision of 23.6%. Screening for rare diseases is extremely difficult compared with other common diseases, for which abundant data exist; however, our results indicate that AI models can show high performance for screening rare diseases.

Conclusions and Future Directions

Considering the prevalence of HAE (1/50,000), the screening performance of this model was 1,000 times greater than that achieved through random searching using US data. Owing to their prevalence and recognition rates, identifying undiagnosed patients with rare diseases is an arduous task. This study suggests that patient screening for HAE may become significantly more efficient if this AI model is used. This approach is particularly valuable for the diagnosis and treatment of rare diseases.

In addition, during the validation phase using the Japanese data, the model was effective at a single university hospital. Although only the diagnosis codes recorded in the EMR were available, the model could detect patients with typical symptoms of HAE. The performance of the model can likely be improved further if this model is applied to the data from city hospitals or medical claims, which contain diagnostic histories of patients in multiple medical institutions. This can provide more comprehensive information on the symptoms and diagnostic histories of each patient.

In this study, only patients with a diagnostic history of HAE within the dataset were defined as correct answers. By providing a diagnosis rate, these data may include patients with undetected HAE. The model performance cannot be strictly calculated in such situations. Therefore, further studies are needed to determine whether patients with undiagnosed HAE should be included in the predicted group. This is because identifying undiagnosed patients with HAE is a critical issue, especially in Japan; we will implement a prospective clinical study using our AI model.

The constructed model may help researchers, physicians, and other health care professionals identify undiagnosed HAE cases. Eventually, if this strategy can identify undiagnosed patients and provide them with proper treatment, their quality of life will likely be improved.

Acknowledgments

We acknowledge the support received from DISCOVERY (Diagnostic Consortium to Advance the Ecosystem for Hereditary Angioedema), which covered the publication costs of the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table of the explanatory variables.
[PNG File , 147 KB - medinform_v12i1e59858_app1.png]

References

1. Bork K, Meng G, Staubach P, Hardt J. Hereditary angioedema: new findings concerning symptoms, affected organs, and course. Am J Med 2006;119(3):267-274. [doi: 10.1016/j.amjmed.2005.09.064] [Medline: 16490473]
2. Zuraw BL. Clinical practice. Hereditary angioedema. N Engl J Med 2008;359(10):1027-1036. [doi: 10.1056/NEJMcp0803977] [Medline: 18768946]
3. Ohsawa I. Nanbyo Iden-sei kekkann-sei fushu HAE(An intractable disease: Hereditary Angioedema (HAE)). Osaka: Iyaku Jānarusha; 2016.
4. Iwamoto K, Yamamoto B, Ohsawa I, Honda D, Horiuchi T, Tanaka A, et al. The diagnosis and treatment of hereditary angioedema patients in Japan: a patient reported outcome survey. Allergol Int 2021;70(2):235-243 [FREE Full text] [doi: 10.1016/j.alit.2020.09.008] [Medline: 33168485]
5. Bellanti JA, Settipane RA. The Floralia: a festive time for Romans and a demanding time for the allergist/immunologist. Allergy Asthma Proc 2018;39(3):167-168 [FREE Full text] [doi: 10.2500/aap.2018.39.4141] [Medline: 29669662]
6. Zanichelli A, Magerl M, Longhurst H, Fabien V, Maurer M. Hereditary angioedema with C1 inhibitor deficiency: delay in diagnosis in Europe. Allergy Asthma Clin Immunol 2013;9(1):29 [FREE Full text] [doi: 10.1186/1710-1492-9-29] [Medline: 23937903]

7. DISCOVERY (Diagnostic Consortium to Advance the Ecosystem for Hereditary Angioedema). URL: https://discovery0208.or.jp/en/top/ [accessed 2024-06-07]
8. Merative. Real-world evidence solutions for life sciences. URL: https://www.merative.com/content/dam/merative/documents/brief/real-world-evidence-solution-brief.pdf [accessed 2024-06-07]
9. WHO. International Statistical Classification of Diseases and Related Health Problems (ICD). URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2024-06-07]
10. Efron B. Bootstrap methods: another look at the jackknife. Ann Stat 1979;7(1):1-26 [FREE Full text] [doi: 10.1214/aos/1176344552]
11. Bradley E, Tibshirani RJ. An Introduction to the Bootstrap. New York, NY: Chapman & Hall; 1993.
12. McCullagh P, Nelder J. Generalized Linear Models. 2nd edition. London, UK: Chapman & Hall/CRC; 1989.
13. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B (Methodological) 1996;58(1):267-288 [FREE Full text]
14. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55-67 [FREE Full text]
15. Medical Information System Development Center (MEDIS-DC). URL: http://www.medis.or.jp/ [accessed 2024-06-07]
16. Medis. Standard disease name master for ICD-10. URL: http://www2.medis.or.jp/stdcd/byomei/byomei.html [accessed 2024-06-07]
17. Lin S, Nateqi J, Weingartner-Ortner R, Gruarin S, Marling H, Pilgram V, et al. An artificial intelligence-based approach for identifying rare disease patients using retrospective electronic health records applied for pompe disease. Front Neurol 2023;14:1108222 [FREE Full text] [doi: 10.3389/fneur.2023.1108222] [Medline: 37153672]

Abbreviations

AI: artificial intelligence
DISCOVERY: Diagnostic Consortium to Advance the Ecosystem for Hereditary Angioedema
DWH: data warehouse
EMR: electronic medical record
HAE: hereditary angioedema
ICD: International Classification of Diseases
KUHP: Kyoto University Hospital
MEDIS-DC: Medical Information System Development Center
ROC: receiver operating characteristic

Edited by C Lovis; submitted 26.04.24; peer-reviewed by T Takemura, R Zhu, J Wu; comments to author 25.05.24; revised version received 13.06.24; accepted 06.08.24; published 13.09.24.

Please cite as:
Yamashita K, Nomoto Y, Hirose T, Yutani A, Okada A, Watanabe N, Suzuki K, Senzaki M, Kuroda T
Early Diagnosis of Hereditary Angioedema in Japan Based on a US Medical Dataset: Algorithm Development and Validation
JMIR Med Inform 2024;12:e59858
URL: https://medinform.jmir.org/2024/1/e59858
doi:10.2196/59858
PMID:

Original Paper

Personalized Prediction of Long-Term Renal Function Prognosis Following Nephrectomy Using Interpretable Machine Learning Algorithms: Case-Control Study

Lingyu Xu[1], PhD; Chenyu Li[1,2], PhD; Shuang Gao[3], PhD; Long Zhao[1], PhD; Chen Guan[1], PhD; Xuefei Shen[1], PhD; Zhihui Zhu[4], PhD; Cheng Guo[5], PhD; Liwei Zhang[6], PhD; Chengyu Yang[1], PhD; Quandong Bu[1], MSM; Bin Zhou[1], PhD; Yan Xu[1], PhD

[1]Department of Nephrology, The Affiliated Hospital of Qingdao University, Qingdao, China

[2]Medizinische Klinik und Poliklinik IV, Klinikum der Universität, Munich, Germany

[3]Ocean University of China, Qingdao, CN, Qingdao, China

[4]Center of Structural Heart Disease, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

[5]Allianz Technology, Allianz, Munich, Germany

[6]Institute of Diabetes and Regeneration Research, Helmholtz Diabetes Center, Helmholtz Center Munich, Neuherberg, Germany

Corresponding Author:
Yan Xu, PhD
Department of Nephrology, The Affiliated Hospital of Qingdao University
16 Jiangsu Road
Qingdao, 266003
China
Phone: 86 0532 82911668
Email: xuyan@qdu.edu.cn

Abstract

Background: Acute kidney injury (AKI) is a common adverse outcome following nephrectomy. The progression from AKI to acute kidney disease (AKD) and subsequently to chronic kidney disease (CKD) remains a concern; yet, the predictive mechanisms for these transitions are not fully understood. Interpretable machine learning (ML) models offer insights into how clinical features influence long-term renal function outcomes after nephrectomy, providing a more precise framework for identifying patients at risk and supporting improved clinical decision-making processes.

Objective: This study aimed to (1) evaluate postnephrectomy rates of AKI, AKD, and CKD, analyzing long-term renal outcomes along different trajectories; (2) interpret AKD and CKD models using Shapley Additive Explanations values and Local Interpretable Model-Agnostic Explanations algorithm; and (3) develop a web-based tool for estimating AKD or CKD risk after nephrectomy.

Methods: We conducted a retrospective cohort study involving patients who underwent nephrectomy between July 2012 and June 2019. Patient data were randomly split into training, validation, and test sets, maintaining a ratio of 76.5:8.5:15. Eight ML algorithms were used to construct predictive models for postoperative AKD and CKD. The performance of the best-performing models was assessed using various metrics. We used various Shapley Additive Explanations plots and Local Interpretable Model-Agnostic Explanations bar plots to interpret the model and generated directed acyclic graphs to explore the potential causal relationships between features. Additionally, we developed a web-based prediction tool using the top 10 features for AKD prediction and the top 5 features for CKD prediction.

Results: The study cohort comprised 1559 patients. Incidence rates for AKI, AKD, and CKD were 21.7% (n=330), 15.3% (n=238), and 10.6% (n=165), respectively. Among the evaluated ML models, the Light Gradient-Boosting Machine (LightGBM) model demonstrated superior performance, with an area under the receiver operating characteristic curve of 0.97 for AKD prediction and 0.96 for CKD prediction. Performance metrics and plots highlighted the model's competence in discrimination, calibration, and clinical applicability. Operative duration, hemoglobin, blood loss, urine protein, and hematocrit were identified as the top 5 features associated with predicted AKD. Baseline estimated glomerular filtration rate, pathology, trajectories of renal function, age, and total bilirubin were the top 5 features associated with predicted CKD. Additionally, we developed a web application using the LightGBM model to estimate AKD and CKD risks.

Conclusions: An interpretable ML model effectively elucidated its decision-making process in identifying patients at risk of AKD and CKD following nephrectomy by enumerating critical features. The web-based calculator, found on the LightGBM model, can assist in formulating more personalized and evidence-based clinical strategies.

(JMIR Med Inform 2024;12:e52837) doi:10.2196/52837

KEYWORDS

nephrectomy; acute kidney injury; acute kidney disease; chronic kidney disease; machine learning

Introduction

Renal tumors rank as the second most prevalent neoplasms in urology, succeeding bladder cancer, and their annual incidence is on the rise [1,2]. Nephrectomy remains the preferred therapeutic modality for localized renal tumors [3], and patients who are eligible for nephrectomy generally favor longer life span [4]. Nevertheless, a decline in kidney function frequently ensues following nephrectomy. It has been proven that the nephron reduction stemming from radical nephrectomy (RN) or partial nephrectomy (PN) can result in postoperative acute kidney injury (AKI), subsequently heightening the risk of chronic kidney disease (CKD) and mortality [5,6]. Therefore, it is crucial to discern the predicted risk factors associated with renal function decline and precisely forecast postoperative renal impairment, enabling timely intervention.

AKI and CKD are not 2 separate clinical syndromes but often manifest as a continuum of disease [7]. The 16th Acute Disease Quality Initiative meeting has defined acute kidney disease (AKD) as the occurrence of acute or subacute damage or loss of kidney function for a duration of 7 to 90 days after the onset of an AKI-initiating event [8]. Within the AKD time frame, interventions like patient education, medication adjustments, and regular follow-up can be initiated, potentially leading to disease reversal [9]. According to AKD definition, renal recovery is classified into 3 primary groups: transient AKI, subacute AKD, and persistent AKI [8].

Recent, noteworthy strides in machine learning (ML) have given rise to remarkable breakthroughs, encompassing fields like autonomous driving, recommending products, and surpassing human expertise in intricate games such as chess [10-12]. These advancements have increasingly impacted the health care domain, particularly in clinical decision support systems, aiding in clinical decision-making, forecasting disease progression, and enhancing the distribution of medical resources [13,14]. ML offers significant advantages in clinical decision-making by analyzing large datasets, facilitating high-throughput and real-time predictions, and identifying complex patterns. However, considering the challenges related to decision-making transparency, individual patient variability, and ethical concerns, ML should be considered a complementary tool to enhance physicians' diagnostic capabilities rather than substituting their expertise. One of the prominent challenges faced by ML in the health care domain is the enigma referred to as the "black-box phenomenon," indicating the deficiency in interpretability experienced by both patients and health care providers [15,16]. The absence of interpretability in predictive models can erode trust in these models, particularly in health care, where numerous decisions directly involve matters of life and death. Recent advancements, however, have introduced algorithms that effectively extract crucial variables and elucidate model decisions [17].

Currently, there is limited research on the risk prediction of AKD following nephrectomy, and the impact of postnephrectomy AKD on CKD remains unclear. Additionally, there is a lack of interpretable ML models and web-based prediction tools for both AKD and CKD. Therefore, this study aimed to achieve the following objectives: (1) assess the postoperative occurrence rates of AKI, AKD, and CKD in patients who underwent nephrectomy; (2) contrast the long-term renal prognosis across AKI recover, subacute AKD, and patients with AKD and AKI; (3) formulate risk prediction models for both AKD and CKD through the use of diverse ML algorithms; (4) determine the optimal models, evaluate their predictive efficacy, and explain via Shapley Additive Explanations (SHAP) values and Local Interpretable Model-Agnostic Explanations (LIME) algorithms; (5) use directed acyclic graphs (DAGs) to explore potential associations and causal pathways between features; and (6) devise an easily accessible web-based prediction tool tailored to estimating the likelihood of AKD and CKD after nephrectomy. We hypothesized that patients with acute or subacute renal impairment are more susceptible to CKD progression compared to those with normal renal function. Furthermore, we expected significant differences in the development of CKD among patients recovering from AKI, those with subacute AKD, and those experiencing AKD with AKI.

Methods

Study Design

We conducted a retrospective review of medical records for 2637 patients who underwent nephrectomy between July 2012 and June 2019. Ultimately, the study included 1559 eligible patients. The patient data were sourced from a prominent tertiary hospital known for its comprehensive services and ranked among the top 60 nationwide in terms of overall performance. Patients were followed up for a duration ranging from 3.0 to 62.8 months until December 2019, with the primary focus being on the development of CKD as a long-term outcome. The patient data were randomly stratified into training, validation, and test sets, using Python's stratified random sampling method, maintaining a ratio of 76.5:8.5:15. Internal validation was performed through 10-fold cross-validation, involving the partitioning of the training and validation sets into 10 subsets. A majority of 9 of these subsets were used for model training, and the remaining 1 was dedicated to model evaluation. The exclusion criteria for this study encompassed the following characteristics: (1) patients younger than 18 years of age or with

hospitalization duration <24 hours, (2) patients with inadequate serum creatinine (Scr) monitoring interval, (3) patients with anatomical kidney malformations, (4) patients undergoing renal cyst unroofing or donor nephrectomies, (5) patients with pre-existing CKD or undergoing dialysis prior to nephrectomy, and (6) patients lacking essential features such as Scr.

Ethical Considerations

This study received approval from the Ethics Committee of the Affiliated Hospital of Qingdao University (approval QDFY WZ 2018-9-13). Informed consent was waived due to the retrospective nature of the data and the large number of patients involved, making it impractical to seek consent from each patient. All data were deidentified. No compensation was provided to the participants as the study did not involve direct participant interaction.

Data Collection

Clinical and demographic data were extracted through the application of natural language processing and parsing methods on structured information within the electronic health record. Preoperative complete blood counts, coagulation markers, blood chemistry analyses, urine tests, and echocardiography were performed within 3 days of admission. Comorbidities were defined based on the *International Statistical Classification of Diseases, Tenth Revision*. Comprehensive data on concomitant medications were meticulously collected, with particular attention to instances where these medications were administered prior to the occurrence of kidney injury. The surgical details encompass the surgical approach (laparotomy, laparoscopy, and da Vinci surgery), procedure type (RN and PN), duration of the surgery, pathological findings, maximum excision diameter, and blood loss.

Outcome Definitions

The primary outcome of our study was postoperative AKI. The secondary outcomes were AKD and CKD. AKI was defined as an increase in Scr to ≥0.3 mg/dL within 48 hours or ≥1.5 times the baseline value within 7 days, following the 2012 Kidney Disease Improving Global Outcomes guideline [18]. According to the 2017 Acute Disease Quality Initiative, AKD was defined as persistent renal damage or renal dysfunction for a duration of 7 to 90 days after exposure to an AKI initiating event [8]. CKD was defined as abnormalities of kidney structure or function for at least 3 months [19]. Based on the diagnostic criteria for AKI and AKD, patients exhibited three distinct trajectories of renal function following kidney injury: (1) AKI recover, if Scr returned to baseline value within 7 days (AKI without AKD); (2) AKD with AKI, if stage 1 or greater AKI persisted for ≥7 days after an AKI initiating event (continuous AKI progressing to AKD); and (3) subacute AKD, if Scr levels increased slowly but lasted more than 7 days (AKD without AKI). The final classification consisted of four categories: (1) no kidney disease (NKD), (2) AKI recover, (3) AKD with AKI, and (4) subacute AKD.

Baseline Scr was defined as the most recent Scr level measured before nephrectomy. The diagnosis time of AKI, AKD, and CKD was established when patients first met the respective diagnostic criteria. All patients underwent at least 3 Scr tests, including 2 during hospitalization and 1 at the first follow-up. If elevated Scr levels did not return to baseline, additional tests were conducted once a week during hospitalization or at the next follow-up. The estimated glomerular filtration rate (eGFR) was calculated by using the Chronic Kidney Disease Epidemiology Collaboration creatinine formula [20].

Model Development and Interpretation

The Light Gradient-Boosting Machine (LightGBM) algorithm was used to construct predictive models. LightGBM, a tree-based gradient-boosting framework, adeptly manages high-dimensional and extensive datasets [21]. By integrating gradient-based 1-side sampling and exclusive feature bundling, LightGBM effectively mitigates overfitting and notably outperforms the computational speed and memory use of Extreme Gradient-Boosting and stochastic gradient-boosting techniques [22]. In our comparative analysis, we trained various ML models, including LightGBM, Gradient-Boosting Machine, k-nearest neighbors, multilayer perceptron, logistic regression (LR), naive Bayes, random forest (RF), and support vector machine, using the same dataset and applying consistent imputation and scaling techniques. We initially used the default hyperparameters of each ML algorithm to establish our models. Subsequently, we conducted manual parameter tuning by grid search to optimize the performance. The process of parameter optimization was facilitated through 10-fold cross-validation, aiding in the identification of the most suitable hyperparameter configurations [23].

For discerning significant features that influenced the algorithm and ensuring the appropriateness of the optimal model, we used SHAP and LIME to interpret the model from both global and instance-based perspectives. SHAP values, rooted in the Shapley value from coalitional game theory, quantify the influence of each feature variable on the target outcome, elucidating the derivation of a sample's predicted result [24]. LIME uses local surrogate models for explaining individual predictions. Its core method perturbs an input instance to generate interpretable samples, upon which a linear model approximates the complex model's decision-making process near the instance [25]. The SHAP summary plots exhibit the relative significance of individual features in predictions, along with their corresponding positive or negative impact directions. The SHAP interaction plots reveal the interactions among multiple features and illustrate how their combined influence impacts model predictions. We separately used the top 10 features from the AKD and CKD models and created SHAP dependence plots through pairwise combinations to elucidate the influence of individual features on the model's predictions and the correlations among them. Additionally, we highlighted features with significant correlations in Figure S5 in Multimedia Appendix 1. The SHAP force plots and LIME bar plots were used to clarify individualized forecasts, demonstrating each feature's contribution to the prediction of individual samples. Finally, we used AKD and CKD as outcomes and applied the PC algorithm to construct DAGs, facilitating the exploration of potential associations and causal pathways among the top 20 features [26,27].

Web-Based Prediction Tool

A web-based calculator for predicting AKD and CKD among those patients was developed using the "Streamlit" application in terms of the optimal model. Streamlit, an open-source Python framework, aids developers in swiftly constructing web-based and responsive applications [28]. To improve the user-friendliness of the web calculator, this study implemented 2 panels: one for inputting model parameters and acquiring AKD or CKD probabilities and another for providing a model introduction.

Statistical Analysis

Features with a missing proportion exceeding 15% (n=234) are removed, while those with missing proportions less than 15% (n=234) are imputed using an RF model. Using LR to calculate the required sample size with CKD as the outcome, we determined that a minimum of 1171 patients is necessary to attain a statistical power of 90% for detecting an effect size of 0.10 at a 2-side α=.05. Categorical features were presented using frequencies and percentages, while continuous features were presented as mean (SD) or median (IQR). Comparative analyses were performed to assess patient characteristics between individuals with and without CKD as well as among various trajectories of renal function postkidney injury. Quantile-quantile plots were generated to visually inspect the distribution patterns of continuous features. The independent 2-tailed *t* test was used for normally distributed continuous features, the Mann-Whitney *U* test for nonnormally distributed continuous features, and the Pearson chi-square test for categorical features. We used a weight rebalancing technique to adjust the weights of both the majority and minority classes in the training dataset [29]. The validation dataset underwent balancing, whereas the test datasets remained unaltered to assess model performance with representative data. The scikit-learn Python library (Python Software Foundation) includes a built-in parameter called "class weight" or "weights" for LR, RF, LightGBM, support vector machine, and k-nearest neighbors. For AKD, the class weight was set to 3.3; and for non-AKD cases, it was set to 0.6. Similarly, the class weight for CKD was set to 10.0; and for non-CKD cases, it was set to 0.5. In the case of the naive Bayes classifier, we established a prior probability of .50 for each class to achieve group balance, and we adjusted class weights in the multilayer perceptron classifier by modifying the loss function's weights. The area under the receiver operating characteristic curve (AUROC) was used for optimal model selection. The model underwent evaluation through graphical techniques, encompassing the receiver operating characteristic curve and decision curve analysis, in addition to quantitative metrics such as AUROC, average precision, precision, recall, accuracy, F_1-score, Brier score loss, and Matthew correlation coefficient. A *P* value of less than .05 was considered as significant (2-tailed). Python programming language (version 3.9.13 and integrated development environment Visual Studio Code 1.81.1) was applied to our analysis.

Results

Study Cohort

The entire study process is illustrated in Figure 1. Among the participants, 1131 (72.6%) underwent RN, and 1152 (73.9%) underwent laparotomy. The incidence rates of AKI, AKD, and CKD were 21.7% (330/1559), 15.3% (238/1559), and 10.6% (165/1559), respectively. In total, there were 451 (28.9%) patients who developed acute or subacute kidney dysfunction (AKI or AKD criterion), with 117 (7.5%) meeting both AKI and AKD criteria, 121 (7.8%) developed subacute AKD, and 213 (13.7%) experienced recovery from AKI. The quantile-quantile plots show that features including blood loss, Scr, and operative duration exhibit skewed distributions, potentially due to the distinct condition of nephrectomy patients (Figure S1 in Multimedia Appendix 1). Increased CKD rates were observed in older patients (mean age 69, SD 9.6 vs mean age 58, SD 12.3 years), male patients (n=118, 12.9% vs n=47, 7.3% in female patients), those who underwent RN (n=143, 12.6% vs n=22, 5.1% in PN), AKD with AKI (n=42, 35.9% vs n=32, 26.4% in subacute AKD, n=24, 11.3% in AKI recovery, and n=67, 6% in NKD), and individuals with 1 or more chronic complications such as hypertension, diabetes mellitus, and coronary heart disease. The demographic and clinical characteristics of the patient cohort, both within different groups and as a whole, are detailed in Table 1 and Table S1 in Multimedia Appendix 2.

Figure 1. Flow diagram of patients' enrollment. AKD: acute kidney disease; AKI: acute kidney injury; CKD: chronic kidney disease; ML: machine learning; Scr: serum creatinine.

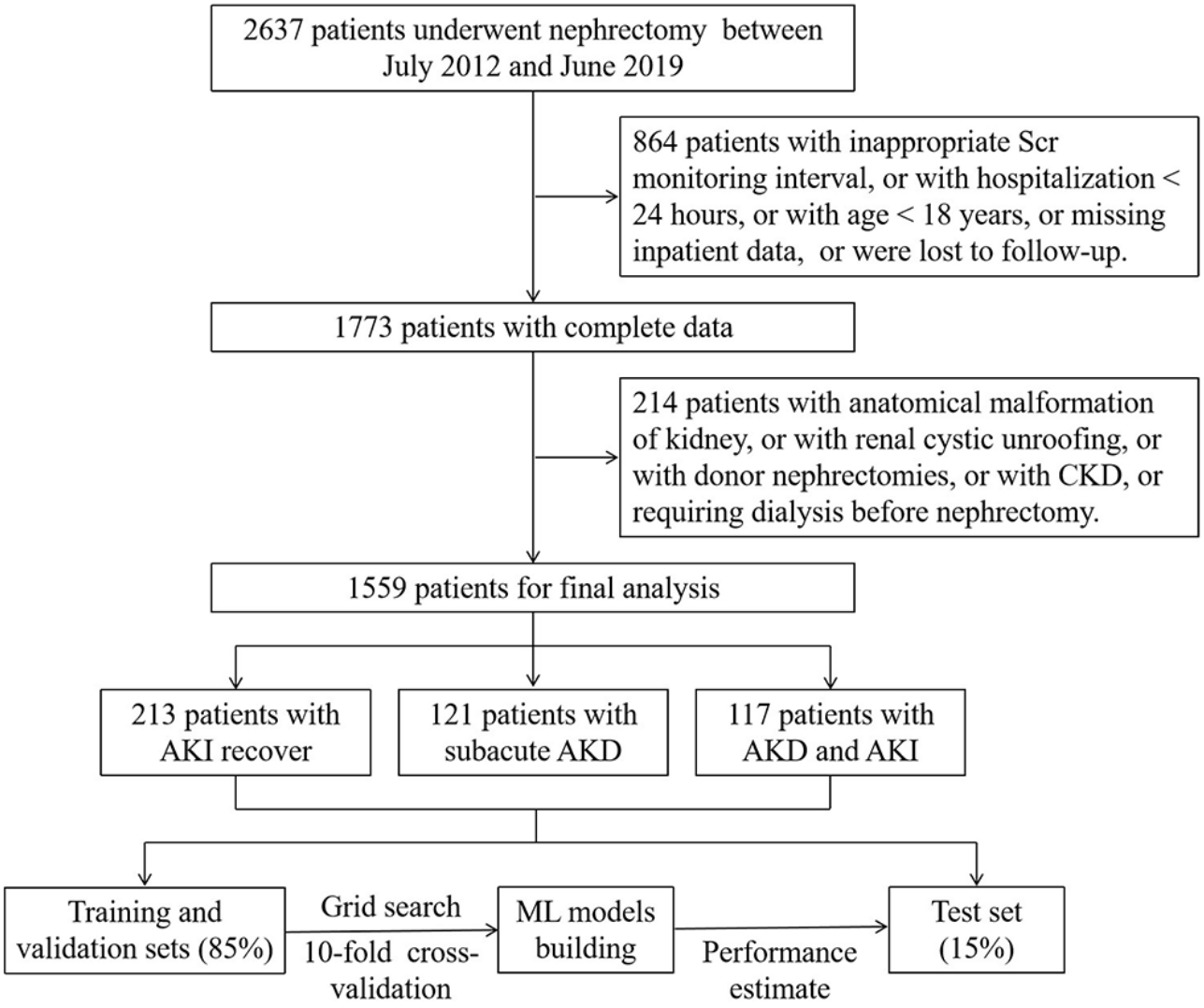


Table 1. Baseline characteristics of patients with and without CKD[a].

Features	Total (N=1559)	CKD-free (n=1394)	CKD (n=165)	*P* value
Age (years), mean (SD)	59.1 (12.5)	58 (12.3)	68.5 (9.6)	<.001
Male, n (%)	914 (58.6)	796 (57.1)	118 (71.5)	<.001
BMI (kg/m^2), mean (SD)	25 (3.4)	24.9 (3.4)	25.7 (3.8)	.01
Systolic blood pressure (mm Hg), mean (SD)	131.9 (18.5)	130.9 (18.2)	139.8 (19.3)	<.001
Smokers, n (%)	570 (36.6)	490 (35.1)	80 (48.5)	<.001
Drinkers, n (%)	464 (29.8)	404 (29)	60 (36.4)	.06
Fever, n (%)	56 (3.6)	52 (3.7)	4 (2.4)	.53
Procedure, n (%)				
Radical nephrectomy	1131 (72.5)	988 (70.9)	143 (86.7)	<.001
Partial nephrectomy	428 (27.4)	406 (29.1)	22 (13.3)	—[b]
Approach, n (%)				
Laparotomy	1152 (73.9)	1038 (74.5)	114 (69.1)	.29
Laparoscopy	316 (20.3)	275 (19.7)	41 (24.8)	—
da Vinci surgery	91 (5.8)	81 (5.8)	10 (6.1)	—
Pathology, n (%)	1973 (126.6)	1721 (123.5)	252 (152.7)	—
Benign	771 (73. 9)	673 (48.3)	98 (59.4)	<.001
Malignant (nonclear)	431 (20.3)	375 (26.9)	56 (33.9)	—
Clear cell	357 (5.8)	346 (24.8)	11 (6.7)	—
Blood loss (mL), median (IQR)	50 (20-150)	50 (20-150)	100 (50-200)	<.001
Excision diameter (cm), median (IQR)	11 (7-13)	12 (10-14)	11 (6-13)	<.001
Operative duration (hours), median (IQR)	2.5 (2-3)	2.8 (2.3-3.3)	2.4 (2-3)	<.001
Laboratory tests				
White blood cell (×10^9/L), median (IQR)	6.1 (5.1-7.4)	6.6 (5.6-7.8)	6 (5-7.4)	<.001
Red blood cell (×10^{12}/L), mean (SD)	4.5 (0.6)	4.5 (0.6)	4.4 (0.6)	.01
Platelet (×10^9/L), median (IQR)	232 (193-278)	221 (190-262)	234 (193-280)	.02
Hemoglobin (g/L), mean (SD)	133.9 (20.4)	134.3 (20.2)	130 (21.3)	.01
Fibrinogen (g/L), median (IQR)	3 (2.6-3.6)	3.1 (2.7-3.7)	3 (2.5-3.6)	.01
Serum creatinine (μmol/L), median (IQR)	85 (73-97)	102 (91-121)	83 (72-95)	<.001
Blood urea nitrogen (mmol/L), median (IQR)	5.7 (4.7-6.8)	6.6 (5.6-7.9)	5.6 (4.6-6.7)	<.001
Uric acid (μmol/L), mean (SD)	316.5 (89.2)	312 (87.8)	354.2 (92.5)	<.001
Baseline estimated glomerular filtration rate (mL/min/1.73 m^2), mean (SD)	79.3 (18.7)	81.8 (17.4)	57.6 (15)	<.001
Alanine transaminase (U/L), median (IQR)	17 (13-24)	17 (13-22)	18 (13-24)	.19
Aspartate transaminase (U/L), median (IQR)	17 (14-20)	16 (13-19)	17 (14-20)	.11
Total bilirubin (μmol/L), median (IQR)	13.1 (10-17.7)	11.8 (9.4-15.5)	13.4 (10.1-17.9)	<.001
Alkaline phosphatase (U/L), median (IQR)	69 (57-84)	68 (57-78)	69 (57-85)	.23
Triglyceride (mmol/L), median (IQR)	1.1 (0.8-1.6)	1.2 (0.9-1.7)	1.1 (0.8-1.6)	.01
Low-density lipoprotein cholesterol (mmol/L), mean (SD)	2.9 (0.8)	2.9 (0.8)	2.8 (0.8)	.40
Albumin (g/L), mean (SD)	39.5 (5)	39.7 (5)	38.4 (5.1)	<.001
Blood glucose (mmol/L), median (IQR)	5.1 (4.6-5.8)	5.2 (4.8-6.2)	5.1 (4.6-5.8)	.01
Urinalysis, n (%)				

Features	Total (N=1559)	CKD-free (n=1394)	CKD (n=165)	*P* value
Protein	307 (19.7)	230 (16.5)	77 (46.7)	<.001
Glucose	206 (13.2)	177 (12.7)	29 (17.6)	.10
Hematuria	1002 (64.3)	857 (61.5)	145 (87.8)	<.001
Echocardiography, median (IQR)				
Ejection fraction	63 (61-65)	63 (61-65)	63 (61-65)	.62
Comorbidities, n (%)				
Diabetes mellitus	202 (13)	166 (11.9)	36 (21.8)	<.001
Coronary heart disease	124 (7.9)	94 (6.7)	30 (18.2)	<.001
Hypertension	488 (31.3)	396 (28.4)	92 (55.8)	<.001
Obesity	302 (19.4)	255 (18.3)	47 (28.5)	<.001
Medications, n (%)				
β-Blocker	630 (40.4)	557 (40)	73 (44.2)	.33
ACEI or ARB[c]	163 (10.5)	129 (9.2)	34 (20.6)	<.001
Calcium channel blocker	378 (24.2)	311 (22.3)	67 (40.6)	<.001
Antibiotics	1042 (66.8)	918 (65.8)	124 (75.1)	.02
Nonsteroidal anti-inflammatory drugs	416 (26.7)	367 (26.3)	49 (29.7)	.41
Diuretics	435 (27.9)	373 (26.8)	62 (37.6)	<.001
Trajectories of renal function, n (%)				
AKI[d] recover	213 (13.6)	189 (13.6)	24 (14.5)	<.001
Subacute AKD[e]	121 (7.8)	89 (6.4)	32 (19.4)	—
AKD with AKI	117 (7.5)	75 (5.4)	42 (25.4)	—
Outcome				
AKI, n (%)	330 (21.2)	264 (18.9)	66 (40)	<.001
AKD, n (%)	238 (15.3)	164 (11.8)	74 (44.8)	<.001
CKD, n (%)	165 (10.6)	0 (0)	165 (100)	<.001
Length of stay, median (IQR)	11 (9-13)	11 (9-13)	11 (9-14)	<.001

[a]CKD: chronic kidney disease.

[b]Not available.

[c]ACEI or ARB: angiotensin-converting enzyme inhibitor or angiotensin receptor blocker.

[d]AKI: acute kidney injury.

[e]AKD: acute kidney disease.

Model Performance

A comprehensive set of over 90 features was served as features for both AKD and CKD and were integrated into the ML models. Among the assessed ML models, the LightGBM model demonstrated superior performance (Figure S2 in Multimedia Appendix 1 and Tables S2 and S3 in Multimedia Appendix 2). In the test set, LightGBM achieved the highest AUROC of 0.97 for AKD and 0.96 for CKD prediction. The F_1-scores, 0.75 for AKD and 0.70 for CKD, indicate a balanced trade-off between precision and recall. Additionally, Brier score loss was maintained at 0.05 for both AKD and CKD predictions, demonstrating the model's impressive calibration. To create a user-friendly web-based calculator, we simplified the model by reducing the number of input features. The inclusion of the top 10 and top 5 features for the AKD and CKD models, respectively, negligibly affected the LightGBM model's AUROC (achieving 0.94 vs 0.97 for AKD prediction and 0.94 vs 0.96 for CKD prediction, as detailed in Figure S2 in Multimedia Appendix 1 and Table 2). Notably, it outperformed all other ML algorithms in terms of AUROC (Tables S4 and S5 in Multimedia Appendix 2), while maintaining an optimal balance between precision, recall, and error rates (both false positives and negatives). Subsequently, we used the LightGBM model for result interpretation and the development of a web-based calculator. Comprehensive insights into performance metrics and visualizations are provided in Figures S2 and S3 in Multimedia Appendix 1, Table 2, and Tables S2-S5 in Multimedia Appendix 2.

Table 2. Performance of Light Gradient-Boosting Machine models in predicting AKD[a] and CKD[b] on the test set.

Outcome	AUROC[c]	Precision	Recall	Accuracy	False positive rate	False negative rate	F_1-score	MCC[d]	BSL[e]
AKD									
Top 5 features	0.87	0.43	0.66	0.82	0.16	0.34	0.52	0.42	0.12
Top 10 features	0.94	0.67	0.80	0.91	0.07	0.20	0.73	0.68	0.07
Top 15 features	0.95	0.74	0.80	0.93	0.05	0.20	0.77	0.73	0.06
Top 20 features	0.95	0.71	0.71	0.92	0.05	0.29	0.71	0.66	0.06
All features	0.97	0.83	0.69	0.93	0.03	0.31	0.75	0.72	0.05
CKD									
Top 5 features	0.94	0.43	0.72	0.91	0.08	0.28	0.54	0.51	0.07
Top 10 features	0.93	0.43	0.67	0.91	0.07	0.33	0.52	0.49	0.07
Top 15 features	0.91	0.42	0.56	0.91	0.07	0.44	0.48	0.43	0.08
Top 20 features	0.92	0.44	0.61	0.91	0.07	0.39	0.51	0.47	0.07
All features	0.96	0.64	0.78	0.95	0.04	0.22	0.70	0.68	0.05

[a]AKD: acute kidney disease.

[b]CKD: chronic kidney disease.

[c]AUROC: area under the receiver operating characteristic curve.

[d]MCC: Matthew correlation coefficient.

[e]BSL: Brier score loss.

Model Interpretation

The SHAP summary plots of the LightGBM models are depicted in Figure 2. Operative duration, hemoglobin (Hb), blood loss, urine protein, and hematocrit were the top 5 features associated with predicted AKD. Baseline eGFR, pathology, trajectories of renal function, age, and total bilirubin were the top 5 features associated with predicted CKD. The SHAP interaction plots visually elucidate the interplays among the top 10 features in both the AKD and CKD models (Figure S4 in Multimedia Appendix 1). The SHAP dependence plots offer detailed insights into the correlations among the top 10 features, as depicted in Figures S6 and S7 in Multimedia Appendix 1, with representative examples showcased in Figure S5 in Multimedia Appendix 1. For instance, the influence of AKI grade on the probability of AKD varies across Hb levels. Among patients with lower Hb levels, higher AKI grades are associated with a significant increase in the risk of AKD. Conversely, this correlation is less pronounced in patients with higher Hb levels. For patients presenting with a baseline eGFR below 80, postoperative complications, such as AKD with AKI, subacute AKD, or AKI recover, markedly elevate the risk of developing CKD. This observation underscores the significance of encompassing factors like trajectories of renal function within a comprehensive clinical framework, particularly one that integrates a patient's eGFR.

Sample individualized predictions with their explanations are shown in Figure 3. The AKD and CKD models, respectively, present the top 10 and top 5 features. We selected 4 random samples from the test set and analyzed them using both the SHAP and LIME algorithms. For example, Figure 3D presents an individualized explanation for a case where the actual and predicted outcomes are both CKD. The notably high predicted probability for CKD (P=.97) primarily stemmed from several incremental factors, including a low baseline eGFR level (39.36 mL/min/1.73 m^2), postoperative complications of AKD with AKI, clear cell pathology, and a history of antibiotic use, despite a normal white blood cell level (3.94×10^9/L). The SHAP force plot revealed minor deviations in the top 5 features for predicting this patient, highlighting the greater significance of γ-glutamyl transferase over albumin-globulin (AG) ratio.

DAG is a type of causal diagram comprising nodes representing features and arrows representing causal relationships between the features. Since the importance of features does not necessarily reflect causality, we designated only AKD and CKD as end points (nodes with only inward-pointing arrows) in the DAGs, without designating source nodes (nodes with only outward-pointing arrows). Given that analyzing all features (over 90) would lead to an excessively complex causal structure, we limited the analysis to the top 20 features for AKD and CKD. During the investigation of AKD as the outcome, we observed direct links from features such as AKI grade, operative duration, systolic blood pressure, Hb, antibiotic, baseline eGFR, urine protein, and hematocrit to AKD, indicating potential direct causality (Figure S8 in Multimedia Appendix 1). All these features, except for antibiotics, are among the top 10 features for AKD prediction. We discovered that the trajectories of renal function, pathology, and baseline eGFR exhibit potential direct causal relationships with CKD, and they also rank among the top 5 features in the CKD model (Figure S9 in Multimedia Appendix 1). Age did not exert a direct effect on CKD but influenced it indirectly through its impact on pathology and baseline eGFR. Despite AG and procedure being within the top 10 features for CKD, our analysis did not reveal a causal link to CKD, suggesting that while there was a correlation between AG or procedure and CKD, they were causally independent.

Figure 2. SHAP summary plots of the top 10 features in the Light Gradient Boosting Machine model for (A and B) AKD and (C and D) CKD prediction. (A) The ranking of feature importance within the AKD prediction model. Features with higher mean absolute SHAP values signify increased predictive influence. (B) Each dot represents the SHAP value of a specific feature for an individual, with red and blue indicating high and low feature values, respectively. On the x-axis, a positive or negative SHAP value signifies that the feature positively or negatively influenced the AKD prediction for the individual. (C) The ranking of feature importance within the CKD prediction model. (D) The distribution of the impacts of the top 10 features on the CKD model output. AG: albumin-globulin; AKD: acute kidney disease; AKI: acute kidney injury; ALB: albumin; CKD: chronic kidney disease; eGFR: estimated glomerular filtration rate; GGT: γ-glutamyl transferase; Hb: hemoglobin; Hct: hematocrit; MPV: mean platelet volume; SBP: systolic blood pressure; SHAP: Shapley Additive Explanations; TBIL: total bilirubin; WBC: white blood cell.

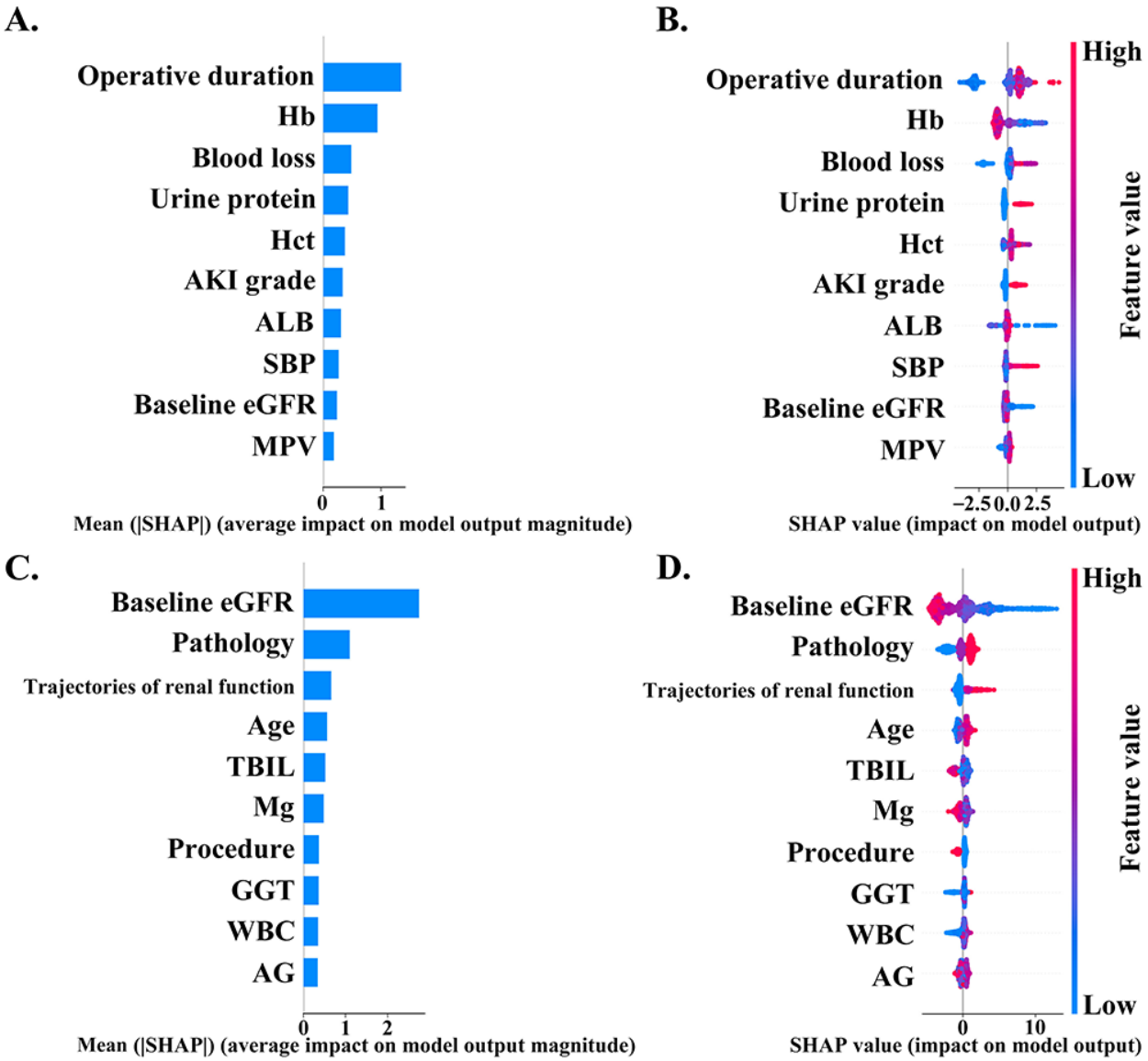
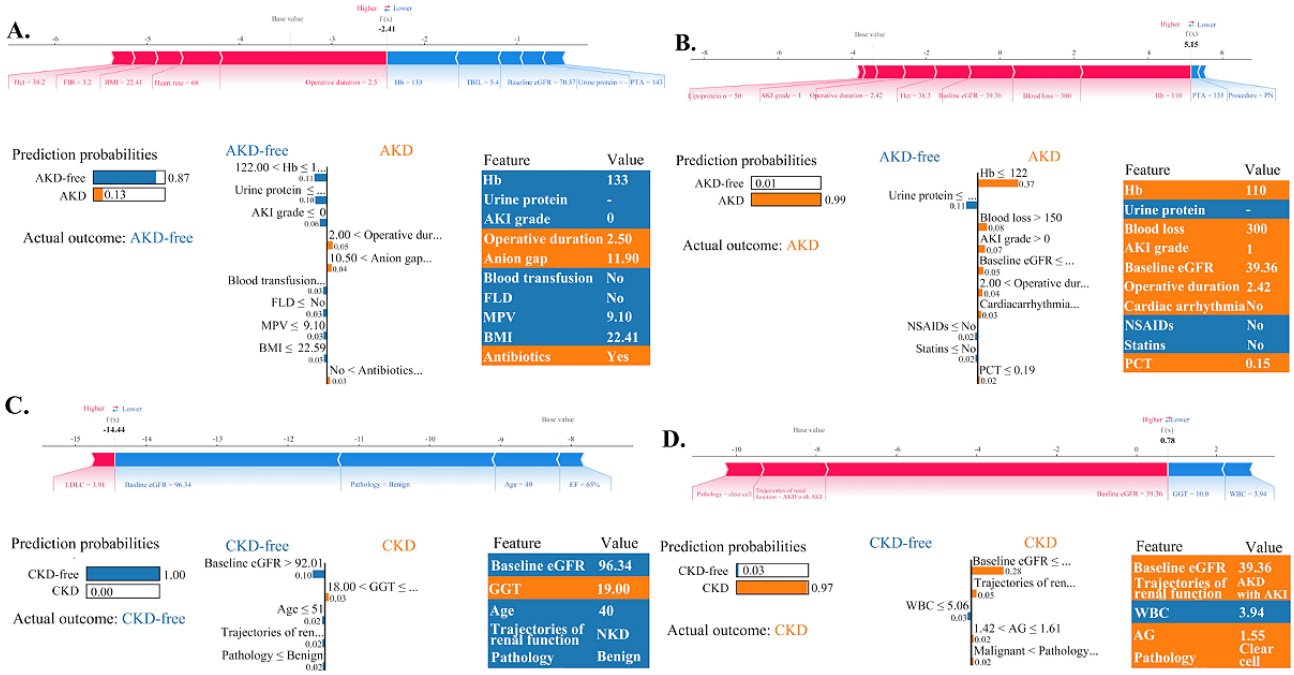


Figure 3. SHAP force plots and LIME bar plots for explaining individual predictions for (A and B) AKD and (C and D) CKD. (A) The SHAP force plot (upper section) and the LIME bar plot (lower section) are used to illustrate a case where both actual and predicted outcomes indicate AKD-free status. The SHAP force plot outlines the top 10 features for the prediction, where red feature values positively impact the AKD outcome, while blue values have a negative impact. The importance of each feature is reflected by the length of its corresponding arrow, with longer arrows highlighting more significant influences. In the LIME bar plot, the left section shows an 87% predicted probability of the patient being AKD-free. The central section lists the top 10 features for predicting AKD-free or AKD status, with the length of each bar indicating its importance. Blue bars indicate positive influences, whereas yellow bars signify negative impacts. The right panel presents the specific values at which these top 10 features have the most substantial impact on the AKD-free or AKD prediction. (B) The SHAP force plot and the LIME bar plot, emphasizing the top 10 features, depict a case where both actual and predicted outcomes align with AKD status. (C) The SHAP force plot and the LIME bar plot, emphasizing the top 5 features, depict a case where both actual and predicted outcomes align with CKD-free status. (D) The SHAP force plot and the LIME bar plot, emphasizing the top 5 features, depict a case where both actual and predicted outcomes align with CKD status. AG: albumin-globulin; AKD: acute kidney disease; AKI: acute kidney injury; ALB: albumin; CKD: chronic kidney disease; EF: ejection fraction; eGFR: estimated glomerular filtration rate; FIB: fibrinogen; FLD: fatty liver disease; GGT: γ-glutamyl transferase; Hb: hemoglobin; Hct: hematocrit; LDLC: low-density lipoprotein cholesterol; LIME: Local Interpretable Model-Agnostic Explanations; MPV: mean platelet volume; NKD: no kidney disease; NSAID: nonsteroidal anti-inflammatory drug; PCT: procalcitonin; PN: partial nephrectomy; PTA: prothrombin activity; SHAP: Shapley Additive Explanations; TBIL: total bilirubin; WBC: white blood cell.



Web-Based Calculator

Since the LightGBM model proved to be the most effective in our study, we developed a web-based calculator using the "Streamlit" application to predict both AKD and CKD with this model. Restricting the LightGBM model to only the top 10 and top 5 features did not diminish predictive performance for the AKD and CKD models (AUROC: 0.94 vs 0.97 in AKD prediction and 0.94 vs 0.96 in CKD prediction). For ease of use, we constructed a web-based calculator using the top 10 and top 5 features to predict AKD and CKD, respectively. You can access this calculator at Streamlit [30].

Discussion

Principal Findings

Our exploration into the use of ML techniques to predict and elucidate outcomes in patients undergoing nephrectomy was instigated by an amplified emphasis on the long-term renal functional prognosis, the accessibility of intricate data within the electronic health record system, and the maturation of interpretable predictive models. Among patients who underwent nephrectomy, 28.9% (n=451) developed AKI or AKD. Specifically, 7.5% (n=117) of patients developed AKD in conjunction with AKI, 13.7% (n=213) experienced recovery from AKI, and 7.8% (n=121) developed subacute AKD. The incidence rate of CKD was 10.6% (n=165). We formulated a diverse array of ML models with a focus on AKD and CKD prognosis. Among these models, LightGBM exhibited the most robust predictive prowess, achieving an AUROC of 0.97 for AKD prediction and 0.96 for CKD prediction. Our research used SHAP values and the LIME algorithm to interpret the decision-making process from both global and instance-based perspectives. Additionally, we used DAG to further visualize the potential causal relationships between features and outcomes. In consideration of clinical applicability, we further developed a web application that uses the final prediction model to estimate AKD and CKD risks.

Comparison to Prior Work and Implications

Assessment of renal injury risk following nephrectomy has predominantly concentrated on AKI and CKD, with limited attention to the recovery of renal function within 7-90 days post-AKI and its enduring consequences [5,31-33]. Our prior research has unveiled discernible distinctions in the predicted risk factors between AKI and AKD. Specifically, AKD is associated with a notably elevated risk of de novo CKD

development when contrasted with AKI [34]. This study encompassed all patients who were hospitalized, with no specific subgroup analysis conducted for those undergoing nephrectomy. Hu et al [35] initially developed a predictive model for postnephrectomy AKD, using an LR model to assess predicted risk factors associated with renal injury within 3 months following nephrectomy. Nevertheless, this study did not differentiate between AKI recover, subacute AKD, and AKD with AKI. Furthermore, it did not investigate the long-term outcomes for patients experiencing these distinct renal function trajectories. In our study, we found a significant association between trajectories of renal function and the onset and progression of CKD. Specifically, the coexistence of AKD with AKI led to a CKD incidence rate of 35.9% (n=42), nearly 1.5 times higher than that observed in patients with subacute AKD (AKD without AKI). Meanwhile, the CKD incidence rate was 11.3% (n=24) for individuals who had recovered from AKI and 6% (n=67) for those with normal kidney function. Among patients with kidney injury, nearly one-third experienced subacute AKD, which did not meet the criteria for either AKI or CKD diagnosis. These individuals are frequently overlooked in the early stages due to the modest changes in renal function they exhibit; however, their risk of developing CKD is significantly elevated when compared to both patients with NKD and those who have recovered from AKI. As such, the presence of AKD serves as a critical link between AKI and CKD, aiding in the assessment of declining renal function and prognosis.

Currently, LR is the most widely used model for predicting kidney injury risk in patients undergoing nephrectomy, with limited application of ML algorithms [36,37]. Lee et al [38] used various ML algorithms to formulate a risk prediction model for AKI after nephrectomy, identifying that the LightGBM model outperforms others in terms of predictive accuracy. Compared to LR, LightGBM demonstrates enhanced speed, more efficient memory use, and superior parallel processing capabilities, which allow it to more effectively manage nonlinear relationships, large datasets, and high-dimensional data [39]. Our study has undertaken a thorough evaluation of the predictive abilities of several ML models, with LightGBM emerging as the most effective in forecasting high-risk AKD and CKD cases, alongside precisely pinpointing individual predicted risk factors. Early alerts assist in promptly notifying clinicians to undertake vigilant monitoring of patients at high risk. Addressing manageable predicted risk factors early, such as Hb, systolic blood pressure, and total bilirubin, presents a considerable opportunity to lower the occurrence of AKD and CKD, thereby enhancing patient outcomes.

Given our emphasis on interpretability, our methodology entails a thorough interpretation of the entire predictive algorithm, exploring the potential causal relationships between major features. First, we generate global-level diagrams that elucidate the contributions of each feature to the model's output along with interactions among key features. Features denoting acute injury, such as surgical factors and AKI grade, exert a significant influence on AKD. Baseline eGFR and trajectories of renal function constitute pivotal features affecting CKD. Features such as advanced age or clear cell carcinoma may be associated with an elevated CKD risk. While these attributes are generally nonmodifiable, augmenting the frequency of follow-up visits for individuals with these characteristics can effectively facilitate the early detection of renal function deterioration. Second, this study delineates the decision-making process for each patient. The examples depicted in Figure 3 elucidate the predominant feature compositions among patients exhibiting diverse predicted probabilities of AKD or CKD. Using SHAP force plots and LIME plots amplifies the individualization and transparency of the decision-making process, thereby alleviating the black-box issue inherent in the model's prediction process. Finally, DAGs were used to delve deeper into the potential causal relationships between features and outcomes. It was found that most of the top 10 features identified by SHAP values have the potential to directly or indirectly influence the occurrence of AKD or CKD.

For the sake of enhancing user convenience, we have developed web-based prediction tools for both AKD and CKD. Users can effortlessly input the values of their chosen features to calculate the probabilities of AKD and CKD following nephrectomy. Our research marks a pioneering effort in constructing web-based prediction tools for postnephrectomy AKD and CKD, which can assist clinicians in identifying high-risk individuals and risk factors. Given the clinical feasibility and straightforward accessibility of features derived from routine medical records, our models are eminently suitable for seamless integration into daily clinical practice.

Limitations and Future Directions

The study exhibits several limitations. First, the web-based prediction tool is crafted to assist clinicians in discerning individuals with elevated risk of AKD and CKD rather than serving as a replacement for clinical diagnosis. Due to the retrospective nature of data collection, it is crucial to undertake additional validation using an independent population to ensure robust predictive validity across diverse usage scenarios. Second, the collection of urine output data is subjective, and a significant number of values are missing. Consequently, this study refrained from using urine output as a diagnostic criterion for AKI. Third, our study lacks time-variant monitored values among its features. Moving forward, we intend to collect longitudinally monitored data from patients undergoing nephrectomy to enable dynamic prediction of AKD and CKD before their clinical identification. Finally, DAGs visually represent the potential causal relationships between features and outcomes. This underscores the need to further explore and quantify the causal mechanisms in future work.

Conclusions

This study has developed prediction models that accurately estimate the risk of AKD and CKD following nephrectomy. These models provide interpretability from both global and instance-based perspectives. We recommend the use of the AKD criterion in clinical practice due to its superior accuracy in predicting prognosis, particularly the development of CKD.

Acknowledgments

This work was supported by the Taishan Scholar Program of Shandong Province (tstp20230665), the National Natural Science Foundation of China (grants 81970582 and 82270724), the Qingdao Key Health Discipline Development Fund, and the Qingdao Key Clinical Specialty Elite Discipline.

Data Availability

The datasets analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions

LX and CL contributed equally to this work and should be considered co–first authors. LX was involved in the study's design, manuscript drafting, statistical analysis, and manuscript revision. CL and YX contributed to the study's design, manuscript drafting, and manuscript revision. SG participated in the study's design, statistical analysis, and manuscript revision. L Zhao, L Zhang, and CY were responsible for the statistical analysis. C Guan, C Guo, XS, and ZZ were contributed toward the use of artificial intelligence. QB and BZ were responsible for data collection and interpretation. All authors critically reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quantile-quantile plots, model performance, Shapley Additive Explanations plots, and directed acyclic graph of the study.
[DOCX File , 4052 KB - medinform_v12i1e52837_app1.docx]

Multimedia Appendix 2
Descriptive statistics for the cohorts; performance of models.
[DOCX File , 46 KB - medinform_v12i1e52837_app2.docx]

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021 Jan;71(1):7-33 [FREE Full text] [doi: 10.3322/caac.21654] [Medline: 33433946]
2. Li K, Chen S, Wang C, Yang L. Comparison between minimally invasive partial nephrectomy and open partial nephrectomy for complex renal tumors: a systematic review and meta-analysis. Int J Surg 2023;109(6):1769-1782 [FREE Full text] [doi: 10.1097/JS9.0000000000000397] [Medline: 37094827]
3. Hora M, Albiges L, Bedke J, Campi R, Capitanio U, Giles RH, et al. European Association of Urology Guidelines Panel on renal cell carcinoma update on the new World Health Organization classification of kidney tumours 2022: the urologist's point of view. Eur Urol 2023;83(2):97-100. [doi: 10.1016/j.eururo.2022.11.001] [Medline: 36435661]
4. Hsu RCJ, Barclay M, Loughran MA, Lyratzopoulos G, Gnanapragasam VJ, Armitage JN. Time trends in service provision and survival outcomes for patients with renal cancer treated by nephrectomy in England 2000-2010. BJU Int 2018;122(4):599-609 [FREE Full text] [doi: 10.1111/bju.14217] [Medline: 29603575]
5. Yang X, Zhang T, Zhou H, Ni Z, Wang Q, Wu J, et al. Acute kidney injury as an independent predicting factor for stage 3 or higher chronic kidney disease after nephrectomy. Urol Oncol 2023;41(3):149.e1-149.e9. [doi: 10.1016/j.urolonc.2022.10.011] [Medline: 36463084]
6. Bravi CA, Vertosick E, Benfante N, Tin A, Sjoberg D, Hakimi AA, et al. Impact of acute kidney injury and its duration on long-term renal function after partial nephrectomy. Eur Urol 2019;76(3):398-403 [FREE Full text] [doi: 10.1016/j.eururo.2019.04.040] [Medline: 31080127]
7. Chawla LS, Eggers PW, Star RA, Kimmel PL. Acute kidney injury and chronic kidney disease as interconnected syndromes. N Engl J Med 2014;371(1):58-66 [FREE Full text] [doi: 10.1056/NEJMra1214243] [Medline: 24988558]
8. Chawla LS, Bellomo R, Bihorac A, Goldstein SL, Siew ED, Bagshaw SM, et al. Acute kidney disease and renal recovery: consensus report of the Acute Disease Quality Initiative (ADQI) 16 workgroup. Nat Rev Nephrol 2017;13(4):241-257 [FREE Full text] [doi: 10.1038/nrneph.2017.2] [Medline: 28239173]
9. Nagata K, Horino T, Hatakeyama Y, Matsumoto T, Terada Y, Okuhara Y. Effects of transient acute kidney injury, persistent acute kidney injury and acute kidney disease on the long-term renal prognosis after an initial acute kidney injury event. Nephrology (Carlton) 2021;26(4):312-318. [doi: 10.1111/nep.13831] [Medline: 33207040]
10. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 2018;362(6419):1140-1144. [doi: 10.1126/science.aar6404] [Medline: 30523106]

11. Ebel P, Lingenfelder C, Vogelsang A. On the forces of driver distraction: explainable predictions for the visual demand of in-vehicle touchscreen interactions. Accid Anal Prev 2023;183:106956. [doi: 10.1016/j.aap.2023.106956] [Medline: 36681017]
12. Jannach D, Abdollahpouri H. A survey on multi-objective recommender systems. Front Big Data 2023;6:1157899 [FREE Full text] [doi: 10.3389/fdata.2023.1157899] [Medline: 37034435]
13. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
14. Menzies SW, Sinz C, Menzies M, Lo SN, Yolland W, Lingohr J, et al. Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial. Lancet Digit Health 2023;5(10):e679-e691 [FREE Full text] [doi: 10.1016/S2589-7500(23)00130-9] [Medline: 37775188]
15. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol 2022;38(2):204-213 [FREE Full text] [doi: 10.1016/j.cjca.2021.09.004] [Medline: 34534619]
16. Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical AI: can the black-box present deliver? Artif Intell Med 2022;124:102158. [doi: 10.1016/j.artmed.2021.102158] [Medline: 34511267]
17. Gupta R, Zhang L, Hou J, Zhang Z, Liu H, You S, et al. Review of explainable machine learning for anaerobic digestion. Bioresour Technol 2023;369:128468. [doi: 10.1016/j.biortech.2022.128468] [Medline: 36503098]
18. Palevsky PM, Liu KD, Brophy PD, Chawla LS, Parikh CR, Thakar CV, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for acute kidney injury. Am J Kidney Dis 2013;61(5):649-672. [doi: 10.1053/j.ajkd.2013.02.349] [Medline: 23499048]
19. Stevens PE, Levin A, Kidney Disease: Improving Global Outcomes Chronic Kidney Disease Guideline Development Work Group Members. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. Ann Intern Med 2013;158(11):825-830 [FREE Full text] [doi: 10.7326/0003-4819-158-11-201306040-00007] [Medline: 23732715]
20. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med 2009;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]
21. Yan J, Xu Y, Cheng Q, Jiang S, Wang Q, Xiao Y, et al. LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol 2021;22(1):271 [FREE Full text] [doi: 10.1186/s13059-021-02492-y] [Medline: 34544450]
22. Fu XY, Mao XL, Wu HW, Lin JY, Ma ZQ, Liu ZC, et al. Development and validation of LightGBM algorithm for optimizing of Helicobacter pylori antibody during the minimum living guarantee crowd based gastric cancer screening program in Taizhou, China. Prev Med 2023;174:107605 [FREE Full text] [doi: 10.1016/j.ypmed.2023.107605] [Medline: 37419420]
23. Adnan M, Alarood AAS, Uddin MI, Rehman IU. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. PeerJ Comput Sci 2022;8:e803 [FREE Full text] [doi: 10.7717/peerj-cs.803] [Medline: 35494796]
24. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
25. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. Comput Biol Med 2023;166:107555 [FREE Full text] [doi: 10.1016/j.compbiomed.2023.107555] [Medline: 37806061]
26. Lipsky AM, Greenland S. Causal directed acyclic graphs. JAMA 2022;327(11):1083-1084. [doi: 10.1001/jama.2022.1816] [Medline: 35226050]
27. Peng Z, Apfelbacher C, Brandstetter S, Eils R, Kabesch M, Lehmann I, et al. Directed acyclic graph for epidemiological studies in childhood food allergy: construction, user's guide, and application. Allergy 2024;79(8):2051-2064. [doi: 10.1111/all.16025] [Medline: 38234010]
28. Haque UM, Kabir E, Khanam R. Early detection of paediatric and adolescent obsessive-compulsive, separation anxiety and attention deficit hyperactivity disorder using machine learning algorithms. Health Inf Sci Syst 2023;11(1):31 [FREE Full text] [doi: 10.1007/s13755-023-00232-z] [Medline: 37489154]
29. Ren Y, Wu D, Tong Y, López-DeFede A, Gareau S. Issue of data imbalance on low birthweight baby outcomes prediction and associated risk factors identification: establishment of benchmarking key machine learning models with data rebalancing strategies. J Med Internet Res 2023;25:e44081 [FREE Full text] [doi: 10.2196/44081] [Medline: 37256674]
30. Streamlit. URL: https://xuly94-akd-app-app-ma1epw.streamlit.app/ [accessed 2024-09-05]
31. Plamm A, Vijayan M, Marinelli B, Vassalotti JA, Winston J, Rein JL. Acute kidney injury from post-nephrectomy renal cell carcinoma thrombus. Kidney Int 2022;102(6):1431. [doi: 10.1016/j.kint.2022.08.034] [Medline: 36411024]
32. Wang S, Liu Z, Zhang D, Xiang F, Zheng W. The incidence and risk factors of chronic kidney disease after radical nephrectomy in patients with renal cell carcinoma. BMC Cancer 2022;22(1):1138 [FREE Full text] [doi: 10.1186/s12885-022-10245-8] [Medline: 36335288]

33. Chae D, Kim NY, Kim KJ, Park K, Oh C, Kim SY. Predictive models for chronic kidney disease after radical or partial nephrectomy in renal cell cancer using early postoperative serum creatinine levels. J Transl Med 2021;19(1):307 [FREE Full text] [doi: 10.1186/s12967-021-02976-2] [Medline: 34271916]
34. Xu L, Li C, Li N, Zhao L, Zhu Z, Zhang X, et al. Incidence and prognosis of acute kidney injury versus acute kidney disease among 71 041 inpatients. Clin Kidney J 2023;16(11):1993-2002 [FREE Full text] [doi: 10.1093/ckj/sfad208] [Medline: 37915910]
35. Hu XY, Liu DW, Qiao YJ, Zheng X, Duan JY, Pan SK, et al. Development and validation of a nomogram model to predict acute kidney disease after nephrectomy in patients with renal cell carcinoma. Cancer Manag Res 2020;12:11783-11791 [FREE Full text] [doi: 10.2147/CMAR.S273244] [Medline: 33235506]
36. Hua YB, Li X, Wang DX. Prevalence and risk factors of myocardial and acute kidney injury following radical nephrectomy with vena cava thrombectomy: a retrospective cohort study. BMC Anesthesiol 2021;21(1):243 [FREE Full text] [doi: 10.1186/s12871-021-01462-y] [Medline: 34641781]
37. Hu J, Jin D, Fan R, Xie X, Zhou Z, Chen Y, et al. The relationships of acute kidney injury duration and severity with long-term functional deterioration following partial nephrectomy. Int Urol Nephrol 2022;54(7):1623-1628. [doi: 10.1007/s11255-021-03033-z] [Medline: 34718932]
38. Lee Y, Ryu J, Kang MW, Seo KH, Kim J, Suh J, et al. Machine learning-based prediction of acute kidney injury after nephrectomy in patients with renal cell carcinoma. Sci Rep 2021 Aug 03;11(1):15704. [doi: 10.1038/s41598-021-95019-1] [Medline: 34344909]
39. Rufo DD, Debelee TG, Ibenthal A, Negera WG. Diagnosis of diabetes mellitus using Gradient Boosting Machine (LightGBM). Diagnostics (Basel) 2021 Sep 19;11(9):1714 [FREE Full text] [doi: 10.3390/diagnostics11091714] [Medline: 34574055]

Abbreviations

AG: albumin-globulin
AKD: acute kidney disease
AKI: acute kidney injury
AUROC: area under the receiver operating characteristic curve
CKD: chronic kidney disease
DAG: directed acyclic graph
eGFR: estimated glomerular filtration rate
Hb: hemoglobin
LightGBM: Light Gradient-Boosting Machine
LIME: Local Interpretable Model-Agnostic Explanations
LR: logistic regression
ML: machine learning
NKD: no kidney disease
PN: partial nephrectomy
RF: random forest
RN: radical nephrectomy
Scr: serum creatinine
SHAP: Shapley Additive Explanations

Edited by C Lovis; submitted 18.09.23; peer-reviewed by X Liu, AF Näher, N Koizumi; comments to author 30.01.24; revised version received 08.04.24; accepted 21.07.24; published 20.09.24.

Please cite as:
Xu L, Li C, Gao S, Zhao L, Guan C, Shen X, Zhu Z, Guo C, Zhang L, Yang C, Bu Q, Zhou B, Xu Y
Personalized Prediction of Long-Term Renal Function Prognosis Following Nephrectomy Using Interpretable Machine Learning Algorithms: Case-Control Study
JMIR Med Inform 2024;12:e52837
URL: https://medinform.jmir.org/2024/1/e52837
doi:10.2196/52837
PMID:

Original Paper

Medication Prescription Policy for US Veterans With Metastatic Castration-Resistant Prostate Cancer: Causal Machine Learning Approach

Deepika Gopukumar[1,2*], PhD; Nirup Menon[3*], PhD; Martin W Schoen[2,4*], MD, MPH

[1]Richard A Chaifetz School of Business, Saint Louis University, St. Louis, MO, United States

[2]School of Medicine, Saint Louis University, St. Louis, MO, United States

[3]Costello College of Business, George Mason University, Fairfax, VA, United States

[4]St Louis Veteran Affairs Medical Center, St. Louis, MO, United States

[*]all authors contributed equally

Corresponding Author:
Deepika Gopukumar, PhD
Richard A Chaifetz School of Business
Saint Louis University
3674 Lindell Blvd, 417
St. Louis, MO, 63108
United States
Phone: 1 314 977 3600
Email: deepika.gopukumar.1@slu.edu

Abstract

Background: Prostate cancer is the second leading cause of death among American men. If detected and treated at an early stage, prostate cancer is often curable. However, an advanced stage such as metastatic castration-resistant prostate cancer (mCRPC) has a high risk of mortality. Multiple treatment options exist, the most common included docetaxel, abiraterone, and enzalutamide. Docetaxel is a cytotoxic chemotherapy, whereas abiraterone and enzalutamide are androgen receptor pathway inhibitors (ARPI). ARPIs are preferred over docetaxel due to lower toxicity. No study has used machine learning with patients' demographics, test results, and comorbidities to identify heterogeneous treatment rules that might improve the survival duration of patients with mCRPC.

Objective: This study aimed to measure patient-level heterogeneity in the association of medication prescribed with overall survival duration (in the form of follow-up days) and arrive at a set of medication prescription rules using patient demographics, test results, and comorbidities.

Methods: We excluded patients with mCRPC who were on docetaxel, cabaxitaxel, mitoxantrone, and sipuleucel-T either before or after the prescription of an ARPI. We included only the African American and white populations. In total, 2886 identified veterans treated for mCRPC who were prescribed either abiraterone or enzalutamide as the first line of treatment from 2014 to 2017, with follow-up until 2020, were analyzed. We used causal survival forests for analysis. The unit level of analysis was the patient. The primary outcome of this study was follow-up days indicating survival duration while on the first-line medication. After estimating the treatment effect, a prescription policy tree was constructed.

Results: For 2886 veterans, enzalutamide is associated with an average of 59.94 (95% CI 35.60-84.28) more days of survival than abiraterone. The increase in overall survival duration for the 2 drugs varied across patient demographics, test results, and comorbidities. Two data-driven subgroups of patients were identified by ranking them on their augmented inverse-propensity weighted (AIPW) scores. The average AIPW scores for the 2 subgroups were 19.36 (95% CI –16.93 to 55.65) and 100.68 (95% CI 62.46-138.89). Based on visualization and *t* test, the AIPW score for low and high subgroups was significant (*P*=.003), thereby supporting heterogeneity. The analysis resulted in a set of prescription rules for the 2 ARPIs based on a few covariates available to the physicians at the time of prescription.

Conclusions: This study of 2886 veterans showed evidence of heterogeneity and that survival days may be improved for certain patients with mCRPC based on the medication prescribed. Findings suggest that prescription rules based on the patient characteristics, laboratory test results, and comorbidities available to the physician at the time of prescription could improve survival by providing personalized treatment decisions.

(JMIR Med Inform 2024;12:e59480) doi:10.2196/59480

KEYWORDS

prostate cancer; metastatic castration resistant prostate cancer; causal survival forest; machine learning; heterogeneity; prescription policy tree; oncology; pharmacology

Introduction

Prostate cancer is the second leading cause of cancer death among men in the United States and the most common cancer affecting men of African descent [1]. While early-stage prostate cancer is curable, around 10% to 50% of cases progress to metastatic castrate-resistant prostate cancer (mCRPC) within 3 years of diagnosis, which is fatal [2]. Multiple therapies exist for treating mCRPC, including androgen-receptor pathway inhibitors (ARPIs) and cytotoxic chemotherapy, such as docetaxel. ARPIs, such as enzalutamide or abiraterone, target the androgen receptor signaling pathways and are administered orally [3]. Enzalutamide is administered at 160 mg orally once daily [4]. In comparison, abiraterone acetate is administered while fasting at a dose of 1000 mg orally with coadministration of the steroid prednisone [3]. Recent prescription trends show that ARPIs are preferred for their tolerable safety profiles and improving survival [5,6]. Prescribing ARPIs instead of docetaxel might also be beneficial for African American men, though the precise mechanisms are yet known [7].

Abiraterone and enzalutamide have different mechanisms of action. Abiraterone requires coadministration of prednisone because it inhibits androgen biosynthesis, leading to mineralocorticoid excess. In contrast, enzalutamide inhibits the androgen receptor by blocking hormone signaling and does not require steroid coadministration. These differences result in varying patient outcomes and survival based on patient demographics, test results, and comorbidities [8]. Clinical trials often do not encompass the full spectrum of patient comorbidities [9]. In the absence of clear clinical evidence favoring one ARPI over the other, retrospective observational data and data analytics techniques can help determine the most suitable drug for individual patients [10].

The methods used to study outcomes using ARPIs included associative and predictive modeling. Association studies identified a few comorbidities, such as cardiovascular diseases and diabetes, as significant predictors of survival in patients with mCRPC and comorbid diseases [10]. In addition, studies of hospitalizations during ARPI treatment showed an increased risk of heart failure, atrial fibrillation, and acute kidney injury with abiraterone [11]. Enzalutamide was found to be better than abiraterone for survival as well [10,12].

Machine learning predictive models effectively predicted survival outcomes and time to treatment discontinuation using tree-based approaches while incorporating laboratory test results such as hemoglobin and albumin and comorbidities such as hypertension and diabetes [13-15]. Multiomic features combined with treatment lines predicted the response types “good, poor, and ambiguous” following ARPI treatment [16]. However, these studies did not identify patient-specific differences in outcomes (heterogeneous treatment effects in subgroups of patients), address selection bias in treatment, or prescribe specific ARPIs based on patient demographics, laboratory test results, and comorbidities.

The main objective of this study was to evaluate patient-level heterogeneity in the association between prescribed medication and overall survival duration in the form of follow-up days and subsequently to identify a set of prescription rules based on patient-specific factors. We used a causal survival forest to measure heterogeneous treatment effects among patients with mCRPC, focusing on the survival duration. Based on these findings, we developed a set of treatment rules, or a policy tree, to prescribe abiraterone and enzalutamide tailored to individual patient factors such as demographics, test results, and comorbidities.

Methods

Dataset and Its Description

The data used for analysis were from the United States Department of Veterans Affairs (VA) centers (compassing VA hospitals and clinics) stored in the Corporate Data Warehouse. A total of 3675 veterans from Corporate Data Warehouse were identified as patients with mCRPC, excluding those with missing values and including only the African American and White populations. We excluded patients with mCRPC who received docetaxel, cabazitaxel, mitoxantrone, or sipuleucel-T before or after starting abiraterone or enzalutamide. This exclusion resulted in a dataset of 2890 veterans who began treatment with abiraterone or enzalutamide between 2014 and 2017. In addition, 4 patients were excluded because they switched to different second-line treatments on the same day. Thus, the final analysis included 2886 patients. The study follow-up concluded in 2020.

Patient Demographics, Test Results, and Comorbidities

In total, 20 covariates’ age, creatinine clearance test result category, albumin result category, bilirubin result category, hemoglobin result category, race, prostate-specific antigen (PSA) test, BMI category, diabetes, hypertension, kidney disease, osteoporosis, fall, fatigue, abnormal gait, peripheral neuropathy, Parkinson’s disease, vision, orchiectomy procedure, and cardiovascular diseases (CVDs) were used. The laboratory values were calculated before treatment and were closest to the start of the first-line treatment. Covariates, such as CVD and diabetes, have been previously used to study the survival of veterans with mCRPC in univariate studies [10]. Other covariates such as abnormal gait, peripheral neuropathy (a common side effect of cancer treatment), and vision are included to capture existing comorbidities better. CVD was calculated based on Charlson and Elixhauser indices for myocardial infarction, heart failure, cardiac arrhythmia, valvular disease, complicated hypertension, peripheral vascular disease, and cerebrovascular disease based on *ICD* (*International Classification of Diseases*) codes [10]. Other comorbidity-related

covariates were based on the standard VA-Frailty health deficits based on diagnosis codes. We used proxies for comorbidities not available in our dataset but have been studied in the past, such as alkaline phosphate level. For example, kidney disease is a proxy for alkaline phosphate levels [17].

Outcome

The primary outcome of our study was overall survival duration, measured in follow-up days from the initiation of first-line treatment with an ARPI (abiraterone or enzalutamide). For a subset of patients with mCRPC (1163 out of 2886) who received the other ARPI drug as second-line treatment following the first-line therapy, we considered follow-up days to the initiation of the second-line treatment as survival duration. This approach assumed survival up to that point without accounting for follow-up days post-second-line treatment, aiming to mitigate treatment switching bias between therapies [18]. In addition, our data were limited to laboratory results and comorbidities recorded on the day of first-line treatment initiation.

Statistical Analysis

We used causal survival forests with 20 covariates to estimate heterogeneous treatment effects in follow-up days [19]. Causal forests and causal survival forests are similar to random forests and random survival forests, respectively, with the primary difference being their data splitting criteria [20]. In a random survival forest, data splitting minimizes differences in prediction errors in follow-up days within each group, considering censoring. In contrast, a causal survival forest splits data to maximize heterogeneity, that is, the difference in the estimated follow-up days between the treated and untreated groups, accounting for censoring. The advantage of using a causal survival forest lies in its robustness to censoring, unlike a causal forest [19]. In addition, compared with a random survival forest, a causal survival forest offers a more comprehensive assessment of heterogeneity. It can outperform linear regression models by measuring heterogeneous treatment effects conditional on a nonlinear function of many covariates, accounting for higher-order interactions [21].

Each observation in the dataset corresponds to a single patient with mCRPC and includes the patient's covariate data, follow-up days (the outcome), treatment assignment (1 for enzalutamide, 0 for abiraterone), and death (event). Survival duration was analyzed based on either abiraterone or enzalutamide treatment without an untreated control group. Consequently, the dataset does not allow for the determination of the individual treatment effects of each drug.

As this is a retrospective observational study, we must account for the nonrandomization of treatment assignment and use the augmented inverse-propensity weighting (AIPW) estimator, a doubly robust method. To do this, the true treatment assignment was fitted as a function of the observed covariates. The predicted value from this model provides a propensity score, that is, an estimate of the probability of treatment assignment conditioned on a set of covariates for each patient. Then, 2 models that estimate the outcome (follow-up) were fitted, one using enzalutamide and the other using abiraterone. Each outcome was then weighted by the estimated propensity score, which yields the weighted average of the two outcome models [22].

We used a causal survival forest for our analysis [19]. To ensure that the assumptions of finite horizon, ignorability, overlap, ignorable censoring, and positivity are satisfied so that the treatment effect of the drugs is identified when using causal survival forests, we made the following analytical choices. The parameter horizon, referring to restricted mean survival time, was set to a threshold so that the estimated censoring probabilities are not below 0.2, satisfying the finite horizon and positivity assumption. We started with a horizon of 2000 days (approximately close to the maximum follow-up days) and decremented by 100 to 1000.

For ignorability, we applied balanced diagnostics by checking the weighted absolute standardized mean difference (ASMD) of variables between patients treated with enzalutamide and patients treated with abiraterone. A number close to 0 indicates that the propensity scores are well-calibrated. We examined the propensity score distribution for both groups to test for overlap and checked if they clustered at 0 or 1.

The honesty fraction was set to 0.7, meaning 70% of the subsample was used for splitting and 30% for populating the leaf nodes. However, for ranking the observations based on covariates and coming up with subgroups, the model must not be fit using the observations being compared. We used 10-fold cross-fitting for this reason. The conditional average treatment effects (CATE) models were fit for 9 folds, and the "unseen" observations in the left-out fold were ranked based on their predictions and split along the median into 2 groups. The same process was repeated 10 times, with each fold serving as a left-out fold once.

We set the number of trees to 15,000. We generated the policy tree after calculating the AIPW scores. A policy tree is a set of treatment rules, that is, rule-based policies, in the form of a decision tree that physicians can use to prescribe abiraterone and enzalutamide. The policy tree was generated using 70% of the training data, and the policy's value was determined using the 30% test data. The policy value is defined as the average difference in follow-up days obtained if the patients with mCRPC are administered enzalutamide or abiraterone. R (version 4.4.1; The R Foundation) with packages *grf* and *policytree* were used for analyses.

Results

We used a causal survival forest to estimate the heterogeneity in the treatment effect on survival for the 20 covariates (refer to Table 1 for the descriptive statistics and description for all patients including based on enzalutamide and abiraterone).

Table 1. Descriptive statistics of metastatic castration-resistant prostate cancer US veterans with androgen receptor pathway inhibitors treatment initiated 2014-2017 (Total sample N=2886; abiraterone n=1649; enzalutamide n=1237).

Predictor and categories	All (entire sample)	Abiraterone (Treatment=0)	Enzalutamide (Treatment=1)
Race, n (%)			
Black	691 (23.94)	380 (23.04)	311 (25.14)
White	2195 (76.06)	1269 (76.96)	926 (74.86)
Age, years			
Minimum-maximum	51-90	51-90	53-90
Median (IQR)	78 (69-84)	78 (69-84)	78 (70-84)
Mean (SD)	77 (9)	77 (9)	77 (9)
PSA[a] test result			
Minimum-maximum	0-7289	0-7289	0-3846
Median (IQR)	32 (10-100)	33 (10-105)	31 (9-90)
Mean (SD)	138 (399)	147 (449)	127 (322)
Creatinine clearance category, n (%)			
≥30	2748 (95.22)	1561 (94.66)	1187 (95.96)
<30	138 (4.78)	88 (5.34)	50 (4.04)
Albumin category, n (%)			
≥3	2723 (94.35)	1560 (94.6)	1163 (94.02)
<3	163 (5.65)	89 (5.4)	74 (5.98)
Bilirubin category, n (%)			
<2	2872 (99.51)	1643 (99.64)	1229 (99.35)
≥2	14 (0.49)	6 (0.36)	8 (0.65)
Hemoglobin category, n (%)			
≥10	2507 (86.87)	1433 (86.9)	1074 (86.82)
<10	379 (13.13)	216 (13.1)	163 (13.18)
BMI category, n (%)			
<18.5	70 (2.43)	38 (2.3)	32 (2.59)
18.5-24.9	791(27.41)	461 (27.96)	330 (26.68)
25-29.9	1055 (36.56)	611 (37.05)	444 (35.89)
≥30	970 (33.61)	539 (32.69)	431 (34.84)
Kidney disease, n (%)			
No	1619 (56.10)	962 (58.34)	657 (53.11)
Yes	1267 (43.9)	687 (41.66)	580 (46.89)
Osteoporosis, n (%)			
No	2618 (90.71)	1491 (90.42)	1127 (91.11)
Yes	268 (9.29)	158 (9.58)	110 (8.89)
Fall, n (%)			
No	2750 (95.29)	1570 (95.21)	1180 (95.39)
Yes	136 (4.71)	79 (4.47)	57 (4.61)
Fatigue, n (%)			
No	2422 (83.92)	1383 (83.87)	1039 (83.99)
Yes	464 (16.08)	266 (16.13)	198 (16.01)
Abnormal gait, n (%)			

Predictor and categories	All (entire sample)	Abiraterone (Treatment=0)	Enzalutamide (Treatment=1)
No	2345 (81.25)	1346 (81.63)	999 (80.76)
Yes	541 (18.75)	303 (18.37)	238 (19.24)
Parkinson disease, n (%)			
No	2834 (98.2)	1623 (98.42)	1211 (97.9)
Yes	52 (1.8)	26 (1.58)	26 (2.1)
Peripheral neuropathy, n (%)			
No	2682 (92.93)	1540 (93.39)	1142 (92.32)
Yes	204 (7.07)	109 (6.61)	95 (7.68)
Vision comorbidity, n (%)			
No	2127 (73.7)	1239 (75.14)	888 (71.79)
Yes	759 (26.3)	410 (24.86)	349 (28.21)
Orchiectomy, n (%)			
No	2881 (99.83)	1648 (99.94)	1233 (99.68)
Yes	5 (0.17)	1 (0.06)	4 (0.32)
Cardiovascular diseases, n (%)			
No	963 (33.37)	568 (34.45)	395 (31.93)
Yes	1923 (66.63)	1081 (65.55)	842 (68.07)
Hypertension, n (%)			
No	532 (18.43)	316 (19.16)	216 (17.46)
Yes	2354 (81.57)	1333 (80.84)	1021 (82.54)
Diabetes, n (%)			
No	1712 (59.32)	1038 (62.95)	674 (54.49)
Yes	1174 (40.68)	611 (37.05)	563 (45.51)

[a]PSA: prostate-specific antigen.

We found the horizon value of 1000 to be the most suitable, with the censoring probability estimates not going below 0.2, satisfying the positivity assumption. The weighted ASMD of covariates, along with their interactions, for the abiraterone and enzalutamide-treated groups is close to zero, satisfying the balance between the 2 groups (Figure 1). Thus, the propensity score estimates from causal survival forests were well-calibrated. The plot of the estimated propensity scores showed that the scores did not cluster at zero or one and were unimodal, exhibiting overlap (Figure 2).

Figure 1. Absolute standardized mean differences of covariates and their interactions between abiraterone and enzalutamide-treated groups. The covariates closeness to zero after adjusting for selection bias (the orange dots) indicates the balance between the 2 groups.

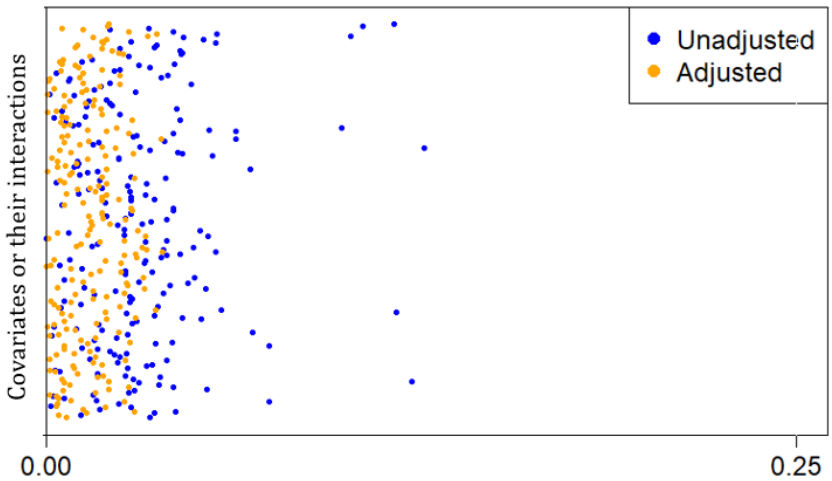
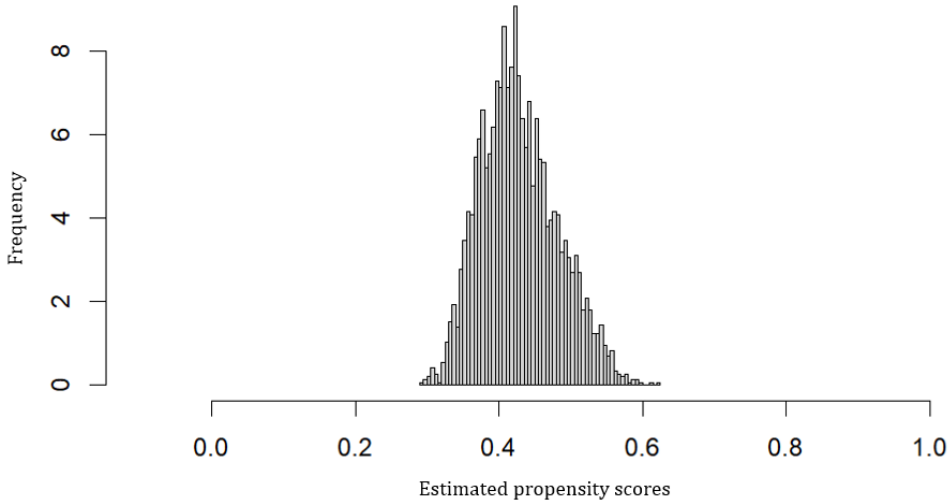
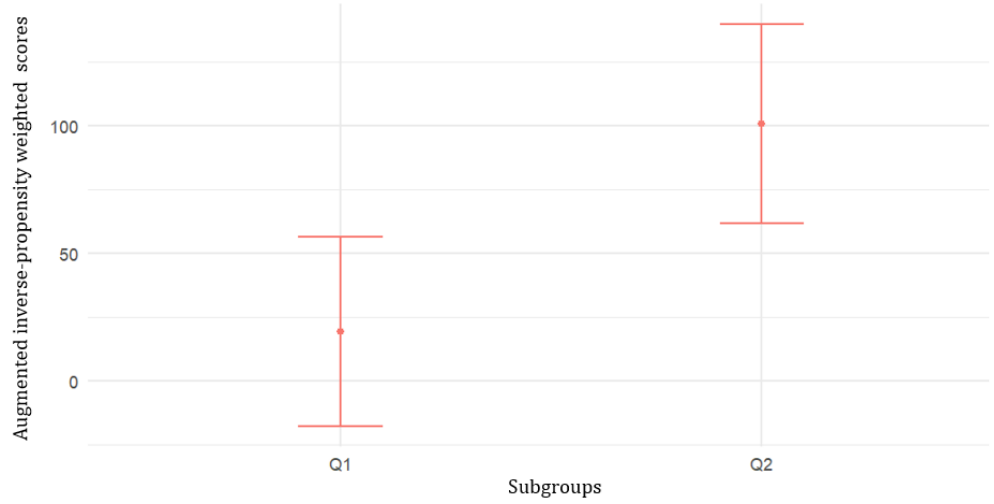


Figure 2. Estimated propensity scores from causal survival forest for metastatic castration-resistant prostate cancer US veterans with androgen receptor pathway inhibitors treatment initiated 2014-2017. The scores do not cluster at 0 or 1, indicating overlap.



For the 2886 veterans, enzalutamide is associated with an average of 59.94 (95% CI 35.6-84.28) more days of survival than abiraterone. The CATE scores computed from the causal survival forest were used to rank and split the observations into 2 subgroups, above and below the median [23]. The average AIPW scores for the 2 subgroups were 19.36 (95% CI -16.93 to 55.65) and 100.68 (95% CI 62.46-138.89). Based on visualization (Figure 3) and *t* test, the difference between the AIPW scores for low and high subgroups was significant (*P*=.003), thereby supporting heterogeneity.

Figure 3. Augmented inverse-propensity weighted scores (mean and 95% CI) for the 2 subgroups above and below the median for all metastatic castration-resistant prostate cancer US veterans with androgen receptor pathway inhibitors treatment initiated 2014-2017.



The heatmap in Figure 4 shows the average value of each covariate within each subgroup (refer to Table 2 for descriptive statistics for the 2 subgroups). The subgroups are ordered from lowest to highest difference in treatment effects; Q2, the second subgroup, has a higher treatment effect than Q1. The color in the heatmap of each covariate is the normalized distance of the average of the covariate in the subgroup from the average of the covariate in the full data. The covariates of the heatmap are arranged in descending order of variation, comparing the variance of the covariate in a subgroup with its variance in the sample. The top 5 covariates contributing to the highest variation were peripheral neuropathy, bilirubin category, osteoporosis, PSA test result, and Parkinson's disease.

Figure 4. Average covariate values for the 2 subgroups above and below the median augmented inverse-propensity weighted score for all metastatic castration-resistant prostate cancer US veterans with androgen receptor pathway inhibitors treatment initiated 2014-2017.

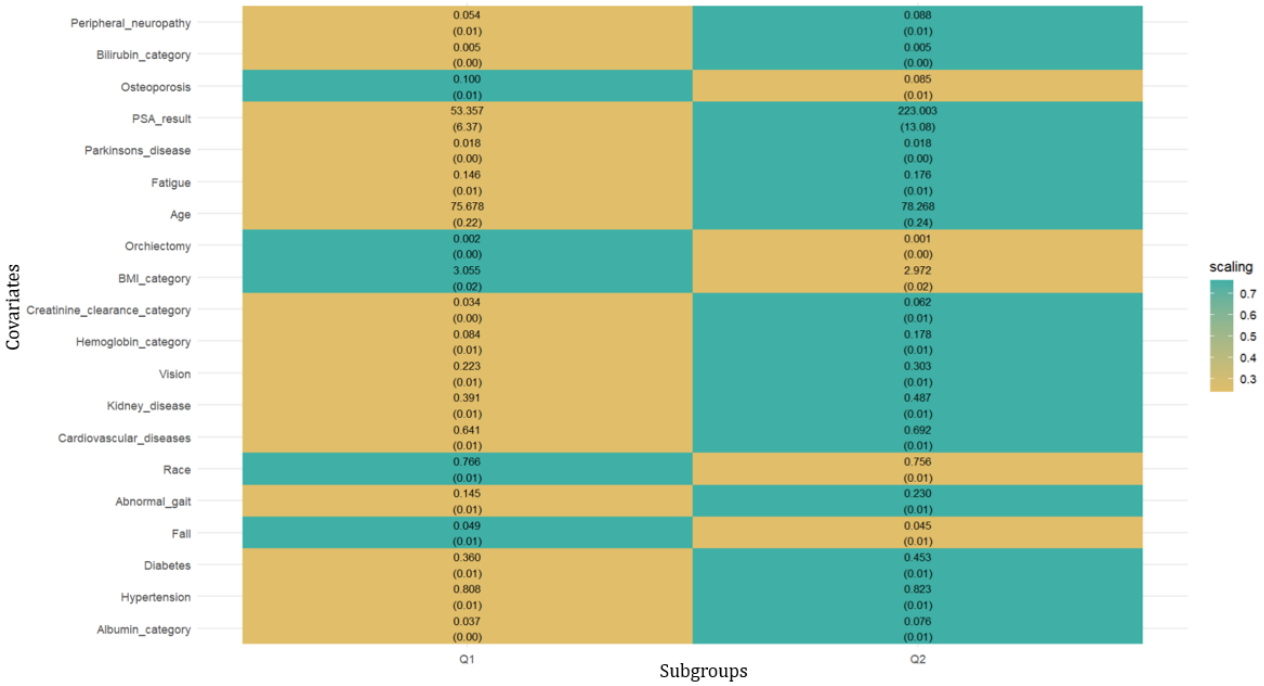


Table 2. Descriptive statistics for the subgroup of US veterans with metastatic castration-resistant prostate cancer below and above the median conditional average treatment effects (Total Sample N=2886).

Predictor and categories	Below median CATE[a] (Ranking=1)	Above median CATE[a] (Ranking=2)
Race, n (%)		
Black	339 (23.44%)	352 (24.44%)
White	1107 (76.56%)	1088 (75.56%)
Age		
Minimum-maximum	51-90	51-90
Median (IQR)	76 (69-83)	80 (71-86)
Mean (SD)	76 (8)	78 (9)
PSA[b] test result		
Minimum-maximum	0-7289	0-6240
Median (IQR)	17 (8-38)	89 (21-203)
Mean (SD)	53 (242)	223 (496)
Creatinine clearance category, n (%)		
≥30	1397 (96.61)	1351 (93.82)
<30	49 (3.34)	89 (6.18)
Albumin category, n (%)		
≥3	1392 (96.27)	1331 (92.43)
<3	54 (3.73)	109 (7.57)
Bilirubin category, n (%)		
<2	1439 (99.52)	1433 (99.51)
≥2	7 (0.48)	7 (0.49)
Hemoglobin category, n (%)		
≥10	1324 (91.56)	1183 (82.15)
<10)	122 (8.44)	257 (17.85)
BMI category, n (%)		
<18.5	36 (2.49)	34 (2.36)
18.5-24.9	374 (25.86)	417 (28.96)
25-29.9	510 (35.27)	545 (37.85)
≥30	526 (36.38)	444 (30.83)
Kidney disease, n (%)		
No	881 (60.93)	738 (51.25)
Yes	565 (39.07)	702 (48.75)
Osteoporosis, n (%)		
No	1301 (89.97)	1317 (91.46)
Yes	145 (10.03)	123 (8.54)
Fall, n (%)		
No	1375 (95.09)	1375 (95.49)
Yes	71 (4.91)	65 (4.51)
Fatigue, n (%)		
No	1235 (85.41)	1187 (82.43)
Yes	211 (14.59)	253 (17.57)

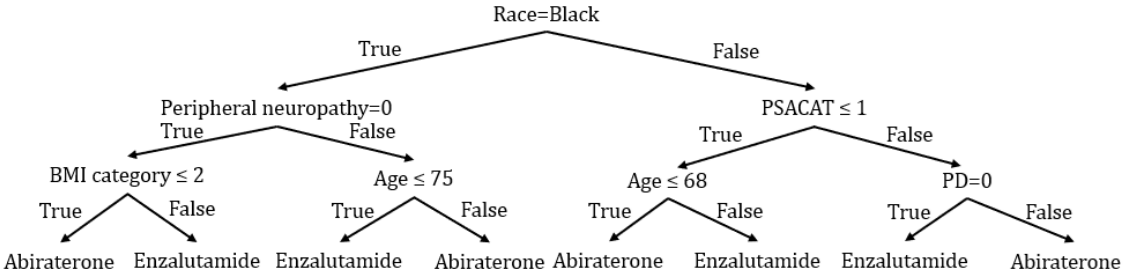
Predictor and categories	Below median CATE[a] (Ranking=1)	Above median CATE[a] (Ranking=2)
Abnormal gait, n (%)		
No	1236 (85.48)	1109 (77.01)
Yes	210 (14.52)	331 (22.99)
Parkinson disease, n (%)		
No	1420 (98.20)	1414 (98.19)
Yes	26 (1.80)	26 (1.81)
Peripheral neuropathy, n (%)		
No	1368 (94.61)	1314 (91.25)
Yes	78 (5.39)	126 (8.75)
Vision comorbidity, n (%)		
No	1124 (77.73)	1003 (69.65)
Yes	322 (22.27)	126 (8.75)
Orchiectomy, n (%)		
No	1443 (99.79)	1438 (99.86)
Yes	3 (0.21)	2 (0.14)
Cardiovascular diseases, n (%)		
No	519 (35.89)	444 (30.83)
Yes	927 (64.11)	996 (69.17)
Hypertension, n (%)		
No	277 (19.16)	255 (17.71)
Yes	1169 (80.84)	1185 (82.29)
Diabetes, n (%)		
No	925 (63.97)	787 (54.65)
Yes	521 (36.03)	653 (45.35)

[a]CATE: conditional average treatment effects.

[b]PSA: prostate-specific antigen.

We estimated the AIPW scores for treatment effects, in terms of follow-up days, for each individual if abiraterone and enzalutamide were administered. Next, we generated prescription policy trees for various depth levels. We used 70% of the data for training to generate the prescription policy tree. To arrive at the value (survival duration) that the policy bestows on a patient, we used the estimated AIPW scores for an individual if administered either abiraterone or enzalutamide based on the predictions from the prescription policy tree. We found significance in the prescription policy tree with a depth of 3. We used a minimum node size of 4. For depth 3 (Figure 5), the average number of follow-up days obtained by administering enzalutamide and abiraterone based on the generated prescription policy tree was 56.38 (95% CI 8.89 to 103.87).

Figure 5. Androgen receptor pathway inhibitors prescription policy tree estimated for US veterans with metastatic castration-resistant prostate cancer.



LEGEND

BMI category mapping

BMI Value	BMICAT
≤ 18.5	1
> 18.5 & ≤ 24.9	2
> 24.9 & ≤ 29.9	3
≥ 30	4

Prostate-specific antigen (PSA) test category mapping

PSA Result	PSACAT
≥ 0 & ≤ 4	0
> 4 & ≤ 10	1
> 10 & ≤ 20	2
> 20 & ≤ 50	3
> 50 & ≤ 100	4
> 100 & ≤ 200	5
> 200	6

Binary variable mapping

1. PD: Veterans Administration (VA)–based Frailty Category of Parkinson's disease coded as 1; 0 otherwise
2. Race: Race White coded as 1; Race Black coded as 0
3. Peripheral neuropathy: VA-based Frailty category of peripheral neuropathy coded as 1; 0 otherwise

Discussion

Principal Findings and Comparison With Previous Work

In this study, we applied a causal survival forest to assess the heterogeneous treatment effects of the mCRPC medications enzalutamide and abiraterone on overall survival. We found that, on average, for 2886 veterans, 59.94 (95% CI 35.60-84.28) more days of survival than abiraterone. The value of our policy is 56.38 (95% CI 8.89-103.87), which is not statistically different from 59.94. This means that the simple, transparent treatment rules using the policy tree method to prescribe abiraterone and enzalutamide can achieve a policy value similar to the conditional average treatment effect. Schoen et al [10] used an adjusted hazard model and found that age, BMI, and PSA level were positively associated with mortality. Our policy tree elicited the decision checks for age, PSA level, and BMI to make prescription decisions on which of the 2 medications is likely to prolong survival.

Interestingly, even though in univariate analyses in previous studies, the median treatment duration was longer in patients with hypertension or diabetes, the policy tree did not choose both hypertension and diabetes and did not contribute to the highest variation between the 2 subgroups. In previous work, increased hemoglobin A_{1c}, a marker of diabetes severity, was associated with more prolonged survival in veterans prescribed enzalutamide compared with abiraterone [24]. Instead, our policy tree chose peripheral neuropathy. However, peripheral neuropathy is a complication of diabetes (in addition to being a predictor of severe diabetes, it could be caused by factors such as injury and exposure to toxins and is also a common side effect of cancer treatment) [7,25]. Vision loss can be a complication of diabetes and is associated with increased hospitalizations [26], impaired cognition [27], and higher mortality [28]. To delay diabetic neuropathy, apart from good glycemic control, it has also been found beneficial to control blood pressure [29]. All these could have contributed to choosing peripheral neuropathy instead of diabetes and hypertension by our policy tree.

In addition, our policy tree incorporated osteoporosis, which was also one of the top covariates contributing to the variation between the 2 subgroups. Previous studies show an increased risk of falls and fractures among patients with mCRPC taking ARPIs [30,31]. Osteoporosis and abnormal gait may identify impaired functional status or increased frailty, which is associated with survival outcomes in mCRPC [9]. Osteoporosis is a hidden nonmotor symptom of Parkinsons' disease, which contributed to variation in subgroups and was chosen by the policy tree [32]. Bilirubin levels could contribute to osteoporosis, which is also shown as one of the top variations among the two subgroups [33]. Our policy tree chose race, and existing studies have demonstrated that abiraterone has improved survival when given as the first line of treatment for the African American population [34]. Together, these results show that our policy tree could elicit a variety of comorbidities and suggest therapy.

In addition to identifying the covariates, selecting treatments for patients with mCPRC using the generated policy tree considering their demographics, laboratory measures, and comorbidities may help mitigate complications and prolong overall survival. As no generic version of enzalutamide is available in the US, the drug remains expensive. Therefore, from a cost perspective, it is prudent to administer enzalutamide only to patients most likely to benefit from it in terms of survival duration.

Limitations and Future Studies

One of the limitations of this study is that the data are only for US veterans, so that the generalization of the prescription rules to the other populations is not advised. Another limitation is that we excluded patients with mCRPC who had docetaxel, cabaxitaxel, mitoxantrone, and sipuleucel-T before or after prescribing abiraterone or enzalutamide. This was performed to identify patients with higher comorbid disease burdens with fewer therapeutic options. However, other treatments are prescribed for patients with mCRPC. A third limitation of this study was the lack of data on time-varying covariates. Even though causal survival forest is robust to censoring, it will be interesting to include time-varying covariates in the future and to check for differences in the generated policy tree. The fourth limitation of this study was that we used only 70% of the training data to generate a policy tree, because the causal survival forest used cross-fitting and honesty approaches. Furthermore, we estimated the overall benefit of the policy tree (ie, the average difference in the follow-up days obtained if the patients with mCRPC are administered enzalutamide or abiraterone by adhering to the policy tree) in this study without considering the benefits to subgroups of patients based on race or other demographics. Future research should estimate the benefit of such policy trees to see if they are equally beneficial to various subgroups of demographics. Similarly, this study used only mCRPC patient data. Future studies should develop treatment selection for earlier stages of prostate cancer, such as the metastatic hormone-sensitive stage.

Conclusions

We used a machine learning-based survival approach, that is, causal survival forest model, to estimate the variations in the survival of patients with mCRPC who were administered either enzalutamide or abiraterone. Our estimation revealed that patients with mCRPC who were administered enzalutamide had longer survival than patients who were administered abiraterone. We were able to use a data-driven approach to identify heterogeneity and subgroups of patients. We then created policy trees to aid physicians administer personalized treatment, that is, abiraterone versus enzalutamide, based on patient characteristics, including demographics, laboratory values, and comorbidities, to improve survival duration.

Acknowledgments

This study was supported by the Department of Defense Physician Research Award (W81XWH-22-1-0602) to MWS and the Igor Tulchinsky, Robert Taubman, and Richard Sandler – Prostate Cancer Foundation Valor Young Investigator Award. The funders had no role in the design, conceptualization, preparation, review, approval, or decision to submit the manuscript for publication. DG was previously affiliated with the School of Medicine at Saint Louis University during the initial submission or revision. DG completed the subsequent work and revision (including rerunning analysis) at her current affiliation with the Richard A Chaifetz School of Business at Saint Louis University.

Data Availability

Scientific community members can request the dataset by e-mailing MWS at martin.schoen@va.gov. It will be shared upon request and execution of an agreement of use. The investigator should state the reason for requesting the data and analysis plans.

Authors' Contributions

All authors had full access to the study’s data. MWS takes responsibility for the integrity of data. DG, NM, and MWS contributed to concept, design, analysis or interpretation of data, writing of manuscript, critical manuscript review, and approval.

Conflicts of Interest

MWS reports honoraria from Pfizer for a lecture on Real-World Data and consulting for ConcertAI. DG and NM do not have any conflicts of interest.

References

1. McHugh J, Saunders EJ, Dadaev T, McGrowder E, Bancroft E, Kote-Jarai Z, et al. Prostate cancer risk in men of differing genetic ancestry and approaches to disease screening and management in these groups. Br J Cancer 2022 Jun 01;126(10):1366-1373 [FREE Full text] [doi: 10.1038/s41416-021-01669-3] [Medline: 34923574]
2. Akaza H, Procopio G, Pripatnanont C, Facchini G, Fava S, Wheatley D, et al. Metastatic castration-resistant prostate cancer previously treated with docetaxel-based chemotherapy: treatment patterns from the PROXIMA prospective registry. J Glob Oncol 2018;4:1-12 [FREE Full text] [doi: 10.1200/JGO.18.00009] [Medline: 30260754]
3. Daniels VA, Luo J, Paller CJ, Kanayama M. Therapeutic approaches to targeting androgen receptor splice variants. Cells 2024;13(1):104 [FREE Full text] [doi: 10.3390/cells13010104] [Medline: 38201308]
4. Hara I, Yamashita S, Nishizawa S, Kikkawa K, Shimokawa T, Kohjimoto Y. Enzalutamide versus abiraterone as a first-line endocrine therapy for castration-resistant prostate cancer: protocol for a multicenter randomized phase 3 trial. JMIR Res Protoc 2018;7(7):e11191 [FREE Full text] [doi: 10.2196/11191] [Medline: 30054264]

5. Halwani AS, Rasmussen KM, Patil V, Li CC, Yong CM, Burningham Z, et al. Real-world practice patterns in veterans with metastatic castration-resistant prostate cancer. Urol Oncol 2020;38(1):1.e1-1.e10. [doi: 10.1016/j.urolonc.2019.09.027] [Medline: 31704142]
6. Thurin NH, Rouyer M, Jové J, Gross-Goupil M, Haaser T, Rébillard X, et al. Abiraterone acetate versus docetaxel for metastatic castration-resistant prostate cancer: a cohort study within the French nationwide claims database. Expert Rev Clin Pharmacol 2022;15(9):1139-1145. [doi: 10.1080/17512433.2022.2115356] [Medline: 35984212]
7. Shah YB, Shaver AL, Beiriger J, Mehta S, Nikita N, Kelly WK, et al. Outcomes following abiraterone versus enzalutamide for prostate cancer: a scoping review. Cancers (Basel) 2022;14(15):3773 [FREE Full text] [doi: 10.3390/cancers14153773] [Medline: 35954437]
8. Tan YY, Papez V, Chang WH, Mueller SH, Denaxas S, Lai AG. Comparing clinical trial population representativeness to real-world populations: an external validity analysis encompassing 43 895 trials and 5 685 738 individuals across 989 unique drugs and 286 conditions in England. Lancet Healthy Longev 2022;3(10):e674-e689 [FREE Full text] [doi: 10.1016/S2666-7568(22)00186-6] [Medline: 36150402]
9. Deol ES, Sanfilippo KM, Luo S, Fiala MA, Wildes T, Mian H, et al. Frailty and survival among veterans treated with abiraterone or enzalutamide for metastatic castration-resistant prostate cancer. J Geriatr Oncol 2023;14(5):101520. [doi: 10.1016/j.jgo.2023.101520] [Medline: 37263065]
10. Schoen MW, Carson KR, Eisen SA, Bennett CL, Luo S, Reimers MA, et al. Survival of veterans treated with enzalutamide and abiraterone for metastatic castrate resistant prostate cancer based on comorbid diseases. Prostate Cancer Prostatic Dis 2023;26(4):743-750 [FREE Full text] [doi: 10.1038/s41391-022-00588-5] [Medline: 36104504]
11. Riekhof F, Yan Y, Bennett CL, Sanfilippo KM, Carson KR, Chang SH, et al. Hospitalizations among veterans treated for metastatic prostate cancer with abiraterone or enzalutamide. Clin Genitourin Cancer 2024;22(2):18-26.e3. [doi: 10.1016/j.clgc.2023.07.006] [Medline: 37495480]
12. George DJ, Ramaswamy K, Yang H, Liu Q, Zhang A, Greatsinger A, et al. Real-world overall survival with abiraterone acetate versus enzalutamide in chemotherapy-naïve patients with metastatic castration-resistant prostate cancer. Prostate Cancer Prostatic Dis 2024:1-9. [doi: 10.1038/s41391-024-00816-0] [Medline: 38538879]
13. Jain P, Hathi DK, Honarvar H, Das RK. Machine learning (ML) on real-world data (RWD) of front-line (1L) metastatic castration resistance prostate cancer (mCRPC) patients for dynamic prediction of time to tx discontinuation (TTD). Cancer Research; 2023;83(7_Supplement) 2023;83(7_Supplement):4400-4400. [doi: 10.1158/1538-7445.am2023-4400]
14. Lim H, Yoo JW, Lee KS, Lee YH, Baek S, Lee S, et al. Toward precision medicine: development and validation of a machine learning based decision support system for optimal sequencing in castration-resistant prostate cancer. Clin Genitourin Cancer 2023;21(4):e211-e218.e4. [doi: 10.1016/j.clgc.2023.03.012] [Medline: 37076338]
15. Veen KM, de Angst IB, Mokhles MM, Westgeest HM, Kuppen M, de Groot CAU, et al. A clinician's guide for developing a prediction model: a case study using real-world data of patients with castration-resistant prostate cancer. J Cancer Res Clin Oncol 2020;146(8):2067-2075 [FREE Full text] [doi: 10.1007/s00432-020-03286-8] [Medline: 32556680]
16. de Jong AC, Danyi A, van Riet J, de Wit R, Sjöström M, Feng F, et al. Predicting response to enzalutamide and abiraterone in metastatic prostate cancer using whole-omics machine learning. Nat Commun 2023;14(1):1968. [doi: 10.1038/s41467-023-37647-x]
17. Lee J, Yang SW, Lee S, Hyon YK, Kim J, Jin L, et al. Machine learning approaches for the prediction of prostate cancer according to age and the prostate-specific antigen level. Korean J Urol Oncol 2019;17(2):110-117. [doi: 10.22465/kjuo.2019.17.2.110]
18. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials--an economic evaluation context: methods, limitations, and recommendations. Med Decis Making 2014;34(3):387-402. [doi: 10.1177/0272989X13520192] [Medline: 24449433]
19. Cui Y, Kosorok MR, Sverdrup E, Wager S, Zhu R. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. Journal of the Royal Statistical Society Series B: Statistical Methodology 2023;85(2):179-211. [doi: 10.1093/jrsssb/qkac001]
20. Credit K, Lehnert M. A structured comparison of causal machine learning methods to assess heterogeneous treatment effects in spatial data. J Geogr Syst 2023:1-28. [doi: 10.1007/s10109-023-00413-0]
21. Athey S, Imbens GW. The econometrics of randomized experiments. In: Handbook of Economic Field Experiments 2017;1:73-140. [doi: 10.1016/bs.hefe.2016.10.003]
22. Kurz CF. Augmented inverse probability weighting and the double robustness property. Med Decis Making 2022;42(2):156-167 [FREE Full text] [doi: 10.1177/0272989X211027181] [Medline: 34225519]
23. Athey S, Wager S. Estimating treatment effects with causal forests: an application. Observational Studies 2019;5(2):37-51. [doi: 10.1353/obs.2019.0001]
24. Govindan S, Cheranda N, Riekhof F, Luo S, Schoen MW. Effect of BMI and hemoglobin A1c on survival of veterans with metastatic castration-resistant prostate cancer treated with abiraterone or enzalutamide. Prostate 2024;84(3):245-253. [doi: 10.1002/pros.24644] [Medline: 37909677]
25. Staff NP, Grisold A, Grisold W, Windebank AJ. Chemotherapy-induced peripheral neuropathy: a current review. Ann Neurol 2017;81(6):772-781 [FREE Full text] [doi: 10.1002/ana.24951] [Medline: 28486769]

26. Chen SP, Bhattacharya J, Pershing S. Association of vision loss with cognition in older adults. JAMA Ophthalmol 2017;135(9):963-970 [FREE Full text] [doi: 10.1001/jamaophthalmol.2017.2838] [Medline: 28817745]
27. Ehrlich JR, Ramke J, Macleod D, Burn H, Lee CN, Zhang JH, et al. Association between vision impairment and mortality: a systematic review and meta-analysis. Lancet Glob Health 2021;9(4):e418-e430 [FREE Full text] [doi: 10.1016/S2214-109X(20)30549-0] [Medline: 33607015]
28. Ruppert LM, Cohn ED, Keegan NM, Bacharach A, Woo S, Gillis T, et al. Spine pain and metastatic prostate cancer: defining the contribution of nonmalignant etiologies. JCO Oncol Pract 2022;18(6):e938-e947 [FREE Full text] [doi: 10.1200/OP.21.00816] [Medline: 35175783]
29. Sethi Y, Uniyal N, Vora V, Agarwal P, Murli H, Joshi A, et al. Hypertension the 'Missed Modifiable Risk Factor' for diabetic neuropathy: a systematic review. Curr Probl Cardiol 2023;48(4):101581. [doi: 10.1016/j.cpcardiol.2022.101581] [Medline: 36584725]
30. Myint ZW, Momo HD, Otto DE, Yan D, Wang P, Kolesar JM. Evaluation of fall and fracture risk among men with prostate cancer treated with androgen receptor inhibitors: a systematic review and meta-analysis. JAMA Netw Open 2020;3(11):e2025826 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.25826] [Medline: 33201234]
31. Graff JN, Baciarello G, Armstrong AJ, Higano CS, Iversen P, Flaig TW, et al. Efficacy and safety of enzalutamide in patients 75 years or older with chemotherapy-naive metastatic castration-resistant prostate cancer: results from PREVAIL. Ann Oncol 2016;27(2):286-294 [FREE Full text] [doi: 10.1093/annonc/mdv542] [Medline: 26578735]
32. Metta V, Sanchez TC, Padmakumar C. Osteoporosis: a hidden nonmotor face of parkinson's disease. Int Rev Neurobiol 2017;134:877-890. [doi: 10.1016/bs.irn.2017.05.034] [Medline: 28805587]
33. Jurado S, Parés A, Peris P, Combalia A, Monegal A, Guañabens N. Bilirubin increases viability and decreases osteoclast apoptosis contributing to osteoporosis in advanced liver diseases. Bone 2022;162:116483. [doi: 10.1016/j.bone.2022.116483] [Medline: 35787483]
34. Marar M, Long Q, Mamtani R, Narayan V, Vapiwala N, Parikh RB. Outcomes among African American and non-hispanic white men with metastatic castration-resistant prostate cancer with first-line abiraterone. JAMA Netw Open 2022;5(1):e2142093 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.42093] [Medline: 34985518]

Abbreviations

AIPW: augmented inverse-propensity weighting
ARPI: androgen-receptor pathway inhibitor
ASMD: absolute standardized mean difference
CATE: conditional average treatment effects
CVD: cardiovascular diseases
ICD: International Classification of Diseases
mCRPC: metastatic castrate-resistant prostate cancer
PSA: prostate-specific antigen
VA: Veterans Affairs

Edited by Q Chen; submitted 12.04.24; peer-reviewed by S Ser, S Cohen, H Demir; comments to author 04.05.24; revised version received 26.05.24; accepted 10.10.24; published 19.11.24.

Please cite as:
Gopukumar D, Menon N, Schoen MW
Medication Prescription Policy for US Veterans With Metastatic Castration-Resistant Prostate Cancer: Causal Machine Learning Approach
JMIR Med Inform 2024;12:e59480
URL: https://medinform.jmir.org/2024/1/e59480
doi:10.2196/59480
PMID:

Original Paper

Predicting Pain Response to a Remote Musculoskeletal Care Program for Low Back Pain Management: Development of a Prediction Tool

Anabela C Areias[1], PhD; Robert G Moulder[2], PhD; Maria Molinos[1], PhD; Dora Janela[1], PT; Virgílio Bento[1], PhD; Carolina Moreira[1,3], MD; Vijay Yanamadala[1,4,5], MD; Steven P Cohen[6,7,8,9,10,11], MD; Fernando Dias Correia[1,12], MD, PhD; Fabíola Costa[1], PhD

[1]Sword Health Inc, Draper, UT, United States

[2]Institute for Cognitive Science, University of Colorado Boulder, Boulder, CO, United States

[3]Instituto de Ciências Biomédicas Abel Salazar, Porto, Portugal

[4]Department of Surgery, Quinnipiac University Frank H Netter School of Medicine, Hamden, CT, United States

[5]Department of Neurosurgery, Hartford Healthcare Medical Group, Westport, CT, United States

[6]Department of Anesthesiology and Critical Care Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States

[7]Department of Physical Medicine and Rehabilitation, Johns Hopkins School of Medicine, Baltimore, MD, United States

[8]Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD, United States

[9]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, United States

[10]Department of Anesthesiology, Uniformed Services University of the Health Sciences, Bethesda, MD, United States

[11]Department of Physical Medicine and Rehabilitation, Uniformed Services University of the Health Sciences, Bethesda, MD, United States

[12]Neurology Department, Centro Hospitalar e Universitário do Porto, Porto, Portugal

Corresponding Author:
Fabíola Costa, PhD
Sword Health Inc
13937 Sprague Lane, Suite 100
Draper, UT, 84020
United States
Phone: 1 385 308 8034
Fax: 1 801 206 3433
Email: f.costa@swordhealth.com

Abstract

Background: Low back pain (LBP) presents with diverse manifestations, necessitating personalized treatment approaches that recognize various phenotypes within the same diagnosis, which could be achieved through precision medicine. Although prediction strategies have been explored, including those employing artificial intelligence (AI), they still lack scalability and real-time capabilities. Digital care programs (DCPs) facilitate seamless data collection through the Internet of Things and cloud storage, creating an ideal environment for developing and implementing an AI predictive tool to assist clinicians in dynamically optimizing treatment.

Objective: This study aims to develop an AI tool that continuously assists physical therapists in predicting an individual's potential for achieving clinically significant pain relief by the end of the program. A secondary aim was to identify predictors of pain nonresponse to guide treatment adjustments.

Methods: Data collected actively (eg, demographic and clinical information) and passively in real-time (eg, range of motion, exercise performance, and socioeconomic data from public data sources) from 6125 patients enrolled in a remote digital musculoskeletal intervention program were stored in the cloud. Two machine learning techniques, recurrent neural networks (RNNs) and light gradient boosting machine (LightGBM), continuously analyzed session updates up to session 7 to predict the likelihood of achieving significant pain relief at the program end. Model performance was assessed using the area under the receiver operating characteristic curve (ROC-AUC), precision-recall curves, specificity, and sensitivity. Model explainability was assessed using SHapley Additive exPlanations values.

Results: At each session, the model provided a prediction about the potential of being a pain responder, with performance improving over time (P<.001). By session 7, the RNN achieved an ROC-AUC of 0.70 (95% CI 0.65-0.71), and the LightGBM achieved an ROC-AUC of 0.71 (95% CI 0.67-0.72). Both models demonstrated high specificity in scenarios prioritizing high precision. The key predictive features were pain-associated domains, exercise performance, motivation, and compliance, informing continuous treatment adjustments to maximize response rates.

Conclusions: This study underscores the potential of an AI predictive tool within a DCP to enhance the management of LBP, supporting physical therapists in redirecting care pathways early and throughout the treatment course. This approach is particularly important for addressing the heterogeneous phenotypes observed in LBP.

Trial Registration: ClinicalTrials.gov NCT04092946; https://clinicaltrials.gov/ct2/show/NCT04092946 and NCT05417685; https://clinicaltrials.gov/ct2/show/NCT05417685

(JMIR Med Inform 2024;12:e64806) doi:10.2196/64806

KEYWORDS

telerehabilitation; predictive modeling; personalized medicine; rehabilitation; clinical decision support; machine learning; artificial intelligence

Introduction

Low back pain (LBP), the leading cause of disability worldwide [1], has a complex and multifaceted etiology [2,3], resulting in varied phenotypes even among individuals with the same diagnosis. Precision medicine supports tailored interventions to address this heterogeneity, with data-driven strategies playing a crucial role [4,5]. These strategies range from simple, pragmatic rules-based methods [6-8] (eg, STARTback, Orebro) to more sophisticated machine learning (ML) tools trained on large data sets [9-11]. In the context of LBP management, ML models have been developed to assist with screening [12-14], assess the probability of a patient transitioning from acute to chronic cases [8,15-18], predict surgical outcomes [19,20], and forecast clinical recovery prognosis [6,7,9-11,21]. Although promising, these models lack sustained, dynamic real-time patient data collection, relying heavily on patient-reported outcome measures (PROMs) and other information that is not easily scalable. This reliance, in turn, limits the ability to deliver real-time predictions throughout treatment or hinders large-scale implementation in real-world settings [9,10].

The emergence of digital care programs (DCPs) as an effective alternative for delivering LBP care [22-24] has facilitated the seamless, automatic collection of abundant data from various sources. This diverse and comprehensive patient data set, which includes passively collected real-time updates of multiple variables via the Internet of Things devices (eg, wearables) enriched with population data, enables the capture of more complex patterns, providing a more accurate representation of patients. These extensive volumes of data can be used to predict outcomes and optimize treatments, with the resulting outcomes feeding back into predictive tools in a virtuous cycle.

Previously, we demonstrated the efficacy of a remote DCP for LBP management [22], which combines exercise, education, and cognitive behavioral therapy under physical therapist (PT) supervision. In this study, we leverage all stored data (both passively and actively collected from patient and public sources) in the cloud portal, actively monitored by the assigned PT, to develop a predictive tool that can assist in optimizing treatment. This study aims to develop an artificial intelligence (AI) tool to assist PTs in predicting an individual's potential to achieve clinically significant pain relief by the end of the program. Such predictions could enable timely adjustments to the program, increasing the likelihood of success and, in turn, providing data to further refine the tool's recommendations. As a secondary aim, we sought to identify the predictors of pain nonresponse, as these factors indicate the need for greater attention from PTs. Thus, the overall objective of the study is to develop ML models to predict the likelihood that a patient will achieve clinically significant pain relief by the end of the program, with ongoing updated predictions after each completed session.

Methods

Study Design

This secondary analysis utilized data from 2 prospective studies evaluating clinical and engagement-related outcomes in patients with musculoskeletal (MSK) conditions. Data were collected from June 2020 to July 2023. This prognostic study adhered to the TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-AI) reporting guidelines (Table S1 in Multimedia Appendix 1).

Ethical Considerations

All research was conducted in accordance with the relevant guidelines and regulations set forth by the Declaration of Helsinki. This study received prospective approval from the New England institutional review board (120190313) and Advarra institutional review board (Pro00063337) and was registered on ClinicalTrials.gov (NCT04092946 and NCT05417685) on September 17, 2019, and June 14, 2022, respectively.

All participants provided electronic informed consent to participate in the study, with a waiver of documentation of consent approved by the New England and Advarra institutional review boards.

All collected data underwent a rigorous anonymization process to safeguard the privacy of the individuals involved in the research. The data collection and analysis methods complied with relevant guidelines and regulations. Participants were not offered any form of compensation.

Population

Beneficiaries of employer health plans from all US states applying to Sword Health's DCPs were included. The inclusion criteria were as follows: acute or chronic LBP, an average Numerical Pain Rating Scale (NPRS) score of ≥4/10, and at least one NPRS reassessment during the intervention. The exclusion criteria were as follows: (1) health conditions (eg, cardiac or respiratory) that are incompatible with at least 20 minutes of light-to-moderate exercise; (2) cancer-related back pain or current cancer treatment for a non-MSK condition; and (3) serious neurological signs or symptoms, including bowel or bladder dysfunction. All participants provided informed consent.

Intervention

The DCP consisted of exercise, education, and cognitive behavioral therapy for up to 12 weeks, with a default recommendation of 3 sessions per week, as described elsewhere [22], after enrolling through a dedicated website. Patients completed a baseline condition form that included information on their demographic and clinical characteristics and selected their PT based on their preferences. Subsequently, an onboarding video call was conducted in which the PTs gathered additional medical history and established goals through shared decision-making. Each patient received a Food and Drug Administration–listed class II medical device, which included a mobile app on a dedicated tablet (containing the program) combined with motion tracking, as well as a cloud-based portal that enabled asynchronous remote monitoring and treatment prescription by the PT. A tailored educational program and cognitive behavioral therapy were also provided [25-27]. Bidirectional communication with the PT was facilitated through a secure chat feature on the smartphone app and video calls.

Outcome

Considering the overarching intervention goal of promoting improvements in pain levels in daily living, pain response was selected as the primary outcome. Pain was assessed using the 11-point NPRS with a 7-day recall period during sessions 9, 18, or 27.

Pain response, defined as at least a 30% reduction in the NPRS by the end of the program, aligns with clinical recommendations from the Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) [28]. This definition is supported by a consensus meeting that interpreted the clinical significance of treatment outcomes in clinical trials for chronic pain treatments. Additionally, we included the criterion of concluding the program with an NPRS score equal to or below 3 points (the threshold for mild pain) [29] to account for patients who started with lower pain levels and ended the program with acceptable pain levels. The widespread use of this binary outcome, along with its practicality in real-world settings, supported its inclusion.

Predictors: Variables Used for Model Development

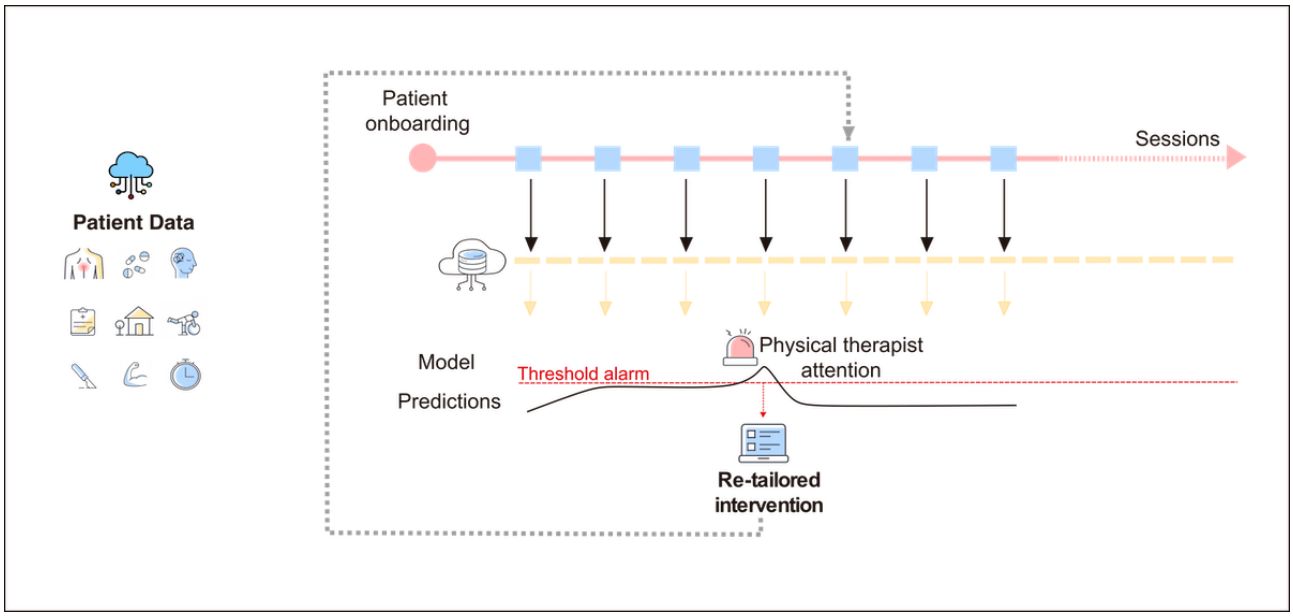
The variables used for model development were selected based on existing literature and their clinical relevance to MSK pain. A detailed description of the variable domains and their acceptable ranges can be found in Table 1. The data set variables included (1) demographic characteristics (eg, race and ethnicity, social deprivation index) [30]; (2) baseline clinical presentations (eg, acuity, disability, mental health, productivity, fear-avoidance beliefs, presence of red flags); (3) time series of mean active range of motion (ROM) throughout all repetitions of each exercise during each session; (4) session-related usability (eg, movement errors, number of exercises, sets and repetitions, the time between sessions, session duration); and (5) self-reported fatigue and pain experienced during exercises, measured on a 0-10 scale (Figure 1).

Table 1. Features domains included in the study.

Domain and subdomains	Description
Demographic	
Individual characteristics	Comprises information about gender, age, race/ethnicity, body mass index, education level, among others
Geographical location	Urban or rural location and time zone
Work characteristics	Employment status and job type
Socioeconomic	
Social deprivation	Assessed by the Social Deprivation Index (score range 0-100) [30]; socially deprived individuals were those with an index ranging from 60 to 100
Productivity	
Participation—work	Assessed by the overall work productivity (score range 0-100), presenteeism (score range 0-100), and absenteeism (score range 0-100) subscales from the Work Productivity Impairment Questionnaire for general health
Participation—social	Assessed by the non-work–related activities impairment subscale from the Work Productivity Impairment Questionnaire for general health (score range 0-100), and scores of items 8 (range 0-5), 9 (range 0-5), and 10 (range 0-5) of the Oswestry Disability Index
Clinical	
Condition	Includes musculoskeletal condition diagnosis, anatomical pain region, presence of leg symptoms (numbness or tingling), past surgery, among others
Acuity	Acute (<12 weeks) or chronic (persistent or recurring pain for ≥12 weeks)
Pain	Pain intensity assessed by the Numerical Pain Rating Scale (score range 0-10)
Functionality	Assessed by the Oswestry Disability Index (score range 0-100) considering the total score and the individual scores of items 2-7
Mental health	Assessed by the 7-item General Anxiety Disorder Scale (score range 0-21) and the 9-item Patient Health Questionnaire (score range 0-27), considering the total score and the individual score for each item; features considering moderate and severe depression or anxiety were calculated using the thresholds described in Spitzer et al [31] and Kroenke et al [32]
Fear-avoidance	Assessed by the Fear-Avoidance Beliefs Questionnaire for Physical Activity (score range 0-24), considering the total score and the individual score for each item
Surgery intent	Assessed by the question “On a scale of 0 to 100, where 0 is not at all and 100 is extremely interested, how interested are you in undergoing shoulder surgery in the next 12 months?” (score range 0-100)
Medication consumption	Considers nonopioid analgesic and opioid consumption (yes or no)
Red flags	Presence/absence of at least one red flag identified during onboard screening, which was cleared before entering the study by a physician (yes or no) [33]
Prescription	
Support level	Includes support-level perception assessed by physical therapist interactions during the intervention
Exercise Performance	
Biomechanics	Active range of motion collected by motion trackers
Exercise accuracy	Correct and wrong repeats; performance in all prescribed exercises (total correct repeats by total repeats); average stars (0-5)
Exercise-induced pain acute response	Pain felt during exercise sessions (assessed by the question “How did you feel during your session: my pain during today’s session?” (score range 0-10)
Exercise-induced fatigue intensity	Self-reported fatigue in response to the question “How did you feel during your session: My fatigue during today’s session?” (score range 0-10)
Exercise adherence	Session duration time spent on sessions (minutes) and time interval between sessions (days)
Motivation and Compliance	

Domain and subdomains	Description
Motivation	Patient's individual goals with the intervention and session commitment (through the question "How many 20-minute sessions can you commit to per week?"), among others
Enrollment aspects	The time span from enrollment to (1) onboarding and (2) session 1
Days of the week	Days of the week the patient performed the exercise session

Figure 1. Machine Learning (ML) tool development within the digital care program. All patient data collected both passively and actively regarding demographic and clinical characteristics, range of motion, and session usability, as well as collected from public sources (eg, social deprivation index) are continuously stored from onboarding to the program end, enabling the creation of a data repository within the physical therapists (PTs) cloud portal. At the end of each session, data are processed through an ML model to predict an individual's potential to achieve clinically significant pain relief at the program end. In the case of the high probability of an unfavorable outcome, an alarm is set in the PT portal for further examination and re-tailoring of the intervention.



All data were automatically stored in the cloud portal. Demographic and clinical characteristics were collected through an onboarding form using validated PROMs. ROM and session usability were passively collected, while fatigue and pain experienced during exercise sessions were recorded at the end of each session.

Sample Size

Given that there is no established method to calculate sample sizes for prognostic models using ML, we followed the recommended guideline for standard model development, which suggests obtaining 10-20 events per predictor parameter. As we estimated including up to 300 predictor variables in the models, a sample size range of 3000 to 6000 was calculated.

Missing Data and Data Preparation

The overall missingness was 18.45% (324,385/1,757,875; see Table S2 in Multimedia Appendix 1), comprising the following: 17.07% (300,117/1,757,875) missing completely at random, where the missing data points were attributed to system settings; 0.29% (5169/1,757,875) missing at random, where the missingness could be explained by other variables (eg, specific protocol prescriptions that did not include one of the studied exercises); and 1.09% (19,099/1,757,875) missing not at random, where the reasons for the missing data were related to unobserved factors. Categorical variable imputation was performed by creating a new category labeled "not available," followed by dummy encoding. Continuous variables were imputed using multiple imputations by chained equations or by assigning a value outside the distribution (eg, –1), depending on the nature of the missingness. Multiple imputations can provide unbiased estimations for data that are missing completely at random and at random. As the rate of missingness for not-at-random data is very small, we used multiple imputation techniques to handle the missing data overall. The cleaned data set included 6125 patients and contained a number of variables ranging from 235 at session 1 to 275 at session 7.

Analytical Methods

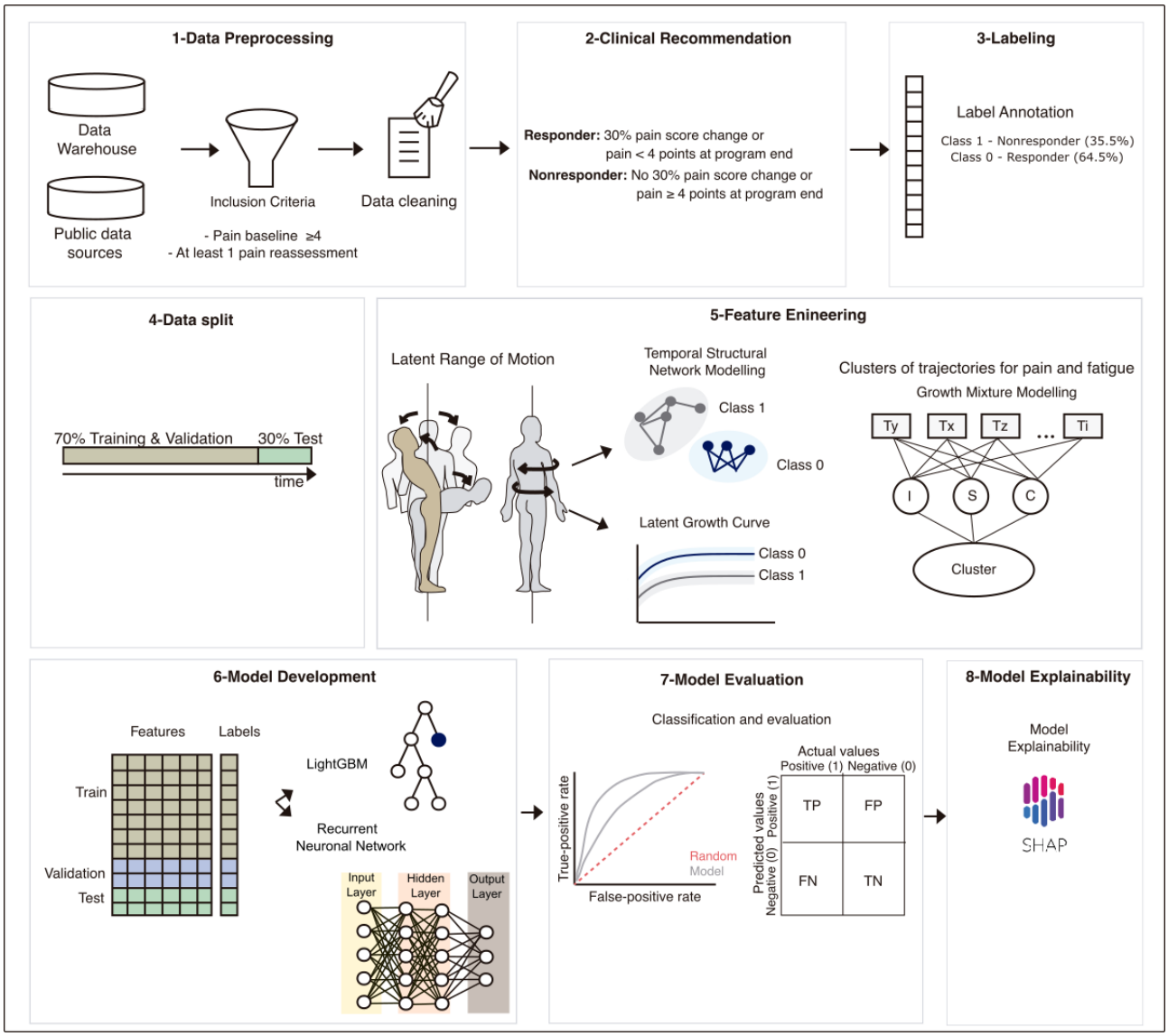
Feature Engineering

Feature engineering was utilized to capture meaningful patterns and insights from multiple variables over time, creating new features that represent the underlying relationships and dynamics within the data (as detailed in Figure S1 in Multimedia Appendix 1). A minimum of 4 sessions was required to conduct a longitudinal analysis.

ROM data from 4 exercises were encoded into a unified composite value that illustrates the longitudinal progression of trunk motion [34], utilizing both latent growth curve analysis (LGCA) [35,36] and temporal structural network modeling (Figure 2) [37]. While the former captures ROM patterns over

time to estimate individual trajectories among the 2 classes (pain nonresponders and pain responders), the latter illustrates how improvements in one exercise may be correlated with improvements in other exercises. Among a diverse array of prescribed exercises (N=117), only those that incorporated trunk motion—extension, flexion, rotation, and side bending—were included [34]. This subset was chosen due to its frequent prescription among patients and its superior performance in model comparisons, as determined by likelihood-ratio tests.

Figure 2. Model development pipeline. Data from patients with low back pain (LBP) who underwent the digital intervention were collected from the data warehouse within a specific time frame. Inclusion criteria: average pain level equal to or greater than 4 on the Numerical Pain Rating Scale (NPRS) at baseline considering a 7-day recall period; and at least one pain reassessment during the intervention. Patient data were preprocessed for the tool. Patients who did not experience a pain (NPRS) reduction of at least 30% or reported pain level higher or equal to 4 at the program end were categorized as nonresponders (labeled as "1" in the data set). Considering the overarching intervention goal of promoting improvement in pain levels, a flag was triggered when patients were likely to be nonresponders to assist PTs' clinical judgment; a split of 70/30 was performed on the data set followed by feature engineering: range of motion (ROM) was computed through temporal structural network modeling and latent growth curves, whereas pain and fatigue experienced during exercises (reported at the end of each exercise session), exercise accuracy, and time between sessions were computed using growth mixture modeling to depict changes over time. Model development consisted of a tree-based binary classifier and a recurrent neural network that were optimized using the receiver operating characteristic (ROC)-area under the curve (AUC) as the target metric. Precision-recall curves were used as evaluation metrics, and model explainability was assessed using SHAP (Shapley Additive Explanations) values. FN: false negative; FP: false positive; LightGBM: light gradient boosting machine; TN: true negative; TP: true positive.



Thus, we first constructed the LGCA and temporal structural network model for both responders and nonresponders (P<.001) to estimate an intercept, slope, curve, and latent ROM, as well as individual covariance matrices for each patient (see Equations S1-S3 in Multimedia Appendix 1). The proximity of an individual covariance matrix to the expected covariance matrices from responders and nonresponders was assessed using the chi-square measure, which was then incorporated as a feature.

Mixture clustering was used to create new features, using the probabilities of belonging to different clusters to leverage the underlying characteristics of those clusters in making predictions, following a mixture of experts approach. A growth

mixture model [38] was applied to cluster longitudinal trajectories of pain and fatigue profiles during exercise, as well as the performance of the exercises and the time spent in sessions. The intent was to capture subpopulations with a high probability of sharing similar patterns (eg, high, moderate, slight, or no improvement). The Bayesian information criterion indicated that 5 clusters provided the optimal fit for the data (see Figure S2 in Multimedia Appendix 1). The individual probability of a patient being assigned to each trajectory class was calculated using likelihood ratios (see Equation S4 in Multimedia Appendix 1), yielding 5 probabilities per participant for each variable at each session.

For the early sessions (ie, sessions 1-3), longitudinal models could not be calculated due to the requirement of at least four time points. Therefore, the raw values or their deltas between sessions were calculated for the outcomes mentioned in this section.

Labeling

Pain responders (as defined in the "Outcome" subsection) were labeled as 0, while nonresponders were labeled as 1. This labeling serves as a flag to assist PTs during monitoring and clinical decision-making, thereby representing a binary classification scheme (see Figure 2). Depending on the discharge time point (which occurred no later than the 30th treatment session), the corresponding latest survey (9th, 18th, or 27th) was used to create the labels.

Model Pipeline

Two different approaches were used to predict pain response from session 1 to session 7: a tree-based model and a recurrent neural network (RNN) model.

For the tree-based model, a light gradient boosting machine (LightGBM) was utilized, with hyperparameter optimization performed using the Optuna library [39]. This involved 50 trials per algorithm, using the area under the receiver operating characteristic curve (ROC-AUC) as the evaluation metric. LightGBM was trained with 5-fold cross-validation on the training data from each session. As these models are not sequence models, features from previous sessions were concatenated as sessions progressed, allowing models in later sessions to access features from historical sessions. An AUC of 0.5 indicates a random predictor, while an AUC of 1 signifies a perfect predictor. For each fold in the cross-validation, the synthetic minority oversampling technique (SMOTE) was applied to oversample the training data for both classes, reaching 5000 samples per class, as the absence of SMOTE negatively affected model performance. Using the hyperparameters that yielded the best validation performance, a final model was fitted on the entire training set.

An RNN model applied across the 7 sessions can be viewed as a time series (using the PyTorch Python library [40]; Python Foundation). Similar to the tree-based method, the Optuna library was utilized for hyperparameter optimization. However, unlike the tree-based setup, SMOTE was not applied due to its unsuitability for sequential data. RNN models were trained using binary cross-entropy loss and early stopping, with a patience of 10 epochs, utilizing the validation ROC-AUC as the early stopping metric. Using the hyperparameters that yielded the best validation performance, a final model was fitted on the entire training set. This final model was trained with a fixed number of epochs, defined as the average number of early stopping epochs across the cross-validation folds.

The performances of both models were assessed on the test set by calculating the ROC-AUC, precision-recall AUC, F_1-score (at a 0.5 threshold), sensitivity, specificity, and negative predictive value (Figure 2).

Model Explainability

Shapley Additive Explanations (SHAP) was applied on the best model to assess model explainability by investigating feature importance [41] (the Python SHAP library [41] for tree-based and the Python timeshare library [42] for RNN models; Figure 2). Positive SHAP values indicate higher prediction scores, meaning a higher probability of nonresponse. To understand the overall impact of different domains, SHAP values for each feature were cumulatively aggregated within their respective domains for each session. As LightGBM models were trained with an increasing number of features over time, SHAP values were normalized per session to examine the relationship between SHAP values and feature values throughout the intervention. Patient SHAP values were visualized using SHAP dependence plots, utilizing a base value of 0.028 and a contribution threshold of 0.05.

Recursive feature elimination, which involved removing features with low SHAP values after hyperparameter optimization, progressively deteriorated model performance.

Fairness

Demographic and clinical characteristics identified as statistically different between the training and test sets were subjected to subgroup analysis to assess potential bias, including factors such as gender, social deprivation index, age, acuity, and pain levels.

Model Output

The model output classified patients as responders or nonresponders, based on the previously defined outcome, and was contingent on the applied threshold (eg, 0.95).

Train and Test Sets

A data split was conducted in which 70% (4313/6125) of the patients were randomly assigned for training the model, while the remaining 30% (1812/6125) were reserved for testing.

Statistical Analysis

Demographic and clinical presentations at baseline were analyzed using means and proportions. The minimum and maximum ROC-AUC values were calculated by bootstrapping the training and test sets separately (1000 bootstrapped sets). Model comparisons involved ROC-AUCs calculated using the DeLong algorithm [43]. Feature engineering models were developed using R (version 4.2.2; R Foundation for Statistical Computing; packages: lavaan, mlVAR, xgboost, mice, and tidySEM). All other analyses were conducted using Python (version 3.9.7; package: TreeSHAP). The level of significance was set at $P<.05$, considering 2-sided hypothesis tests.

Results

Study Population

A total of 6125 patients met the inclusion criteria (see the flowchart of the study cohort in Figure S3 in Multimedia Appendix 1). Among them, 2172 (35.46%) were classified as nonresponders, while 3953 (64.54%) were responders. The analysis encompassed various domains and subdomains, including demographic, socioeconomic, productivity, clinical, prescription, exercise performance, and motivation/compliance data, up to session 7 (see further details in Table 1). Baseline demographic and clinical characteristics for the overall cohort, as well as stratified by training and test data sets, are presented in Table 2. The overall cohort primarily consisted of women (3445/6125, 56.24%), patients aged 41-60 years (3488/6125, 56.95%), fully employed individuals (5119/6125, 83.58%), and college graduates (4037/6125, 65.91%). Additionally, 2335 (38.12%) individuals were classified as obese, 1931 (31.53%) identified as racial or ethnic minorities (ie, Asian, Black, Hispanic, and other), and 1585 (25.88%) came from socially deprived backgrounds.

Table 2. Baseline demographics and clinical data.

Demographics	Overall cohort (n=6125)	Train data set (n=4313)	Test data set (n=1812)	*P* value
Age (years), mean (SD)	48.5 (11.3)	48.6 (11.4)	48.2 (10.9)	.25
Age category (years), n (%)				.02
<25	51 (0.83)	42 (0.97)	9 (0.50)	
25-40	1559 (25.45)	1078 (24.99)	481 (26.55)	
41-60	3488 (56.95)	2438 (56.53)	1050 (57.95)	
>60	1027 (16.77)	755 (17.51)	272 (15.01)	
Gender, n (%)				.03
Women	3445 (56.24)	2378 (55.14)	1067 (58.89)	
Men	2667 (43.54)	1925 (44.63)	742 (40.95)	
Nonbinary	9 (0.15)	6 (0.14)	3 (0.17)	
Prefers not to answer	4 (0.07)	4 (0.09)	0 (0)	
BMI (kg/m^2), mean (SD)[a]	29.5 (6.9)	29.4 (6.9)	29.7 (6.9)	.24
BMI category, n (%)[a]				.39
Underweight (<18.5 kg/m^2)	48 (0.78)	38 (0.88)	10 (0.55)	
Normal (18.5-25 kg/m^2)	1572 (25.67)	1125 (26.08)	447 (24.67)	
Overweight (≥25-30 kg/m^2)	2163 (35.31)	1521 (35.27)	642 (35.43)	
Obese (>30-40 kg/m^2)	1847 (30.16)	1278 (29.63)	569 (31.40)	
Morbidly obese (40 kg/m^2)	488 (7.97)	346 (8.02)	142 (7.84)	
Race and ethnicity, n (%)				.005
Asian	567 (9.26)	388 (9.0)	179 (9.88)	
Black	564 (9.21)	395 (9.16)	169 (9.33)	
Hispanic	653 (10.66)	469 (10.87)	184 (10.15)	
Non-Hispanic White	3524 (57.53)	2442 (56.62)	1082 (59.71)	
Other	147 (2.40)	105 (2.43)	42 (2.32)	
Not available/prefers not to specify	670 (10.94)	514 (11.92)	156 (8.61)	
Employment status, n (%)				<.001
Full-time employed	5119 (83.58)	3641 (84.42)	1478 (81.57)	
Part-time employed	440 (7.18)	265 (6.14)	175 (9.66)	
Not employed	482 (7.87)	345 (8.00)	137 (7.56)	
Not available/prefers not to answer	84 (1.37)	62 (1.44)	22 (1.21)	
Education level, n (%)				.13
Less than a high school diploma	61 (1.0)	48 (1.11)	13 (0.72)	
High school diploma	512 (8.36)	355 (8.23)	157 (8.66)	
Some college	1515 (24.73)	1089 (25.25)	426 (23.51)	
Bachelor's degree	2328 (38.01)	1650 (38.26)	678 (37.42)	
Graduate degree	1709 (27.90)	1171 (27.15)	538 (29.69)	
Social Deprivation Index[b], n (%)				.81
Category 1 (0-20)	1887 (30.81)	1316 (30.51)	571 (31.51)	
Category 2 (21-40)	1495 (24.41)	1052 (24.39)	443 (24.45)	
Category 3 (41-60)	1136 (18.55)	796 (18.46)	340 (18.76)	

Demographics	Overall cohort (n=6125)	Train data set (n=4313)	Test data set (n=1812)	*P* value
Category 4 (61-80)	961 (15.69)	686 (15.91)	275 (15.18)	
Category 5 (81-100)	624 (10.19)	449 (10.41)	175 (9.66)	
Clinical presentation, mean (SD)				
Pain	5.7 (1.4)	5.7 (1.4)	5.7 (1.4)	.36
Acuity, n (%)				.24
Acute	1168 (19.07)	839 (19.45)	329 (18.16)	
Chronic	4957 (80.93)	3474 (80.55)	1483 (81.84)	
Oswestry Disability Index, mean (SD)	23.1 (12.7)	22.6 (12.5)	24.4 (13.2)	<.001
Fear-Avoidance Beliefs Questionnaire ≥15 (n=1391), mean (SD)	2.8 (18.1)	2.8 (18.1)	2.8 (18.0)	.52
Anxiety (7-item General Anxiety Disorder Scale) ≥5 (n=2144), mean (SD)	3.8 (8.7)	3.9 (8.6)	3.8 (8.7)	.66
Depression (9-item Patient Health Questionnaire) ≥5 (n=1590), mean (SD)	4.2 (9.4)	4.2 (9.4)	4.2 (9.4)	.87
Overall work (Work Productivity and Activity Impairment Questionnaire for General Health overall) score >0 (n=3296), mean (SD)	22.9 (34.0)	22.8 (34.0)	23.0 (33.9)	.85
Activities (Work Productivity and Activity Impairment Questionnaire for General Health) >0 (n=4900), mean (SD)	22.5 (38.8)	22.5 (39.0)	22.4 (38.3)	.32

[a]Denotes 7 missing values.

[b]Denotes 22 missing values (higher quantiles indicate higher social deprivation).

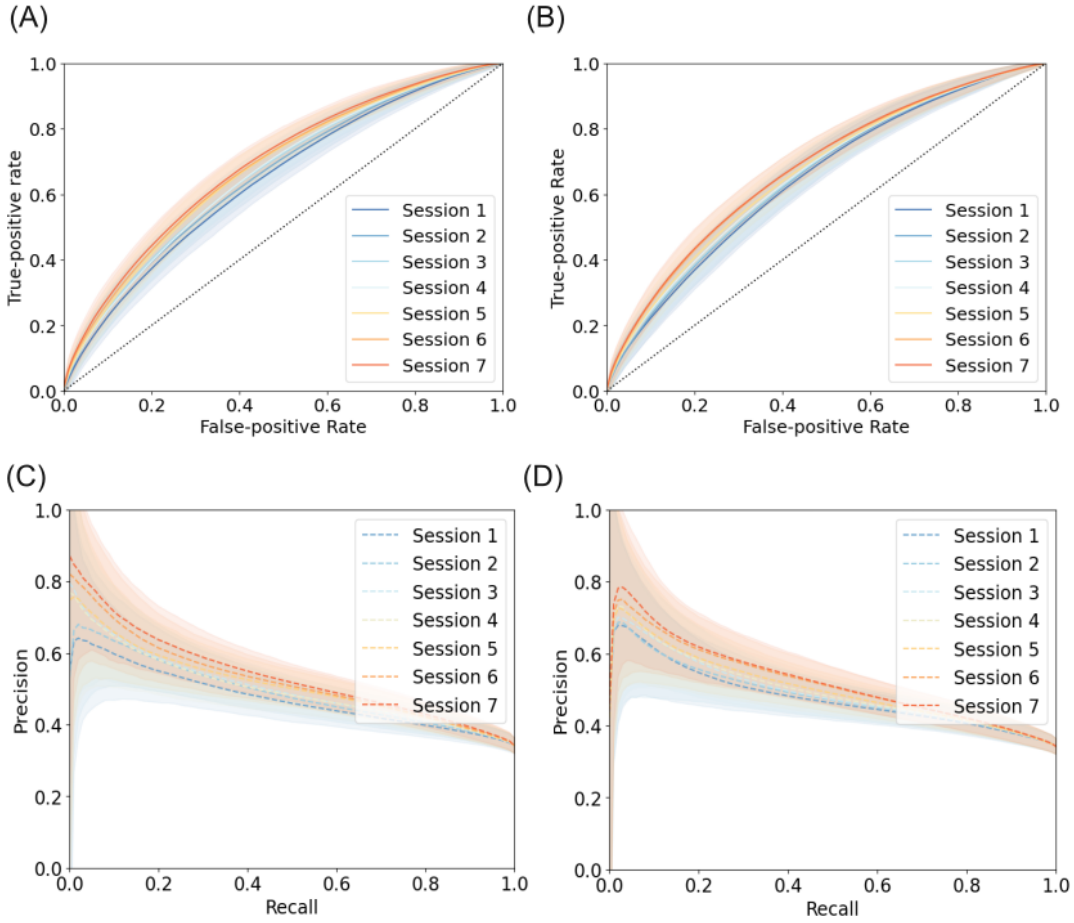
Overall, the cohort reported moderate pain, with an NPRS score of 5.7 out of 10.0 (SD 1.4), and a disability score on the Oswestry Disability Index (ODI) [44] of 23.1 (SD 12.7) at baseline. This was primarily associated with chronic pain, defined as pain lasting 3 months or longer (4957/6125, 80.93%). Severe pain (NPRS: 8-10) was predominantly reported by patients with acute LBP (168/1168, 14.38% vs 519/4957, 10.47%, in chronic cases, *P*<.001). High fear-avoidance beliefs (Fear-Avoidance Beliefs Questionnaire score ≥15 [45]) were reported by 1391 of the 6125 (22.71%) patients of the cohort, while 2144 (35.0%) and 1590 (25.96%) patients reported at least mild anxiety (7-item General Anxiety Disorder Scale [GAD-7] score ≥5 [31]) or depression (9-item Patient Health Questionnaire [PHQ-9] score ≥5 [32]), respectively. A large majority (4900/6125, 80.0%) reported impairment in daily activities (Work Productivity and Activity Impairment Questionnaire for General Health [WPAI]—activity score >0), while 3296 of the 6125 (53.81%) patients reported either presenteeism or absenteeism (WPAI overall score >0). Clinical outcomes were similar across both the training and test data sets, with the exception of the ODI, which was statistically higher in the test data set (22.6, SD 12.5 vs 24.4, SD 13.2), although this difference was not clinically meaningful [46].

Model Development

The LGCA revealed a significantly different latent representation of trunk motion over time for responders and nonresponders (*P*<.001; Figure S4 in Multimedia Appendix 1). Similarly, the temporal structural network model showed 2 significantly different temporal correlation patterns of trunk motion for responders and nonresponders (*P*<.001; Figure S5 in Multimedia Appendix 1).

The best models from both methodologies (LightGBM and RNN), following hyperparameter optimization, resulted in similar ROC-AUC scores in the test set across all time points (*P*=.98, .52, .69, .81, .38, .68, and .22, respectively, for sessions 1-7; Figure 3A and B). Model performance improved over time for both models (*P*<.001 for both, DeLong algorithm), with predictions at session 7 reaching an ROC-AUC of 0.70 (95% CI 0.65-0.71, based on bootstrap resampling; Table S3 in Multimedia Appendix 1) for RNN and 0.71 (95% CI 0.67-0.72) for LightGBM, compared with 0.66 (95% CI 0.61-0.68) and 0.67 (95% CI 0.61-0.67) obtained in session 1. These results corresponded to a precision-recall AUC at session 7 of 0.56 (LightGBM: 95% CI 0.49-0.58; RNN 95% CI 0.48-0.57; Figure 3C and D) and a weighted F_1-score of 0.68 (95% CI 0.64-0.69) for both models.

Figure 3. Model performance on the test data set. Test receiver operating characteristic curve (ROC) for both (A) light gradient boosting machine (LightGBM) and (B) recurrent neural network (RNN) across the 7 sessions. True positive rate (true positives over true positives and false negatives); false positive rate (false positives over false positives and true negatives); dashed line denotes an area under the curve (AUC) of 0.5 corresponding to a random predictor; precision-recall curve for (C) LightGBM and (D) RNN models across the 7 sessions. Precision denotes true positives over true positives and false positives; recall denotes true positives over true positives and false negatives; shaded areas denote the 95% CIs.



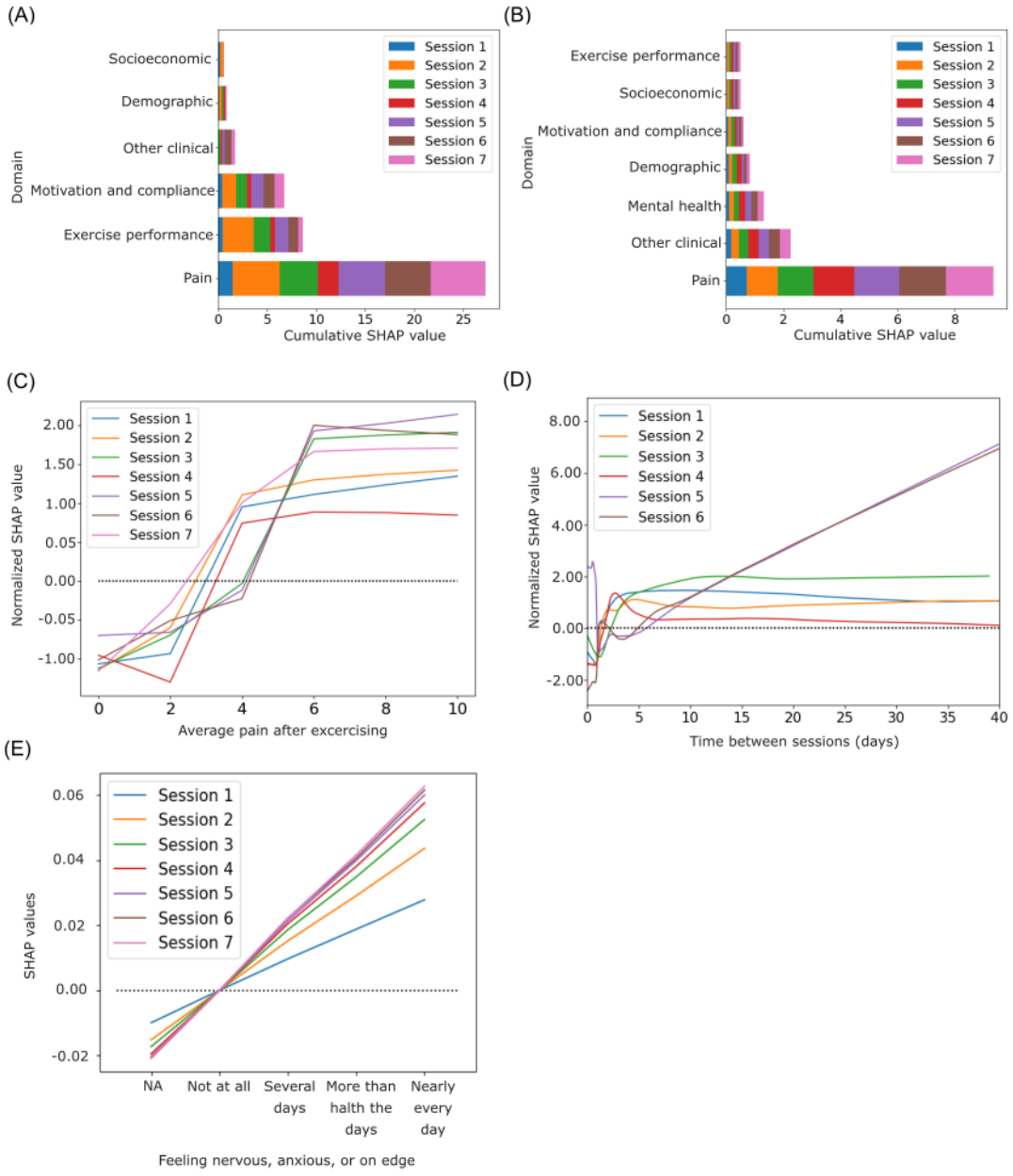
For the models described, considering a precision of 70%, the LightGBM achieved a specificity of 0.97 (95% CI 0.95-0.98), a sensitivity of 0.13 (95% CI 0.02-0.25), and a negative predictive value of 0.68 (95% CI 0.68-0.68). Similarly, the RNN achieved a specificity of 0.97 (95% CI 0.94-0.97), a sensitivity of 0.12 (95% CI 0.00-0.24), and a negative predictive value of 0.68 (95% CI 0.65-0.71).

Predictive Factors in Digital Care

Considering the best models, we analyzed the top 20 outcome predictors at each session based on absolute mean SHAP values [41]. Figure 4 depicts cumulative values of features aggregated by domains over time for the LightGBM (Figure 4A) and RNN (Figure 4B). Detailed feature stratification for both models is available in Figures S6 and S7 in Multimedia Appendix 1, respectively.

Although both models incorporated features from comparable domains, the importance of each feature varied depending on the model. The flagged domains included pain-associated metrics, exercise performance, and motivation and compliance data, which utilized real-time passive data collection.

Figure 4. Model explainability. Cumulative Shapley Additive Explanations (SHAP) values per domain considering the top 20 features at each session for both (A) light gradient boosting machine (LightGBM) and (B) recurrent neural network (RNN). SHAP values depicting the relationship between the outcome (ie, pain response) and the feature of interest: (C) average pain felt during exercising, (D) time between sessions (days), and (E) feeling nervous, anxious, or on edge (7-item General Anxiety Disorder Scale [GAD-7] scale). As LightGBM models differ in the number of features across time, and because we are interested in qualitative comparisons, SHAP values were normalized at each session by dividing by the SD of SHAP values.



Although the pain domain was highly predictive for both models (Figures 4A and B), LightGBM placed greater emphasis on exercise performance, motivation, and compliance data than the RNN. Within the exercise performance domain, the most informative features included time spent exercising, exercise accuracy, and the execution of specific exercises related to trunk motion. Notably, lower training time or poorer performance in particular sessions was associated with an increased likelihood of nonresponse (Figure S8A in Multimedia Appendix 1). Low motivation and compliance, specifically characterized by low exercise consistency—evaluated by longer intervals between sessions (>3 days, Figure 4D)—was associated with a higher likelihood of being a nonresponder. This observation was consistent across other variables reflecting motivation toward compliance, such as the time between registration and program start, as well as the reasons preventing patients from completing the program.

The RNN was more reliant on the temporal patterns of mental health and other clinical variables (eg, medication intake, intent to undergo surgery) for its predictions. The presence of mental distress was associated with a higher likelihood of being a nonresponder across all sessions, as exemplified by the item "Feeling nervous, anxious, or on edge" from the GAD-7 scale (Figure 4E). Even though LightGBM did not rank mental health features among the top 20 predictors, higher GAD-7 and PHQ-9 scores were correlated with poorer outcomes, particularly in the later sessions (eg, Figure S8B in Multimedia Appendix 1). Additionally, other clinical features, such as prescribed

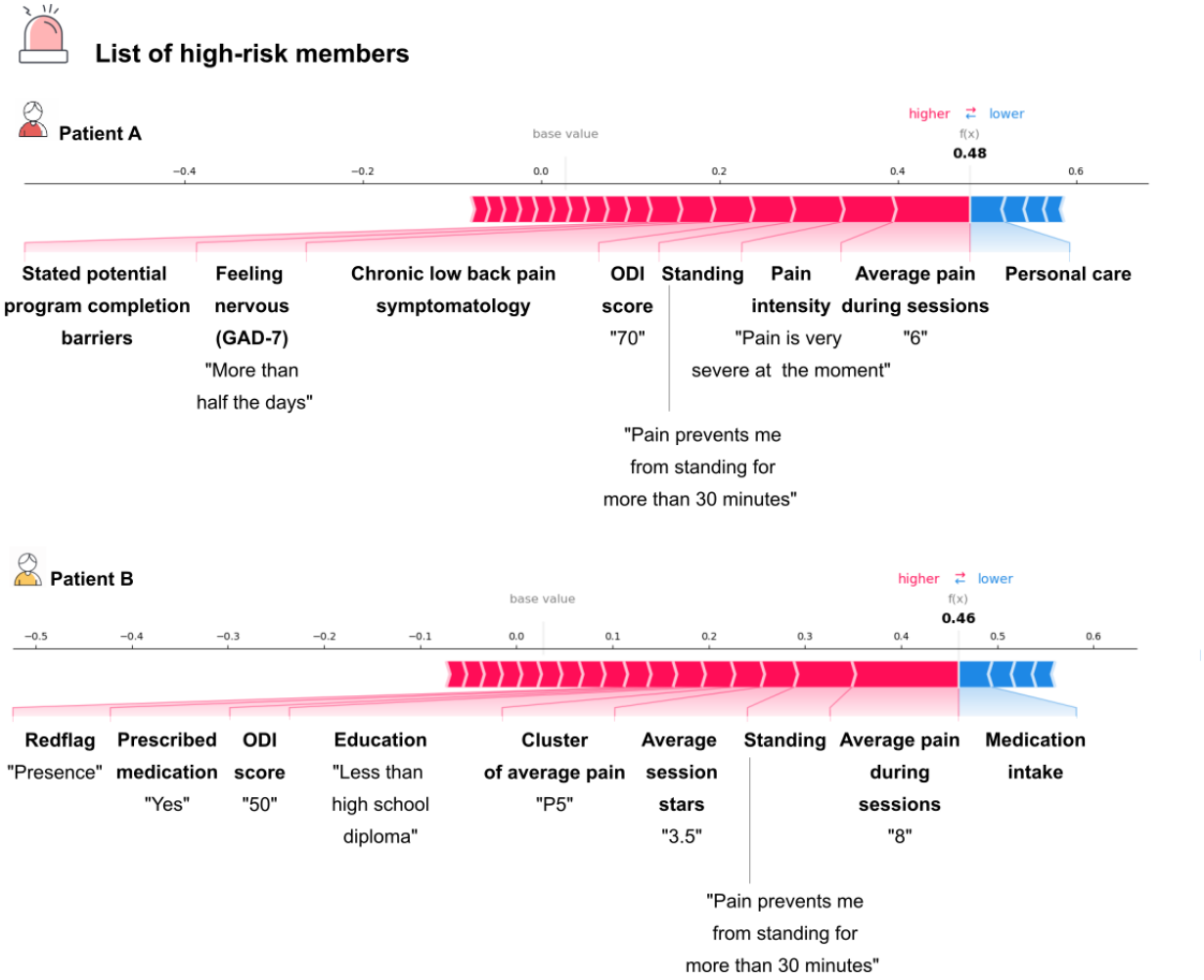
medication, were associated with a lower probability of being a responder.

Both models utilized features related to individual characteristics, including demographics and socioeconomic status. Subgroup analyses evaluating potential imbalances between the training and test sets indicated that similar results could be obtained regardless of demographic or clinical characteristics (Table S4 in Multimedia Appendix 1). Overall, women and individuals from more socially deprived areas, particularly those from regions with higher unemployment rates, had a worse prognosis compared with their counterparts (Figure S8C and S8D in Multimedia Appendix 1).

Model Applicability

Figure 5 illustrates an example of the output from the tool provided on the PT portal for 2 patients during session 4. This output includes the model's prediction of nonresponders (ie, true positives, with an RNN precision of 70% on the test set) and the key drivers behind this classification. For patient A, the classification is primarily influenced by high pain reported during exercise in session 4 (NPRS = 6 out of 10), baseline characteristics indicating severe clinical presentation (such as high ODI and mental distress), and self-reported potential barriers to motivation. Patient B also reported high pain during exercise in session 4 (NPRS = 8 out of 10) and exhibited low exercise performance for that session (3.5 out of 5 stars). Additionally, patient B indicated prescribed medication intake (eg, opioids) and the presence of at least one clinical red flag, despite being cleared by a physician to participate in the program. With an ODI score of 50 and low educational attainment, this information is subsequently analyzed by the PT, allowing for necessary refinements to the intervention. This may involve adjusting exercise intensity or type or prioritizing specific components of the DCP, such as education and cognitive behavioral therapy. These adjustments will subsequently impact future model predictions in a dynamic closed-loop optimization that contributes to the refinement of the tool.

Figure 5. Example of the integration of the model in the physical therapist portal. Example of model explainability for 2 patients, A and B, classified as high-risk nonresponders (recurrent neural network [RNN] precision 70%) at session 4. In the physical therapist portal, each patient's health record includes the prediction of the tool along with the variables that sustained the model's classification, providing insights into the specific factors affecting the patient's prognosis. The most influential factors contributing to a nonresponder prediction are highlighted in red, with factors positioned toward higher "base values" (ie, to the right) indicating a stronger impact. For example in the case of patient A, the pain felt during exercise and the baseline pain are the most influential for the model output. After analyzing the case, the physical therapist reevaluates the patient's program and the potential need for further support. Those adjustments will be subsequently fed into the tool reinforcing the dynamic optimization close loop. GAD-7: 7-item General Anxiety Disorder Scale; ODI: Oswestry Disability Index.



Discussion

Principal Findings

This study leveraged passively and automatically collected patient data from a digital MSK intervention to develop a predictive tool that assists PTs in optimizing treatment. These models utilize continuously updated data from each enrolled patient to identify early those who are less likely to achieve a clinically meaningful reduction in pain. Model performance significantly increased over time, achieving an ROC-AUC of 0.70 (95% CI 0.65-0.71, from bootstrap resampling) for the RNN and 0.71 (95% CI 0.67-0.72) for the LightGBM at session 7. These models identified features from domains related to pain, exercise performance, motivation, and compliance as the most predictive of outcomes. This information can inform PTs for continuous treatment adjustments to maximize response rates. This predictive tool operates through a dynamic loop, and when applied in clinical practice, it continuously enhances treatment refinement and improves tool performance over time.

Comparison Prior Work

Previous ML models designed to support LBP management without real-time data collection have reported performances (ROC-AUCs) ranging from 0.49 to 0.71 [6,7,9,10,21], with 1 model achieving a higher AUC of 0.84, but only near the end of treatment [11]. Additionally, the currently used STarT Back prediction tool has reported AUCs of 0.63 and 0.60-0.62 when predicting pain intensity at 12 weeks and 1 year, respectively [6,7]. The wide range of reported model performances reflects not only the challenges associated with each proposed objective but also the level of uncertainty that these tools generate. In this study, both models' performances significantly improved over time, achieving an AUC of 0.70-0.71 at session 7 using real-world data. This AUC suggests that pain response at an early stage of the intervention is partially explained by the variables used. However, as treatment progresses, it is plausible that model performance will continue to improve toward the end. As an example, Brennan et al [11] demonstrated a 1.3 times increase in model performance when using PROMs collected at session 3 to predict outcomes at session 6. Our study aimed to be less reliant on PROMs and more focused on a scalable solution capable of improving over time [47,48]. While specificity and sensitivity depend on the threshold applied, the results presented here align with those previously reported [6,7,9-11,21]. The choice of threshold to balance precision and recall must align with the clinical setting's needs, taking into account the costs and implications of false positives and false negatives.

In addition to real-time predictions, these models may offer timely insights that enable PTs to determine the need for referring patients to specialized care (eg, escalation to psychologists) through shared decision-making with the patient. Both models identified similar feature domains, including pain-associated metrics, exercise performance, motivation, and compliance, thereby surpassing the predictive value of individual demographic prognostic factors. These domains cannot be captured solely through PROMs, further emphasizing the relevance of such models in managing LBP.

To predict the pain that patients experience during daily activities, the model incorporates pain levels reported during exercise, alongside baseline clinical presentations (including the NPRS, the "pain intensity" item from the ODI, and acuity)—features previously established as poor prognostic indicators [49-51]. This underscores the importance of pain experienced during exercise as a predictor of outcomes. Specifically, exercising with pain levels exceeding 4 out of 10 has been shown to predict negative outcomes [52], suggesting that these patients may benefit from adjustments to their exercise prescriptions. The use of analgesic medications, such as opioids, has also been linked to a poorer prognosis, as noted in previous studies [53-55].

Aside from pain, both models prioritized motivation, compliance, and exercise performance data for predictions. Regarding motivation, indicators such as self-reported resistance to commit to the intervention, delayed program initiation, and extended intervals between sessions (>3 days) were associated with a lower likelihood of response, reinforcing previous findings [56]. In the exercise performance domain, the models emphasized exercise execution data, including the time dedicated to sessions, performance scores, and the number of incorrect exercises, with lower scores linked to worse outcomes. Challenges associated with compliance frequently contribute to increased frustration and decreased adherence to treatment [57,58] and have been previously described as risk factors for poor outcomes [5,59]. Co-occurring with low adherence rates, anxiety and depression were associated with a lower likelihood of response in one of the models (RNN). This finding is unsurprising, considering the well-established negative feedback loop between mental health and MSK pain [60,61].

Although the demographic and socioeconomic profiles of patients are relevant, they may be of lower importance than the previously mentioned domains. Certain groups, such as women and the unemployed, may encounter systemic barriers that hinder their ability to achieve pain relief.

In summary, implementing this tool would enhance care coordination and patient management by identifying individuals at higher risk for poor pain outcomes. This, in turn, can streamline workflows, fostering informed decision-making and tailored treatment programs. The continuous stream of data generated by patients enables refined predictions at each interaction, allowing PTs to make timely adjustments to the program. These adjustments could then be fed back into the system to further enhance its recommendations. To assess potential biases, the subgroup analyses conducted in this study focused on demographic characteristics and socioeconomic factors, indicating a low risk of bias. However, further evaluation of algorithmic bias will be necessary once the model is fully operational and deployed in real-world settings. Continuous monitoring will be essential to ensure that the model does not disproportionately impact specific populations or reinforce existing health disparities.

While there may be concerns among PTs regarding the "black box" nature of AI, model explainability will be crucial in overcoming this barrier. Before full implementation, training for PTs (or clinicians) is essential to ensure they understand and

can effectively use the tool, thereby mitigating fears and preventing erroneous use of AI. Nevertheless, the design of this AI tool ensures that final clinical decisions remain with the PT, preserving human judgment and critical reasoning. This approach can foster trust in the system and promote the safe integration of AI into clinical workflows.

Future research on the development of this predictive tool should concentrate on clinically validating the closed feedback loop. Additionally, optimizing the performance of the AI tool may be further enhanced by incorporating supplementary data sources, such as real-life context data variables obtainable from smartwatches (eg, pedometry and sleep quality) and clinical context information (eg, presence of specific comorbidities and clinical records). This could reveal additional data patterns and, consequently, contribute to improved model accuracy. With the nearly exponential increase in data, it is plausible that RNNs or other deep sequence models will continue to enhance their performance and eventually surpass LightGBM, given their natural ability to handle sequential data [47,48]. Additionally, the integration of continuous volumes of data may expand the model's predictive capabilities to assess long-term outcomes. In particular, for LBP, predicting long-term outcomes will help determine which types of interventions lead to more sustained recovery with functional improvement, and reduce the likelihood of relapse or chronicity over time.

Finally, as the AI model relies on protected health information, implementing robust safeguards and establishing data stewardship protocols are essential to maintain confidentiality, integrity, and security.

Other future considerations include exploring the application of such tools to other MSK conditions.

Strengths and Limitations

This study has important strengths. First is the novelty of exploring how data collected from a DCP can be applied in an ML model to continuously predict patient outcomes and foster personalized care. Second, this cohort's demographic and clinical characteristics mirror that of the US population reporting LBP [62,63], with the percentage of nonresponders for pain being consistent with previous studies evaluating in-person and digital LBP management [11,22,64,65]. Finally, the nature of the data that are automatically and passively collected, which are less reliant on PROMs, makes it easy to obtain and less burdensome for both patients and clinicians. However, this study also has limitations that warrant noting: (1) Although the cohort was large and decentralized, we cannot dismiss the possibility of overfitting; additionally, as the population focuses on beneficiaries of health plans, this may limit generalizability to populations not covered by health insurance. Therefore, extended testing on new cohorts is critical for the external validation of the AI tool. Such validation would provide a more accurate assessment of the model's performance, ensuring that its effectiveness and robustness are maintained across different populations that may enroll in the program. Additionally, it could help identify areas in need of improvement. Furthermore, external validation studies should be followed by real-world applications to evaluate the usefulness of the deployed model in clinical settings; (2) despite being highly used and recommended by clinical guidelines, transforming the pain outcome into a binary classification may obscure subtle variations in patient responses. This simplification could lead to misclassification of patients who are close to the threshold. Future approaches could consider modeling the outcome as a continuous variable. Additionally, developing a composite score that reflects improvements across multiple core outcome domains—including Patient Global Impression of Change and Activities of daily living functionality—could better capture the multifactorial nature of LBP and provide a more comprehensive assessment of patient improvement in a single metric; (3) we used an extensive number of features that capture critical domains in the rehabilitation realm; however, other potential features, such as measures of self-efficacy and indicators depicting transitional states between clusters of trajectories, which could hold high predictive power, were not included. Incorporating these features in future designs could enhance the model's predictive accuracy; (4) finally, the early stage of this tool's development hampers the ability to assess its impact on clinical outcomes. Future steps should focus on implementing and evaluating the tool's dynamic loop to better understand its effectiveness and feasibility, as well as to ascertain its net benefits.

Conclusions

This study underscores the potential of a predictive tool leveraging ML models to enhance the management of LBP in a digital setting, redirecting care pathways early on and throughout the treatment course, thereby fostering personalized care. The application of such a strategy is particularly important for managing a heterogeneous population with diverse phenotypes across different cultural contexts.

Acknowledgments

The authors acknowledge the team of PTs responsible for managing the participants, Guilherme Freches for data validation, and Daniel Joplin for his contributions to revising the data analysis and the manuscript, all of whom are current or former employees of Sword Health, Inc. The authors also acknowledge funding from the NextGeneration EU funding program under project no. 62 - "Responsible AI," which is financed by European funds.

Data Availability

The data sets used or analyzed during this study are available from the corresponding author upon reasonable request.

Code Availability

The underlying code for this study and the training/validation data sets are not publicly available but may be made accessible to qualified researchers upon reasonable request from the corresponding author.

Authors' Contributions

All authors made significant contributions to the work reported as follows: FC was responsible for the study concept and design; MM acquired the data; ACA prepared the data; RGM performed the feature engineering, while ACA developed the machine learning models and corresponding model explainability; ACA, MM, DJ, and FC interpreted the data; ACA, DJ, and FC drafted the work; CM, VY, SPC, and FDC provided critical revisions of the manuscript; and VB was responsible for funding. All authors approved the final submitted version and agreed to be accountable for all aspects of the work.

Conflicts of Interest

The authors declare the following competing financial interests: ACA, RGM, MM, DJ, CM, VY, FDC, and FC are employees of Sword Health Inc, the sponsor of this study. FDC, VY, and VB also hold equity in Sword Health Inc. SPC is an independent scientific and clinical consultant who received advisory honoraria from Sword Health.

Multimedia Appendix 1
TRIPOD-AI guidelines and additional analysis and results. TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—artificial intelligence.
[PDF File (Adobe PDF File), 2725 KB - medinform_v12i1e64806_app1.pdf]

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2021 Dec 19;396(10267):2006-2017 [FREE Full text] [doi: 10.1016/S0140-6736(20)32340-0] [Medline: 33275908]
2. Cholewicki J, Breen A, Popovich JM, Reeves NP, Sahrmann SA, van Dillen LR, et al. Can biomechanics research lead to more effective treatment of low back pain? A point-counterpoint debate. J Orthop Sports Phys Ther 2019 Jun;49(6):425-436 [FREE Full text] [doi: 10.2519/jospt.2019.8825] [Medline: 31092123]
3. Cholewicki J, Popovich JM, Aminpour P, Gray SA, Lee AS, Hodges PW. Development of a collaborative model of low back pain: report from the 2017 NASS consensus meeting. Spine J 2019 Jun;19(6):1029-1040. [doi: 10.1016/j.spinee.2018.11.014] [Medline: 30508588]
4. Lin I, Wiles L, Waller R, Goucke R, Nagree Y, Gibberd M, et al. What does best practice care for musculoskeletal pain look like? Eleven consistent recommendations from high-quality clinical practice guidelines: systematic review. Br J Sports Med 2020 Jan;54(2):79-86. [doi: 10.1136/bjsports-2018-099878] [Medline: 30826805]
5. Booth J, Moseley GL, Schiltenwolf M, Cashin A, Davies M, Hübscher M. Exercise for chronic musculoskeletal pain: a biopsychosocial approach. Musculoskeletal Care 2017 Dec;15(4):413-421. [doi: 10.1002/msc.1191] [Medline: 28371175]
6. Kendell M, Beales D, O'Sullivan P, Rabey M, Hill J, Smith A. The predictive ability of the STarT Back Tool was limited in people with chronic low back pain: a prospective cohort study. J Physiother 2018 Apr;64(2):107-113 [FREE Full text] [doi: 10.1016/j.jphys.2018.02.009] [Medline: 29602747]
7. Toh I, Chong H, Suet-Ching Liaw J, Pua Y. Evaluation of the STarT Back screening tool for prediction of low back pain intensity in an outpatient physical therapy setting. J Orthop Sports Phys Ther 2017 Apr;47(4):261-267. [doi: 10.2519/jospt.2017.7284] [Medline: 28257616]
8. Linton SJ, Boersma K. Early identification of patients at risk of developing a persistent back problem: the predictive validity of the Orebro Musculoskeletal Pain Questionnaire. Clin J Pain 2003;19(2):80-86. [doi: 10.1097/00002508-200303000-00002] [Medline: 12616177]
9. Tschuggnall M, Grote V, Pirchl M, Holzner B, Rumpold G, Fischer M. Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. Informatics in Medicine Unlocked 2021 Jan;24:100598 [FREE Full text] [doi: 10.1016/j.imu.2021.100598]
10. Zmudzki F, Smeets RJEM. Machine learning clinical decision support for interdisciplinary multimodal chronic musculoskeletal pain treatment. Front Pain Res (Lausanne) 2023;4:1177070 [FREE Full text] [doi: 10.3389/fpain.2023.1177070] [Medline: 37228809]
11. Brennan GP, Snow G, Minick KI, Stevans JM. Significant clinical improvement was predicted in a cohort of patients with low back pain early in the care process. Phys Ther 2023 Sep 01;103(9):pzad082. [doi: 10.1093/ptj/pzad082] [Medline: 37402701]
12. Tagliaferri SD, Wilkin T, Angelova M, Fitzgibbon BM, Owen PJ, Miller CT, et al. Chronic back pain sub-grouped via psychosocial, brain and physical factors using machine learning. Sci Rep 2022 Sep 07;12(1):15194 [FREE Full text] [doi: 10.1038/s41598-022-19542-5] [Medline: 36071092]

13. Aggarwal N. Prediction of low back pain using artificial intelligence modeling. J Med Artif Intell 2021 Mar 30;4:2-2. [doi: 10.21037/jmai-20-55]
14. Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. Arthritis Rheum 2008 May 15;59(5):632-641 [FREE Full text] [doi: 10.1002/art.23563] [Medline: 18438893]
15. Traeger AC, Henschke N, Hübscher M, Williams CM, Kamper SJ, Maher CG, et al. Estimating the risk of chronic pain: development and validation of a prognostic model (PICKUP) for patients with acute low back pain. PLoS Med 2016 May;13(5):e1002019 [FREE Full text] [doi: 10.1371/journal.pmed.1002019] [Medline: 27187782]
16. Knoop J, van Lankveld W, Beijer L, Geerdink FJB, Heymans MW, Hoogeboom TJ, et al. Development and internal validation of a machine learning prediction model for low back pain non-recovery in patients with an acute episode consulting a physiotherapist in primary care. BMC Musculoskelet Disord 2022 Sep 03;23(1):834 [FREE Full text] [doi: 10.1186/s12891-022-05718-7] [Medline: 36057717]
17. Tanguay-Sabourin C, Fillingim M, Guglietti GV, Zare A, Parisien M, Norman J, PREVENT-AD Research Group, et al. A prognostic risk score for development and spread of chronic pain. Nat Med 2023 Jul;29(7):1821-1831 [FREE Full text] [doi: 10.1038/s41591-023-02430-4] [Medline: 37414898]
18. Shim JG, Ryu KH, Cho EA, Ahn JH, Kim HK, Lee YJ, et al. Machine learning approaches to predict chronic lower back pain in people aged over 50 years. Medicina (Kaunas) 2021 Nov 11;57(11):1230 [FREE Full text] [doi: 10.3390/medicina57111230] [Medline: 34833448]
19. Langenberger B, Schrednitzki D, Halder AM, Busse R, Pross CM. Predicting whether patients will achieve minimal clinically important differences following hip or knee arthroplasty. Bone Joint Res 2023 Sep 01;12(9):512-521 [FREE Full text] [doi: 10.1302/2046-3758.129.BJR-2023-0070.R2] [Medline: 37652447]
20. Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. BMC Med Inform Decis Mak 2019 Jan 08;19(1):3 [FREE Full text] [doi: 10.1186/s12911-018-0731-6] [Medline: 30621670]
21. da Silva T, Macaskill P, Kongsted A, Mills K, Maher C, Hancock M. Predicting pain recovery in patients with acute low back pain: updating and validation of a clinical prediction model. Eur J Pain 2019 Feb;23(2):341-353. [doi: 10.1002/ejp.1308] [Medline: 30144211]
22. Cui D, Janela D, Costa F, Molinos M, Areias AC, Moulder RG, et al. Randomized-controlled trial assessing a digital care program versus conventional physiotherapy for chronic low back pain. NPJ Digit Med 2023 Jul 07;6(1):121 [FREE Full text] [doi: 10.1038/s41746-023-00870-3] [Medline: 37420107]
23. Raiszadeh K, Tapicer J, Taitano L, Wu J, Shahidi B. In-clinic versus web-based multidisciplinary exercise-based rehabilitation for treatment of low back pain: prospective clinical trial in an integrated practice unit model. J Med Internet Res 2021 Mar 18;23(3):e22548 [FREE Full text] [doi: 10.2196/22548] [Medline: 33734088]
24. Toelle TR, Utpadel-Fischler DA, Haas K, Priebe JA. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: a randomized controlled trial. NPJ Digit Med 2019 May 03;2:34. [doi: 10.1038/s41746-019-0109-x] [Medline: 31304380]
25. Wong JJ, Côté P, Sutton DA, Randhawa K, Yu H, Varatharajan S, et al. Clinical practice guidelines for the noninvasive management of low back pain: a systematic review by the Ontario Protocol for Traffic Injury Management (OPTIMa) Collaboration. Eur J Pain 2017 Feb;21(2):201-216. [doi: 10.1002/ejp.931] [Medline: 27712027]
26. Joypaul S, Kelly F, McMillan SS, King MA. Multi-disciplinary interventions for chronic pain involving education: a systematic review. PLoS One 2019;14(10):e0223306 [FREE Full text] [doi: 10.1371/journal.pone.0223306] [Medline: 31577827]
27. Williams ACDC, Eccleston C, Morley S. Psychological therapies for the management of chronic pain (excluding headache) in adults. Cochrane Database Syst Rev 2012 Aug 12;11:CD007407. [doi: 10.1002/14651858.CD007407.pub3] [Medline: 23152245]
28. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 2008 Feb;9(2):105-121. [doi: 10.1016/j.jpain.2007.09.005] [Medline: 18055266]
29. Boonstra AM, Stewart RE, Köke AJA, Oosterwijk RFA, Swaan JL, Schreurs KMG, et al. Cut-off points for mild, moderate, and severe pain on the Numeric Rating Scale for Pain in patients with chronic musculoskeletal pain: variability and influence of sex and catastrophizing. Front Psychol 2016;7:1466 [FREE Full text] [doi: 10.3389/fpsyg.2016.01466] [Medline: 27746750]
30. Butler DC, Petterson S, Phillips RL, Bazemore AW. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. Health Serv Res 2013 Apr;48(2 Pt 1):539-559 [FREE Full text] [doi: 10.1111/j.1475-6773.2012.01449.x] [Medline: 22816561]
31. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
32. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Oct;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

33. Finucane LM, Downie A, Mercer C, Greenhalgh SM, Boissonnault WG, Pool-Goudzwaard AL, et al. International framework for red flags for potential serious spinal pathologies. J Orthop Sports Phys Ther 2020 Jul;50(7):350-372. [doi: 10.2519/jospt.2020.9971] [Medline: 32438853]
34. Troke M, Moore AP, Maillardet FJ, Cheek E. A normative database of lumbar spine ranges of motion. Man Ther 2005 Aug;10(3):198-206. [doi: 10.1016/j.math.2004.10.004] [Medline: 16038855]
35. Iacobucci D. Structural equations modeling: fit indices, sample size, and advanced topics. J Consum Psychol 2010 Jan;20(1):90-98 [FREE Full text] [doi: 10.1016/j.jcps.2009.09.003]
36. Brown TA. In: Kenny DA, Little TD, editors. Confirmatory Factor Analysis for Applied Research (2nd Edition). New York, NY: The Guilford Press; 2015:206-273.
37. Epskamp S, Waldorp LJ, Mõttus R, Borsboom D. The Gaussian graphical model in cross-sectional and time-series data. Multivariate Behav Res 2018;53(4):453-480. [doi: 10.1080/00273171.2018.1454823] [Medline: 29658809]
38. Ram N, Grimm KJ. Growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. Int J Behav Dev 2009;33(6):565-576 [FREE Full text] [doi: 10.1177/0165025409343765] [Medline: 23885133]
39. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. New York, NY: Association for Computing Machinery; 2019 Presented at: The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; August 4-8, 2019; Anchorage, AK p. 2623-2631. [doi: 10.48550/arXiv.1907.10902]
40. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Red Hook, NY: Curran Associates, Inc; 2019 Dec Presented at: The 33rd International Conference on Neural Information Processing Systems; December 8-14, 2019; Vancouver, BC, Canada p. 8026-8037 URL: https://dl.acm.org/doi/10.5555/3454287.3455008
41. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
42. Bento J, Saleiro P, Cruz AF, Figueiredo MAT, Bizarro P. TimeSHAP: explaining recurrent models through sequence perturbations. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York, NY: Association for Computing Machinery; 2021 Presented at: The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining; August 14-18, 2021; Virtual Event, Singapore p. 2565-2573. [doi: 10.48550/arXiv.2012.00073]
43. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett 2014 Nov;21(11):1389-1393. [doi: 10.1109/LSP.2014.2337313]
44. Vianin M. Psychometric properties and clinical usefulness of the Oswestry Disability Index. J Chiropr Med 2008 Dec;7(4):161-163 [FREE Full text] [doi: 10.1016/j.jcm.2008.07.001] [Medline: 19646379]
45. Wertli MM, Rasmussen-Barr E, Weiser S, Bachmann LM, Brunner F. The role of fear avoidance beliefs as a prognostic factor for outcome in patients with nonspecific low back pain: a systematic review. Spine J 2014 May 01;14(5):816-36.e4. [doi: 10.1016/j.spinee.2013.09.036] [Medline: 24412032]
46. Ostelo RWJG, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine (Phila Pa 1976) 2008 Jan 01;33(1):90-94. [doi: 10.1097/BRS.0b013e31815e3a10] [Medline: 18165753]
47. Ahmed DM, Hassan MM, Mstafa RJ. A review on deep sequential models for forecasting time series data. Applied Computational Intelligence and Soft Computing 2022 Jun 3;2022:1-19. [doi: 10.1155/2022/6596397]
48. Wen M, Lin R, Wang H, Yang Y, Wen Y, Mai L, et al. Large sequence models for sequential decision-making: a survey. Front Comput Sci 2023 Aug 05;17(6):176349. [doi: 10.1007/s11704-023-2689-5]
49. Artus M, Campbell P, Mallen CD, Dunn KM, van der Windt DAW. Generic prognostic factors for musculoskeletal pain in primary care: a systematic review. BMJ Open 2017 Jan 17;7(1):e012901 [FREE Full text] [doi: 10.1136/bmjopen-2016-012901] [Medline: 28096253]
50. Cruz EB, Canhão H, Fernandes R, Caeiro C, Branco JC, Rodrigues AM, et al. Prognostic indicators for poor outcomes in low back pain patients consulted in primary care. PLoS One 2020;15(3):e0229265 [FREE Full text] [doi: 10.1371/journal.pone.0229265] [Medline: 32218561]
51. Garcia AN, Costa LOP, Costa LDCM, Hancock M, Cook C. Do prognostic variables predict a set of outcomes for patients with chronic low back pain: a long-term follow-up secondary analysis of a randomized control trial. J Man Manip Ther 2019 Sep;27(4):197-207 [FREE Full text] [doi: 10.1080/10669817.2019.1597435] [Medline: 30946005]
52. Smith BE, Hendrick P, Smith TO, Bateman M, Moffatt F, Rathleff MS, et al. Should exercises be painful in the management of chronic musculoskeletal pain? A systematic review and meta-analysis. Br J Sports Med 2017 Dec;51(23):1679-1687 [FREE Full text] [doi: 10.1136/bjsports-2016-097383] [Medline: 28596288]
53. Yerneni K, Nichols N, Abecassis ZA, Karras CL, Tan LA. Preoperative opioid use and clinical outcomes in spine surgery: a systematic review. Neurosurgery 2020 Jun 01;86(6):E490-E507. [doi: 10.1093/neuros/nyaa050] [Medline: 32271911]
54. Cohen SP, Bhatia A, Buvanendran A, Schwenk ES, Wasan AD, Hurley RW, et al. Consensus guidelines on the use of intravenous ketamine infusions for chronic pain from the American Society of Regional Anesthesia and Pain Medicine, the American Academy of Pain Medicine, and the American Society of Anesthesiologists. Reg Anesth Pain Med 2018 Jul;43(5):521-546 [FREE Full text] [doi: 10.1097/AAP.0000000000000808] [Medline: 29870458]

55. Chen Y, Vu TH, Chinchilli VM, Farrag M, Roybal AR, Huh A, et al. Clinical and technical factors associated with knee radiofrequency ablation outcomes: a multicenter analysis. Reg Anesth Pain Med 2021 Apr;46(4):298-304. [doi: 10.1136/rapm-2020-102017] [Medline: 33558282]
56. Essery R, Geraghty AWA, Kirby S, Yardley L. Predictors of adherence to home-based physical therapies: a systematic review. Disabil Rehabil 2017 Mar;39(6):519-534. [doi: 10.3109/09638288.2016.1153160] [Medline: 27097761]
57. Nicholas M, Asghari A, Corbett M, Smeets R, Wood B, Overton S, et al. Is adherence to pain self-management strategies associated with improved pain, depression and disability in those with disabling chronic pain? Eur J Pain 2012 Jan;16(1):93-104. [doi: 10.1016/j.ejpain.2011.06.005] [Medline: 21705246]
58. Argent R, Daly A, Caulfield B. Patient involvement with home-based exercise programs: can connected health interventions influence adherence? JMIR Mhealth Uhealth 2018 Mar 01;6(3):e47 [FREE Full text] [doi: 10.2196/mhealth.8518] [Medline: 29496655]
59. Young JL, Rhon DI, Cleland JA, Snodgrass SJ. The influence of exercise dosing on outcomes in patients with knee disorders: a systematic review. J Orthop Sports Phys Ther 2018 Mar;48(3):146-161. [doi: 10.2519/jospt.2018.7637] [Medline: 29320945]
60. de Vos Andersen N, Kent P, Hjort J, Christiansen DH. Clinical course and prognosis of musculoskeletal pain in patients referred for physiotherapy: does pain site matter? BMC Musculoskelet Disord 2017 Mar 29;18(1):130 [FREE Full text] [doi: 10.1186/s12891-017-1487-3] [Medline: 28356140]
61. Nieminen LK, Pyysalo LM, Kankaanpää MJ. Prognostic factors for pain chronicity in low back pain: a systematic review. Pain Rep 2021;6(1):e919 [FREE Full text] [doi: 10.1097/PR9.0000000000000919] [Medline: 33981936]
62. Quick facts United States. U.S. Census Bureau. 2022. URL: https://www.census.gov/quickfacts/fact/table/US/PST045222 [accessed 2024-01-29]
63. Zhang T, Liu Z, Liu Y, Zhao J, Liu D, Tian Q. Obesity as a risk factor for low back pain: a meta-analysis. Clin Spine Surg 2018 Feb;31(1):22-27. [doi: 10.1097/BSD.0000000000000468] [Medline: 27875413]
64. Bagg MK, Wand BM, Cashin AG, Lee H, Hübscher M, Stanton TR, et al. Effect of graded sensorimotor retraining on pain intensity in patients with chronic low back pain: a randomized clinical trial. JAMA 2022 Aug 02;328(5):430-439 [FREE Full text] [doi: 10.1001/jama.2022.9930] [Medline: 35916848]
65. Cecchi F, Pasquini G, Paperini A, Boni R, Castagnoli C, Pistritto S, et al. Predictors of response to exercise therapy for chronic low back pain: result of a prospective study with one year follow-up. Eur J Phys Rehabil Med 2014 Apr;50(2):143-151 [FREE Full text] [Medline: 24429917]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
DPC: digital care program
GAD-7: 7-item General Anxiety Disorder Scale
IMMPACT: Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials
LBP: low back pain
LGCA: latent growth curve analysis
LightGBM: light gradient boosting machine
ML: machine learning
MSK: musculoskeletal
NPRS: Numerical Pain Rating Scale
ODI: Oswestry Disability Index
PROM: patient-reported outcome measure
PT: physical therapist
RNN: recurrent neural network
ROC: receiver operating characteristic
ROM: range of motion
SHAP: Shapley Additive Explanations
SMOTE: synthetic minority oversampling technique
TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—artificial intelligence
WPAI: Work Productivity and Activity Impairment Questionnaire for General Health

Edited by C Lovis; submitted 26.07.24; peer-reviewed by M Popovic; comments to author 27.08.24; revised version received 05.09.24; accepted 23.10.24; published 19.11.24.

Please cite as:
C Areias A, G Moulder R, Molinos M, Janela D, Bento V, Moreira C, Yanamadala V, P Cohen S, Dias Correia F, Costa F
Predicting Pain Response to a Remote Musculoskeletal Care Program for Low Back Pain Management: Development of a Prediction Tool
JMIR Med Inform 2024;12:e64806
URL: https://medinform.jmir.org/2024/1/e64806
doi:10.2196/64806
PMID:

Corrigenda and Addenda

Correction: A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study

Zhixiang Zhao[1,2*], MD, PhD; Che-Ming Wu[3*], MEng; Shuping Zhang[1,2], MD, PhD; Fanping He[1,2], MD, PhD; Fangfen Liu[1,2], MD, PhD; Ben Wang[1,2], MD, PhD; Yingxue Huang[1,2], MD, PhD; Wei Shi[1,2], MD, PhD; Dan Jian[1,2], MD, PhD; Hongfu Xie[1,2], MD, PhD; Chao-Yuan Yeh[3*], MD; Ji Li[1,2,4,5*], MD, PhD

[1]Department of Dermatology, Xiangya Hospital of Central South University, Changsha, China

[2]Hunan Key Laboratory of Aging Biology, Xiangya Hospital of Central South University, Changsha, China

[3]aetherAI, Co Ltd, Taipei, Taiwan, China

[4]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital of Central South University, Changsha, China

[5]Key Laboratory of Organ Injury, Aging and Regenerative Medicine of Hunan Province, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Ji Li, MD, PhD
Department of Dermatology
Xiangya Hospital of Central South University
87 Xiangya Rd.
Changsha, 410008
China
Phone: 86 073189753406
Email: liji_xy@csu.edu.cn

Related Article:

Correction of: https://medinform.jmir.org/2021/3/e23415/

(JMIR Med Inform 2024;12:e57654) doi:10.2196/57654

In “A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study” (JMIR Med Inform 2021;9(3):e23415) the authors made one addition.

An “Acknowledgments” section has been added that reads as follows:

> *This work was supported by The Educational Science and Planning Project of Hunan Province (XTK20BGD008).*

The correction will appear in the online version of the paper on the JMIR Publications website on March 8, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 26.02.24; this is a non–peer-reviewed article;accepted 27.02.24; published 08.03.24.

Please cite as:
Zhao Z, Wu CM, Zhang S, He F, Liu F, Wang B, Huang Y, Shi W, Jian D, Xie H, Yeh CY, Li J
Correction: A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study
JMIR Med Inform 2024;12:e57654
URL: https://medinform.jmir.org/2024/1/e57654
doi:10.2196/57654
PMID:38457810

Corrigenda and Addenda

Correction: A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification

Ying Zhang[1*], MA; Xiaoying Li[1*], PhD; Yi Liu[1], MA; Aihua Li[1], PhD; Xuemei Yang[1], PhD; Xiaoli Tang[1], MA

Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Xiaoli Tang, MA
Institute of Medical Information
Chinese Academy of Medical Sciences
No 69, Dongdan North Street
Beijing, 100020
China
Phone: 86 10 52328902
Email: tang.xiaoli@imicams.ac.cn

Related Article:

Correction of: https://medinform.jmir.org/2023/1/e44892

(JMIR Med Inform 2024;12:e62757) doi:10.2196/62757

In "A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification" (JMIR Med Inform 2023;11:e44892), the authors made one addition.

An Acknowledgments section was added to the paper, as follows:

> *This work was supported by the Innovation Fund for Medical Sciences of Chinese Academy of Medical Sciences (grant: 2021-I2M-1-033) and the Fundamental Research Funds for the Central Universities (grant:3332023163).*

The correction will appear in the online version of the paper on the JMIR Publications website on June 5, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 30.05.24; this is a non–peer-reviewed article;accepted 03.06.24; published 05.06.24.

Please cite as:
Zhang Y, Li X, Liu Y, Li A, Yang X, Tang X
Correction: A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification
JMIR Med Inform 2024;12:e62757
URL: https://medinform.jmir.org/2024/1/e62757
doi:10.2196/62757
PMID:38838306

Correction: A Call to Reconsider a Nationwide Electronic Health Record System: Correcting the Failures of the National Program for IT

James Seymour Morris, BA

School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital NHS Foundation Trust, Cambridge, United Kingdom

Corresponding Author:
James Seymour Morris, BA

Related Article:

Correction of: https://medinform.jmir.org/2023/1/e53112

(JMIR Med Inform 2024;12:e56050) doi:10.2196/56050

In "A Call to Reconsider a Nationwide Electronic Health Record System: Correcting the Failures of the National Program for IT" (JMIR Med Inform 2023;11:e53112) the author noted one error.

In the section titled "The Status Quo," the following sentence appears:

> *Clinical research would achieve unprecedented statistical power if physicians were granted access to the full cohort of patients registered with NHS GPs—comprising over 62 people in England alone.*

This has been changed to read as follows:

> *Clinical research would achieve unprecedented statistical power if physicians were granted access to the full cohort of patients registered with NHS GPs—comprising over 62 million people in England alone.*

The correction will appear in the online version of the paper on the JMIR Publications website on January 12, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 03.01.24; this is a non–peer-reviewed article;accepted 03.01.24; published 12.01.24.

Please cite as:
Morris JS
Correction: A Call to Reconsider a Nationwide Electronic Health Record System: Correcting the Failures of the National Program for IT
JMIR Med Inform 2024;12:e56050
URL: https://medinform.jmir.org/2024/1/e56050
doi:10.2196/56050

Viewpoint

Exploring Impediments Imposed by the Medical Device Regulation EU 2017/745 on Software as a Medical Device

Liga Svempe[1], MA

Faculty of Social Sciences, Riga Stradins University, Riga, Latvia

Corresponding Author:
Liga Svempe, MA
Faculty of Social Sciences
Riga Stradins University
Dzirciema 16
Riga, LV1007
Latvia
Phone: 371 67409120
Email: liga.svempe@rsu.edu.lv

Abstract

In light of rapid technological advancements, the health care sector is undergoing significant transformation with the continuous emergence of novel digital solutions. Consequently, regulatory frameworks must continuously adapt to ensure their main goal to protect patients. In 2017, the new Medical Device Regulation (EU) 2017/745 (MDR) came into force, bringing more complex requirements for development, launch, and postmarket surveillance. However, the updated regulation considerably impacts the manufacturers, especially small- and medium-sized enterprises, and consequently, the accessibility of medical devices in the European Union market, as many manufacturers decide to either discontinue their products, postpone the launch of new innovative solutions, or leave the European Union market in favor of other regions such as the United States. This could lead to reduced health care quality and slower industry innovation efforts. Effective policy calibration and collaborative efforts are essential to mitigate these effects and promote ongoing advancements in health care technologies in the European Union market. This paper is a narrative review with the objective of exploring hindering factors to software as a medical device development, launch, and marketing brought by the new regulation. It exclusively focuses on the factors that engender obstacles. Related regulations, directives, and proposals were discussed for comparison and further analysis.

(JMIR Med Inform 2024;12:e58080) doi:10.2196/58080

KEYWORDS

software; artificial intelligence; medical device regulation; rights; digital health

Introduction

Technology has significantly reshaped how we engage with a multitude of products and services, spanning from finance and travel to health care. This progression has introduced a myriad of digital solutions into the market, designed to facilitate the diagnosis, monitoring, and treatment of diverse medical conditions, ultimately enhancing the quality of life of patients. As per MedTech Europe data, there were more than 500,000 medical technologies available in 2023 [1]. Consequently, the role of regulatory measures becomes paramount in safeguarding patient well-being. Nevertheless, regulatory frameworks must adeptly mirror technological advancements to remain pertinent and efficacious, particularly considering the transformative potential of software within the health care domain. Consequently, in the past decades, the novel concept of software as a medical device (SAMD) has been introduced [2].

Historically, Europe has served as the preferred pathway for obtaining medical device approvals. In 2010, Stanford University professor Josh Makower (Makower et al [3]) published survey results of over 200 medical technology companies titled “FDA Impact on US Medical Technology Innovation.” The survey distinctly indicated the predilection for the European Union market over its US counterpart, owing to the expeditious and cost-effective nature of the European Union regulatory process. The European Union regulatory process was noted as significantly more predictable, reasonable, and transparent; 75% of respondents rated their regulatory experience in the European Union as excellent or very good, in stark contrast to the mere 16% who bestowed similar evaluations upon the Food and Drug Administration (FDA). The author aligns with this perspective because Directive 93/42/EEC (hereinafter Medical Device Directive 93/42/EEC [MDD]) had general requirements with fewer regulatory responsibilities.

However, a certain degree of skepticism has arisen concerning the adequacy of the European Union regulation in ensuring patient safety [4]. This concern was addressed by a study conducted in 2016 by Thomas J Hwang (Hwang et al [5]) from Harvard University. The study entailed a cohort analysis, comparing safety issue rates and trial outcome reporting for medical devices approved within the European Union and the United States. The authors concluded that the medical devices approved first in the European Union were associated with an elevated risk and experienced more recalls.

Furthermore, the medical device industry in the past decades has experienced various scandals, casting doubts on the efficacy of the regulatory framework in achieving its overarching goal to ensure patient safety. Foremost among these is the Poly Implant Prothèse (PIP) scandal, an incident that has been prominently cited as a poignant illustration of regulatory deficiencies that can impede the fulfillment of their primary objectives [6].

Also, the MDD, being introduced in 1993, exhibited a notable misalignment with the latest technological advancements and was not sufficient to comprehensively cover SAMD development and launch. The European Union policy makers recognized the shortcomings, namely, outdated regulation, insufficient oversight leading to safety issues, and imprecise requirements, resulting in an "uneven level of protection of the patients, users, and public health," consequently requiring an improved regulatory framework [7]. Therefore, in 2017, the new Medical Device Regulation (EU) 2017/745 (MDR) came into force, replacing the previous MDD as well as the active implantable medical devices Directive 90/385/EEC (however, the latter is not in the scope of this paper) [8]. As per Recital 1, the new MDR is believed to solve emerging problems as well as to provide transparency and strengthen market surveillance and overall quality of medical devices, benefiting the patient. Undoubtedly, the new regulation clarifies rules for SAMD that were not clearly defined under the MDD. Furthermore, the new regulatory framework applies to all the manufacturers equally, ensuring the same high standards, irrespective of their geographic location. Such standardization and clarification of requirements eliminate disparities that have existed under the MDD, where the interpretation of regulations had varied between different member states, notified bodies (NBs), and manufacturers. However, it is too early to judge if the new regulation has achieved its main goals; as of now, no quantifiable data or research exist to demonstrate if device safety has improved or if the European Union is experiencing fewer product recalls.

Methods

The paper is a narrative literature review, and it seeks to provide a comprehensive summary of the problems, given the absence of a comparable analysis. A search in Scopus and PubMed databases was performed in November 2022 for articles written in English and published since 2017, the year when the MDR was adopted. The search included two strings: (1) ("medical" AND "device" AND "regulation") in the title, abstract, and keywords; and (2) ("medical" AND "device" AND "regulation") in the title. The specific term "software" was intentionally omitted from the search parameters, as its infrequent occurrence in titles, abstracts, or keywords would substantially limit the available pool of literature. Furthermore, it is important to acknowledge that papers can encompass analysis of the MDR in a broader context, yet not explicitly focused on SAMD, and can provide valuable insights, analysis, and conclusions for the research. First, non-English and duplicate articles were excluded, and the remaining articles were evaluated by reviewing their titles and abstracts. Then the articles were fully read, and only papers exploring and analyzing the impact of the MDR were included in the analysis. The initial search returned 341 items in total, out of which 307 items were excluded after screening due to their irrelevance. Thus, 34 papers were deemed applicable for this review.

The paper begins with a concise general overview of the impact of the MDR, exploring market data on innovations and communications from major organizations. It is followed by the main section of the paper (*The Challenges of the MDR*), presenting the analysis of the hindering factors that have been identified in academic literature. These factors were then grouped into thematic dimensions and explored in more detail, including their potential impact on the industry. The *Conclusion* section summarizes the main findings from the literature and discussions in the previous sections and suggests directions for future improvements.

The Overall Impact of the MDR

The MDR introduces more detailed requirements for all medical devices in terms of development, quality assurance, and clinical evaluation, as well as postmarket surveillance. However, the novel framework has sparked discussions on how these changes impact innovation. A survey conducted by Climedo Health revealed that 81% of the respondents consider the MDR challenging [9]. As per Stern [10], industry regulation often results in delayed or reduced firm entry into markets due to the increased time and costs, eventually reducing incentives to innovate. Thus, while the regulation is introduced to improve patient safety, the slower innovation could paradoxically result in a reduction of patient safety, since the improvements and novel devices might take longer to enter the market. So, the question remains: how to improve regulation without hindering innovation?

The regulatory framework holds particular importance for the pioneer innovators. Manufacturers who develop a first-of-its-kind product experience several disadvantages, such as the lack of specific guidance, the absence of clinical data, and little knowledge that can result in a longer development and approval process as well as additional costs. Stern [10] notes that the first entrant experiences approximately 34% longer regulatory approval process than the first follow-on entrant, also because the regulatory bodies then can release guidance materials that the later entrants can benefit from. This also means increased costs for the first entrants, although they can potentially gain the largest market share.

The industry has already brought considerable attention to the ramifications and challenges associated with the MDR on

multiple occasions. In an open letter dated April 15, 2019, the CEO of MedTech Europe, Serge Bernasconi, pointed out the lack of NBs and the unpreparedness of the regulatory system that could result in a shortage of medical devices [11]. The latest appeal was at the end of 2022 when many medical technology CEOs addressed the need for changes or else "Europe faces a scenario where a high number of existing medical devices, upon which patients, hospitals and other health institutions rely, will fail to be recertified on time and therefore risk permanently disappearing from the market. At the same time, the certification of new and improved products is also delayed, resulting in delayed patient access to the benefits of innovation" [12]. The medical associations conglomerate, The Standing Committee of European Doctors, in its letter to the President of the European Commission, Ursula von der Leyen, is being more dramatic, stating that "in some countries, up to 75% of medical devices are at risk of becoming unavailable," and that the "situation is unacceptable from the point of view of patient safety and quality of care" [13]. The emerging issue is already noticed as the manufacturers exit the European Union market due to various reasons, including increased costs and certification time, which eventually impact the patients as the devices become unavailable [14]. The expression of concern was duly acknowledged and subsequently addressed by Commissioner Stella Kyriakides at the Employment, Social Policy, Health and Consumers Affairs (EPSCO) Council in December 2022, who proposed to extend the transition period [15], and the extension was adopted with Regulation 2023/607, Article 1, by the European Parliament and the Council [16]. But would the extension, which applies only to the existing medical devices, solve all the problems?

Due to the complexity of the MDR, there are various implications for SAMD development. Therefore, the author conducted research exploring currently identified hampering factors. The paper serves as a comprehensive resource for manufacturers to proactively configure their organizational infrastructure and allocate resources in advance. In addition, it is beneficial for health care policy makers in their endeavors to assess and ameliorate industrial policies and practices. This paper will also be beneficial to the government bodies and policy makers who are responsible for the medical technology industry, as it underscores the barriers impeding the evolution of the digital health sector and can contribute to developing tools to foster and maintain innovation.

The Challenges of the MDR

In the following sections, the hindering factors are grouped into 8 dimensions and are further described and analyzed. Each of the factors has received varying attention in the literature, and their impact on a manufacturer can overlap; however, each one of them can be addressed separately. Table 1 presents an overview of the dimensions and a shortlist of the hindering factors included in each dimension.

Table 1. Hindering factors listed and consolidated into dimensions.

Dimensions	Consolidated factors	References
More complex requirements leading to delays	• More time is needed for development • More financial resources are needed, thus new products will take longer to make available in the market • Slowing industry innovation • Delays also for recertification	[17-28]
More requirements for clinical evaluation	• More manufacturers will need to conduct or repeat clinical trials • Harder to prove technical equivalence • Most small and medium-sized enterprises lack the financial resources to conduct trials	[17,19,21,22,25,26,29-34]
Increased expenses	• Regulatory changes bring additional costs to being compliant • Certification and recertification processes are costly • Postmarket surveillance seeks more resources	[17-19,24-26,30-41]
Classification issues	• Uncertainty if the product is a medical device • Classifying correctly • Up-classification, which means more complex requirements for development and launch	[17,22,29,32,36,39,42-46]
Limited availability of NBs[a]	• The small number of NBs • The capacity of existing NBs • Interdependence between manufacturers and NBs • Poor communication between stakeholders	[21,22,24,26,27,31,33,34,38,39,43,47,48]
Lack of knowledge	• Market entry depends on regulatory knowledge • Additional costs to acquire competences • Lack of knowledge of the European Databank on Medical Devices	[30,31,33,36,37,41]
Lack of guidance	• Lack of guidance materials for specific matters • Uncertainty about the processes as harmonized standards are not published • No provisions for orphan devices	[20,27,28,31,33,40,46,47]
Constraints on software updates	• Complicated process to change or add new features • Limited possibility of software customization	[17,26,34,36]

[a]NB: notified body.

More Complex Requirements Leading to Delays

One of the main challenges for all manufacturers is the complex requirements to develop and market a medical device. While the requirements are updated to ensure patient safety, numerous manufacturers may encounter challenges in meeting these requirements. This can lead to several outcomes. For example, first, the new products will take longer to be deployed [17-21], thus slowing the innovation in health care [22-24]. According to MedTech Europe data, the time to certify a medical device under MDR has now doubled to 13-18 months [49]. Second, delays in certification or recertification procedures could potentially result in a reduction of the available product range [18] and the discontinuation of certain products [25,26]. Third, new product launches could potentially be deferred or even canceled, as the emphasis shifts toward the maintenance of existing medical devices [27]. Fourth, some manufacturers might choose to continue supplying their medical devices while opting to withdraw from the European Union market [28].

Given that one of the plausible outcomes is a delayed launch or even the eventual discontinuation of a medical device, there arises a potential jeopardy to the fundamental objective of the MDR, which is to safeguard patient safety [26], since the devices will no longer be available.

More Requirements for Clinical Evaluation

Another troublesome issue for the manufacturers is the increased need to conduct clinical evaluations [17,19,21,22,29]. Although it derives from the general complexity of the requirements explored in the previous section, this matter is specific and critical for the development thus separated.

For lower-risk medical devices, the manufacturer can provide clinical evaluation without conducting its own clinical trial, yet the device must be proven to be equivalent to the compared device. In those cases, the manufacturer can use other clinical investigations and studies and papers published in peer-reviewed sources. While there are no data available on how many devices are approved based on equivalence in Europe, the data in the United States suggest that it is the majority: 99% of the devices approved between 2015 and 2020 used the 510(k) pathway (the mean number of premarket approvals was 38, compared with a mean of 2982 510(k)s annually) [50].

However, clinical evaluation without conducting own clinical trial is not an option if the developed medical device is innovative, meaning there are no equivalent devices, even if the device in question has low risk. According to the MDR, the equivalence shall be demonstrated in 3 dimensions, namely, technical, clinical, and biological (the latter does not apply to a SAMD).

Technical equivalence means "the device is of similar design; is used under similar conditions of use; has similar specifications and properties including physicochemical properties such as intensity of energy, tensile strength, viscosity, surface characteristics, wavelength, and software algorithms; uses similar deployment methods, where relevant; has similar principles of operation and critical performance requirements" (MDR, Annex XIV Part A, Article 3). The new additional requirement, which is also the most concerning aspect of a SAMD, is to compare software algorithms. While Medical Device Coordination Group 2020-5 [51] suggests the comparison needs to be done only in terms of functionality and clinical performance, not the code itself, algorithms are not public and are the essence (and the unique selling point) of software. Thus, the actual functionality cannot be thoroughly compared, especially if it is an artificial intelligence (AI) and machine learning (ML)–based solution. In the meantime, for implantable and higher-risk devices (Class III), the equivalence can be claimed only if the manufacturer has a contract in place that allows full access to the technical documentation of the equivalent device on an ongoing basis (MDR Article 61(5)). This requirement is almost impossible to fulfill in a competitive market [27].

According to Kearney and McDermott's [31] research about challenges with clinical evaluation, manufacturers also frequently have issues related to obtaining and understanding the level of clinical data required by the MDR, being reluctant to the more stringent requirements. However, the most common challenge here is the lack of skills and knowledge for preparing the clinical evaluation, which has led to an increase in outsourcing the knowledge (and consequently, increasing the costs).

An apparent consequence of the increased need for clinical trials is increased research and development costs [22,30,32], which also complicates the development process [33] and increases the maintenance costs [34]. In addition, it is no secret that most small- and medium-sized enterprises (SMEs) lack the financial resources to conduct large clinical trials [26].

Consequently, manufacturers are already deciding to exit the European Union market or at least postpone the launch of innovative products [31].

Increased Expenses

While most of the identified implications result in increased expenses, it is important to single it out and explore its impact on the development and launch of SAMDs. Also, increased costs were highlighted in the Climedo Health survey [9], where 70% of the respondents named it as their greatest challenge, 44% of respondents projected that the MDR will incur additional costs of 5% of their annual turnover, while the same number of respondents plan cost increase by 1% to 5% of their annual turnover.

The regulatory improvements will clearly bring more development process costs [17,31] as well as conformity assessment, certification, and recertification costs [25,26,32-38]. It will also require changes in organizational structure and a need to acquire new competencies, including bringing more talent on board [18,24,30,32,33,39] and seeking additional funding, which is challenging, especially for SMEs.

The new regulatory updates also establish more requirements for the postmarket surveillance process, which brings more costs [32,40]. In addition, the new unique device identification (UDI) system requires adaptation of the existing information systems, also resulting in cost increase [41].

The increased expenses, if not covered by additional investments, can eventually lead to price increases, which means decreased availability of medical devices [17,19,32,39], or in worst-case scenarios it can lead to an exit from the European Union market [32,35]. Some SMEs might need to seek other exit solutions, such as merging with large companies to keep the products running [17,32,33], resulting in less competition and few dominant companies in the market.

Classification Issues

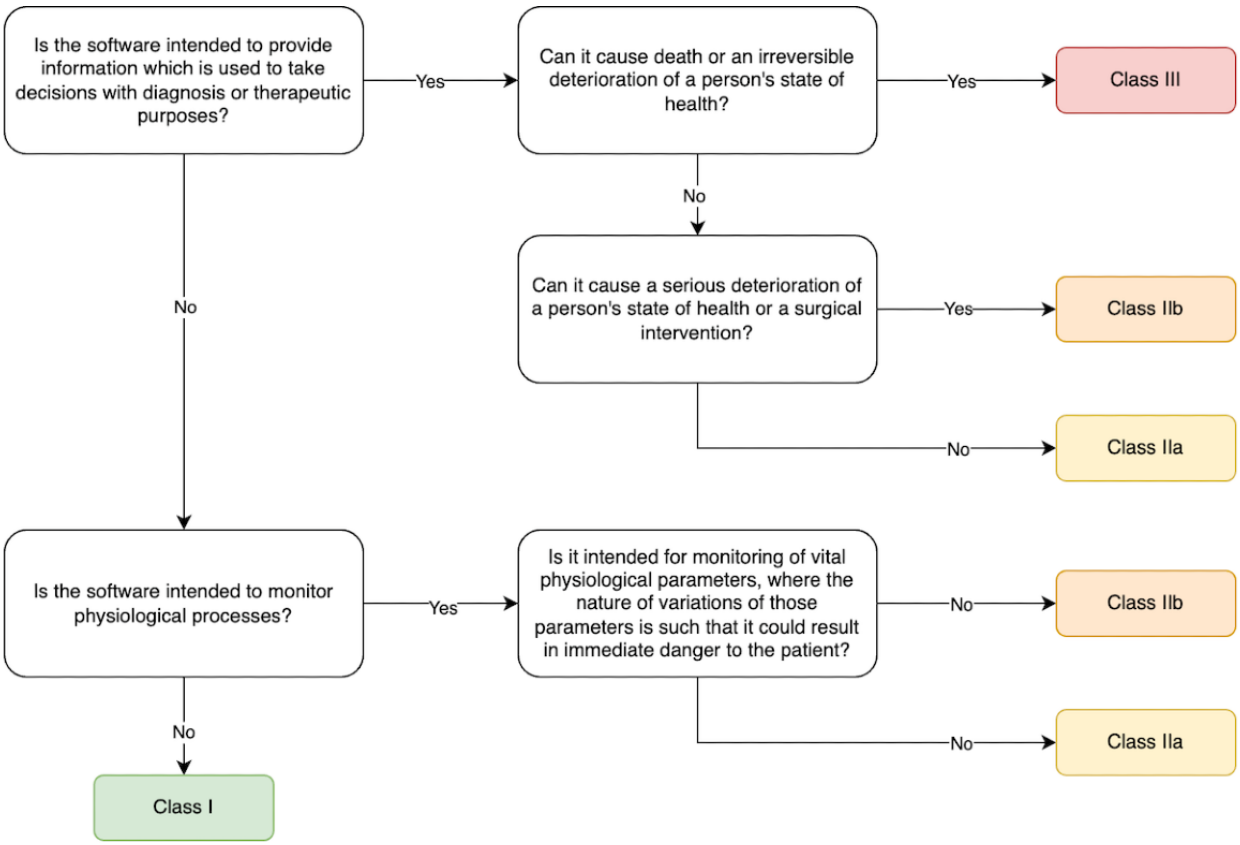
Before specifying the device class, the manufacturer must define if the developed product is a medical device: software with a medical purpose acting to benefit individual patients. Then the next step is to choose the right class according to the risk level.

In comparison to the previous directive (MDD), the MDR separates a rule for SAMDs (Rule 11, Annex VIII, Chapter III, 6.3.; Figure 1).

Yet the classification process has its struggles, such as (1) uncertainty if the developed product is a medical device at all [32,36], thus if it falls within the scope of MDR; (2) choosing the correct class [32,36,42,43]; and (3) for most SAMDs, the new MDR means up-classification [29,32,39,44,45].

The first 2 issues can be addressed with a rigorous regulation study and consultations. This is a one-time issue not impacting the manufacturers in the long term.

Figure 1. Software as a medical device classification flow.



However, considering the complexity of the regulatory requirements, SMEs might opt to position their products as health and wellness devices. This strategic approach could allow them to avoid rigorous regulatory scrutiny and lessen the challenges to launch and market their products [46]. In this context, the manufacturer's envisaged intended use of the device plays the role, rather than its actual use. This perspective is supported by the Court of Justice of the European Union with its decision on Brain Products, wherein it was established that if a product is not intended to be used for medical purposes, Conformité Européenne certification is not required [52]. Yet, the intentional falling out of the regulatory scope means the uncertain quality of the digital health tool, as well as it potentially carrying safety risks to human health, which was the initial cause of the regulatory changes. This concern is important, particularly in light of the increasing reliance of individuals on digital health technologies [46].

The third factor, up-classification, can become troublesome, as most SAMDs will be classified in at least the IIa risk class, and software in the lowest risk class I will remain as an exception [45]. This up-classification entails more complex requirements and certification processes and will result in more time needed for development and subsequent delays in obtaining market access [17,22]. Some manufacturers might choose to eliminate or limit some of the device features to keep the device in a lower-risk class [32], and this might in turn limit the availability of innovative solutions.

The complex requirements as an impediment are explored earlier in this paper.

Limited Availability of NBs

An NB is an organization that is authorized to perform conformity assessments in accordance with the MDR. Annex VII sets out general, organizational, resource, and other requirements to be met by the NBs, thus the process of establishing an NB can be lengthy and resource-intensive as well.

Consequently, the limited availability of NBs is one of the hindering factors identified, and it has two dimensions: (1) the absolute number of NBs [22,23,25,28,34,39,43,47,48] and (2) the capacity of existing NBs [33,38,43].

As of May 2024, according to the European Commission database, there are 46 authorized NBs to perform assessments following the MDR [53]. Germany and Italy have the highest concentration of NBs, with each country hosting 10. The Netherlands follows with 4 NBs. Finland, Czechia, Turkey, Poland, France, and Sweden each host 2 NBs. Meanwhile, Belgium, Croatia, Hungary, Ireland, Norway, Slovakia, Slovenia, Spain, Cyprus, and Denmark each have 1 NB.

Although the MDR entered into force in 2017, seven years have not been enough to establish a sufficient number of bodies. It is worth pointing out that not all European countries host an NB. For example, Switzerland is the second largest medical technology employer in Europe per capita, or fifth largest in absolute number of people employed [1]; however, it does not host a single NB. Ireland, which has the largest number of people directly employed in the medical technology industry per 10,000 inhabitants, hosts 1 NB. Similarly, France, which is fourth in Europe with the highest direct employment in terms

of absolute number of people employed, hosts only 2 NBs. While the international environment does not limit the applicants based on location, and manufacturers can freely choose an NB in any other country, Peter et al [34] suggested that the lack of NBs and poor communication between the manufacturers and NBs can lead to lost market opportunities. However, the authors do not mention any particular or potential examples.

The capacity of existing bodies is also being explored. Because of the low number of NBs and the increased need to acquire a certification, this can result in delays and shortages of medical devices or some manufacturers even leaving the European Union market [48]. The increased time for the certification process has been already mentioned in this paper earlier: according to MedTech Europe data, the time to certify a medical device under MDR has now doubled to 13-18 months. As per the preliminary results of the NBs survey led by Gesundheit Österreich GmbH (Austrian National Public Health Institute) with Areté and Civic Consulting [54], in March 2023, in 86% of cases, the time to acquire the certificates was more than 13 months.

Another factor identified is the interdependence of the NBs with the manufacturers [26], as the NBs would want to ensure their turnover and eventually profit, thus prioritizing the clients who would bring the largest revenue. This argument is especially troublesome for SMEs, as large companies might have the leverage to be certified first, thus leading to delays for SMEs [39]. As per the MedTech Europe report conducted in April 2022, at least 15% and up to 30% of SMEs have no access to an MDR-designated NB [49]. In addition, poor communication between the NBs and manufacturers can lead to potential setbacks during the certification process [31].

Interestingly, Fink and Akra [55] delve into an aspect concerning the regulatory frameworks in the European Union compared with the United States. While this does not pertain specifically to the Regulation, it remains a noteworthy consideration in a broader context. The distinction lies in the centralized approval process in the United States, where the FDA singularly holds authority, and the decentralized process in the European Union, involving multiple NBs. The authors mention the potential risk of different interpretations of requirements stemming from the decentralized nature of the European Union regulatory system.

Lack of Knowledge

Stern [10] in her paper titled “Innovation under Regulatory Uncertainty: Evidence from Medical Technology” separates 2 factors: technological uncertainty and uncertainty about application content and format. Thus, it is worth exploring both factors separately. (1) Technological uncertainty means the lack of knowledge and understanding of how the innovative product in question works and how it is used in the human body, as well as how the regulator will understand the mechanisms behind it; and (2) the uncertainty about application content and format is related to a lack of guidance for the product assessment phase (including the evaluation of clinical trials and the information needed to submit), and this factor is explored in the next section.

This paper revealed that the lack of knowledge can be an obstacle to entering the market, especially for start-ups [33], and it can be associated with legal risks as well [36]. There is a need to acquire more competencies [31,37], which results in additional costs [30]. Manufacturers need to improve their knowledge and skills to perform postmarket surveillance [33], and better understand the European Databank on Medical Devices system [41].

Lack of Guidance

Chatterji [56] finds evidence that nontechnical knowledge, such as understanding regulation and marketing knowledge, is of greater importance than technical knowledge. Thus, this impediment should draw the attention of the manufacturers and stakeholders.

The MDR consists of 123 articles and 17 annexes on 175 pages. In comparison, the MDD had 23 articles and 12 annexes on 60 pages. While this suggests that the new regulation shall bring clear requirements and explanations, the actual situation is on the contrary. The Climedo Health survey shows that 59% of the respondents name the lack of clarity as one of their greatest challenges [9]. The unclarity is evident in clinical evaluation [28,31], postmarket surveillance processes [46] and activities [33], or the lack of guidance in general [27,40]. Gilbert et al [20] also specify that there is a need for smarter regulation, particularly for the highest-risk (III) class devices.

Melvin [47] draws attention to the fact that the MDR has no provisions for orphan devices that are intended for rare life-threatening or chronically debilitating conditions. Since the market for these devices is small, it may become economically unfeasible for manufacturers to continue supplying them, potentially leading to their exit from the European Union market.

A significant shortcoming of the MDR is the lack of detailed requirements and guidance for AI solutions, which can become a barrier to their clinical adoption [20]. AI solutions experience rapid growth, including in health care, and the lack of regulation can lead to uncertainty for development as well as compliance. While there is the AI Act that also covers the health care industry, the manufacturers today do not have guidance, which leaves room for their interpretation.

The importance of regulatory guidance and its impact on medical devices is proved by Stern [10], showing an average decrease in regulatory approval times of 2.8 to 6.6 months when comparing innovative firstcomers with their followers.

Constraints on Software Updates

Opposite to a common hardware medical device, software can be updated regularly. It can vary from a minor update, such as a new data field or color change, to a significant update, such as a brand new feature, delivering a new type of content, or improving the AI algorithm.

This is a considerable difference if we explore the maintenance of a medical device throughout its lifecycle. The MDR now requires an NB’s involvement, namely, if a manufacturer has “any plan for substantial changes to the quality management system, or the device-range covered” (Annex IX, Chapter I, 2.4), or if technical documentation is being changed and “such changes could affect the safety and performance of the device or the conditions prescribed for use of the device” (Annex IX, Chapter II, 4.10). Thus, this requirement to involve the NB if

the manufacturer plans updates in the device is considered an impediment [17,26,34], which brings constraints on changing the software or adding new features, as well as limits software customization [36].

Nevertheless, the MDR does not specify what changes should be communicated with the NB, which eventually allows the manufacturer's interpretation to some degree. However, the Medical Device Coordination Group has provided a guidance document on significant changes regarding the transitional provision under Article 120 [57] to better understand which changes are considered "significant." It gives the explanation that minor software changes would be adding a new language or fixing bugs. Yet, changes in algorithms shall be considered as a major change, thus requiring the involvement of the NB, and this is burdensome for all AI solutions as the algorithms can change regularly. Furthermore, adding a new therapeutic feature (even if it is only enriching the content base) is considered a significant update, requiring the involvement of the NB.

Conclusions

This paper is scoped to the impact of the new MDR on the development and launch of SAMDs and does not explore its effect on patient safety. Future research should seek to investigate the benefits and impact of the MDR on increasing patient safety and if the regulatory changes have had the desired effect.

The identified hindering factors of the MDR were consolidated into eight dimensions: (1) more complex requirements leading to delays, (2) more requirements for clinical trials, (3) an increase in expenses, (4) classification issues, (5) limited availability of NBs, (6) lack of knowledge, (7) lack of guidance, and (8) constraints on software updates. Each of the factors has received varying attention, yet any single one can have a critical impact on a manufacturer, thus each company shall evaluate its strengths and weaknesses to sufficiently prepare for development.

The results show that the new regulation heavily impacts the European Union medical device industry, which can lead to either price increases or shortage of medical devices, as well as stifling innovation, which can eventually even harm the patients. Some manufacturers might evaluate the costs and potential revenue and decide to discontinue the devices. Some small start-ups may find themselves compelled to shut down their operations, while other enterprises might opt to exit the European Union market or engage in mergers with more sizable corporations. This trend is supported by Kearney and McDermott's [31] research showing the first signs of manufacturers either leaving the European Union market or seeking approval for devices in the United States first.

The fact that Europe has lost its appeal is supported by the Boston Consulting Group and University of California, Los Angeles Biodesign report [58], which shows that 89% of surveyed companies consider prioritizing the United States over the European Union. While the registration of digital technologies is quite uncertain for both United States and European Union markets, still 32% of the respondents considered the US pathway rather predictable, which is more than double that of the European Union pathway (15%) [58]. The market changes are alarming for the European Union; thus, the problems must be addressed at the European Union level. Policy makers should reconsider if all the current regulation requirements bring actual value and ensure patient safety rather than build unnecessary burdens to launch innovative digital solutions. Although the transition period for the existing medical devices has been extended, it gives time for the existing medical devices to fulfill the requirements, yet it does not address the issues with complexity, and hence further actions must be taken.

In the meantime, the growing development and approbation of digital health tools, including AI solutions, currently require a more targeted regulatory framework as we see the lack of guidance and knowledge in the domain. The MDR and respective guidelines exhibit limitations in addressing the complexities inherent to most pioneering technologies, thereby AI and ML–based solution manufacturers have room for interpretation of the applicable regulation. To address the regulatory gaps for AI and ML–driven solutions, in March 2024, the European Commission passed the AI Act [59], which will now lessen the legal uncertainties. This legislation covers various domains, and also applies to health care and medical AI, thereby ensuring more robust regulatory oversight within this landscape. Henceforth, manufacturers of AI and ML–driven solutions will be required to ensure compliance with both the MDR and the AI Act. Although the adoption of the act is commendable, in reality, it has introduced new compatibility challenges. For instance, there is uncertainty surrounding the process of providing clinical evidence for certification under the MDR. It appears that AI medical devices will be required to have Conformité Européenne certification before undergoing testing, potentially creating an infinite loop of unmet requirements, or forcing the manufacturers to conduct trials outside the European Union. Therefore, this seeks further discussions and implementation guidelines from the European Commission to help the manufacturers in their compliance journey.

Last but not the least, each European Union member state that aspires to foster the advancement of the digital health sector is encouraged to consider both monetary and nonmonetary assistance to SMEs. Such support mechanisms hold the potential to facilitate a seamless introduction of cutting-edge innovations to the marketplace. While financial assistance might be subject to budgetary constraints, nonfinancial support can be equally pronounced. This encompasses diverse facets, such as the establishment of digital health hubs to facilitate the exchange of knowledge and experience, endeavors aimed at attracting skilled personnel, and the active promotion of educational initiatives.

Conflicts of Interest

None declared.

References

1. The European medical technology industry in figures. MedTech Europe. 2023. URL: https://www.medtecheurope.org/wp-content/uploads/2023/10/the-european-medical-technology-industry-in-figures_2023.pdf [accessed 2024-02-02]
2. Svempe L. Regulation and its impact on innovation in healthcare: SAMD case. Socrates 2022;1(22):43-52 [FREE Full text] [doi: 10.25143/socr.22.2022.1.043-052]
3. Makower J, Meer A, Denend L. FDA impact on U.S. medical technology innovation: a survey of over 200 medical technology companies. MedTech Europe. URL: https://www.medtecheurope.org/wp-content/uploads/2015/07/01112010_FDA-impact-on-US-medical-technology-innovation_Backgrounder.pdf [accessed 2024-02-02]
4. Cohen D, Billingsley M. Europeans are left to their own devices. BMJ 2011;342:d2748 [FREE Full text] [doi: 10.1136/bmj.d2748] [Medline: 21572130]
5. Hwang TJ, Sokolov E, Franklin JM, Kesselheim AS. Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European union and United States: cohort study. BMJ 2016;353:i3323 [FREE Full text] [doi: 10.1136/bmj.i3323] [Medline: 27352914]
6. Martindale V, Menache A. The PIP scandal: an analysis of the process of quality control that failed to safeguard women from the health risks. J R Soc Med 2013;106(5):173-177 [FREE Full text] [doi: 10.1177/0141076813480994] [Medline: 23761525]
7. Commission staff working document. Impact assessment on the revision of the regulatory - framework for medical devices. European Commission. 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52012SC0273&qid=1714901442078 [accessed 2024-05-05]
8. Regulation (EU) 2017/745 of the European parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. The European Parliament and the Council of the European Union. 2017. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017R0745 [accessed 2024-02-02]
9. Survey results: MDR readiness check 40 days before the deadline. Climedo. 2021. URL: https://climedo.de/wp-content/uploads/2021/04/EU-MDR-Survey-Results-2021-EN.pdf [accessed 2024-02-02]
10. Stern AD. Innovation under regulatory uncertainty: evidence from medical technology. J Public Econ 2017;145:181-200 [FREE Full text] [doi: 10.1016/j.jpubeco.2016.11.010] [Medline: 28652646]
11. Open letter to the European commission on the implementation. MedTech Europe. 2019. URL: https://www.medtecheurope.org/wp-content/uploads/2019/04/MedTech-Europe_VP-Katainen_MDR-implementation-status_15-April-2019.pdf [accessed 2024-02-02]
12. Maxwell A. Whole EU medtech industry behind latest push to have MDR problems tackled at highest levels. Medtech Insight. 2022. URL: https://medtech.pharmaintelligence.informa.com/MT146028/Whole-EU-Medtech-Industry-Behind-Latest-Push-To-Have-MDR-Problems-Tackled-At-Highest-Levels [accessed 2024-02-02]
13. Collis H. Bungled EU medical device rules put lives at risk. Politico. 2022. URL: https://www.politico.eu/article/bungled-eu-medical-device-rules-put-lives-risk/ [accessed 2024-02-02]
14. Fick M. Insight: medical device makers drop products as EU law sows chaos. Reuters. 2022. URL: https://www.reuters.com/business/healthcare-pharmaceuticals/medical-device-makers-drop-products-eu-law-sows-chaos-2022-12-19/ [accessed 2024-02-02]
15. Opening remarks by Commissioner Stella Kyriakides at the EPSCO council-implementation of the medical devices regulation. European Commission. 2022. URL: https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_22_7627 [accessed 2024-02-02]
16. Regulation (EU) 2023/607 of the European parliament and of the council of 15 March 2023 amending regulations (EU) 2017/745 and (EU) 2017/746 as regards the transitional provisions for certain medical devices and in vitro diagnostic medical devices. The European Parliament and the council of the European Union. 2023. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32023R0607 [accessed 2024-02-02]
17. Malvehy J, Ginsberg R, Sampietro-Colom L, Ficapal J, Combalia M, Svedenhag P. New regulation of medical devices in the EU: impact in dermatology. J Eur Acad Dermatol Venereol 2022;36(3):360-364 [FREE Full text] [doi: 10.1111/jdv.17830] [Medline: 34816498]
18. Kaule S, Bock A, Dierke A, Siewert S, Schmitz K, Stiehm M, et al. Medical device regulation and current challenges for the implementation of new technologies. Current Directions in Biomedical Engineering 2020;6(3):334-337 [FREE Full text] [doi: 10.1515/cdbme-2020-3086]
19. Niemiec E. Will the EU medical device regulation help to improve the safety and performance of medical AI devices? Digit Health 2022;8:20552076221089079 [FREE Full text] [doi: 10.1177/20552076221089079] [Medline: 35386955]

20. Gilbert S, Fenech M, Hirsch M, Upadhyay S, Biasiucci A, Starlinger J. Algorithm change protocols in the regulation of adaptive machine learning-based medical devices. J Med Internet Res 2021;23(10):e30545 [FREE Full text] [doi: 10.2196/30545] [Medline: 34697010]
21. Fraser AG, Byrne RA, Kautzner J, Butchart EG, Szymański P, Leggeri I, et al. Implementing the new European regulations on medical devices-clinical responsibilities for evidence-based practice: a report from the regulatory affairs committee of the European society of cardiology. Eur Heart J 2020;41(27):2589-2596 [FREE Full text] [doi: 10.1093/eurheartj/ehaa382] [Medline: 32484542]
22. Letourneur D, Joyce K, Chauvierre C, Bayon Y, Pandit A. Enabling MedTech translation in academia: redefining value proposition with updated regulations. Adv Healthc Mater 2021;10(1):e2001237 [FREE Full text] [doi: 10.1002/adhm.202001237] [Medline: 32935923]
23. Kanti SPY, Csóka I, Adalbert L, Jójárt-Laczkovich O. Analysis of the renewed European medical device regulations in the frame of the non - EU regulatory landscape during the COVID facilitated change. J Pharm Sci 2022;111(10):2674-2686 [FREE Full text] [doi: 10.1016/j.xphs.2022.07.011] [Medline: 35872025]
24. Garzotto F, Comoretto RI, Dorigo L, Gregori D, Zotti A, Meneghesso G, et al. Preparing healthcare, academic institutions, and notified bodies for their involvement in the innovation of medical devices under the new European regulation. Expert Rev Med Devices 2022;19(8):613-621 [FREE Full text] [doi: 10.1080/17434440.2022.2118046] [Medline: 36039712]
25. Martelli N, Eskenazy D, Déan C, Pineau J, Prognon P, Chatellier G, et al. New European regulation for medical devices: what is changing? Cardiovasc Intervent Radiol 2019;42(9):1272-1278 [FREE Full text] [doi: 10.1007/s00270-019-02247-0] [Medline: 31123774]
26. Shatrov K, Blankart CR. After the four-year transition period: is the European union's medical device regulation of 2017 likely to achieve its main goals? Health Policy 2022;126(12):1233-1240 [FREE Full text] [doi: 10.1016/j.healthpol.2022.09.012] [Medline: 36202647]
27. Vasiljeva K, van Duren BH, Pandit H. Changing device regulations in the European union: impact on research, innovation and clinical practice. Indian J Orthop 2020;54(2):123-129 [FREE Full text] [doi: 10.1007/s43465-019-00013-5] [Medline: 32257028]
28. Pazart L, Pelayo S, Chevallier T, Gruionu G, Mabo P, Bayon Y, et al. Threats and opportunities for the clinical investigation of high-risk medical devices in the context of the new European regulations. 2022 Presented at: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies SCITEPRESS - Science and Technology Publications; February 11-13, 2021; Setúbal, Portugal URL: https://hal.science/hal-03657023/document [doi: 10.5220/0010382902740284]
29. Bayrak T, Safak Yilmaz E. What will be the economic impact of the new medical device regulation? An interrupted time-series analysis of foreign trade data. Value Health Reg Issues 2022;29:1-7 [FREE Full text] [doi: 10.1016/j.vhri.2021.07.010] [Medline: 34794046]
30. Tarricone R, Ciani O, D'Acunto S, Scalzo S. The rise of rules: will the new EU regulation of medical devices make us safer? Eur J Intern Med 2020;80:117-120 [FREE Full text] [doi: 10.1016/j.ejim.2020.07.012] [Medline: 32703676]
31. Kearney B, McDermott O. Challenges faced by manufacturers with clinical evaluation under the new European medical device regulations. Cogent Eng 2023;10(2):1-22 [FREE Full text] [doi: 10.1080/23311916.2023.2261236]
32. Agyei EEYF, Pohjolainen S, Oinas-Kukkonen H. Impact of Medical Device Regulation on Developing Health Behavior Change Support Systems. In: Baghaei N, Vassileva J, Ali R, Oyibo K, editors. Persuasive Technology. New York, NY: Springer International Publishing; 2022:1-15.
33. Ben-Menahem SM, Nistor-Gallo R, Macia G, von Krogh G, Goldhahn J. How the new European regulation on medical devices will affect innovation. Nat Biomed Eng 2020;4(6):585-590 [FREE Full text] [doi: 10.1038/s41551-020-0541-x] [Medline: 32203280]
34. Peter L, Hajek L, Maresova P, Augustynek M, Penhaker M. Medical devices: regulation, risk classification, and open innovation. Journal of Open Innovation: Technology, Market, and Complexity 2020;6(2):42 [FREE Full text] [doi: 10.3390/joitmc6020042]
35. Maresova P, Hajek L, Krejcar O, Storek M, Kuca K. New regulations on medical devices in Europe: are they an opportunity for growth? Administrative Sciences 2020;10(1):16 [FREE Full text] [doi: 10.3390/admsci10010016]
36. Blagec K, Jungwirth D, Haluza D, Samwald M. Effects of medical device regulations on the development of stand-alone medical software: a pilot study. Stud Health Technol Inform 2018;248:180-187 [FREE Full text] [Medline: 29726435]
37. Behan R, Watson M, Pandit A. New EU medical device regulations: impact on the medtech sector. Medical Writing 2017;26:20-24 [FREE Full text] [doi: 10.1201/9781003301202-8]
38. Giefing-Kröll C, Laumen G. How the EU Medical Device Regulation is affecting the medical device landscape. An interview with Suzanne Halliday, the Regulatory Head of BSI, Medical Devices Notified Body. Medical Devices 2022;31(2):62-64 [FREE Full text] [doi: 10.1201/9781420033984.ch10]
39. Wagner MV, Schanze T. Challenges of medical device regulation for small and medium sized enterprises. Current Directions in Biomedical Engineering 2018;4(1):653-656 [FREE Full text] [doi: 10.1515/cdbme-2018-0157]
40. Bianchini E, Mayer CC. Medical device regulation: should we care about it? Artery Res 2022;28(2):55-60 [FREE Full text] [doi: 10.1007/s44200-022-00014-0] [Medline: 35378951]

41. Camus D, Thiveaud D, Josseran A, Barthélémy CE, Chambrin PY, Hebbrecht G, et al. New European medical device regulation: how the French ecosystem should seize the opportunity of the EUDAMED and the UDI system, while overcoming the constraints thereof. Therapie 2019;74(1):73-85 [FREE Full text] [doi: 10.1016/j.therap.2018.12.001] [Medline: 30598315]
42. Beckers R, Kwade Z, Zanca F. The EU medical device regulation: implications for artificial intelligence-based medical device software in medical physics. Phys Med 2021;83:1-8 [FREE Full text] [doi: 10.1016/j.ejmp.2021.02.011] [Medline: 33657513]
43. Jarman H, Rozenblum S, Huang TJ. Neither protective nor harmonized: the crossborder regulation of medical devices in the EU. Health Econ Policy Law 2021;16(1):51-63 [FREE Full text] [doi: 10.1017/S1744133120000158] [Medline: 32631465]
44. Py S, Lihoreau T, Puyraveau M, Grosdemouge S, Butterlin N, Francois S, et al. Meeting an end-user need in a collaborative high risk medical device software development in accordance with future European regulations. 2021 Presented at: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies SCITEPRESS - Science and Technology Publications. 2021; February 11-13, 2021; Virtual event p. 265-273. [doi: 10.5220/0010381502650273]
45. Becker K, Lipprandt M, Röhrig R, Neumuth T. Digital health? Software as a medical device in focus of the medical device regulation (MDR). Information Technology 2019;61(5-6):211-218 [FREE Full text] [doi: 10.1515/itit-2019-0026]
46. Yu H. Digital health technologies under the new EU medical devices regulation: monitoring and governing intended versus actual use. BMJ Innov 2021;7(4):637-641 [FREE Full text] [doi: 10.1136/bmjinnov-2021-000713]
47. Melvin T. The European medical device regulation-What biomedical engineers need to know. IEEE J Transl Eng Health Med 2022;10:4800105 [FREE Full text] [doi: 10.1109/JTEHM.2022.3194415] [Medline: 36003070]
48. Ahmed T, Zafar J, Sharif F, Zafar H. Critical analysis of the effect the new medical device regulation will have on the relevant stakeholder. BMJ Innov 2022;8(4):285-290 [FREE Full text] [doi: 10.1136/bmjinnov-2021-000855]
49. MedTech Europe survey report analysing the availability of medical devices in 2022 in connection to the medical device regulation (MDR) implementation. MedTech Europe. 2022 Jul 14. URL: https://www.medtecheurope.org/wp-content/uploads/2022/07/medtech-europe-survey-report-analysing-the-availability-of-medical-devices-in-2022-in-connection-to-the-medical-device-regulation-mdr-implementation.pdf [accessed 2024-02-02]
50. Darrow JJ, Avorn J, Kesselheim AS. FDA regulation and approval of medical devices: 1976-2020. JAMA 2021;326(5):420-432 [FREE Full text] [doi: 10.1001/jama.2021.11171] [Medline: 34342614]
51. Medical Device Coordination Group. MDCG 2020-5. clinical evaluation - equivalence. a guide for manufacturers and notified bodies. European Commission. 2020. URL: https://health.ec.europa.eu/system/files/2020-09/md_mdcg_2020_5_guidance_clinical_evaluation_equivalence_en_0.pdf [accessed 2024-02-02]
52. Judgment of the Court (Third Chamber), 22 November 2012. EUR-Lex. 2022 Nov 22. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62011CJ0219 [accessed 2024-02-02]
53. Bodies. European Commission. URL: https://webgate.ec.europa.eu/single-market-compliance-space/#/notified-bodies/notified-body-list?filter=bodyTypeId:3,legislationId:34 [accessed 2024-05-05]
54. van Raamsdonk A, Loh E. Alarming results and some good news from MDR and IVDR survey of notified bodies. Medical Design and Outsourcing. 2023 Jul 31. URL: https://www.medicaldesignandoutsourcing.com/mdr-ivdr-notified-bodies-survey-alarming-results/ [accessed 2024-02-02]
55. Fink M, Akra B. Comparison of the international regulations for medical devices-USA versus Europe. Injury 2023;54 Suppl 5:110908 [FREE Full text] [doi: 10.1016/j.injury.2023.110908] [Medline: 37365092]
56. Chatterji AK. Spawned with a silver spoon? Entrepreneurial performance and innovation in the medical device industry. Strategic Management Journal 2008;30(2):185-206 [FREE Full text] [doi: 10.1002/smj.729]
57. Medical Device Coordination Group. UPDATE - MDCG 2020-3 Rev.1 - guidance on significant changes regarding the transitional provision under article 120 of the MDR with regards to devices covered by certificates according to MDD or AIMDD. European Commission. 2023. URL: https://health.ec.europa.eu/document/download/800e8e87-d4eb-4cc5-b5ad-07a9146d7c90_en?filename=mdcg_2020-3_en_1.pdf [accessed 2024-02-02]
58. Boston Consulting Group. Interstates and autobahns: global medtech innovation and regulation in the digital age. BCG, UCLA Biodesign Program. 2022. URL: https://web-assets.bcg.com/8c/f0/06744e8848ea9654bbd0765bf285/bcg-interstates-and-autobahns-mar-2022.pdf [accessed 2024-02-02]
59. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). The European Parliament and the Council of the European Union. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [accessed 2024-09-03]

Abbreviations

AI: artificial intelligence

EPSCO: Employment, Social Policy, Health and Consumers Affairs
FDA: Food and Drug Administration
MDD: Medical Device Directive 93/42/EEC
MDR: Medical Device Regulation (EU) 2017/745
ML: machine learning
NB: notified body
PIP: Poly Implant Prothèse
SAMD: software as a medical device
SME: small- and medium-sized enterprise
UDI: unique device identification

Edited by M Focsa; submitted 05.03.24; peer-reviewed by Z Zandesh, A Seitel, JIJ Green; comments to author 03.05.24; revised version received 08.05.24; accepted 25.05.24; published 05.09.24.

Please cite as:
Svempe L
Exploring Impediments Imposed by the Medical Device Regulation EU 2017/745 on Software as a Medical Device
JMIR Med Inform 2024;12:e58080
URL: https://medinform.jmir.org/2024/1/e58080
doi:10.2196/58080
PMID:39235850

Development of a Trusted Third Party at a Large University Hospital: Design and Implementation Study

Eric Wündisch[1], MSc; Peter Hufnagl[2], Prof Dr; Peter Brunecker[3], Dr rer medic; Sophie Meier zu Ummeln[1], CEng; Sarah Träger[1], MA; Marcus Kopp[1], BSc; Fabian Prasser[4,*], Prof Dr; Joachim Weber[1,5,6,*], Dr med

1

2

3

4

5

6

[*]these authors contributed equally

Corresponding Author:
Eric Wündisch, MSc

Abstract

Background: Pseudonymization has become a best practice to securely manage the identities of patients and study participants in medical research projects and data sharing initiatives. This method offers the advantage of not requiring the direct identification of data to support various research processes while still allowing for advanced processing activities, such as data linkage. Often, pseudonymization and related functionalities are bundled in specific technical and organization units known as trusted third parties (TTPs). However, pseudonymization can significantly increase the complexity of data management and research workflows, necessitating adequate tool support. Common tasks of TTPs include supporting the secure registration and pseudonymization of patient and sample identities as well as managing consent.

Objective: Despite the challenges involved, little has been published about successful architectures and functional tools for implementing TTPs in large university hospitals. The aim of this paper is to fill this research gap by describing the software architecture and tool set developed and deployed as part of a TTP established at Charité – Universitätsmedizin Berlin.

Methods: The infrastructure for the TTP was designed to provide a modular structure while keeping maintenance requirements low. Basic functionalities were realized with the free MOSAIC tools. However, supporting common study processes requires implementing workflows that span different basic services, such as patient registration, followed by pseudonym generation and concluded by consent collection. To achieve this, an integration layer was developed to provide a unified Representational state transfer (REST) application programming interface (API) as a basis for more complex workflows. Based on this API, a unified graphical user interface was also implemented, providing an integrated view of information objects and workflows supported by the TTP. The API was implemented using Java and Spring Boot, while the graphical user interface was implemented in PHP and Laravel. Both services use a shared Keycloak instance as a unified management system for roles and rights.

Results: By the end of 2022, the TTP has already supported more than 10 research projects since its launch in December 2019. Within these projects, more than 3000 identities were stored, more than 30,000 pseudonyms were generated, and more than 1500 consent forms were submitted. In total, more than 150 people regularly work with the software platform. By implementing the integration layer and the unified user interface, together with comprehensive roles and rights management, the effort for operating the TTP could be significantly reduced, as personnel of the supported research projects can use many functionalities independently.

Conclusions: With the architecture and components described, we created a user-friendly and compliant environment for supporting research projects. We believe that the insights into the design and implementation of our TTP can help other institutions to efficiently and effectively set up corresponding structures.

(JMIR Med Inform 2024;12:e53075) doi:10.2196/53075

KEYWORDS

pseudonymisation; architecture; scalability; trusted third party; application; security; consent; identifying data; infrastructure; modular; software; implementation; user interface; health platform; data management; data privacy; health record; electronic health record; EHR; pseudonymization

Introduction

Background

Medical research relies on the effective collection, management, and analysis of biomedical data [1]. However, the complexity of associated data flows is increasing constantly due to the rising importance of data-driven approaches from the areas of data science and artificial intelligence [2,3]. These typically require data to be reused and shared to generate the necessary large data sets, for example in neuroscience [4]. At the same time, relevant data are often highly sensitive and require protection against unauthorized use and disclosure [5]. In alignment with this need, various laws, regulations, guidelines, and best practices suggest pseudonymization as a central data protection mechanism, especially in biomedical research [6]. Pseudonymization refers to a process in which data that directly identifies individuals (henceforth denoted as identifying data), such as names and addresses, are stored separately from data and biosamples needed for scientific analyses, and research assets are identified using protected identifiers, known as pseudonyms [7]. This protects the identity of patients or study participants while still allowing the implementation of complex research workflows, for example, data linkage. It is frequently suggested to bundle pseudonymization with other functionalities relevant to data protection and compliance, such as consent management, and that those should be carried out by particularly trusted units, knwon as trusted third parties (TTPs). One example of a concept recommending TTPs is the Guideline for Data Protection in Medical Research Projects by Technology, Methods, and Infrastructure for Networked Medical Research (TMF), the German umbrella organization for networked medical research [8].

Although the general functionalities required by medical research projects may be similar, the way they are combined into workflows often differs significantly. The reason is that due to varying study schedules and (data) modalities, studies often have different requirements concerning the necessary number and types of pseudonyms as well as the research assets that have to be registered. The timing of consent collection can also vary, for example, if reconsenting is required. Another factor that can contribute to heterogeneity is the need for integration of or linkage with data from external systems or institutions. As a result, studies often develop study- or project-specific solutions to fulfill specific registration, pseudonymization, linkage, and consenting requirements [9]. Some open tools, such as Enterprise Identifier Cross-Referencing (E-PIX) [10], Generic Pseudonym Administration Service (gPAS) [11], Generic Informed Consent Service (gICS) [12], or Mainzelliste [13], have been developed and are in widespread use; however, they are usually not integrated with each other, making the implementation of more complex workflows involving different TTP operations challenging and potentially lead to systematic limitations (explained further in the *Discussion* section). Although research exists on the components mentioned above, the literature lacks insights into the design of more comprehensive architectures that support complex research workflows that are actually in production use [14,15].

Objectives

This paper presents the design of a comprehensive architecture for a TTP that aims to support a wide range of different research projects and studies using a unified system. As a first step, we present requirements elicited for this structure and then describe the implementation of a corresponding solution that reuses existing open components. These components are extended with a common application programming interface (API) and a common graphical user interface (GUI). We then present insights into our experiences with piloting this structure and describe our plans for future developments.

Methods

Requirements

TTPs typically offer a range of core functionalities based on their role in supporting research projects and clinical studies with data protection services. Three key functionalities provided are as follows: (1) identity management, through which patients and study participants are registered and their identities are managed across different systems using record linkage; (2) pseudonym management, which provides and manages pseudonyms for different research contexts and is thus critical for data protection compliance; and (3) consent management, to obtain and manage patient and participant consent for various research activities. Further components are usually included to make these core functionalities accessible. An API is necessary for the systematic retrieval of information, the implementation of complex workflows, and integration with further health care and research systems. Moreover, a well-designed GUI is necessary to enable TTP staff and study personnel to perform common tasks efficiently. An audit trail is required to ensure transparency and traceability. Furthermore, data import and export functions are necessary for transferring data from legacy systems and archiving in study-specific contexts. Finally, platform independence is an important nonfunctional requirement to support wide adoption.

A common set of tools providing these core functionalities and features (Table 1) are E-PIX [10], gPAS [11], and gICS [12], which are provided as free web-based software by the MOSAIC project from the University of Greifswald (explained in the following section). They are successfully used in a range of research projects and infrastructures [16]. Table 1 illustrates which of the above-mentioned core requirements are fulfilled by which of the MOSAIC tools.

Table . Core functional requirements and MOSAIC tools that fulfill them.

Core functional requirements		Tools		
		E-PIX[a]	gPAS[b]	gICS[c]
Basic services				
	Identity management	✓	—[d]	—
	Pseudonym management	—	✓	—
	Consent management	—	—	✓
Additional features				
	Application programming interface	✓	✓	✓
	Graphical user interface	✓	✓	✓
	Audit trail	✓	—	✓
	Data import and export	✓	✓	✓

[a]E-PIX: Enterprise Identifier Cross-Referencing.

[b]gPAS: Generic Pseudonym Administration Service.

[c]gICS: Generic Informed Consent Service.

[d]Not applicable.

Although the MOSAIC tools provide the basic functionalities needed, we elicited additional requirements from our extensive experience with supporting research projects. An overview is provided in Table 2. A detailed discussion is available in the section *Comparison With Prior Work*.

Table . Additional functional requirements and core services for which they are relevant.

Additional functional requirements		Identity management	Pseudonym management	Consent management
Programmatic interfaces and workflows				
	Modern REST[a] application programming interface	✓	✓	✓
	Information exchange with other systems (eg, for ingesting consents documented in the EHR[b] system)	✓	✓	✓
	Cross-system workflows (eg, creation of a primary identifier, combined with the creation of all necessary pseudonyms based on the domain tree and preparation of a consent document)	✓	✓	✓
User interfaces and services				
	Integrated user interface across all services	✓	✓	✓
	Common authentication and authorization framework with single-sign-on and associated rights and roles with the ability to connect to institutional directory services	✓	✓	✓
	Sending status messages to users in case of relevant events (eg, when a new patient has been registered)	✓	✓	✓
Specific features				
	Visualization of pseudonyms as QR codes	—[c]	✓	—
	Automated versioning when storing consent updates	—	—	✓
	Kiosk mode for consent documentation	—	—	✓

[a]REST: representational state transfer.

[b]EHR: electronic health record.

[c]Not applicable.

Programmatic Interfaces and Workflows

Representational state transfer (REST) services have become a de facto standard for modern applications over the last couple of years, as they are stateless, lean, and based on open web standards. Hence, we considered a REST API to be an important requirement for all 3 areas—identity management, pseudonym management, and consent management. Together with other common technologies, such asJavaScript Object Notation, this makes the services offered by the TTP accessible to other systems and processes. It also fosters effective information exchange with other systems, for example, to automatically generate primary identifiers and pseudonyms in case a patient is registered in the electronic health record (EHR) system. Moreover, a common API across all services also enables cross-service workflows, which we consider particularly important. An example of this is the automatic creation of pseudonyms linked to the primary identifier when registering a patient or study participant.

User Interfaces and Services

We considered an integrated user interface (UI) together with a shared authentication and authorization mechanism to be central for our TTP infrastructure. Important functionalities that the UI needs to support include depseudonymization, patient and participant registration, consent management and configuration, as well as administration. A tighter integration of the different components also facilitates sending status messages to users in case actions are required on their side.

Specific Features

We further identified requirements in regard to specific management functionalities. For example, representing pseudonyms as QR codes is important for seamless workflows

across different media; this includes printing the codes on accompanying documents or biospecimen tubes and then reading them using QR code readers. This is particularly important for biospecimen management. Moreover, we identified a need for versioning of managed consent documents. In the event of updates to consents, for example, due to wrong information on the consent form, versioning of the various consents in the system is important for traceability. This also requires the system to be able to assign consents or withdrawals to other participants (eg, if a wrong identifier has been used when originally collecting the form). In addition, a kiosk mode that locks the user into the application is needed for the secure collection of consents from patients using tablets.

Nonfunctional Requirements

The most important nonfunctional requirements are as follows: (1) scalability, particularly when executing cross-service operations, and (2) documentation of administration functions.

Building Blocks

In this section, we will describe basic building blocks of the developed application stack.

MOSAIC Tools

As mentioned previously, the application has been developed around the MOSAIC tools [17] as core components. Although these tools do not fulfill all our requirements, they provide a solid basis for implementing the core functionalities. The MOSAIC tools have been positively evaluated by the data protection authority of Mecklenburg-Vorpommern in Germany [18] and have been successfully used in several research projects, for example, the BeLOVE (Berlin Longterm Observation of Vascular Events) [19,20] and NAKO (German National Cohort) studies [21].

The MOSAIC suite consists of 3 tools [22]: E-PIX provides a master patient index following the Integrating the Healthcare Enterprise (IHE) profiles, Patient Identifier Cross-Reference (PIX), and Patient Demographics Query [23,24]; gPAS provides associated pseudonymization functionalities; and gICS supports integrated consent management. More specifically, E-PIX enables the central management of directly identifying master data and supports probabilistic record linkage. The resolution of potential matches between identifying data is supported through the UI. gPAS supports the generation and management of pseudonyms on top of the identities managed by E-PIX using different pseudonym domains that can refer to different systems, locations, or contexts. Finally, gICS supports digitally managing informed consent and supports different consent templates and associated use policies.

Following our requirements, we implemented an authentication and authorization model as well as programmatic interfaces and graphical UIs around E-PIX, gPAS, and gICS to enable integrated workflows across all 3 tools and to improve their interfaces.

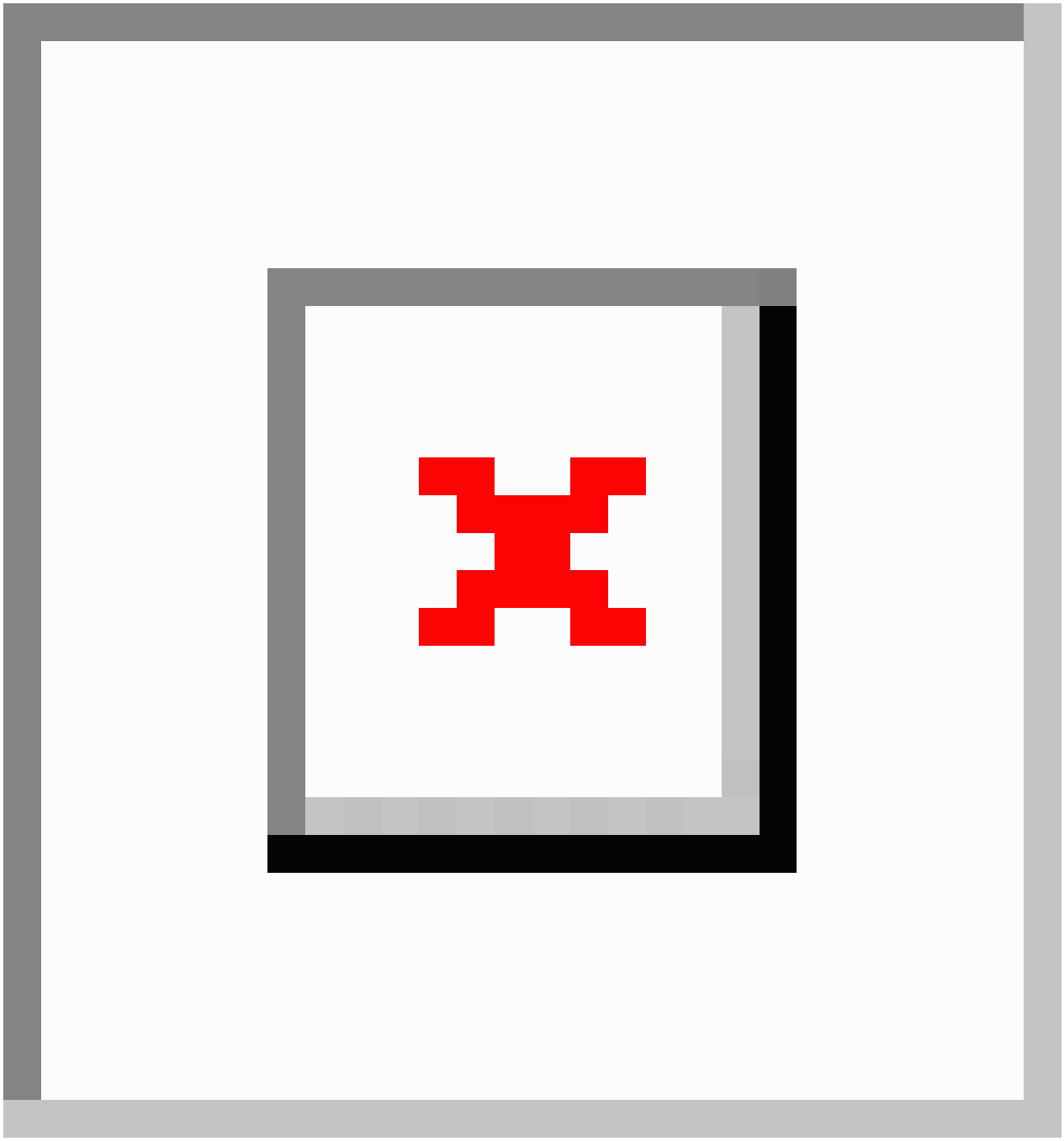
Authorization and Authentication

We designed a simple, yet flexible 3-stage authorization model, which combines permissions for basic object access with permissions regarding the domain of the object to be accessed (with create, read, write, or delete permissions) by a machine or human user of the infrastructure. An overview is provided in Figure 1.

A domain defines the scope of the data managed by the TTP (eg, a research process, a study, a project, or an institute). Multiple domains can be created within a project (eg, to store pseudonyms used in specific subprojects or contexts). Additionally, in gPAS, a domain can have parent and child domains. This results in a tree structure that can be used to tailor permissions to different scopes within individual projects [25].

On the implementation side, we mapped this model to OpenID Connect (OIDC), which is based on OAuth 2.0 [26]. The JavaScript Object Notation Web Token generated in this process contains role names as attributes, which are platform independent and can also be processed on mobile devices. This is important for the additional UIs that we had to develop. As an identity and access management solution, we chose Keycloak, which is in widespread use, has a native administration interface, and is published as open-source software under the Apache License 2.0. Importantly, it can also be connected to a range of directory services usually maintained by hospitals for account and permission management.

Figure 1. Stages of the functional authorization model.



Programmatic Interface

We decided to implement a REST API to extend the programmatic interfaces of E-PIX, gPAS, and gICS and support cross-tool workflows. Due to its stateless nature, this design enables the management and sharing of data across different systems, combined workflows, and calls by external components. One important application of the unified REST API is to combine participant registration with automatic consent checking in gICS, indexing the participant in E-PIX, and generating pseudonyms in gPAS. Furthermore, the REST API can easily be integrated with the developed authentication and authorization model as well as logging and audit trail functionalities. Existing interfaces of MOSAIC tools can also be integrated with the permission model by wrapping them behind REST interfaces.

Graphical Interfaces

Web Interface

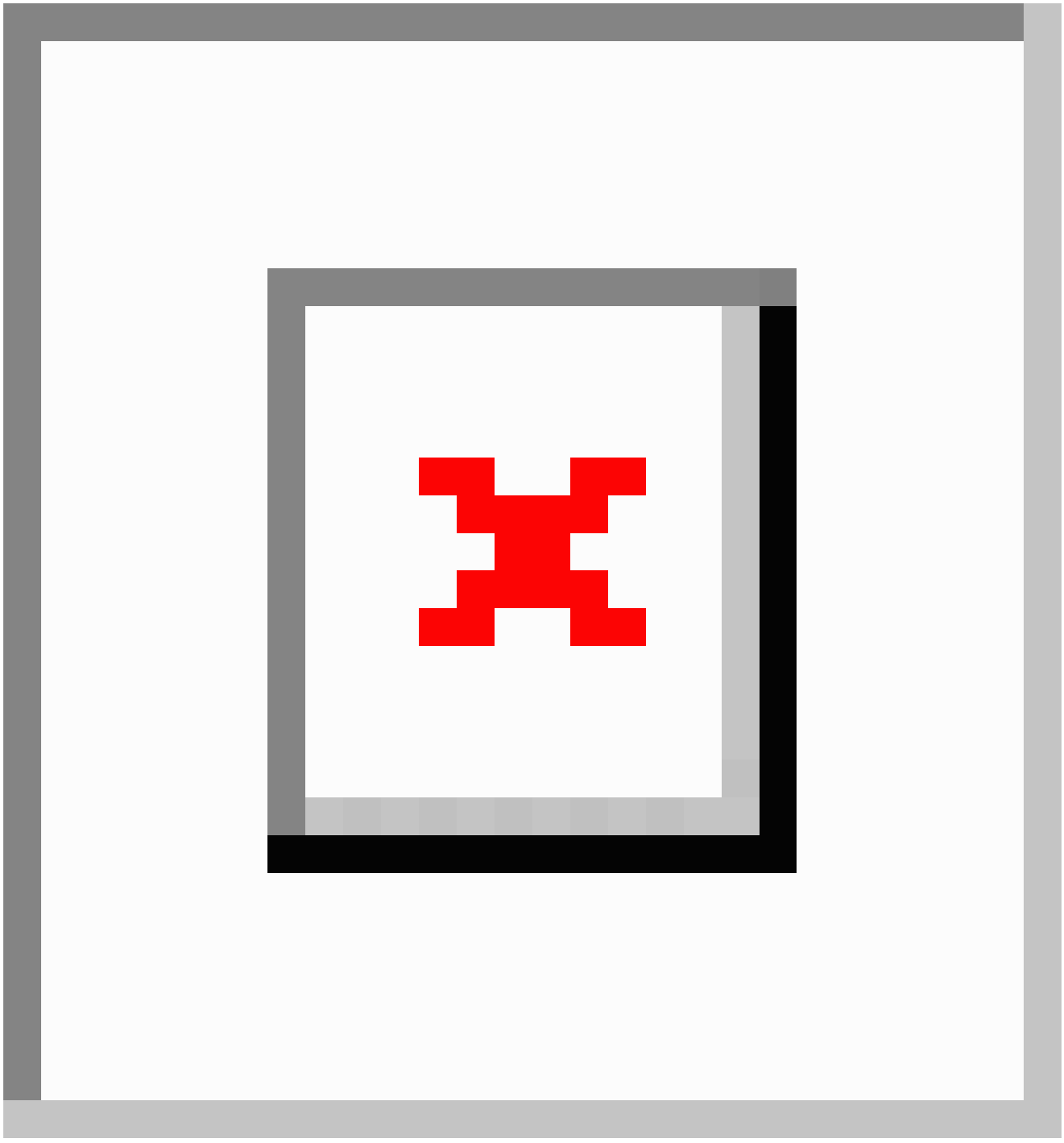
Based on the integrated programmatic API that supports all services, we have also implemented an integrated GUI, which allows accessing all TTP services in a unified manner. Analogously to the programmatic API, the UIs are integrated with the described authentication and authorization model. Users can log into the platform with their account from the connected directory service, which is abstracted way using OIDC with Keycloak. The token generated at log-in contains all assigned permissions, which are used in the UI and sent as a bearer token

with each request to the REST services. A strict content-security-policy workflow blocks the execution of foreign scripts outside the origin domain, thus increasing the level of security. Actions such as participant administration, depseudonymization, or consent administration can be performed through wizards. Users can request essential documents, such as copies of consent, directly from the web application.

Mobile App

The final building block is provided by a mobile app that serves as a direct channel from the TTP services to the participants. The most important application is collecting consent and handling withdrawals. A typical deployment consists of installing the appl on a tablet, which is then configured by study personnel and handed over to the participants (Figure 2).

Figure 2. Workflow of actions in the app.



The study personnel can log into the app using the same log-in data as for the TTP web interface. After the project staff member enters a participant identification code and selects either a consent or a withdrawal form, the selected participant fills out the form. To prevent participants from accessing unauthorized information, the app will be started in kiosk mode. The identification code is either a temporary pseudonym or an already existing pseudonym for the participant, providing direct linkage to the research project managed by the TTP. In the latter case, the app automatically opens the associated consent template. After filling out the form, the participants can enter their name and place of residence, and then, they can put their

signature in a designated field. Afterwards, the staff member provides their signature, confirming that the form has been completed with them as the assigned project staff member.

Supported Pseudonym Algorithms

In our system, generated random numbers are used as pseudonyms. The length is configurable, with a minimum of 6 digits, and is chosen based on the number of pseudonyms that are needed for the respective project. Additionally, we use the Damm algorithm to detect single-digit errors and all adjacent transposition errors with a simple checksum [27]. Moreover, pseudonyms are combined with study- and context-specific prefixes. For example, the pseudonym "BLV-US-123456" could represent an ultrasound ("US") measurement for a study participant in a study called BeLOVE ("BLV"). Finally, our system can also import and manage existing pseudonyms. As those are usually generated using different algorithms and often do not contain a checksum, we mark them as "external" within the system.

Ethical Considerations

This paper covers the design and implementation of a generic research service, which requires no ethics committee approval according to local policies. However, the individual studies that use the service have to apply for ethics approval. For example, the BeLOVE study, which is described as a case study in this paper, was approved by Charité's ethics committee (vote number EA1/066/17).

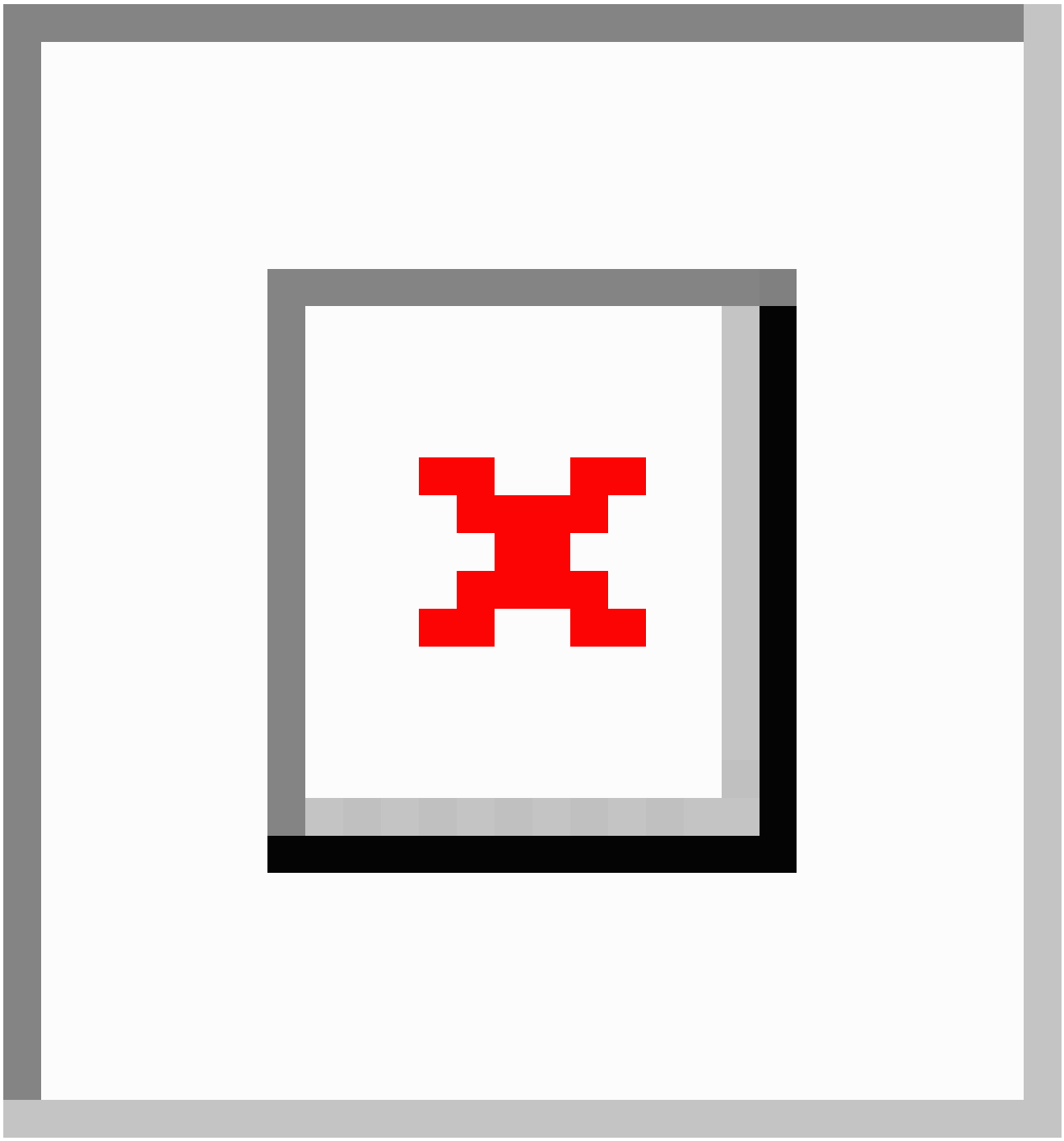
Results

In this section, we will first describe the general architecture of our solution, then cover important implementation details, and finally report on real-world experiences with the platform.

Architecture

The overall architecture is divided into the API, which wraps around the MOSAIC tools, the graphical interfaces oriented toward users, as well as the access and identity management component (Figure 3 presents more details).

Figure 3. Architecture overview, including wrapped MOSAIC stack (core components); systems maintained by the trusted third party (TTP; graphical components as well as access and identity components); systems queried by the TTP (electronic health record [EHR] system and directory services); and systems from which the TTP is queried (Research Electronic Data Capture [REDCap]). E-PIX: Enterprise Identifier Cross-Referencing; gICS: Generic Informed Consent Service; gPAS: Generic Pseudonym Administration Service.



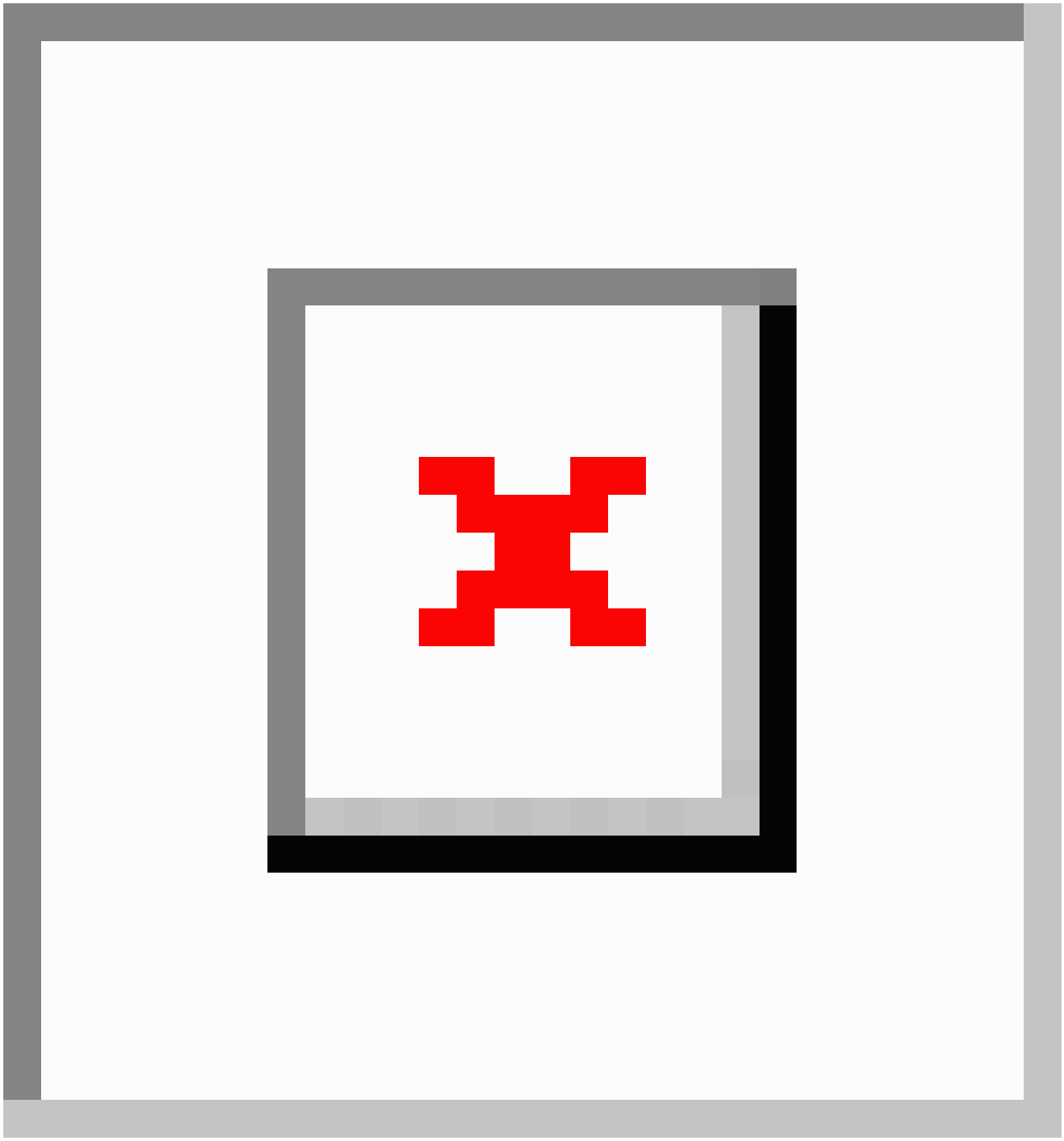
As illustrated, the core components are provided with an interface to the EHR system to support the pseudonymization of patient identities for direct reuse in the respective research context. Other systems that can access the TTP services via the REST API are, for example, electronic data capture systems, such as Research Electronic Data Capture (REDCap), or biobank information systems. All components of the respective interfaces are containerized with Docker [28] and deployed on a Docker swarm [29]. By using OIDC based on OAuth 2.0 as the standard, we were able to integrate other systems via existing packages (eg, Spring-Boot-Security) and allow other applications to access the systems. When modeling the interfaces, we ensured that anything that could be done graphically could also be done programmatically. This keeps the platform open and supports other information systems with the integration of TTP services.

Implementation

The REST API was implemented using Java 13 with the Spring Boot framework [30] by focusing on stable packages, including Spring Security for OIDC, and relying on an established framework. The resulting platform is robust, maintainable, extensible, and flexible. We have implemented 35 generic interfaces so far, most of which are Create-Read-Update-Delete (CRUD) interfaces for the key information objects Domain,

Participant, Identifier, Pseudonym, Consent, and Consent Template (Figure 4), as well as additional directory and search functions for pseudonyms and consents.

Figure 4. Key information objects and their relationships.



The web-based interface (Figures 5 and 6) is implemented using the PHP-based lightweight enterprise web framework Laravel [31]. Laravel uses a Model-View-Controller pattern [32], has a template engine named Blade, and supports agile development processes. By integrating the open-source framework Bootstrap, we were able to implement a responsive front end that could be displayed in browsers on multiple types of devices. The web application directly interfaces with the REST API and does not manage any participant data in a separate database.

Figure 5. Screenshots of the user interface: editing consent information.

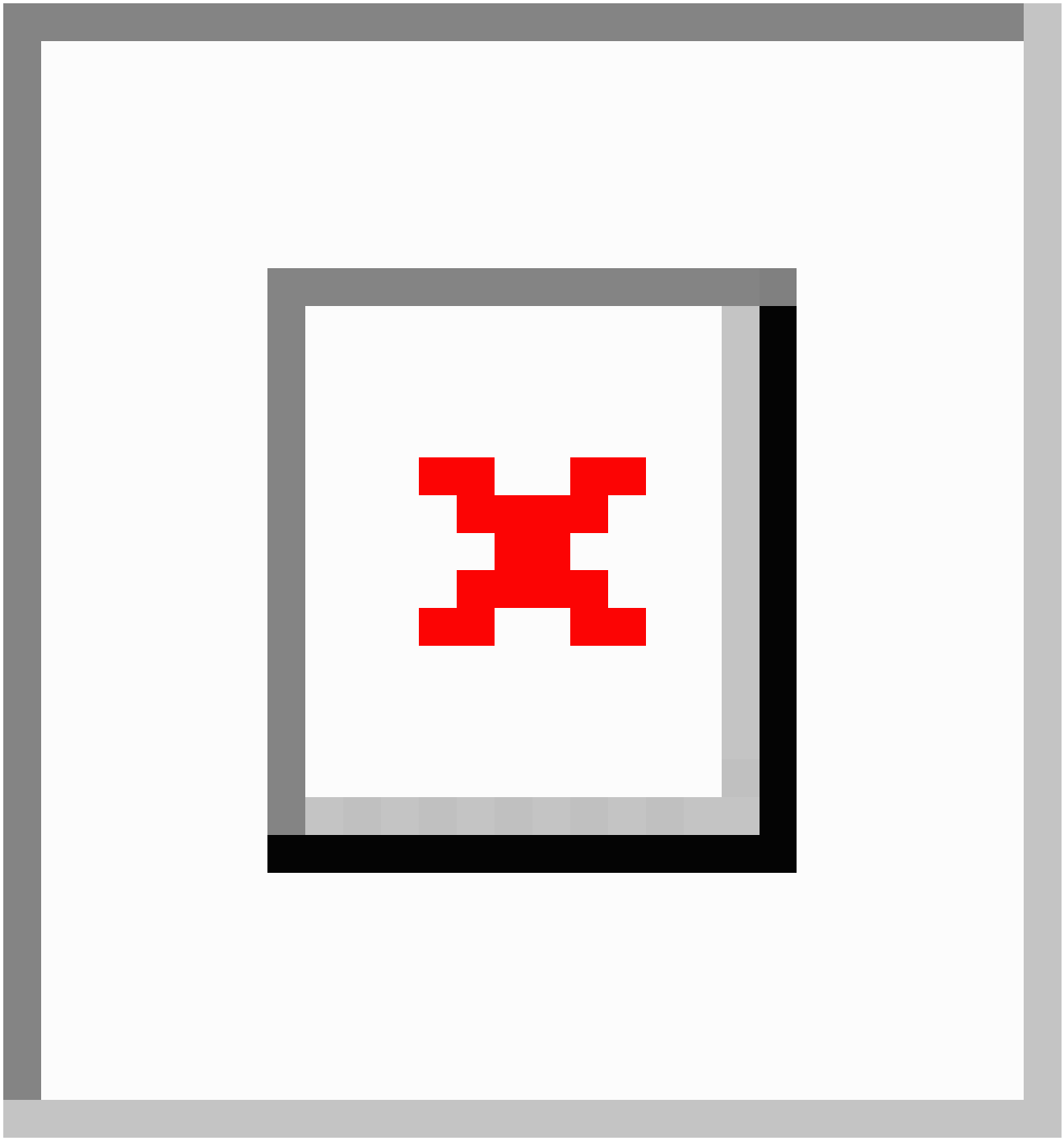
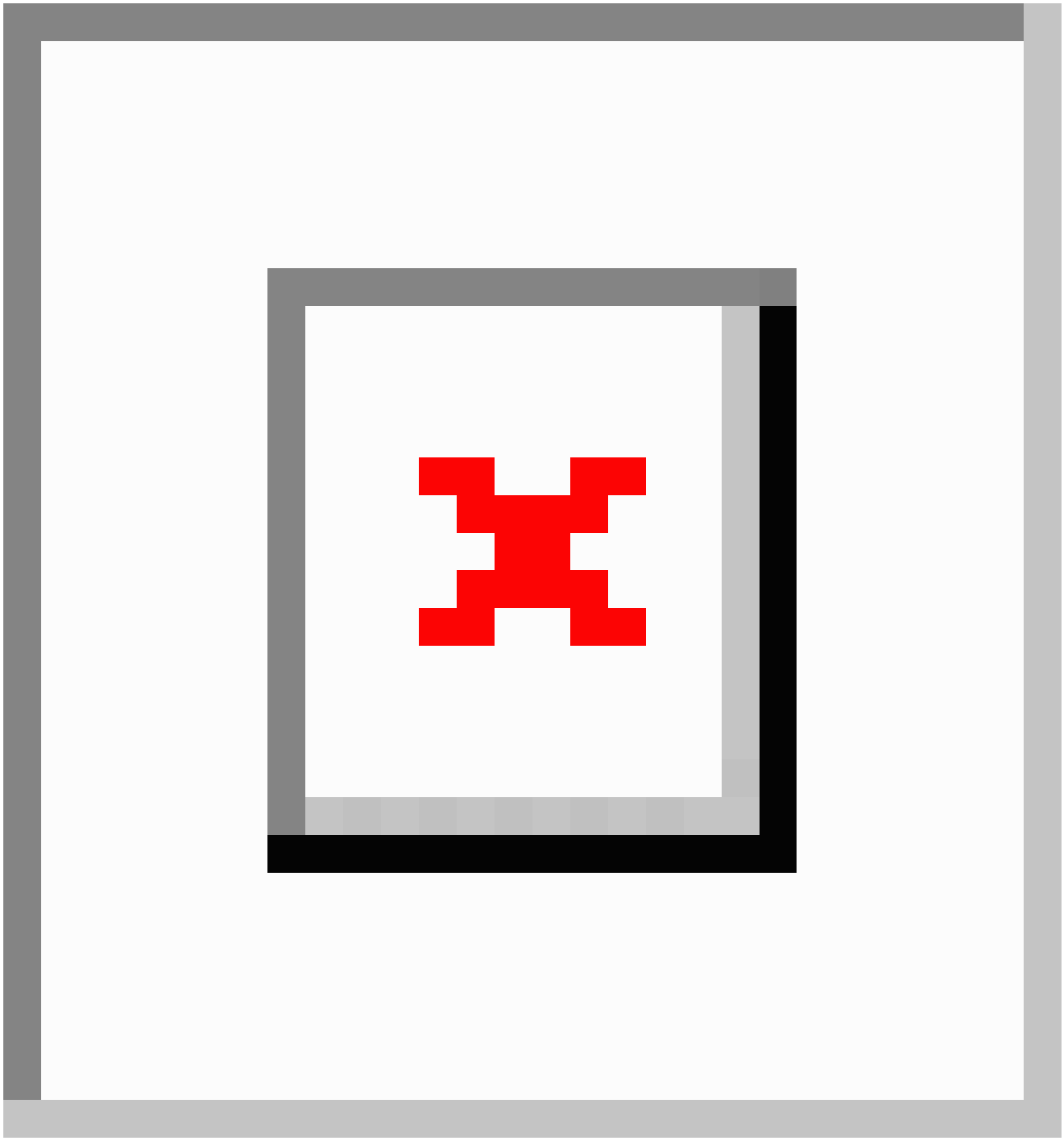


Figure 6. Screenshot of the user interface: overview of consent status.



The app front end (see Figures 7-9) was developed in React Native [33] and then significantly extended to work on tablets integrated into our mobile device management. The application does not permanently store any data on the device, and processing is carried out exclusively via React Native state management.

Figure 7. Screenshot of the consent app: entering or scanning an ID.

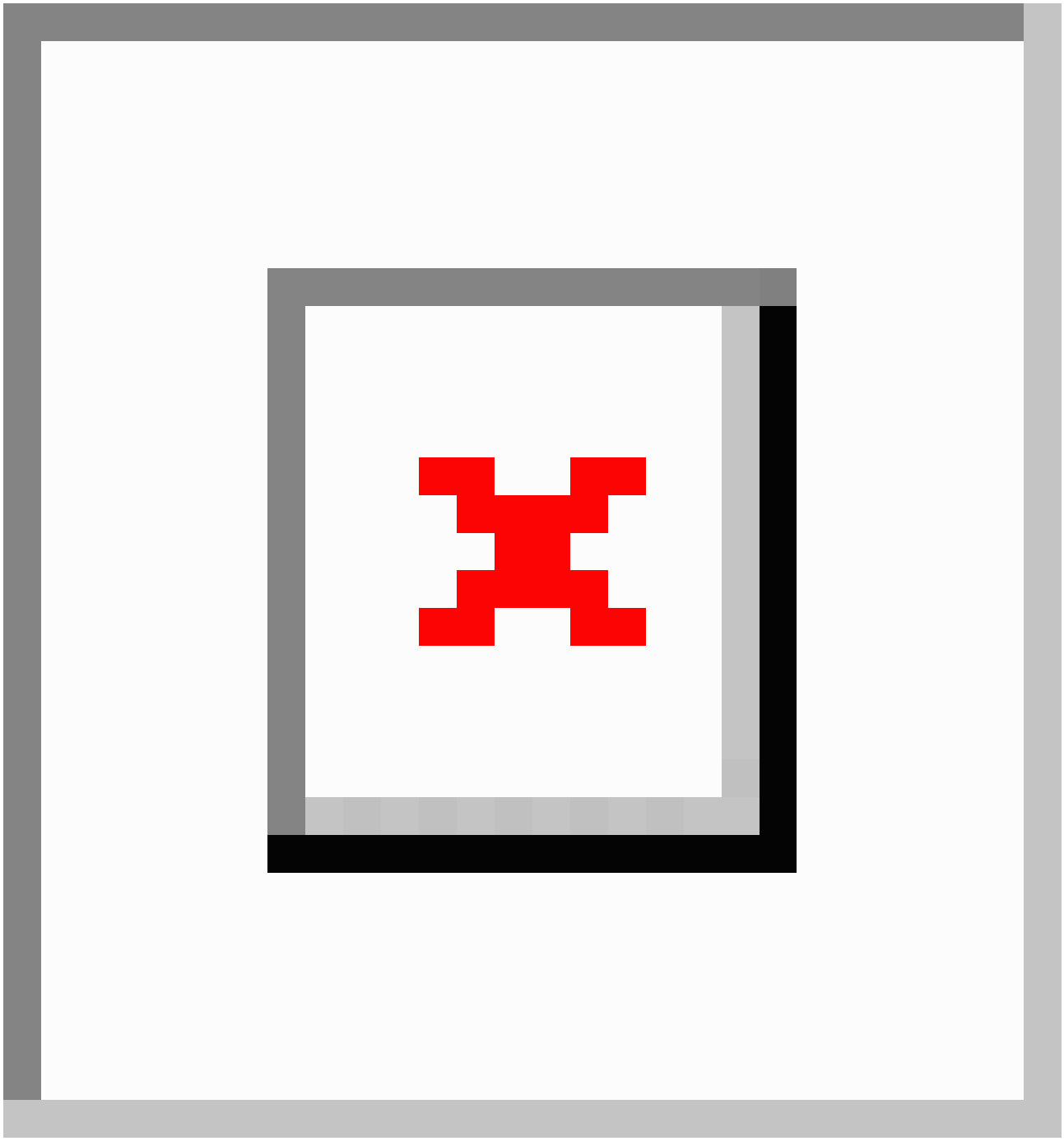


Figure 8. Screenshot of the consent app: filling out consent forms.

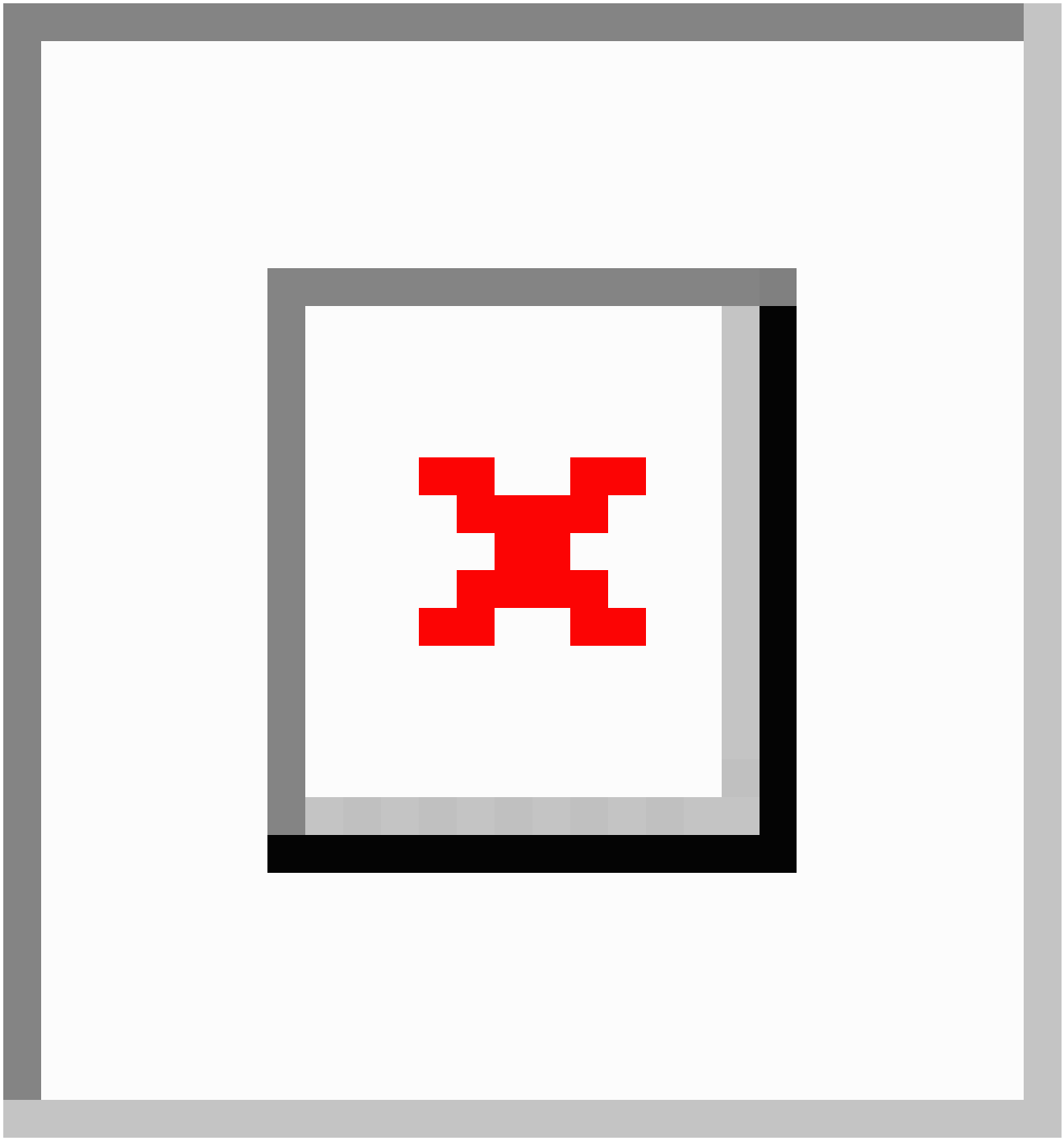
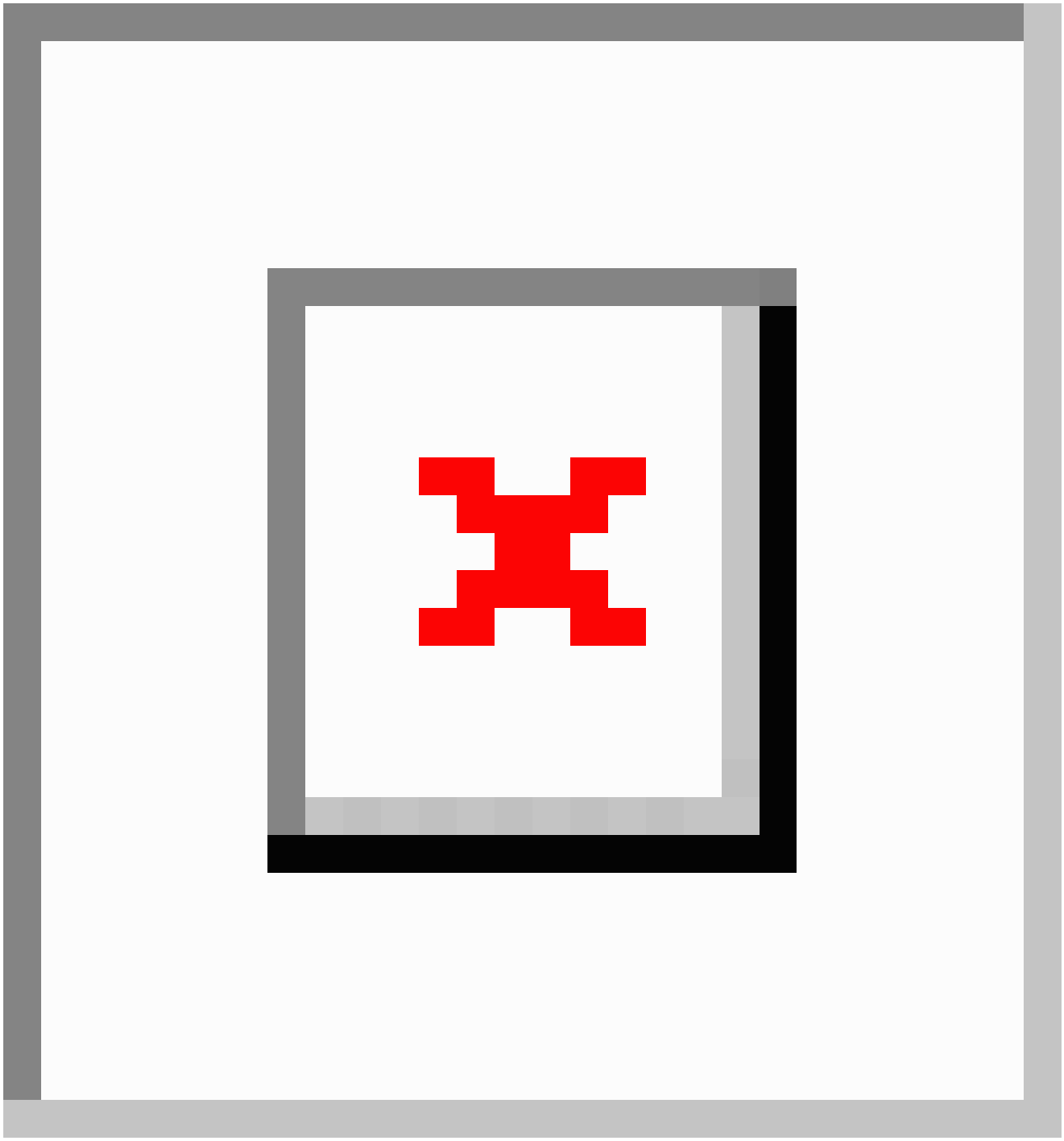


Figure 9. Screenshot of the consent app: sign and submit.



Core Functionalities for Research Projects

As a result of our development efforts, the TTP software stack provides a wide range of functionalities that research projects need. Table 3 provides an overview of frequently used common features.

Table . Essential functionalities provided to research projects.

Component	Process	Description
API[a]	Obtaining a temporary pseudonym	Automated creation of participant placeholders that can be used in third-party systems and later linked to the study identity
App	Electronic consent management	Viewing, completing, saving, and printing an electronic consent template of the respective project under a pseudonym
Web UI[b]	Participant registration	Master data and contact details can be entered manually or imported from the EHR[c] system
Web UI	Participant overview	Provides an overview of the participants and pseudonyms associated with a specific project
API	Integration with other systems	Interface for pseudonymization, depseudonymization, and linkage for third-party systems
Web UI	Depseudonymization	Resolve pseudonym to participant master data
Web UI	Retrieval of usage permissions based on consent information	Retrieve electronic representation of usage permissions from consents associated with a specific patient or participant pseudonym
Web UI	Update participant information	Use pseudonyms to update participant information

[a]API: application programming interface.

[b]UI: user interface.

[c]EHR: electronic health record.

On the API level, these features include integration with other systems to manage pseudonymization, depseudonymization, and data linkage. The app specializes in electronic consent management, specifically viewing, completing, and saving of consent templates. The web-based UI permits registration of participant details; provides an overview of participants, consents, and pseudonyms; supports depseudonymization as well as the retrieval of use permissions based on consent information. CRUD operations for major participant properties and printing consents are also supported.

Experiences in Real-World Operational Settings

The TTP has already supported more than 10 research projects since it was launched in December 2019. As of December 2022, our TTP system manages data of 3610 registered participants with 384,813 pseudonyms and 1762 consent documents. The pseudonyms fall into 2 categories: 40,867 pseudonyms have been assigned to individual participants managed by the TTP and 343,946 pseudonyms to other identifiers (eg, health insurance numbers that are managed by the TTP as part of its support for data linkage). On average, the TTP manages about 11 pseudonyms for each individual participant. As many as 153 research personnel actively engage with the software on a daily basis. Backups of our databases are created every night. These backups are stored for 90 days along with all log files.

As a case study, we will describe how the TTP services are being used by the large-scale BeLOVE study [20], which is carried out as a cooperation between several sites and departments at Charité. BeLOVE uses all services provided, from patient as well as participant registration and consent management, to pseudonym generation for the various diagnostics and phenotyping activities performed during hospitalizations or study visits (about 12 pseudonyms per participant). Compared to the initial planning of the study, which required 2 study staff for the administrative tasks, these staff requirements were in the meantime reduced to zero due to the functionality of our TTP and the associated secure outsourcing of tasks to all study staff. The use of central TTP services has also significantly reduced the efforts required for coordinating BeLOVE and its substudies with the data protection and information security officers. Within Charité's internal data integration platform, consistent pseudonyms and API access to mapping rules are frequently used to link data collected about BeLOVE participants with routine health care data collected during inpatient and outpatient encounters for various types of analyses. Secondary pseudonyms have already been generated for 10 projects in which the data have been analyzed or shared with others.

Discussion

Principal Results

In this paper, we have presented a software stack to support a TTP with its core tasks at a large German academic medical center. Our architecture extends existing systems for key functionalities, identity management, pseudonymization, and consent management with a fine-grained authentication and authorization model, a modern REST API, two types of UIs, and connections to third-party systems. These extensions were necessary to support cross-service workflows on the programmatic as well as the user level and to meet further functional and nonfunctional requirements. Our application is built using various open-source enterprise frameworks and standards (eg, OIDC) to ensure sustainability and integration

with important institutional services (eg, our user directory and leading master patient index). Our experiences with supporting a wide range of research projects with TTP services over a longer period have shown that our approach works and provides functionalities that are generic enough to support a wide range of applications.

Comparison With Prior Work

Our architecture and implementation are based on the MOASIC tools [16], which we have extended with additional components to overcome functional and nonfunctional shortcomings. Most importantly, the publicly available basic versions of the MOSAIC tools are not suitable for handling more complex and flexible workflows with fine-grained authorization. For example, supporting cross-service workflows, like registering a patient, generating pseudonyms, and preparing a consent form as an integrated operation, cannot be implemented without an additional dispatcher component that is currently not publicly available. We solved this by implementing a cross-service REST API. Although the MOSAIC tools already come with an API, it is provided individually for each service and is based on the Simple Object Access Protocol [34], which originates from the IHE web service standards [35] and is complex and slow, requiring managing server-side state. Analogously to an API, the MOSAIC tools also offer GUIs. However, they are provided individually for each service and hence do not enable users to seamlessly perform operations that require interactions with multiple core services. For this reason, we developed a cross-service UI that is based on our API. Additionally, we added functionalities for generating QR codes, versioning consent documents, and starting the system in kiosk mode. Finally, our extensions also improve the system's scalability when executing cross-service operations, such as querying for links between pseudonyms and identifiers, which can be slow when using the MOSAIC tools [36]. We also added comprehensive documentation of administration functions, which is not fully available for the current open-source versions without registration with the vendor [37].

Prior work on TTP-related services usually focused on individual components or algorithms that could support TTP operations, deployments in specific research projects, or high-level architecture overviews.

One well-known example is the one-way hash approach employed by Vanderbilt University Medical Center as part of the ingest process into their deidentified layer within a research data warehouse [38]. Pommering et al [39] describe strategies for how pseudonymization could be used in different contexts, for example, in the secondary use of EHR data or in medical research networks and biobanks. They introduced two models that support repeated depseudonymization as well as one-time use [40]. The former model was later integrated into a concept for sharing large data sets in medical research networks and biobanks [39].

Building on this, Lo Iacono [41] investigated a cryptographic approach for generating consistent pseudonyms in multicentric studies but without describing a specific implementation within a concrete project. Dangl et al [42] describe concepts and requirements for TTP services for a specific biobank of a clinical research group. Heinze et al [43] developed two services based on IHE profiles that have been implemented into the Heidelberg Personal Electronic Health Record. One service is used to capture patient consent, while the other provides a GUI to manage consents. Further components (eg, for pseudonym or identity management) were not described in detail.

Lablans et al [13] introduce the Mainzeliste, which supports managing patient identities and pseudonyms through a web-based front end. Bialke et al [10] introduce the MOSAIC tools, which we also use in our work, as a set of tools supporting central data management for studies or research networks. They also introduce the "dispatcher" as an additional component for building complex workflows [22], which is, as we described above, unfortunately not publicly available.

Aamot et al [44] compare different strategies for depseudonymization in which, among others, the strategy of Pommering et al [39] is compared with alternative approaches. Based on this comparison, they develop a pseudonymization approach using deterministic one-way mappings based on cryptographic protocols. Lautenschläger et al [45] implement and describe a generic and tightly coupled architecture and component for pseudonymization that has been used in several research projects. On the application side, Bahls et al [14] describe a TTP architecture using the MOSAIC tools for the Routine Anonymized Data for Advanced Health Services Research project. Hampf et al [17] benchmark parts of the MOSAIC tools and conclude that it would take several days to register 2 million patients with the hardware setup utilized.

Limitations and Future Work

As the most recent versions of the MOSAIC tools are not distributed as open-source software in a public repository [37], it was not possible for us to make changes to the core tools used. Instead, workarounds had to be implemented at the API or UI level, which is not ideal from an architecture perspective. Moreover, our TTP platform is currently focused on providing intra-institutional services only. In future work, we plan to extend our platform with external interfaces, enabling the TTP to act as a central trustee for multicentric projects. We also aim to implement additional programmatic interfaces following international interoperability standards, in particular, Health Level 7 Fast Healthcare Interoperability Resources [46] and enable study personnel to directly manage the permissions of associated staff. Finally, we plan to introduce a unified pool of consent policy keys to harmonize the permission information that can be queried from our system to enable automated downstream processing that considers consent information.

Conclusions

Scalable and comprehensive TTP services are central to modern data-driven medical research. However, community-based comprehensive platforms that can be used to implement such services are still lacking. We believe that our description of key requirements as well as the insights provided into our flexible architecture that combines core tools with user- and application-oriented workflows and interfaces, including third-party applications, can help other institutions setting up comparable services.

Acknowledgments

The authors would like to thank the BeLOVE (Berlin Longterm Observation of Vascular Events) study team, who have contributed to the improvement of the entire system with their constant feedback. This work was, in part, supported by the German Federal Ministry of Education and Research under grant agreement number 16DTM215 (THS-MED).

Conflicts of Interest

None declared.

References

1. Pommerening K, Sax U, Müller T, Speer R, Ganslandt T, Drepper J. Integrating eHealth and medical research: the TMF data protection scheme. EHealth Comb Health Telemat Telemed Biomed Eng Bioinforma Edge 2008 Jan:5-10 [FREE Full text]
2. Borda A, Gray K, Fu Y. Research data management in health and biomedical citizen science: practices and prospects. JAMIA Open 2019 Dec;3(1):113-125. [doi: 10.1093/jamiaopen/ooz052] [Medline: 32607493]
3. Wang X, Williams C, Liu ZH, Croghan J. Big data management challenges in health research-a literature review. Brief Bioinform 2019 Jan 18;20(1):156-167. [doi: 10.1093/bib/bbx086] [Medline: 28968677]
4. Zhao Z, Chuah JH, Lai KW, et al. Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: a review. Front Comput Neurosci 2023 Feb 6;17:1038636. [doi: 10.3389/fncom.2023.1038636] [Medline: 36814932]
5. Eggert K, Wüllner U, Antony G, et al. Data protection in biomaterial banks for Parkinson's disease research: the model of GEPARD (Gene Bank Parkinson's Disease Germany). Mov Disord 2007 Apr 15;22(5):611-618. [doi: 10.1002/mds.21331] [Medline: 17230444]
6. Bourka A, Drogkaris P, editors. Recommendations on Shaping Technology According to GDPR Provisions - An Overview on Data Pseudonymisation: The European Union Agency for Network and Information Security (ENISA); 2019.
7. Kohlmayer F, Lautenschläger R, Prasser F. Pseudonymization for research data collection: is the juice worth the squeeze? BMC Med Inform Decis Mak 2019 Sep 4;19(1):178. [doi: 10.1186/s12911-019-0905-x] [Medline: 31484555]
8. Pommerening K, Drepper J, Helbing K, Ganslandt T. Leitfaden Zum Datenschutz in Medizinischen Forschungsprojekte: Medizinisch Wissenschaftliche Verlagsgesellschaft (MWV); 2015.
9. Lowrance W. Learning from experience: privacy and the secondary use of data in health research. J Health Serv Res Policy 2003 Jul;8 Suppl 1:S1:2-7. [doi: 10.1258/135581903766468800] [Medline: 12869330]
10. Bialke M, Bahls T, Havemann C, et al. MOSAIC—a modular approach to data management in epidemiological studies. Methods Inf Med 2015;54(4):364-371. [doi: 10.3414/ME14-01-0133] [Medline: 26196494]
11. Geidel L, Bahls T, Hoffmann W. Generische Pseudonymisierung ALS Modul des Zentralen Datenmanagements Medizinischer Forschungsdaten. Universitätsmedizin. 2013. URL: https://www.ths-greifswald.de/wp-content/uploads/2019/09/Poster_DGEpi_PSN_2013_09_27.pdf [accessed 2024-04-10]
12. Rau H, Geidel L, Bialke M, et al. The generic informed consent service gICS: implementation and benefits of a modular consent software tool to master the challenge of electronic consent management in research. J Transl Med 2020 Jul 29;18(1):287. [doi: 10.1186/s12967-020-02457-y] [Medline: 32727514]
13. Lablans M, Borg A, Ückert F. A restful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak 2015 Feb 7;15:2. [doi: 10.1186/s12911-014-0123-5] [Medline: 25656224]
14. Bahls T, Pung J, Heinemann S, et al. Designing and piloting a generic research architecture and workflows to unlock German primary care data for secondary use. J Transl Med 2020 Oct 19;18(1):394. [doi: 10.1186/s12967-020-02547-x] [Medline: 33076938]
15. Bruland P, Doods J, Brix T, Dugas M, Storck M. Connecting healthcare and clinical research: workflow optimizations through seamless integration of EHR, pseudonymization services and EDC systems. Int J Med Inform 2018 Nov;119:103-108. [doi: 10.1016/j.ijmedinf.2018.09.007] [Medline: 30342678]
16. Projekte. Unabhängige Treuhandstelle. URL: https://www.ths-greifswald.de/forscher/projekte/ [accessed 2023-08-09]
17. Hampf C, Geidel L, Zerbe N, et al. Assessment of scalability and performance of the record linkage tool E-PIX in managing multi-million patients in research projects at a large university hospital in Germany. J Transl Med 2020 Feb 17;18(1):86. [doi: 10.1186/s12967-020-02257-4] [Medline: 32066455]
18. Unabhängige Treuhandstelle der Universitätsmedizin Greifswald. Universitätsmedizin. URL: https://www.medizin.uni-greifswald.de/de/forschung-lehre/core-units/treuhandstelle/ [accessed 2023-08-09]
19. Siegerink B, Weber J, Ahmadi M, et al. Disease Overarching mechanisms that explain and predict outcome of patients with high cardiovascular risk: rationale and design of the Berlin long-term observation of vascular events (Belove) study. medRxiv 2019 Jul 15:19001024. [doi: 10.1101/19001024]

20. Weber JE, Ahmadi M, Boldt LH, et al. Protocol of the Berlin long-term observation of vascular events (BeLOVE): a prospective cohort study with deep Phenotyping and long-term follow up of cardiovascular high-risk patients. BMJ Open 2023 Oct 31;13(10):e076415. [doi: 10.1136/bmjopen-2023-076415] [Medline: 37907297]
21. Bozoyan C, Fitzer K, Ostrzinski S, et al. Unabhängige Treuhandstelle (THS). NAKO Treuhandstellenkonzept. 2014. URL: https://nako.de/allgemeines/der-verein-nako-e-v/unabhaengig-treuhandstelle/ [accessed 2023-08-09]
22. Bialke M, Penndorf P, Wegner T, et al. A Workflow-driven approach to integrate generic software modules in a trusted third party. J Transl Med 2015 Jun 4;13:176 [FREE Full text]
23. GmbH GG. Das Sollten SIE Über EAN Nummern Wissen. GS1 Germany. URL: https://www.gs1-germany.de/ean-nummern/ [accessed 2024-01-04]
24. 23 patient identifier cross-referencing Hl7 V3 (Pixv3). IHE International. URL: https://profiles.ihe.net/ITI/TF/Volume1/ch-23.html [accessed 2023-09-25]
25. Hampf C, Bialke M. Unabhängige Treuhandstelle der Universitätsmedizin Greifswald. gPAS Anwenderhandbuch. 2023. URL: https://www.ths-greifswald.de/gpas/handbuch
26. Ma W, Sartipi K, Sharghigoorabi H, Koff D, Bak P. Openid connect as a security service in cloud-based medical imaging systems. J Med Imaging (Bellingham) 2016 Apr;3(2):026501. [doi: 10.1117/1.JMI.3.2.026501] [Medline: 27340682]
27. Damm MH. Total Anti-Symmetrische Quasigruppen [article in German].: Philipps-Universität Marburg; 2004 URL: https://archiv.ub.uni-marburg.de/diss/z2004/0516/ [accessed 2024-04-10]
28. Docker overview. Docker Docs. 2023. URL: https://docs.docker.com/get-started/overview/ [accessed 2023-08-09]
29. Docker swarm overview. Docker Docs. 2023. URL: https://docs.docker.com/engine/swarm/ [accessed 2023-10-09]
30. Spring Boot. URL: https://spring.io/projects/spring-boot/ [accessed 2023-08-14]
31. The PHP framework for web artisans. Laravel. URL: https://laravel.com/ [accessed 2023-08-14]
32. Krasner G, Pope S. A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. JOOP 1988 Jan [FREE Full text]
33. Kopp M. Entwicklung Einer App Zur Erfassung von Einverständniserklärungen Zur Datenverarbeitung Im Rahmen Einer Medizinischen Studie an Der Charité Berlin: HTW Berlin; 2021.
34. SOAP version 1.2 part 1: messaging framework (second edition). W3. URL: https://www.w3.org/TR/soap12/ [accessed 2023-08-10]
35. Appendix V: web services for IHE transactions. URL: https://profiles.ihe.net/ITI/TF/Volume2/ch-V.html [accessed 2023-09-25]
36. Fischer H, Röhrig R, Thiemann VS. A generic IT infrastructure for identity management and pseudonymization in small research projects with heterogeneous and distributed data sources under consideration of the GDPR. Stud Health Technol Inf 2019 Aug 21;264:1837-1838. [doi: 10.3233/shti190673]
37. Community. Unabhängige Treuhandstelle. URL: https://www.ths-greifswald.de/forscher/community/#collapse-1-5454 [accessed 2023-08-11]
38. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014 Dec;52:28-35. [doi: 10.1016/j.jbi.2014.02.003] [Medline: 24534443]
39. Pommerening K, Schröder M, Petrov D, Schlösser-Faßbender M, Semler SC, Drepper J. Pseudonymization service and data custodians in medical research networks- and biobanks. : Gesellschaft für Informatik eV; 2006. URL: https://dl.gi.de/handle/20.500.12116/23646 [accessed 2023-08-09]
40. Pommerening K, Reng M. Secondary use of the EHR via pseudonymisation. Stud Health Technol Inform 2004;103:441-446. [Medline: 15747953]
41. Lo Iacono L. Multi-centric universal pseudonymisation for secondary use of the EHR. Stud Health Technol Inform 2007;126:239-247. [Medline: 17476066]
42. Dangl A, Demiroglu SY, Gaedcke J, et al. The IT-infrastructure of a biobank for an academic medical center. Stud Health Technol Inform 2010;160(Pt 2):1334-1338. [Medline: 20841901]
43. Heinze O, Birkle M, Köster L, Bergh B. Architecture of a consent management suite and integration into IHE-based regional health information networks. BMC Med Inform Decis Mak 2011 Oct 4;11:58. [doi: 10.1186/1472-6947-11-58] [Medline: 21970788]
44. Aamot H, Kohl CD, Richter D, Knaup-Gregori P. Pseudonymization of patient Identifiers for translational research. BMC Med Inform Decis Mak 2013 Jul 24;13(1):75. [doi: 10.1186/1472-6947-13-75] [Medline: 23883409]
45. Lautenschläger R, Kohlmayer F, Prasser F, Kuhn KA. A generic solution for web-based management of pseudonymized data. BMC Med Inform Decis Mak 2015 Nov 30;15:100. [doi: 10.1186/s12911-015-0222-y] [Medline: 26621059]
46. HL7 FHIR. URL: https://www.hl7.org/fhir/ [accessed 2023-08-09]

Abbreviations

API: application programming interface
BeLOVE: Berlin Longterm Observation of Vascular Events
BLV: BeLOVE

CRUD: Create-Read-Update-Delete
E-PIX: Enterprise Identifier Cross-Referencing
EHR: Electronic Health Record
gICS: Generic Informed Consent Service
gPAS: Generic Pseudonym Administration Service
GUI: Graphical user interface
IHE: Integrating the Healthcare Enterprise
NAKO: German National Cohort
OIDC: OpenID Connect
PHP: Hypertext Preprocessor
PIX: Patient Identifier Cross-Reference
REDCap: Research Electronic Data Capture
REST: representational state transfer
TMF: Technology, Methods, and Infrastructure for Networked Medical Research
TTP: trusted third party

Edited by A Benis; submitted 25.09.23; peer-reviewed by HJ Kim, X Wu; revised version received 15.02.24; accepted 17.02.24; published 17.04.24.

Please cite as:
Wündisch E, Hufnagl P, Brunecker P, Meier zu Ummeln S, Träger S, Kopp M, Prasser F, Weber J
Development of a Trusted Third Party at a Large University Hospital: Design and Implementation Study
JMIR Med Inform 2024;12:e53075
URL: https://medinform.jmir.org/2024/1/e53075
doi:10.2196/53075

Original Paper

Privacy-Preserving Prediction of Postoperative Mortality in Multi-Institutional Data: Development and Usability Study

Jungyo Suh[1*], Prof Dr Med; Garam Lee[2*], MS; Jung Woo Kim[2], PhD; Junbum Shin[2], PhD; Yi-Jun Kim[3], Prof Dr Med, PhD; Sang-Wook Lee[4*], Prof Dr Med, PhD; Sulgi Kim[2*], PhD

[1]Department of Urology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[2]CryptoLab Inc, Seoul, Republic of Korea

[3]Department of Environmental Medicine, Ewha Womans University College of Medicine, Seoul, Republic of Korea

[4]Department of Anesthesiology and Pain Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Sang-Wook Lee, Prof Dr Med, PhD
Department of Anesthesiology and Pain Medicine
Asan Medical Center, University of Ulsan College of Medicine
88, Olympic-ro 43-gil, Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 1783
Fax: 82 2 3010 6790
Email: sangwooklee20@gmail.com

Abstract

Background: To circumvent regulatory barriers that limit medical data exchange due to personal information security concerns, we use homomorphic encryption (HE) technology, enabling computation on encrypted data and enhancing privacy.

Objective: This study explores whether using HE to integrate encrypted multi-institutional data enhances predictive power in research, focusing on the integration feasibility across institutions and determining the optimal size of hospital data sets for improved prediction models.

Methods: We used data from 341,007 individuals aged 18 years and older who underwent noncardiac surgeries across 3 medical institutions. The study focused on predicting in-hospital mortality within 30 days postoperatively, using secure logistic regression based on HE as the prediction model. We compared the predictive performance of this model using plaintext data from a single institution against a model using encrypted data from multiple institutions.

Results: The predictive model using encrypted data from all 3 institutions exhibited the best performance based on area under the receiver operating characteristic curve (0.941); the model combining Asan Medical Center (AMC) and Seoul National University Hospital (SNUH) data exhibited the best predictive performance based on area under the precision-recall curve (0.132). Both Ewha Womans University Medical Center and SNUH demonstrated improvement in predictive power for their own institutions upon their respective data's addition to the AMC data.

Conclusions: Prediction models using multi-institutional data sets processed with HE outperformed those using single-institution data sets, especially when our model adaptation approach was applied, which was further validated on a smaller host hospital with a limited data set.

(JMIR Med Inform 2024;12:e56893) doi:10.2196/56893

KEYWORDS

machine learning; privacy; in-hospital mortality; homomorphic encryption; multi-institutional system

Introduction

The demand for combining widespread data from various hospitals to create a larger data set for research from medical researchers is ongoing. However, exchanging or sharing medical data among hospitals is highly challenging because of various regulations and restrictions [1]. Sharing medical data with other institutions is limited owing to concerns over personal data

breach. In other words, most medical data are exclusively accessible to each institution, but data access is mutually exclusive, blocking access from other institutions. Owing to strict data protection policies and privacy regulations, various legal and regulatory barriers to transferring patient-level heterogeneous data among institutions exist. However, as predictive studies using large data have been actively conducted in precision medicine recently, demands to compile multi-institutional data and develop widely applicable models in more diverse clinical environments are increasing. Efforts have been invested to address these challenges using emerging privacy-enhancing technologies (PETs), including homomorphic encryption (HE)—a form of encryption that permits calculations directly on encrypted data, ensuring the security of both input and output of a numerical model [2-5]. It has been demonstrated to be effective in specific "privacy-preserving data sharing and analytics" contexts, for tasks such as delegated computation (wherein data are processed by a third party without revealing its content) or generating summary statistics without exposing individual raw data [6,7]. However, owing to HE's inherent computational constraints, several HE applications have primarily focused on computationally simpler tasks, such as computing summary statistics. Nevertheless, recent advancements in HE technology have evolved to a stage wherein the development or training of predictive models—particularly with large data sets in multi-institutional studies—has become achievable.

Recent advancements in privacy-preserving techniques in medical data analysis have significantly influenced the field, particularly through the use of HE. Several studies have explored the application of HE for privacy-preserving logistic regression and collaborative learning. For example, Kim et al [8] demonstrated the feasibility of training logistic regression models on homomorphically encrypted data, while Bos et al [9] applied HE to enable secure genome-wide association studies. Furthermore, the scalability of HE-based logistic regression has been demonstrated by Crawford et al [10], who successfully trained 30,000 models on encrypted data.

Our study distinguishes itself from previous works by applying HE to enable secure multi-institutional learning for postoperative mortality prediction using real-world clinical data. In addition, we propose a method called "model adaptation" strategy that allows smaller institutions to leverage encrypted data from larger institutions, improving their predictive models' performance without compromising patient privacy. By focusing on developing a predictive model through multi-institutional collaboration and emphasizing the practical applicability of our approach, our study pushes the boundaries of privacy-preserving medical data analysis and offers tangible solutions for enhancing predictive modeling in a secure, collaborative manner.

This study, aiming to verify the feasibility of securely developing a predictive model with multi-institutional data sets, focused on protecting each institution's data set using HE technology. Furthermore, we sought to determine whether the predictive performance can be improved by merging various multi-institutional data sets to project the 30-day postoperative mortality rate. Additionally, we showcased the application of our proposed "model adaptation" strategy. By supplementing and learning from a small amount of data based on an HE-encrypted large-scale data set from external institutions, institutions can construct a predictive model applicable within their clinical setting.

Methods

Ethical Considerations

This study was approved by the Institutional Review Board (IRB) of the Asan Medical Center (AMC) (IRB No. 2021-0186) and Ewha Womans University Medical Center (EUMC) (IRB No. 2020-11-017). The requirement of obtaining written informed consent was waived owing to the retrospective study design. We used the publicly available INSPIRE data set provided by the Seoul National University Hospital (SNUH). The composition, release, and usage of the INSPIRE data set were separately approved by the SNUH's IRB (H-2210-078-1368).

Inclusion and Exclusion Criteria

We retrospectively analyzed data collected from 341,007 patients aged 18 years and older who underwent noncardiac surgeries at 3 independent institutions. The data collection period for SNUH patients who underwent noncardiac surgeries was adjusted to January 2011 to December 2020, resulting in the inclusion of 46,956 patients. Moreover, we obtained data from 162,184 patients who underwent surgeries between January 2017 and April 2021 at the AMC. The apparent disparity in the number of patients between these institutions primarily stems from the mapping of our data set with the pre-existing public database, VitalDB. Additionally, our data set included 131,867 patients who underwent surgeries between January 2001 and December 2019 at the EUMC. Patients who had undergone cardiac procedures, organ transplantations, and neurosurgical operations, as well as those with an indeterminable final clinical outcome because of insufficient follow-up within the study timeframe, were excluded. Our analysis only incorporated the first surgical procedure post-admission for patients who had undergone several surgeries within the study period.

Data Collection and Variable Selection

Data encompassing patient demographics, preoperative laboratory evaluations, details of the surgery, and postoperative clinical outcomes were culled from the electronic medical record system of each respective hospital. Variables for the model were selected in the same manner as in the previous study [11]. Features that consistently ranked high across all hospitals were considered potential candidates for subsequent analyses. The study disregarded variables exhibiting substantial disparities among hospitals, a high incidence of missing values, and susceptibility to subjective measurement and execution by medical personnel. Through these processes, the following 19 variables that served as features for our investigation were selected: (1) demographic data (age, sex, and BMI); (2) preoperative laboratory results (white blood cells, hemoglobin, platelets, sodium, potassium, blood urea nitrogen, creatinine, albumin, aspartate transaminase, alanine transaminase, glucose, prothrombin time, and activated partial thromboplastin clotting time); (3) surgery type (general, otolaryngological, urological,

orthopedic, gynecological, and plastic); (4) anesthesia type (general, neuroaxial, monitored anesthesia care, and regional); and (5) status of emergency surgery.

During the modeling process with encrypted data, we strictly adhered to the principle of complete ignorance of the data's content. This approach, integral to our study design, is not merely a precaution; rather, it is essential for ensuring our analysis' objectivity and reliability. By consciously avoiding any knowledge of the data's nature, we aimed to minimize potential biases from prior data set familiarity, thus bolstering our findings' integrity and validity, particularly in encrypted data scenarios.

Data Preprocessing

All continuous variables underwent scaling using the StandardScaler function from the scikit-learn library, while categorical variables were incorporated into the model using one-hot encoding. We assumed that standardization for each feature had been implemented at every hospital before encryption, using their respective means and standard deviations.

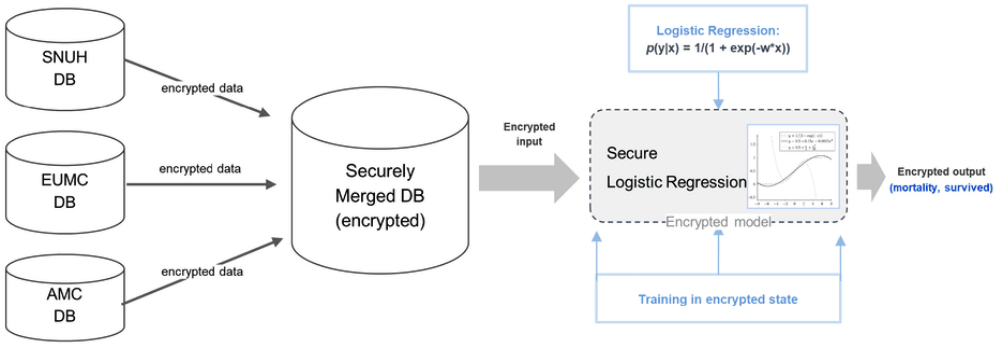
Model Outcomes

The primary outcome of interest in the study was in-hospital mortality within 30 days postoperatively. Data on in-hospital mortality were procured as binary information, derived from the final mortality date in the electronic medical record within 30 days postoperatively ("1" representing mortality and "0" indicating survival, with a threshold of 0.5).

Model Building

For modeling, secure logistic regression was used to compare the models' predictive performance (Figure 1) [8]. Complete data sets of each hospital were partitioned into training, validation, and testing data sets using a 6:2:2 distribution to develop all the models. The Nesterov Accelerated Gradient optimizer was applied to build trained models with a learning rate of 0.01 and batch size of 64 [12]. Binary cross-entropy served as the loss function for the model [13], with parameters being optimized to reduce each model's loss of function to the minimum. To address the imbalance in clinical data and more robustly assess the generalized performance of each model, we use the bootstrap sampling technique [14], which involved repeated processes of resampling training data, creating a new model, and evaluating that model. Thereafter, the model's performance was quantified as the mean performance of separate models developed with the bootstrap approach. Overfitting issues can be mitigated by averaging their results, thus increasing the model's generalizability. Consequently, the bootstrap method can significantly diminish the developed models' overfitting. To validate the predictive performance, the model was evaluated using the test set of the chosen hospital and data gathered from the amalgamations of other hospitals. For a fair comparison, we used the plaintext version of logistic regression for unencrypted data using NumPy from scratch [15], based on a stochastic approach, as opposed to using scikit-learn.

Figure 1. Schematic diagram of external validation of each hospital model. AMC: Asan Medical Center; EUMC: Ewha Womans University Medical Center; DB: database; SNUH: Seoul National University Hospital.



Model Adaptation

We proposed a methodology for "model adaptation," substantiated by the validation results for a host hospital. In the scenario, the host hospital was a small institution that may have not had an extensive data set of its own. The institution aimed to use an external data set, provided by a larger donor hospital, as a foundational training set. We assumed that the donor hospital provided its data set to the recipient hospital in an encrypted state. Moreover, we gradually increased the size of the host hospital's data set when training the predictive model to ascertain the volume at which the utilization of its data—when adapting the donated data set as a foundational training set—became effective. The approach was applied to the comprehensive AMC data set, which acted as a donor hospital, on postoperative 30-day mortality; that is, we initiated our study with all the AMC data set and progressively incorporated an increasing proportion of the EUMC and SNUH data sets. EUMC is assumed to be the AMC data's recipient. The AMC provides its own data in an encrypted state to augment the EUMC data set. Thereafter, the encrypted merged data set is used to train a logistic regression model, and inferences are made based on the EUMC's plaintext data. Over the course of this study, the EUMC data set's volume was incrementally increased by steps of 1000, 2000, 3000, and 4000, resulting in the sizes of 0, 1000, 3000, 6000, 10,000, 15,000, and 20,000. The adaptation process was applied with each increase in data size and integrated with the AMC data set, followed by training of the model and a thorough examination of the ensuing performance metrics—the receiver operating characteristic (AUROC) and area under the precision-recall curve (AUPRC). Moreover, we applied the same experimental process to the SNUH data set, using all the AMC data and gradually increasing the former's proportion.

Model Validation

Secure logistic regression with L2 regularization (called ridge regression) was developed using all the possible combinations of multicenter datasets with all input variables. The predictive performances were compared by assessing both the areas under the AUROC and AUPRC; the comparison was performed both numerically and statistically. Furthermore, AUROC and AUPRC were compared using the DeLong test [16].

Statistical Analysis and Modeling Tools

Continuous variables were expressed as mean (SD), while categorical variables were expressed as count and percentage. Continuous and categorical variables were compared among the 3 institutions using one-way ANOVA and chi-squared test, respectively. Variables with 2-tailed *P* values <.05 were considered to hold statistical significance.

We comparatively analyzed feature importance for each institution's data—as represented by the Shapley additive explanations (SHAP) values within logistic regression models—to investigate potential heterogeneity in data distributions across the 3 hospitals under consideration. Feature importance was evaluated using a logistic regression model and performed on plaintext data at each hospital without using HE. Statistical analyses were performed using Python 3.8.16 [17]. The DeLong test was performed using R 4.2.2 [18]. Secure logistic regression was conducted using scikit-learn 1.2.0 [19] and HEaaN.stat 0.2.0 [20].

Results

Study Population Characteristics

The average age of surgical patients at EUMC was the lowest, compared to the other 2 institutions, at 48.5 years (Table 1). Emergency surgeries occurred most frequently at the SNUH, with a rate of 7.4% (Table 1). Postoperative mortality within 30 days was relatively rare across all hospitals, with rates ranging from 0.2% to 0.4% (Table 1). Specifically, the rates were 0.3% at SNUH, 0.2% at AMC, and 0.2% at EUMC. The data characteristics of each hospital are shown in Table 1. When examining the SHA*P* values across all hospitals, we observed significant variation in data distributions, suggesting inherent biases within each hospital's data set, as presented in Figure S4 in Multimedia Appendix 1. Figure S5 in Multimedia Appendix 1 presents the relative odds ratios of the predictor variables affecting the outcome variable in the logistic regression models trained based on each hospital's data set. These odds ratios offer insights into each predictor variable's effect on the likelihood of the outcome and help interpret associations' magnitude and direction. Evidently, the distribution of the odds ratios of the predictor variables affecting the outcome variable differs among institutions. In the EUMC data set, only the age and general surgery department variables are statistically significant. In contrast, the significance of these and other variables varies across institutions, as demonstrated by the diverse distribution of odds ratios affecting the outcome variable depicted in Figure S2 in Multimedia Appendix 1.

Table 1. Data characteristics of the 3 medical institutions.

	SNUH[a] (n=46,956)	AMC[b] (n=162,184)	EUMC[c] (n=131,867)	*P* value
Demographic data				
Age (years), mean (SD)	55.9 (16.1)	54.2 (15.9)	48.5 (17.1)	<.001
Sex (female), n (%)	26,236 (55.9)	94,413 (58.2)	79,232 (60.1)	<.001
BMI (kg/m^2), mean (SD)	24.6 (3.9)	24.2 (3.7)	23.8 (3.8)	<.001
Preoperative laboratory results, mean (SD)				
White blood cell (10^3/μL)	6.6 (3.0)	6.7 (2.4)	7.5 (3.9)	<.001
Hemoglobin (g/dL)	13.1 (1.8)	12.8 (1.9)	13.1 (1.9)	<.001
Platelet (10^3/μL)	239.8 (73.5)	247.1 (72.7)	245.6 (72.0)	<.001
Sodium (mmol/L)	140.2 (2.7)	139.8 (2.4)	140.7 (3.0)	<.001
Potassium (mmol/L)	4.2 (0.4)	4.3 (0.3)	4.2 (0.4)	<.001
BUN[d] (mg/dL)	15.5 (8.1)	14.8 (6.8)	13.7 (6.9)	<.001
Creatinine (mg/dL)	1.0 (1.1)	0.9 (0.7)	0.9 (0.7)	<.001
Albumin 9g/dL)	4.2 (0.5)	3.8 (0.5)	4.1 (0.6)	<.001
GOT[e] (IU/L)	24.4 (36.7)	25.0 (33.7)	26.5 (95.0)	<.001
GPT[f] (IU/L)	23.4 (32.5)	22.7 (32.3)	25.1 (50.9)	<.001
Glucose (mg/dL)	110.8 (30.5)	113.3 (36.9)	198.3 (243.9)	<.001
PT[g] (INR[h])	1.0 (0.1)	1.0 (0.1)	1.0 (0.4)	<.001
aPTT[i] (s)	31.6 (4.6)	27.0 (3.3)	26.9 (5.4)	<.001
Type of surgery, n (%)				
General surgery	13,487 (28.7)	60,6130 (36.4)	40,611 (30.8)	<.001
Otolaryngologic surgery	4537 (9.7)	15,270 (10.8)	14,279 (10.8)	<.001
Urologic surgery	4738 (10.1)	20,551 (12.7)	9117 (6.9)	<.001
Orthopedic surgery	6736 (14.3)	24,480 (14.7)	23,486 (17.8)	<.001
Gynecological surgery	4956 (14.5)	31,691 (19.5)	26,509 (20.1)	<.001
Plastic surgery	1862 (4.0)	6213 (1.4)	9788 (7.4)	<.001
Type of anesthesia, n (%)				
General anesthesia	36,060 (76.8)	149,542 (92.2)	100,223 (76.0)	<.001
Neuroaxial anesthesia	5052 (16.5)	11,281 (7.0)	10,716 (8.1)	<.001
MAC[j]	5792 (12.3)	0 (0.0)	4985 (3.8)	<.001
Regional anesthesia	52 (0.1)	1361 (0.8)	509 (0.4)	<.001
Emergency surgery	3456 (7.4)	8876 (5.5)	4208 (3.2)	<.001
30-day mortality	156 (0.3)	306 (0.2)	316 (0.2)	<.001

[a]SNUH: Seoul National University Hospital.
[b]AMC: Asan Medical Center.
[c]EUMC: Ewha Womans University Medical Center.
[d]BUN: blood urea nitrogen.
[e]GOT: glutamate oxaloacetate transaminase.
[f]GPT: glutamate pyruvate transaminase.
[g]PT: prothrombin time.
[h]INR: international normalized ratio.
[i]aPTT: activated partial thromboplastin time.
[j]MAC: monitored anesthesia care.

Data Preprocessing: Missing Value Characteristics and Standardization

Herein, the average rates of missing values were 0.00% to 7.63% for various features (Table S4 in Multimedia Appendix 1). This discrepancy—particularly in EUMC data for BMI, type of anesthesia, and preoperative glucose—may reflect distinct characteristics inherent to the databases of each hospital (refer to Figure S2 in Multimedia Appendix 1). There was a substantial correlation (absolute correlation value of 0.7 or greater) between variables that were part of collective testing procedures, such as laboratory tests. Conversely, the correlation between other variables was relatively weak, with absolute correlation values below 0.7 (Figure S1 in Multimedia Appendix 1). Variables with a higher incidence of missing values, such as BMI and type of anesthesia at EUMC, did not exhibit significant correlations with the missing values in other variables (absolute correlation value <0.7). The analysis did not reveal any consistent pattern in the occurrence of missing values across the hospitals, implying a random nature of missing data for individual patients at each facility (Figure S2 in Multimedia Appendix 1). Considering this randomness and the low intervariable correlation of missing values, we opted to impute the missing data using the respective variables' median values.

Model Performance

Table 2 presents the validation results of the bootstrapping performance of the secure logistic regression model using various single- and multicenter combinations. Typically, the AMC and EUMC models that already had abundant data delivered superior performance when externally validated using data from other institutions. However, regarding the AMC data set's internal validation, the merged model using the entire data set demonstrated the highest performance, as indicated by AUROC of 0.941. Similarly, the AUPRC signified peak performance in the AMC data set's internal validation when the model merged with the AMC and SNU data sets, reaching 0.132. Figure S3 in Multimedia Appendix 1 provides a comparative visualization of AUROC and AUPRC. Table S3 in Multimedia Appendix 1 presents *P* values, indicative of statistical significance via the DeLong test, when comparing the predictive performance of the plaintext single-institution model and encrypted multi-institution model based on AUROC and AUPRC. Small *P* values signify a statistically significant difference in the 2 models' predictive performance.

Table 2. Validation results of single and merged secure logistic regression models for postoperative 30-day mortality on AMC, EUMC, and SNUH test data sets.

Train	Test		
	AMC[a] (n=32,437)	SNUH[b] (n=9392)	EUMC[c] (n=26,373)
Single (plaintext)			
AMC, mean AUROC[d] (95% CI)	0.939 (0.927-0.950)	0.915 (0.902-0.928)	0.890 (0.867-0.912)
SNUH, mean AUROC (95% CI)	0.925 (0.913-0.936)	*0.942 (0.9300.953)* [e]	0.937 (0.926-0.947)
EUMC, mean AUROC (95% CI)	0.880 (0.853-0.906)	0.906 (0.890-0.921)	0.952 (0.943-0.961)
Merged (ciphertext)			
AMC + EUMC, mean AUROC (95% CI)	0.931 (0.914-0.947)	0.919 (0.907-0.931)	0.952 (0.942-0.962)
AMC + SNUH, mean AUROC (95% CI)	0.940 (0.925-0.955)	0.927 (0.902-0.952)	0.934 (0.920-0.947)
SNUH + EUMC, mean AUROC (95% CI)	0.931 (0.916-0.945)	0.925 (0.903-0.946)	0.956 (0.950-0.962)
AMC + SNUH + EUMC, mean AUROC (95% CI)	*0.941 (0.927-0.955)*	0.929 (0.905-0.953)	*0.957 (0.951-0.963)*
Single (plaintext)			
AMC, mean AUPRC[f] (95% CI)	0.125 (0.088-0.161)	0.089 (0.071-0.107)	0.072 (0.051-0.093)
SNUH, mean AUPRC (95% CI)	0.070 (0.044-0.095)	0.123 (0.099-0.146)	0.060 (0.075-0.072)
EUMC, mean AUPRC (95% CI)	0.087 (0.055-0.118)	0.085 (0.066-0.104)	*0.120 (0.090-0.150)*
Merged (ciphertext)			
AMC + EUMC, mean AUPRC (95% CI)	0.107 (0.078-0.136)	0.089 (0.074-0.104)	0.107 (0.081-0.133)
AMC + SNUH, mean AUPRC (95% CI)	*0.132 (0.094-0.169)*	*0.171 (0.112-0.230)*	0.081 (0.060-0.102)
SNUH + EUMC, mean AUPRC (95% CI)	0.098 (0.069-0.126)	0.098 (0.069-0.126)	0.116 (0.089-0.143)
AMC + SNUH + EUMC, mean AUPRC (95% CI)	0.113 (0.082-0.144)	0.113 (0.082-0.144)	0.114 (0.089-0.139)

[a]AMC: Asan Medical Center.

[b]SNUH: Seoul National University Hospital.

[c]EUMC: Ewha Womans University Medical Center.

[d]AUROC: area under the receiver operating characteristic curve.

[e]Italics indicate significant results.

[f]AUPRC: area under the precision-recall curve.

For unencrypted data, we used a plaintext version of the logistic regression model, developed from scratch using NumPy and featuring an architecture identical to that of the HE-based logistic regression model. Further, we evaluated the discrepancies between the results computed in ciphertext and subsequently decrypted, compared to those calculated directly in plaintext. The mean absolute difference was 2.02×10^{-5}, indicating a marginal difference. The minimum absolute difference was remarkably low at 6.56×10^{-10}, while the maximum absolute difference reached 7.71×10^{-4}. This observation suggests that the outcomes—irrespective of whether they are computed in plaintext or ciphertext—demonstrate an almost indistinguishable difference.

Model Adaptation Results

We investigated the scenario of model adaptation wherein we gradually incorporated the data set from another institute. Using the EUMC data set, starting with an AUROC of 0.930, there was an initial temporary decline to 0.906 when the first 1000 records from the EUMC data set were incorporated into the complete AMC data set (Figure 2, Table S2A in Multimedia Appendix 1). As more EUMC records were progressively added, a significant improvement in AUROC was observed, eventually reaching 0.954 (Figure 2, Table S2A in Multimedia Appendix 1). Similarly, the AUPRC initially decreased from 0.09 to 0.075 with the addition of the initial 1000 EUMC data to the total AMC data (Figure 2, Table S2A in Multimedia Appendix 1). However, as we continued introducing more EUMC data, the AUPRC began increasing (Figure 2, Table S2A in Multimedia Appendix 1). Upon the inclusion of 30,000 EUMC records, the AUPRC ascended to 0.11 (Figure 2, Table S2A in Multimedia Appendix 1). Using the SNUH data set, we began with an AUROC of 0.916. The increase was less pronounced than that observed with the EUMC data set (Figure 3, Table S2B in Multimedia Appendix 1). However, as we progressively included segments of the SNUH data set, the AUROC exhibited a moderate trend of progressive improvement, eventually reaching 0.926 (Figure 3, Table S2B in Multimedia Appendix 1). A decrease in the AUPRC from 0.151 to 0.131 was observed

when the first 1000 SNUH data were added to the AMC data set (Figure 3, Table S2B in Multimedia Appendix 1). As more SNUH data were added, the AUPRC gradually increased, improving to approximately 0.149, compared to the AMC single-institution model's performance, once 30,000 data points were included (Figure 3, Table S2B in Multimedia Appendix 1).

Figure 2. Validation results of bootstrap samples using increased EUMC data size with the AMC whole data set for postoperative 30-day mortality. (a) Boxplot analysis of AUROC using bootstrap samples; (b) boxplot analysis of AUPRC using bootstrap samples. AMC: Asan Medical Center; AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; EUMC: Ewha Womans University Medical Center; SNUH: Seoul National University Hospital.

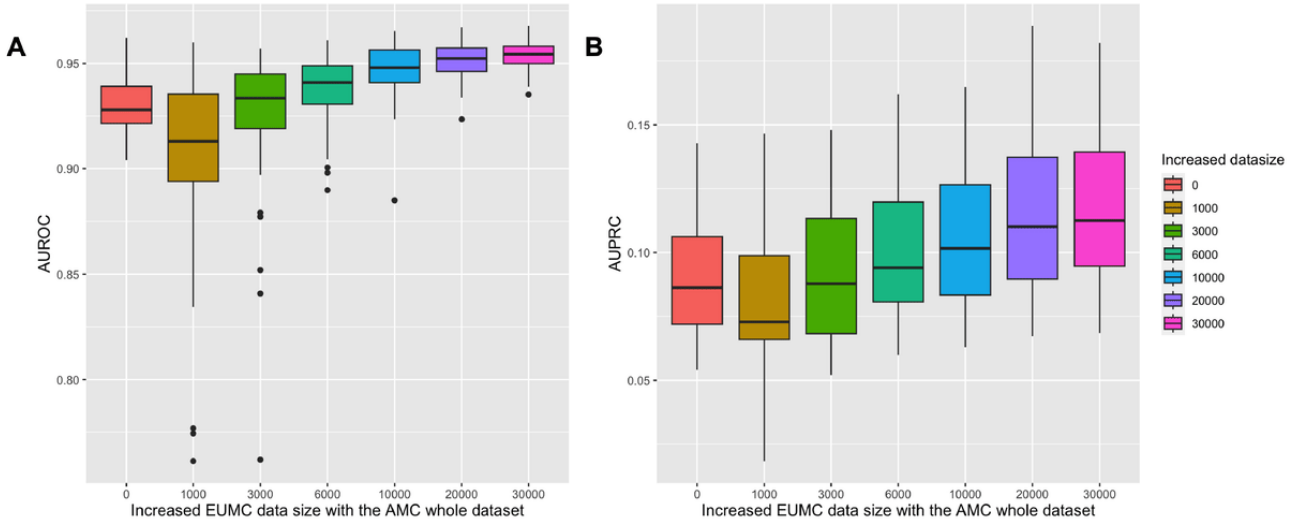
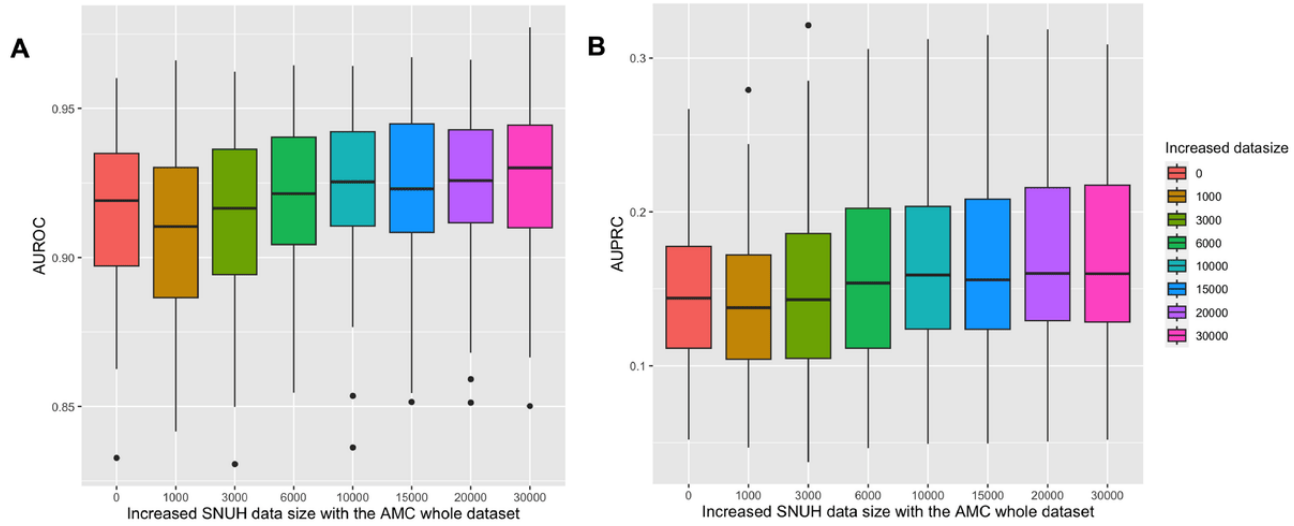


Figure 3. Validation results of bootstrap samples using increased SNUH data size with the AMC whole data set for postoperative 30-day mortality. (a) Boxplot analysis of AUROC; (b) boxplot analysis of AUPRC. AMC: Asan Medical Center; AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; EUMC: Ewha Womans University Medical Center; SNUH: Seoul National University Hospital.



Discussion

Overview of Multi-Institutional Model Performance and Implications

This study demonstrated the potential of overcoming limitations associated with single-institution models, such as reduced external predictive power and data overfitting, through secure multi-institutional data integration using HE technology. Our approach effectively adapts predictive models to specific hospital environments, indicating a substantial improvement in model performance across different data sets. The results suggest that small- and medium-sized hospitals with limited data can enhance the predictive performance of their AI models by adopting data from larger hospitals and conducting additional combined learning using HE technology. The significance of this study lies in its practical application and validation of HE technology using real-world, multi-institutional clinical data, laying the groundwork for its potential applicability to various forms of multi-institutional clinical data in future research.

Advantages and Challenges of Multicenter Studies

The transition from single-center to multicenter studies generates large data sets ("big data"), enhancing the robustness and generalizability of research findings. These larger and more diverse data sets increase the accuracy and applicability of the results. However, multicenter studies introduce challenges such as secure and legal data sharing, inherent incompatibility

between data security and research efficacy, and potential biases from selective participant inclusion [21]. To address these issues, researchers are exploring innovative PETs like HE, federated learning, and multiparty computation, which enable secure data analyses while preserving patient confidentiality.

Federated Learning and Comparison to HE

Federated learning has been proposed as a security solution in multi-institutional data environments, as it only shares each local model's weights or parameters. The strength of federated learning—a more decentralized approach than ours—is that no patient-level data are transferred to third parties with or without encryption. However, even with aggregate-level data, such as weights of a model, patient-level information can potentially be inferred [22-25].

In this multicenter study, we used cutting-edge HE to protect personal information leakage and data security. Additionally, HE enables operations and predictive modeling on encrypted data, providing an ultimate solution that can completely resolve issues related to personal information leakage and data security. Furthermore, HE provides the maximum (strongest) security when used appropriately, such as in outsourced computation, wherein HE secures data breaches in computation. In medical fields, HE has been applied to numerous cases for fulfilling privacy requirements [26]. Previous computational inefficiency of HE may have limited its application in computation-intensive steps, such as in developing a prediction model; however, recent advancements have led the technology to be used in practice. The present multicenter study demonstrated that a prediction model can be developed completely without a data breach risk in training using HE.

Limitations and Future Directions

While HE technology allowed secure data integration, it introduced several challenges. Notably, encrypting data led to a marked increase in data size compared to plaintext, intensifying data storage requirements. Additionally, the encrypted model necessitated longer training time. Furthermore, in multi-institutional contexts, such as health care data sharing, key management in multiparty HE becomes a complex, practical challenge. These factors—expanded data size, prolonged training periods, and intricacies of key management—are essential considerations in the effective deployment and ongoing development of secure logistic regression models within encrypted data frameworks.

The study also highlighted limitations in predictive performance when models trained on diverse data sets were applied to individual hospitals. In some data sets, the merged data model's predictive performance fell short of the single-institution data model. This discrepancy indicates a complex interplay between data heterogeneity and model performance, suggesting that predictive performance may not always be enhanced through data augmentation alone, as evidenced in this study. A prediction model may lose prediction power in some institutions when trained using data from institutions with disparate data distributions. Consequently, when implementing a trained model on individual hospital data, we occasionally observed a deviation from our initial expectation that a model trained on the merged set would invariably outperform others.

Another limitation of the study was the reliance on retrospectively collected data, featuring varying extraction periods across institutions. The effects of temporal changes, including advancements in medical technology, were not fully adjusted for, potentially reducing the results' discernibility.

To address these limitations, future research should explore methods for data integration that adjust for heterogeneity. This can be achieved by prospectively collecting data from multiple institutions and conducting comparative studies on predictive performance using HE technology. Such methodologies would help to mitigate the impact of varying data extraction periods and temporal changes in medical technology. Additionally, the use of advanced statistical methods to better account for data heterogeneity might be explored as a promising avenue for further research. These studies would undoubtedly offer invaluable insights into potential strategies for enhancing predictive performance in multi-institutional settings while preserving data privacy and security.

Conclusions

In conclusion, this study demonstrated the practicality of using HE technology to combine data from real-world multi-institutional sources and develop predictive models for in-hospital mortality within 30 days postoperatively. Additionally, we showcased the implementation of privacy-preserving artificial intelligence prediction models. The findings highlight the potential for both practical applications and protection of personal information in the realm of predictive modeling. HE technology should be applied to diverse forms of multi-institutional clinical data in future endeavors to replicate, validate, and extend this study's findings.

Acknowledgments

The work was supported by the National IT Industry Promotion Agency (NIPA) grant funded by the Korean government (MSIT; A0122-23-1194, development of adoption and application technologies of AI solutions specialized in sensitive medical data protection). The research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant HR20C0026). Moreover, the study was supported by a grant (2022IF0020, 2023IP0132) from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea. We would like to thank everyone who participated in the study. Specifically, we would like to express our gratitude to Professor Hyung Chul Lee from the Department of Anaesthesia and Pain Medicine at Seoul National University Hospital (SNUH). His assistance in granting us access to the INSPIRE open data set from SNUH for our research was invaluable.

Data Availability

The data set used in the study is not publicly available. However, it can be provided upon reasonable request to the corresponding authors. For the SNUH data set, we used the deidentified open data set, INSPIRE, which was compiled and made publicly available by SNUH. The construction and release of the INSPIRE data set were separately approved by the institutional review board of SNUH (H-2210-078-1368). The codes that support the findings of the study are available for download on GitHub [27].

Authors' Contributions

J Suh contributed to the design and execution of the study, the analysis and interpretation of the data, and funding acquisition. GL contributed to the analysis and interpretation of the data and the writing of the manuscript. JWK contributed to the analysis and interpretation of the data. J Shin contributed to the analysis and interpretation of the data. YJK contributed to the acquisition of the data. SWL contributed to the design and execution of the study; the acquisition, analysis, and interpretation of the data; the writing of the manuscript; and funding acquisition. SK contributed to the design and execution of the study, the analysis and interpretation of the data, the writing of the manuscript, and funding acquisition. All authors contributed to the review and editing of the manuscript and confirmed the final version of the submitted manuscript.

SWL (sangwooklee20@gmail.com) and SK (sulgik@cryptolab.co.kr) contributed equally as co-corresponding authors of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[DOCX File , 1952 KB - medinform_v12i1e56893_app1.docx]

References

1. EU Commission. The EU General Data Protection Regulation (GDPR). 2018 May 25. URL: https://gdpr.eu/ [accessed 2024-06-20]
2. Brakerski Z. Fully homomorphic encryption without modulus switching from classical GapSVP. Advances in Cryptology – CRYPTO 2012 2012;7417:868-886. [doi: 10.1056/NEJM196803212781204]
3. Cheon JH, Kim A, Kim M, Song Y. Homomorphic encryption for arithmetic of approximate numbers. In: Takagi T, Peyrin T, editors. Advances in Cryptology – ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I. Cham: Springer; 2017:409-437.
4. Gentry C, Sahai A, Waters B. Homomorphic encryption from learning with errors: conceptually-simpler, asymptotically-faster, attribute-based. In: Canetti R, Garay JA, editors. Advances in Cryptology – CRYPTO 2013: 33rd Annual Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2013. Proceedings, Part I. Berlin, Heidelberg: Springer; 2013:75-92.
5. Weed LL. Medical records that guide and teach. N Engl J Med 1968 Mar 21;278(12):652-7 concl. [doi: 10.1056/NEJM196803212781204] [Medline: 5637250]
6. Munjal K, Bhatia R. A systematic review of homomorphic encryption and its contributions in healthcare industry. Complex Intell Systems 2022:1-28 [FREE Full text] [doi: 10.1007/s40747-022-00756-z] [Medline: 35531323]
7. Son Y, Han K, Lee YS, Yu J, Im Y, Shin S. Privacy-preserving breast cancer recurrence prediction based on homomorphic encryption and secure two party computation. PLoS One 2021;16(12):e0260681 [FREE Full text] [doi: 10.1371/journal.pone.0260681] [Medline: 34928973]
8. Kim M, Song Y, Wang S, Xia Y, Jiang X. Secure logistic regression based on homomorphic encryption: design and evaluation. JMIR Med Inform 2018 Apr;6(2):e19 [FREE Full text] [doi: 10.2196/medinform.8805] [Medline: 29666041]
9. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform 2014;50:234-243 [FREE Full text] [doi: 10.1016/j.jbi.2014.04.003] [Medline: 24835616]
10. Crawford JLH, Gentry C, Halevi S, Platt D, Shoup V. Doing real work with FHE: the case of logistic regression. In: Proceedings of the 6th Workshop on Encrypted Computing & Applied Homomorphic Cryptography. 2018 Presented at: WAHC '18; October 15-19, 2018; Toronto, ON.
11. Lee SW, Lee H, Suh J, Lee KH, Lee H, Seo S, et al. Multi-center validation of machine learning model for preoperative prediction of postoperative mortality. NPJ Digit Med 2022 Jul 12;5(1):91. [doi: 10.1038/s41746-022-00625-6] [Medline: 35821515]
12. Botev A, Lever G, Barber D. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. 2017 Presented at: 2017 International Joint Conference on Neural Networks; May 14-19, 2017; Anchorage, AK p. 1899-1903. [doi: 10.1109/ijcnn.2017.7966082]

13. Dong Y, Li J, Wang Z, Jia W. CoDC: accurate learning with noisy labels via disagreement and consistency. Biomimetics (Basel) 2024 Feb 03;9(2):92 [FREE Full text] [doi: 10.3390/biomimetics9020092] [Medline: 38392138]
14. Hesterberg TC. What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. Am Stat 2015;69(4):371-386 [FREE Full text] [doi: 10.1080/00031305.2015.1089789] [Medline: 27019512]
15. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357-362. [doi: 10.1038/s41586-020-2649-2]
16. Sun X, Xu W. Fast implementation of De Long's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett 2014 Nov;21(11):1389-1393. [doi: 10.1109/lsp.2014.2337313]
17. Guido VR, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; Mar 20, 2009.
18. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Computational Graphical Statistics 1996 Sep;5(3):299-314. [doi: 10.1080/10618600.1996.10474713]
19. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014;8:14 [FREE Full text] [doi: 10.3389/fninf.2014.00014] [Medline: 24600388]
20. Lee Y, Seo J, Nam Y, Chae J, Cheon JH, Lee Y. HEaaN-STAT: a privacy-preserving statistical analysis toolkit for large-scale numerical, ordinal, and categorical data. IEEE Trans Dependable Secure Comput 2024 May;21(3):1224-1241. [doi: 10.1109/tdsc.2023.3275649]
21. Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. BMJ 2009 Mar 12;338:b866 [FREE Full text] [doi: 10.1136/bmj.b866] [Medline: 19282440]
22. Carlini N, Liu C, Erlingsson U, Kos J, Song D. The secret sharer evaluating and testing unintended memorization in neural networks. In: Proceedings of the 28th USENIX Conference on Security Symposium. 2019 Presented at: SEC'19; August 14-16, 2019; Santa Clara, CA p. 267-284.
23. Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017 Presented at: CCS '17; October 30-November 3, 2017; Dallas, TX p. 603-618. [doi: 10.1145/3133956.3134012]
24. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning passive and active white-box inference attacks against centralized and federated learning. 2019 Presented at: 2019 IEEE Symposium on Security and Privacy; May 19-23, 2019; San Francisco, CA p. 739-753. [doi: 10.1109/sp.2019.00065]
25. Melis L, Song C, De CE, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. 2019 Presented at: 2019 IEEE Symposium on Security and Privacy; May 19-23, 2019; San Francisco, CA p. 691-706. [doi: 10.1109/sp.2019.00029]
26. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform 2014;50:234-243 [FREE Full text] [doi: 10.1016/j.jbi.2014.04.003] [Medline: 24835616]
27. Sang-Wook Lee, Jungyo Suh, Garam Lee, Jung Woo Kim, Junbum Shin, Sulgi Kim. Privacy-Preserving Prediction of Postoperative Mortality in Multi-Institutional Data: Development and Usability Study. 2024. URL: https://github.com/CryptoLabInc/secure_LR [accessed 2024-06-20]

Abbreviations

AMC: Asan Medical Center
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
DB: database
EUMC: Ewha Womans University Medical Center
HE: homomorphic encryption
PETs: privacy-enhancing technologies
SHAP: shapley additive explanations
SNUH: Seoul National University Hospital

Edited by C Lovis; submitted 29.01.24; peer-reviewed by A Qaisar Al Badawi, Y Yan; comments to author 27.04.24; revised version received 07.05.24; accepted 08.06.24; published 05.07.24.

Please cite as:
Suh J, Lee G, Kim JW, Shin J, Kim YJ, Lee SW, Kim S
Privacy-Preserving Prediction of Postoperative Mortality in Multi-Institutional Data: Development and Usability Study
JMIR Med Inform 2024;12:e56893
URL: https://medinform.jmir.org/2024/1/e56893
doi:10.2196/56893
PMID:

Evaluation of AI-Driven LabTest Checker for Diagnostic Accuracy and Safety: Prospective Cohort Study

Dawid Szumilas[1], MD; Anna Ochmann[1], MD; Katarzyna Zięba[1], MD; Bartłomiej Bartoszewicz[2], PhD; Anna Kubrak[2], MD; Sebastian Makuch[3], PhD; Siddarth Agrawal[2], PhD, MD; Grzegorz Mazur[2,4], Prof Dr, MD; Jerzy Chudek[1], Prof Dr

1

2

3

4

Corresponding Author:
Jerzy Chudek, Prof Dr

Abstract

Background: In recent years, the implementation of artificial intelligence (AI) in health care is progressively transforming medical fields, with the use of clinical decision support systems (CDSSs) as a notable application. Laboratory tests are vital for accurate diagnoses, but their increasing reliance presents challenges. The need for effective strategies for managing laboratory test interpretation is evident from the millions of monthly searches on test results' significance. As the potential role of CDSSs in laboratory diagnostics gains significance, however, more research is needed to explore this area.

Objective: The primary objective of our study was to assess the accuracy and safety of LabTest Checker (LTC), a CDSS designed to support medical diagnoses by analyzing both laboratory test results and patients' medical histories.

Methods: This cohort study embraced a prospective data collection approach. A total of 101 patients aged ≥18 years, in stable condition, and requiring comprehensive diagnosis were enrolled. A panel of blood laboratory tests was conducted for each participant. Participants used LTC for test result interpretation. The accuracy and safety of the tool were assessed by comparing AI-generated suggestions to experienced doctor (consultant) recommendations, which are considered the gold standard.

Results: The system achieved a 74.3% accuracy and 100% sensitivity for emergency safety and 92.3% sensitivity for urgent cases. It potentially reduced unnecessary medical visits by 41.6% (42/101) and achieved an 82.9% accuracy in identifying underlying pathologies.

Conclusions: This study underscores the transformative potential of AI-based CDSSs in laboratory diagnostics, contributing to enhanced patient care, efficient health care systems, and improved medical outcomes. LTC's performance evaluation highlights the advancements in AI's role in laboratory medicine.

Trial Registration: ClinicalTrials.gov NCT05813938; https://clinicaltrials.gov/study/NCT05813938

(JMIR Med Inform 2024;12:e57162) doi:10.2196/57162

KEYWORDS

LabTest Checker; CDSS; symptom checker; laboratory testing; AI; assessment; accuracy; artificial intelligence; health care; medical fields; clinical decision support systems; application; applications; diagnoses; patients; patient; medical history; tool; tools

Introduction

In recent times, the implementation of artificial intelligence (AI) within diverse medical domains has garnered significant attention and practical application [1]. AI-driven technology has sparked a transformative wave in health care, introducing inventive solutions to enhance patient care, diagnosis, and decision-making processes [2]. A notable instance of AI's application is evident in the emergence of clinical decision support systems (CDSSs), direct tools designed to streamline health care decision-making [3].

Laboratory tests are essential in modern health care, providing valuable insight into a patient's health status and improving the accuracy of diagnosing medical conditions. The interpretation of laboratory test results is a complex process requiring medical expertise and knowledge. However, the mounting reliance on laboratory testing poses a formidable challenge for health care systems, particularly in regions where tests are often administered without direct medical oversight, as seen in Poland [4].

The significance of this challenge is highlighted by the substantial volume of inquiries related to laboratory test result interpretation, which include identifying potential causes or implications of certain findings and seeking guidance on the next steps or actions based on the test results. Data indicate that, in Poland alone, there are approximately 7 million monthly searches concerning the significance of laboratory test results. On a larger scale, within the European Union, this number escalates to around 82 million monthly searches based on data from SENUTO and Google AdWords [5]. These cases emphasize the need for effective strategies to manage laboratory test interpretation in modern health care settings.

Given the widespread use of laboratory diagnostics, which includes a wide range of tests that analyze blood, urine, tissues, and other bodily fluids to diagnose and monitor diseases, assess organ function, and guide treatment decisions, there is a growing interest in exploring the potential of CDSSs within this realm. The inherent complexities tied to test result interpretation underscore this interest. While the efficacy and safety of CDSSs have been demonstrated in various medical contexts, such as symptom assessment tools [6,7], the integration of CDSSs into laboratory diagnostics remains underexplored.

Several studies have assessed the effectiveness and safety of AI-driven symptom checkers, tools designed to aid patients in self-diagnosing symptoms and making informed health care choices [8-10]. These tools use algorithms and databases to generate potential diagnoses based on user inputs.

A notable study conducted by Semigran et al [11] scrutinized the diagnostic precision of 23 distinct symptom checkers, comparing their outcomes against physician diagnoses. The investigation disclosed that symptom checkers achieved accurate diagnoses in 34% of instances, while physicians achieved 58% accuracy. Despite relatively lower accuracy, the study underscored the potential of symptom checkers in offering reasonable differential diagnoses and supporting patient decision-making.

A more recent study by Hennemann et al [12] evaluated the performance of an app-based symptom checker within the realm of mental disorders. Results revealed that the studied symptom checker demonstrated moderate-to-good accuracy in suggesting conditions for mental disorders concerning formal diagnosis, albeit with variations across disorder categories and interrater reliability. The symptom checker's primary condition suggestion corresponded with interview-based diagnoses in 51% (25/49) of cases, with at least 1 of the initial 5 condition suggestions aligning in 69% (34/49) of cases across the patient cohort. Accuracy fluctuated across disorder categories, ranging from 82% precision for somatoform and related disorders, 65% for affective disorders, to 53% for anxiety disorders. The study concluded that symptom checkers hold promise as supplementary screening tools in the diagnostic process. Still, their diagnostic efficacy requires assessment in more extensive samples and comparison with alternative diagnostic methods.

This paper addresses the status of AI-based technologies in health care, specifically focusing on implementing CDSSs in direct-to-patient tools. After emphasizing the importance of laboratory diagnostics in contemporary health care and the challenges tied to test result interpretation, we examine the existing but limited literature concerning CDSSs' role in laboratory diagnostics, underscoring the need for further research and advancement in this domain. The objective of this study is to evaluate the performance of a novel CDSS named LabTest Checker (LTC) in a cohort of adult patients requiring laboratory testing. The main question it aims to answer pertains to the accuracy and safety of LTC.

Methods

Description of LTC Technology

LTC is an intricate medical software designed to provide assistance in the preliminary medical diagnosis process through the analysis of laboratory test results and comprehensive medical history. By leveraging advanced white-box machine learning algorithms and data analytics, LTC empowers patients and health care practitioners to derive insightful conclusions and make informed decisions. The AI models were trained on a comprehensive dataset encompassing clinical data from electronic health records, public repositories, documented case studies, and expert medical knowledge. LTC seamlessly integrates with existing electronic health record systems, automatically importing patients' latest laboratory results. This triggers a dynamic medical questionnaire presented on a user-friendly tablet interface, typically completed within 90 - 120 seconds, which delves into the patient's medical history, symptoms, and pertinent risk factors. Through this methodical scrutiny and correlation of pivotal data, LTC effectively evaluates an individual's health status and detects potential medical issues, empowering patients and health care providers to establish more accurate diagnoses and improve patient care and outcomes.

Study Setting and Population

The study was conducted at the Emergency Department (ED) of Andrzej Mielecki Public Clinical Hospital in Katowice, Poland, with 8 specialized departments and a total bed capacity of 351. This cohort study embraced a prospective data collection approach. A total of 101 self-referred patients aged ≥18 years, in stable condition, but requiring comprehensive diagnosis were enrolled between December 22, 2022, and March 31, 2023. Comprehensive diagnosis refers to cases where diagnosis based solely on subjective evaluation and physical examination is unattainable, necessitating in-depth assessment through laboratory tests. Inclusion criteria encompassed (1) age ≥18 years and (2) requirement of in-depth laboratory test investigation. The only exclusion criteria was pregnancy. Trained research staff identified and invited eligible patients to participate following an initial medical evaluation to assess eligibility based on the predefined inclusion and exclusion criteria (further details are available on ClinicalTrials.gov [13]). The study achieved a high response rate of 84.9% (101/119), indicating strong participant willingness.

Study Design

This prospective cohort study involved 101 patients, all requiring comprehensive diagnosis beyond subjective evaluation and physical examination. A panel of blood laboratory tests

exceeding the routine diagnostic work-up at the ED, including a lipid profile, erythrocyte sedimentation rate, high-sensitivity C-reactive protein, creatinine, urea, iron, liver enzymes (alanine transaminase, aspartate transferase, and gamma-glutamyl transferase), sodium, potassium, glucose, uric acid, thyroid-stimulating hormone, and complete blood count, was conducted for each participant. LTC was used to interpret these results, and its performance was compared to that of an internal medicine specialist (JC), who reviewed the urgency categorizations assigned by ED physicians and the AI-generated suggestions without prior knowledge of the LTC results. It is important to clarify that JC did not directly assess patients in the ED. Instead, attending physicians in the ED assigned the initial urgency category for each patient based on their clinical judgment. JC then reviewed these urgency classifications assigned by the ED physicians, alongside the assessments generated by the model under study. This 2-pronged approach aimed to ensure the accuracy of the urgency categorizations and provide an additional layer of validation.

Patients presenting at the ED underwent laboratory tests and provided health-related information under a doctor's supervision. This encompassed biometric details, medical history, medications, substances used, family history, symptoms, and prior test results. Based on these data and test outcomes, AI algorithms suggested underlying pathology and diagnostic-therapeutic guidance.

Accuracy and safety were assessed by comparing AI-generated suggestions to experienced doctor (consultant) recommendations, which are considered the gold standard. The consultant, blinded to the LTC results, categorized the urgency of physician interaction for each test (emergency, urgency, routine, and self-care; Table 1). Following this assessment, the LTC results were disclosed to the consultant to evaluate if adhering to LTC recommendations could avoid needless medical visits and whether LTC correctly identified the underlying causes of any abnormal results.

Table . Diagnostic and therapeutic recommendations generated by LabTest Checker (LTC) and specialist recommendations were categorized to assess LTC's precision. Sensitivity for the emergency category was computed as the ratio of LTC's correct emergency identifications to the physician's emergency identifications: *A / (A + B + C + D)*. Similarly, sensitivity for the urgency category was calculated as *F / (E + F + G + H)*. Triage accuracy was calculated as *(A + F + L + R) / total number of patients in the study*. Triage safety as calculated as *(A + E + F + J + K + L + N + O + P + R) / total number of patients in the study*.

Urgency category of contact with doctor, assigned by the consultant	Urgency category of contact with doctor, assigned by LTC			
	Emergency	Urgency	Routine	Self-care
Emergency	A	B	C	D
Urgency	E	F	G	H
Routine	J	K	L	M
Self-care	N	O	P	R

Owing to the technology's design, certain variables were excluded from determining pathology identification accuracy: (1) interpretations labeled as urgent or requiring immediate contact with a doctor were omitted to ensure patient safety and prioritize triage in emergencies, and (2) interpretations categorized as "end of diagnostic - no need for doctor contact" were omitted when results were valid or deviations were insignificant and did not signify pathology.

Ethical Considerations

The study protocol was registered on ClinicalTrials.gov (NCT05813938), and ethical approval was granted by the Bioethics Committee of the Medical University of Silesia (approval code: PCN/CBN/0052/KB1/115/I/22; approval date: November 8, 2022). All patients provided written informed consent before undergoing screening for study eligibility. To ensure privacy and confidentiality, all data collected during the study were anonymized and deidentified before analysis. Participants received no compensation for their involvement in the study. The study involved noninvasive procedures, and the primary intervention was the use of LTC to interpret laboratory test results. Participants were informed that they could seek clarification or assistance from medical professionals while using LTC.

Statistics

A power analysis was performed to determine the statistical power of this study, considering a total sample size of 101 participants in a single group of patients, which was predetermined in the study design. The power analysis was conducted using the G*Power software (version 3.1.9.7; Heinrich-Heine-Universität Düsseldorf). The power analysis was based on a 1-tailed test with an α level of .05. The effect size was calculated at 0.36. Using these parameters and the total sample size of 101, the power analysis indicated that the study would have moderate statistical power to detect a significant effect size within a single group of patients. The estimated power achieved with the given sample size was 0.82, indicating that the study had a reasonable likelihood of detecting meaningful differences within the group.

Outcome measures were prespecified and calculated with 95% CIs. The Wilson score method was used to produce CIs for sensitivity to emergency, sensitivity to urgency, accuracy of triage, safety of triage, and reduction of unnecessary visits. Calculations were performed using the statistical software package *Statistica* (version 13.0 PL; TIBCO Software Inc). Analytic data are presented as point estimates and 95% CIs, with a *P* value <.05 being considered significant.

Results

In the context of this study, the triage accuracy in the 101-patient cohort was 74.3%, with a safety sensitivity of 100% for identifying emergency cases and a sensitivity of 92.3% for detecting urgent cases. The implementation of the system led to a noteworthy 41.6% (42/101) reduction in unnecessary medical visits, and its accuracy in identifying the underlying pathology was 82.9%.

The system classified patients based on urgency: 9 patients required immediate contact; 41 needed urgent contact; 50 warranted routine contact; and 1 did not necessitate doctor contact, falling into the self-care category. Analysis by the consultant revealed disparities in urgency category assignments for 26 patients. Notably, the technology overestimated urgency for 25 patients, including cases where the consultant recommended urgent contact, but the technology indicated immediate or scheduled contact. However, the technology inaccurately assessed the urgency for 1 patient, failing to align with the specialist's urgent contact suggestion, instead proposing scheduled contact. These findings collectively underscore the triage system's effective urgency categorization while also pinpointing areas for enhancement to improve precision, diminish disparities, and prevent false negatives. These findings are detailed in Table 2.

Table . Classification outcomes of diagnostic-therapeutic recommendations proposed by LabTest Checker (LTC) and those provided by the consultant.

Urgency category of contact with doctor, assigned by the consultant	Urgency category of contact with doctor, assigned by LTC			
	Emergency, n	Urgency, n	Routine, n	Self-care, n
Emergency	7	0	0	0
Urgency	1	24	1	0
Routine	1	15	43	0
Self-care	0	2	6	1

Discussion

Principal Findings

The promising results obtained from the evaluation of LTC show the potential of AI-driven tools in assisting patients and medical professionals in navigating the complexities of laboratory test result interpretation. An accuracy rate of 74.3% demonstrates LTC's capability to furnish dependable medical recommendations grounded in blood test results, a development that holds promise for enhancing operational efficiency in the medical domain. Particularly noteworthy is LTC's impressive safety sensitivity of 100% for identifying emergency cases and a high sensitivity of 92.3% for detecting urgent cases. These results imply the system's adeptness in identifying critical scenarios, aligning with its intended role of providing secure and precise medical counsel.

Comparison to Prior Work and Broader Implications

The clinical implications of our findings extend beyond acute care settings. LTC has the potential to revolutionize health care delivery across various domains, including outpatient clinics, preventive care, and direct-to-consumer health management.

In outpatient settings, LTC could streamline triage processes by providing rapid, accurate assessments of laboratory results, allowing health care providers to prioritize patients more effectively and potentially reducing the burden on overstretched health care systems. Additionally, integrating LTC into preventive care programs could empower individuals to monitor their health proactively, fostering early detection and intervention for potential health issues.

The potential impact of LTC on the direct-to-consumer health landscape is equally promising. By providing individuals with accessible and easily understandable interpretations of their laboratory results, LTC could facilitate informed decision-making, leading to earlier detection and more effective management of health conditions. This shift toward proactive health care could encourage individuals to take greater ownership of their well-being.

While the potential of AI-driven CDSSs such as LTC is evident, it is important to acknowledge the nascent nature of this field. One such study by Gräf et al [14] compared physician and AI-based symptom checker diagnostic accuracy, where the AI achieved a diagnostic accuracy of 70%. Furthermore, a systematic review of 10 studies revealed consistently low diagnostic accuracy (range 19% - 37.9%), while triage accuracy (range 48.8% - 90.1%) was relatively higher but displayed variability among different symptom checkers [9]. Our study attempted to adhere to established reporting guidelines for machine learning models [15], but more robust research is necessary to fully understand the capabilities and limitations of AI-powered CDSSs in diverse health care settings.

Strengths and Limitations

While the study yielded promising results, several inherent limitations should be acknowledged when assessing the accuracy and safety of LTC. First, the sample size was relatively small, comprising only 101 participants. Although efforts were taken to ensure analytical strength, a larger and more diverse sample would enhance the generalizability of findings to the broader population. Furthermore, the study allowed participants to seek guidance from medical professionals when faced with uncertainties while filling out the questionnaire, which might not mirror real-world use where such guidance might not be readily accessible. While this provision was aimed at optimizing data quality, it could have potentially introduced an artificial

element, warranting caution when considering the practical implications of the technology's recommendations.

Unexpected results during the experiments included occasional discrepancies between LTC and consultant assessments, indicating potential areas for model improvement. These discrepancies could stem from complex interactions between predictor variables that were not fully captured by the model, suggesting the need for further refinement of the AI algorithms.

These limitations underscore the necessity for future research involving more representative samples and real-world use scenarios to validate the robustness and effectiveness of emerging CDSS technologies. By exploring the intersection of AI and laboratory diagnostics, we aim to lay the groundwork for future progress and foster a deeper comprehension of AI-based CDSSs' potential in reshaping laboratory medicine.

Conclusions

In conclusion, our study demonstrates the transformative potential of integrating AI into laboratory diagnostics through the LTC software. The high accuracy and safety sensitivity achieved underscore the ability of AI-driven CDSSs to identify medical conditions and provide tailored recommendations, enhancing health care decision-making.

As AI continues to evolve within health care, this study validates the promise of AI in medical diagnostics and highlights the need for continued research to refine and expand such tools. By fostering collaboration between AI experts and laboratory medicine specialists, we can unlock the full potential of AI-powered CDSSs, paving the way for a more efficient, personalized, and patient-centered approach to health care.

The results presented here offer a compelling glimpse into a future where AI-driven tools play a pivotal role in optimizing patient care and revolutionizing laboratory diagnostics.

Acknowledgments

We would like to acknowledge Karol Dobrzyński for his valuable assistance in preparing this manuscript. This research was funded by the National Center for Research and Development under submeasure 1.1.1 Industrial Research and Development Works, the Intelligent Development Operational Program 2014-2020, co-financing agreement POIR.01.01.01-00-0297/19-00 of November 13, 2019.

Data Availability

All data produced in this study are available upon reasonable request to the authors.

Conflicts of Interest

SA is the chief executive officer of Labplus, the company that owns and develops the artificial intelligence (AI)–based LabTest Checker (LTC). BB is an associate of Labplus. Both SA and BB receive compensation from Labplus and hold equity in the company. AK is an employee, also receiving compensation from Labplus. These financial connections could be perceived as a potential conflict of interest. Every effort has been made to ensure that the study design, data analysis, and interpretation have been conducted objectively and rigorously. The rest of the authors declare no conflicts of interest.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019 Jun;6(2):94-98. [doi: 10.7861/futurehosp.6-2-94] [Medline: 31363513]
2. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021 Jan;14(1):86-93. [doi: 10.1111/cts.12884] [Medline: 32961010]
3. Castaneda C, Nalley K, Mannion C, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 2015 Mar 26;5:4. [doi: 10.1186/s13336-015-0019-3] [Medline: 25834725]
4. Nagel A. Polacy leczą się sami. 90 proc. zażywa leki bez recepty [Article in Polish]. WP abcZdrowie. 2019 Jul 26. URL: https://portal.abczdrowie.pl/polacy-lecza-sie-sami-90-proc-zazywa-leki-bez-recepty [accessed 2023-08-18]
5. Pokańko K. Polskie AI ma w kilka sekund zweryfikować wyniki badań. wszystko online i bez wychodzenia z domu [Article in Polish]. Spider's Web. 2022 Jul 17. URL: https://bizblog.spidersweb.pl/analiza-wynikow-badan-online [accessed 2023-08-18]
6. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018 Nov 24;392(10161):2263-2264. [doi: 10.1016/S0140-6736(18)32819-8] [Medline: 30413281]
7. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 6;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
8. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. BMJ Open 2019 Aug 1;9(8):e027743. [doi: 10.1136/bmjopen-2018-027743] [Medline: 31375610]

9. Wallace W, Chan C, Chidambaram S, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022 Aug 17;5(1):118. [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
10. Nateqi J, Lin S, Krobath H, et al. From symptom to diagnosis—symptom checkers re-evaluated: are symptom checkers finally sufficient and accurate to use? an update from the ENT perspective [Article in German]. HNO 2019 May;67(5):334-342. [doi: 10.1007/s00106-019-0666-y] [Medline: 30993374]
11. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015 Jul 8;351:h3480. [doi: 10.1136/bmj.h3480] [Medline: 26157077]
12. Hennemann S, Kuhn S, Witthöft M, Jungmann SM. Diagnostic performance of an app-based symptom checker in mental disorders: comparative study in psychotherapy outpatients. JMIR Ment Health 2022 Jan 31;9(1):e32832. [doi: 10.2196/32832] [Medline: 35099395]
13. Assessment of accuracy and safety of LabTest Checker (LTC-AI). ClinicalTrials.gov. URL: https://clinicaltrials.gov/study/NCT05813938 [accessed 2024-08-02]
14. Gräf M, Knitza J, Leipe J, et al. Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. Rheumatol Int 2022 Dec;42(12):2167-2176. [doi: 10.1007/s00296-022-05202-4] [Medline: 36087130]
15. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323. [doi: 10.2196/jmir.5870] [Medline: 27986644]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
ED: emergency department
LTC: LabTest Checker

Edited by C Lovis; submitted 06.02.24; peer-reviewed by A Al-Asadi, J Walsh, R Campbell; revised version received 22.05.24; accepted 25.05.24; published 14.08.24.

Please cite as:
Szumilas D, Ochmann A, Zięba K, Bartoszewicz B, Kubrak A, Makuch S, Agrawal S, Mazur G, Chudek J
Evaluation of AI-Driven LabTest Checker for Diagnostic Accuracy and Safety: Prospective Cohort Study
JMIR Med Inform 2024;12:e57162
URL: https://medinform.jmir.org/2024/1/e57162
doi:10.2196/57162

Original Paper

Interpretable Deep Learning System for Identifying Critical Patients Through the Prediction of Triage Level, Hospitalization, and Length of Stay: Prospective Study

Yu-Ting Lin[1*], MSc; Yuan-Xiang Deng[1*], MSc; Chu-Lin Tsai[2*], MD, SCD; Chien-Hua Huang[2*], MD, PhD; Li-Chen Fu[1*], PhD

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[2]Department of Emergency Medicine, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan

[*]all authors contributed equally

Corresponding Author:
Li-Chen Fu, PhD
Department of Computer Science and Information Engineering
National Taiwan University
CSIE Der Tian Hall No. 1, Sec. 4, Roosevelt Road
Taipei, 106319
Taiwan
Phone: 886 935545846
Email: lichen@ntu.edu.tw

Abstract

Background: Triage is the process of accurately assessing patients' symptoms and providing them with proper clinical treatment in the emergency department (ED). While many countries have developed their triage process to stratify patients' clinical severity and thus distribute medical resources, there are still some limitations of the current triage process. Since the triage level is mainly identified by experienced nurses based on a mix of subjective and objective criteria, mis-triage often occurs in the ED. It can not only cause adverse effects on patients, but also impose an undue burden on the health care delivery system.

Objective: Our study aimed to design a prediction system based on triage information, including demographics, vital signs, and chief complaints. The proposed system can not only handle heterogeneous data, including tabular data and free-text data, but also provide interpretability for better acceptance by the ED staff in the hospital.

Methods: In this study, we proposed a system comprising 3 subsystems, with each of them handling a single task, including triage level prediction, hospitalization prediction, and length of stay prediction. We used a large amount of retrospective data to pretrain the model, and then, we fine-tuned the model on a prospective data set with a golden label. The proposed deep learning framework was built with TabNet and MacBERT (Chinese version of bidirectional encoder representations from transformers [BERT]).

Results: The performance of our proposed model was evaluated on data collected from the National Taiwan University Hospital (901 patients were included). The model achieved promising results on the collected data set, with accuracy values of 63%, 82%, and 71% for triage level prediction, hospitalization prediction, and length of stay prediction, respectively.

Conclusions: Our system improved the prediction of 3 different medical outcomes when compared with other machine learning methods. With the pretrained vital sign encoder and repretrained mask language modeling MacBERT encoder, our multimodality model can provide a deeper insight into the characteristics of electronic health records. Additionally, by providing interpretability, we believe that the proposed system can assist nursing staff and physicians in taking appropriate medical decisions.

(JMIR Med Inform 2024;12:e48862) doi:10.2196/48862

KEYWORDS

emergency department; triage system; hospital admission; length of stay; multimodal integration

Introduction

Background

Emergency services are an essential aspect of the health care system in hospitals, and the demand for these services has increased exponentially in recent years. For instance, due to a rising number of elderly patients, a high volume of low-acuity patients waiting for the emergency department (ED), and limited access to medical resources in the community, it may take a long time for patients to receive medical treatment in the ED. Additionally, the situation has worsened with the shortage of experienced health care providers. In the ED, this can cause many severe clinical outcomes, such as delayed diagnosis, longer patient wait times, and increased mortality rates. Moreover, the patient and the standard health care operation procedure may be disturbed. Therefore, prioritizing ED visits and maintaining the regular operation of the health care system are essential.

Triage is the process of accurately assessing patients' symptoms and providing them with proper clinical treatment in the ED. Patients are assigned different priorities depending on their vital signs and chief complaints, and the judgment description from the nursing staff [1]. Many countries have developed their triage process to stratify the clinical severity of patients and thus distribute medical resources. For instance, the US Emergency Severity Index (ESI), Canadian Triage and Acuity Scale (CTAS) [2], and Taiwan Triage Acuity Scale (TTAS) are designed to improve the triage prioritizing process [3-5]. In terms of personnel, hospitals employ dedicated nurses who have been certified by the authorities to undertake the triage process. It is also essential to maintain the quality of education, training, and evaluation of those professionals, which is more difficult nowadays with the increase in the complexity of emergency care and the increase in the number of patients visiting the ED nationwide [6]. Although many standardized scales have been adopted to improve the process, there are still some limitations of the current triage system [7-9]. Among these issues, the lack of capability to prioritize patients and assign patients to appropriate triage levels is the most serious problem. According to records collected in Taiwan from 2009 to 2015, 167,598 out of 268,716 (nearly 60%) visits in the ED were assigned to level 3 in the triage process. In addition, 5-level triage mainly relies on an experienced nurse's diagnosis that is based on a mix of subjective and objective criteria. Any human judgement errors or even inaccurate measurements that occur during the triage assessment can severely affect the outcome.

Related Work

Contextualized Word Embedding

A word vector is an attempt to mathematically capture the syntactic and semantic features of a word and represent its meaning simultaneously. Computers calculate how often words appear next to each other by going through a large corpus. For instance, with GloVe [10] or word2vector [11], the word can be projected into a high-dimensional vector for further tasks.

Although these traditional word embedding methods are easy to understand and simple to implement, some limitations still need to be addressed. For example, after applying word vectors, it would be tough to train systems equipped with the softmax function owing to a large number of categories. On the other hand, the GloVe word embedding involves a numeric representation of a word regardless of where the word occurs in the sentence and the different meanings the word may have. Hence, several language models have been proposed to address these limitations, including embeddings from language models (ELMo) [12], bidirectional encoder representations from transformers (BERT) [13], and generative pretrained transformer (GPT) [14]. These celebrated language models generate general contextualized sentence embeddings by using a large scale of unlabeled corpora.

Among these famous models, BERT is the most popular model commonly used in solving natural language processing (NLP) tasks. BERT is a language model trained bidirectionally, which means that as compared to single-direction language models, it can provide a more profound sense of language context and flow. Moreover, instead of predicting the next word in the sentence, BERT also uses a novel method called "mask language modeling" (MLM). This novel algorithm randomly masks the words and then predicts them. BERT relies on the transformer architecture; however, since BERT aims to generate a language representation model, it only uses the transformer encoder by stacking them up. Later, with the help of MLM and "next sentence prediction" (NSP), BERT can achieve significant performance on lots of NLP downstream tasks by further fine-tuning on specific domains.

Deep Learning for Tabular Data

In statistics, tabular data refer to data organized in a table. Within the table, the rows and columns represent observations and attributes for those observations, respectively. Although many domains like vision, NLP, and speech enjoy the benefit of deep learning models, tabular data using deep learning methods remain questionable. On the other hand, when it comes to handling tabular data, the traditional machine learning method dominates most of the benchmarks and is commonly used in competitions, such as Kaggle, around the world. The conventional machine learning methods include methods based on decision tree (DT) such as extreme gradient boosting (XGBoost) [15], category boosting (CatBoost) [16], and light gradient boosting machine (LightBGM) [17]. The strength of these DT-based methods is that their output is easy to understand and available to provide interpretability without requiring any statistical knowledge. However, there are still some limitations of DT-based methods. Among these limitations, the most serious is that DT-based methods do not allow efficient learning with image or text encoders. Hence, many experts turn to deep learning methods instead of DT-based methods. Deep learning models enable end-to-end learning for tabular data and have many benefits at the same time. First, they can achieve better performance in a bigger data set. Second, they can alleviate the need for feature engineering. Finally, they encode multiple data types efficiently, like images along with tabular data.

However, the shortcoming of most deep learning methods is that they cannot provide interpretability. Fortunately, researchers have been aware of the problem in recent years, and several

deep learning models with interpretability have been proposed, such as TabNet [18], neural oblivious decision ensembles (NODE) [19], and TabTransformer [20].

Current Work in the Triage System

Although current triage systems, such as the ESI and TTAS, follow clear guidelines to assign patient acuity, it implicitly leaves room for clinician interpretation. Hence, the diagnosis still depends heavily on the judgment and experience of individual nursing staff. Several studies have shown that cognitive biases can influence clinical judgments [6]. In written case scenarios at multiple EDs, the average accuracies of nurses were 56.2%, 59.2%, and 59.6% in Taiwan, Brazil, and Switzerland, respectively [21]. In view of this, some studies [6,21,22] have turned to the use of artificial intelligence (AI) systems to assist with decision-making in triage. They also demonstrate the system's effectiveness with higher accuracy from the assisted means.

Numerous studies have attempted to use traditional machine learning methods in their approaches. Choi et al [6] used 3 types of conventional machine learning methods, including logistic regression, random forest, and XGBoost, to predict the Korea Triage Acuity Scale (KTAS) level. They used patients' chief complaints as categorical features, meaning that they assigned a key code to each symptom. Their best model using random forest achieved precision, recall, and area under the receiver operating characteristic curve values of 0.737, 0.730, and 0.917, respectively. Liu et al [22] used CatBoost as their model; however, the study focused on distinguishing the mis-triage of patients in levels 3 and 4 since they believed that the under-triage of critically ill patients could be life-threatening. Their model was able to reduce the life-threatening mis-triage rate from 1.2% to 0.9% prospectively. Ivanov et al [21] carried out a series of experiments to demonstrate the effectiveness of their novel idea "clinical natural language processing (C-NLP)." To cope with free-text data, C-NLP uses sentence tokenization, word tokenization, and part-of-speech tagging to extract the meaning behind free-text data. Their best model included C-NLP and XGBoost, and it was able to achieve an accuracy of 75.7%, which is 26.9% higher than the average nurse's accuracy.

The previously mentioned studies [6,21,22] achieved great performance in dealing with triage-level problems; however, these methods still have some limitations. Our proposed model aims to address these limitations and alleviate them. Multimedia Appendix 1 presents comparisons between earlier work and our study in different aspects.

Goal of This Study

Although the studies mentioned in the previous section successfully demonstrated that AI improved the triage system for predicting triage level, they unfortunately had some serious drawbacks. In this study, we attempted to overcome these drawbacks while developing an appropriate prediction system based on triage information, including demographics, vital signs, and chief complaints. We propose a system that can handle the collected heterogeneous data, including tabular data and free-text data. The proposed system is capable of providing precise suggestions for ED staff in hospitals, and it has interpretability for better acceptance by users. Moreover, it is applicable to real-world situations.

Methods

System Overview

In this study, we have proposed a system comprising 3 subsystems, with each of them handling 1 task. As shown in Figure 1, these tasks include triage level prediction, hospitalization prediction, and length of stay prediction, which are important outcomes in the ED of a hospital. Since these subsystems are developed in a similar training process, we will first introduce the conceptual level of the typical training process of each model in each subsystem and then provide further information. Finally, we will show the detailed design of each model in each subsystem.

Our study focuses on establishing an effective and precise AI system to predict the criticality of patients waiting in the ED of hospitals. By leveraging a model trained on a data set where data labels include different scales, we look forward to developing a robust model that can provide more information to the physician and nursing staff. Moreover, to assist them in making precise medical decisions, our proposed system offers multiple prediction outcomes, including triage-level classification, hospitalization estimation, and length of stay.

The system flowchart is shown in Figure 2. The system can be divided into 3 stages: pretraining stage, fine-tuning stage, and testing stage. Additionally, 2 data sets were used in our study. One was the National Taiwan University Hospital (NTUH) retrospective data set, and the other was the NTUH prospective data set collected from May 26, 2020, till February 21, 2022. These 2 data sets will be elaborated in the following sections.

In the pretraining stage, a large amount of retrospective data were used to pretrain the encoders to learn the basic information of the medical data. In addition, the pretrained encoders were transferred to the second stage. In the fine-tuning stage, we used prospective data with golden labels to fine-tune the pretrained encoder. Therefore, when the diagnosis outcomes from the physician are treated as the ground truth label, the model is more applicable to real-world situations. Finally, in the testing stage, we implemented our system in the hospital and assessed the effectiveness of the system.

Figure 1. The proposed system comprising 3 subsystems that are responsible for different tasks. AI: artificial intelligence.

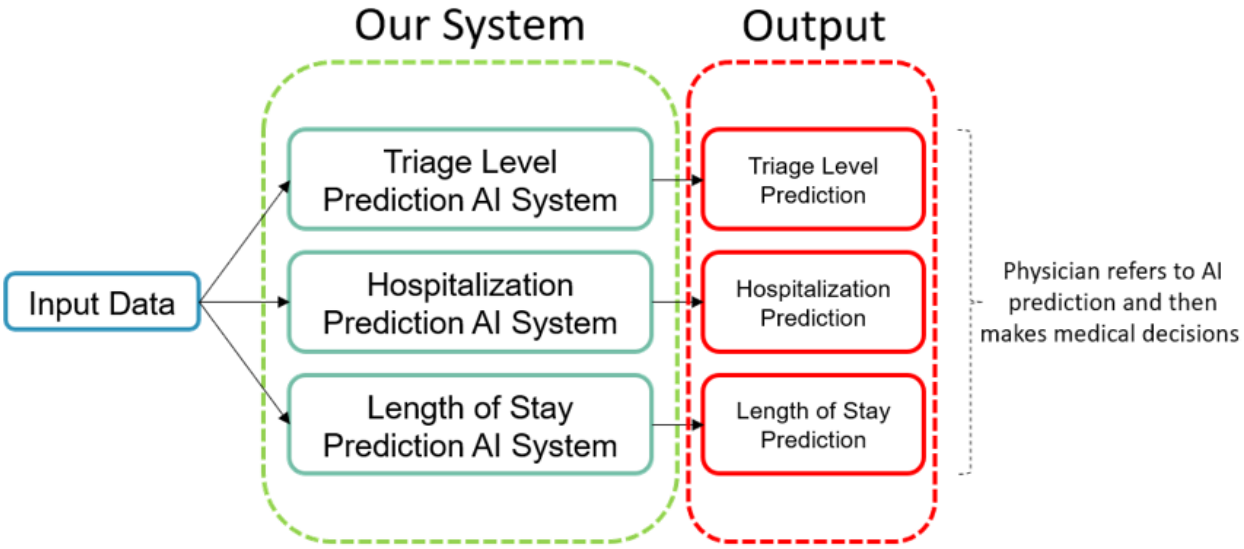
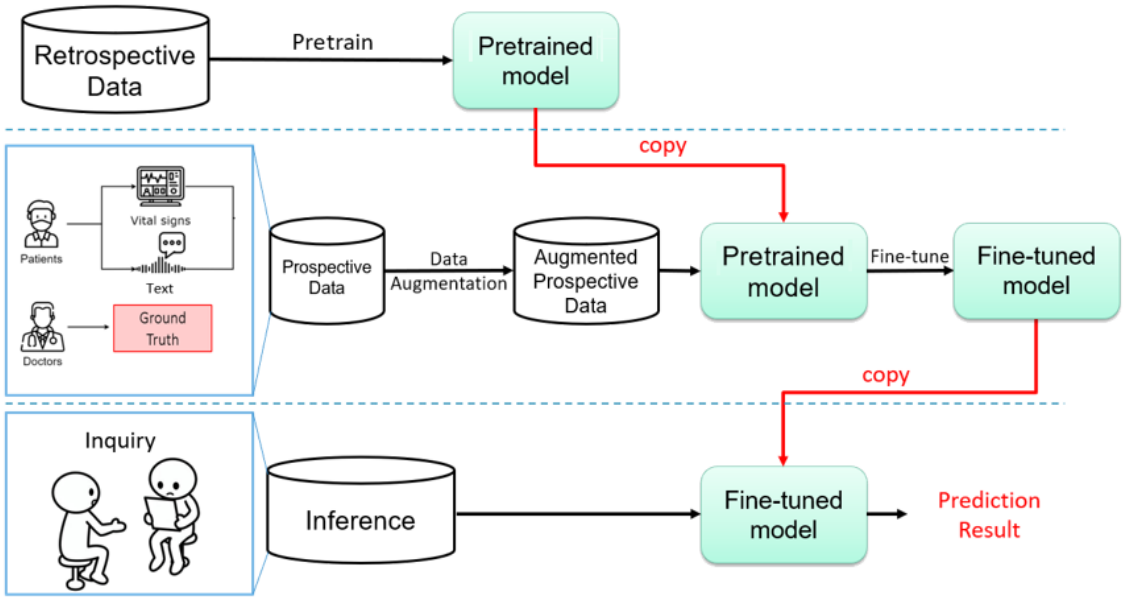


Figure 2. System flowchart.



Ethical Considerations

This study has been approved by the NTUH Institutional Review Board (201606072RINA, 201911054RINA, 202108090RINC).

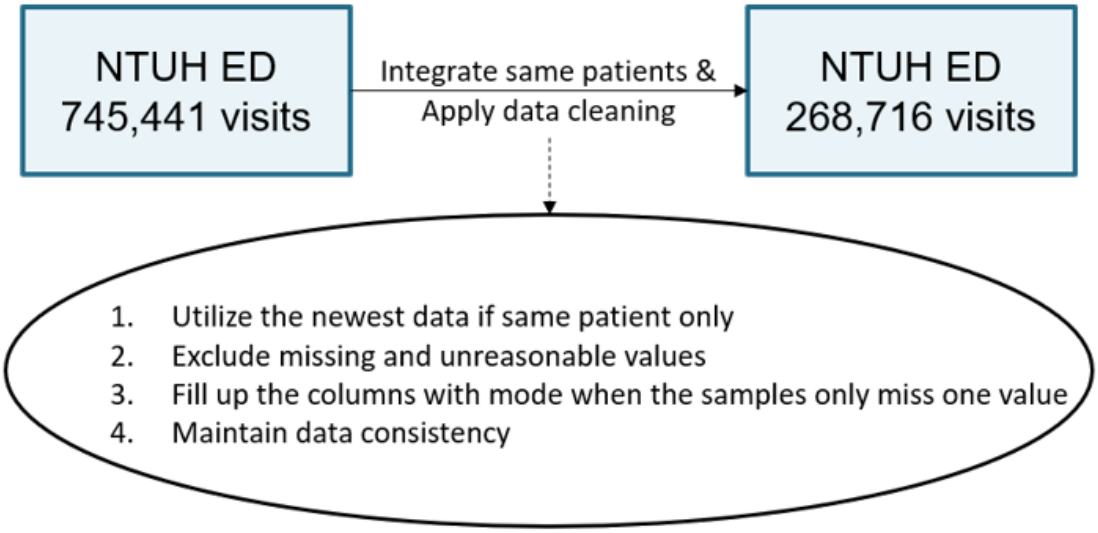
Data Preparation

NTUH Retrospective Data Set

The NTUH is a tertiary academic medical center that has almost 2400 beds and 100,000 emergency room visits per year. After receiving approval from the NTUH Institutional Review Board, we obtained the NTUH retrospective data set, which contained a total of 745,441 electronic health records (EHRs) of patients who visited the ED from the years 2009 to 2015. Since triage is the starting point of care for the ED, it is essential to ensure consistent and precise estimation of patients. The records were evaluated by dedicated personnel who were certified by the Taiwan Union of Nurses Association (TUNA), following a standard protocol.

As shown in Figure 2, in the first stage, we used the NTUH retrospective data set to pretrain our model. However, in the NTUH retrospective data set, we needed to unify the uncleaned data (Multimedia Appendix 2) initially as the members of the nursing staff have their own ways to record the estimation. We included all patients aged 20 years or older who attended the ED and excluded patients whose EHR data contained missing or unreasonable values. Unreasonable data had unreasonable values, which may have resulted from typing errors. For instance, the diastolic pressure and systolic pressure may be typed in reverse, or a nurse may accidentally omit a digit when entering values on the computer. In such a scenario, even though we may be able to infer the original intended values by examining individual data, we cannot consider this a correct sample for use. After data cleaning and merging, only 268,716 patients were enrolled in our program (Figure 3).

Figure 3. Preprocessing of the National Taiwan University Hospital (NTUH) retrospective data set. ED: emergency department.



NTUH Prospective Data Set

Each patient who visits the ED will have a PDF document form generated (triage examination and evaluation record). These records are kept for the physician to make a diagnosis. The records comprise 2 types of information. The first is structural data, including patient demographics, triage information, and vital signs, and the second is textual data, including chief complaints, historical medical information, and drug allergy.

In general, it is impossible to directly use the aforementioned records to train the model, and thus, data preprocessing is needed to extract the data from the records. We used the PDFMiner library in Python code to extract the information from the document forms as "structural data" and applied a transformation function to generate "textual data."

The information extracted from the forms and records can be divided into 2 groups: target prediction and patient feature. Detailed explanations of the patient features are provided in Table 1. On the other hand, the target ground truth contains 3 different tasks. The first task is triage level prediction, which is a 4-class classification problem, where the physician's suggestion is considered (golden standard label that is obtained from the physician by observing the process of patient diagnosis) instead of the traditional triage level. A lower level indicates that the patient more urgently requires immediate attention. The second task is hospitalization prediction, which is a 2-class classification problem, where "0" represents that the patient needs to be discharged by the hospital and "1" represents that the patient needs to be admitted. The last task is length of stay, which is a 3-class classification problem, where "0" represents that the patient will stay in the ED for less than 6 hours, "1" represents that the patient will stay in the ED for 6 to 24 hours, and "2" represents that the patient will stay in the ED for more than 24 hours.

Table 1. Detailed explanation of structural variables.

Variable	Explanation
Demographics	
Age	Patient age
Sex	Patient gender
Triage information	
Session	Patient arrival time
Return in 24 hours	Number of times the patient revisited the ED[a] in 24 hours
Clinic visit mode	Patient arrival mode
Work related	Whether the patient visited the ED because of a work accident
On the way to work	Whether the patient was on the way to work before visiting the ED
Vital sign information	
Systolic pressure	Systolic blood pressure
Diastolic pressure	Diastolic blood pressure
Pulse	Pulse
Oxygen	Oxygen saturation
Respiration	Respiration
Body temperature	Body temperature
Acute change	Any acute changes before entering the ED
Fever	Whether the patient has fever
Pain index	Self-evaluated pain score
GCS-E	Glasgow Coma Scale score of the patient (eye opening)
GCS-V	Glasgow Coma Scale score of the patient (verbal response)
GCS-M	Glasgow Coma Scale score of the patient (motor response)
Major disease	Whether the patient has an IC[b] card for severe illness
Admission count	The number of times the patient went to the hospital in 1 year
Judgement code	The judgement code for describing the patient's condition
Textual data	
Chief complaint	The patient's description of the symptoms
Judgment description	The record that describes the patient's symptoms written by the nursing staff

[a]ED: emergency department.

[b]IC: integrated circuit.

Data Augmentation

After analyzing our prospective data set, we observed an imbalanced data distribution. As machine learning algorithms tend to increase accuracy by reducing errors, most of them are biased toward the majority class and tend to ignore the minority class. For instance, 758 out of 901 (84.1%) ED patients were discharged from the hospital in our prospective data set, and the system could achieve 85% accuracy if it kept on predicting discharge. However, we did not want the system to only indicate discharge. Therefore, to avoid the above situation, we used the "synthetic minority oversampling technique" (SMOTE) to generate some synthesized data to ensure that the system could learn the different patterns between each class. In our study, the iteration of the SMOTE algorithm started by selecting 1 minority sample and finding its top 5 nearest neighbors. These 5 neighbors were chosen to generate new synthesized data by the interpolation method. Finally, the iteration was repeated several times until we obtained the minority class where the number was the same as that of the majority class. However, as the synthesized data may be too diverse, some of the data can have negative influences on the model. Therefore, we used the Tomek Links algorithm to remove some ambiguous data that may hurt model performance by pairing samples and removing the pairs with different labels. An example of the augmentation process is shown in Multimedia Appendix 3. In the original data set, we can observe that only 143 patients are admitted. After applying the SMOTE algorithm on our data set, the number of admitted patients increases to 758. We then use the Tomek Links algorithm to remove some samples that are regarded as

ambiguous samples by the algorithm. Finally, in this example, a total of 1294 patients are included in our new augmented prospective data set.

As for text data, since the SMOTE algorithm cannot generate text, we set up a mapping relation to add the text feature for each synthesized sample. First, we created a number of lists, each of which stores the chief complaints from data samples sharing the same class label. After these lists and the synthesized data were ready, for each synthesized sample, we randomly selected 1 chief complaint from the list according to its label and added it as a text feature of the synthesized sample.

Pretraining of the Vital Sign Encoder

The TabNet architecture is composed of feature transformers and attentive transformers. In TabNet's design, the mask from the attentive transformer can select the most vital feature from several features, eliminating noise caused by irrelevant features. Furthermore, the mask can be calculated to provide some interpretable information about the feature's importance. Therefore, considering the objective of this study, our work takes advantage of the encoder-decoder architecture of TabNet, which is inspired by Arik [18], and we adopted this architecture to construct our vital sign encoder (Figure 4).

Before training on the prospective data set, the vital sign encoder was pretrained on retrospective data by unsupervised learning to learn some basic information about such structural data. Structural features of demographics, triage information, and vital sign information (Table 1) were used in this step.

Figure 5 shows the process used for pretraining our vital sign encoder. In triage level prediction and length of stay prediction, since we did not have a triage golden label and length of stay label for pretraining the vital sign encoder, we used only unsupervised learning. On the other hand, both unsupervised learning and supervised learning were used for hospitalization prediction. The reason why we used the unsupervised learning algorithm is that the model can discover hidden data patterns without human intervention by analyzing and clustering the unlabeled information.

Figure 4. Vital sign encoder architecture (adapted from TabNet). FC: fully connected networks; ReLU: rectified linear unit.

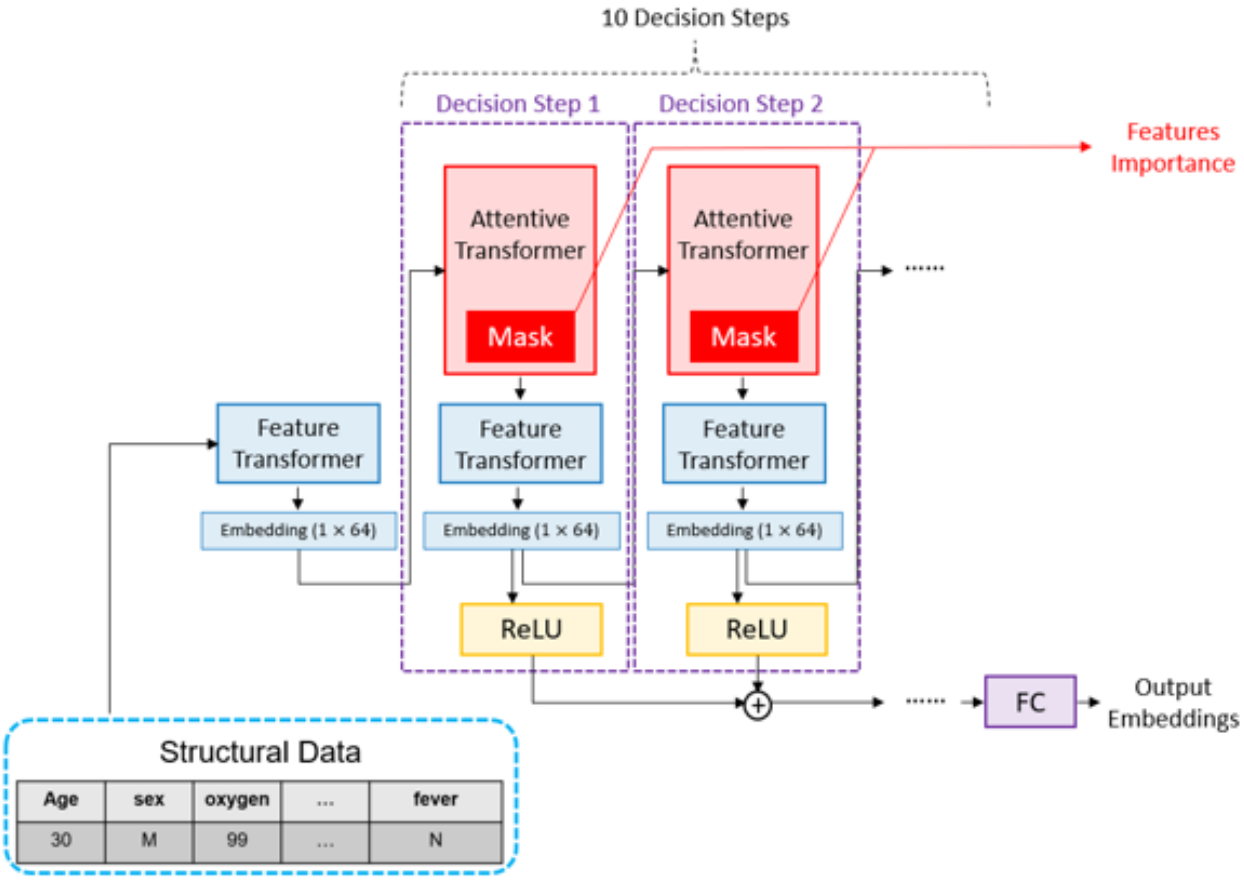
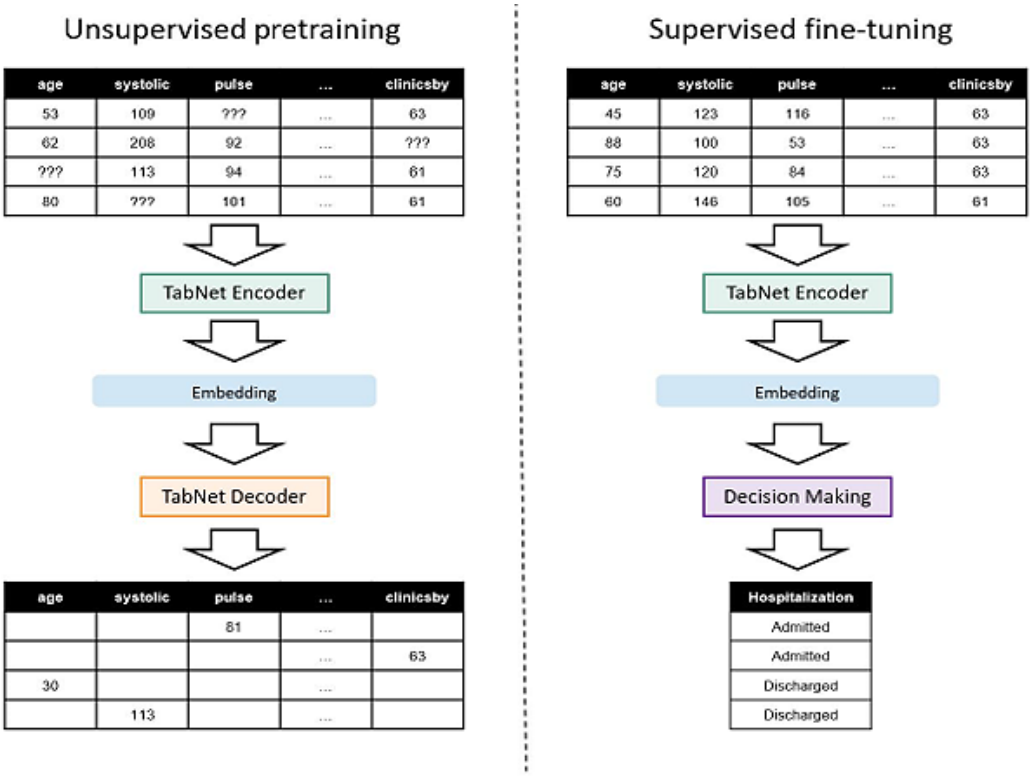


Figure 5. The flow of pretraining the vital sign encoder.



In our study, we used the encoder-decoder architecture. We masked some of the input features in our data and then reconstructed the masked features. The reconstruction loss during unsupervised learning is described as follows:

where B denotes the batch size, D denotes the dimension of features (number of features), S is a binary mask sparse matrix with size B×D for masking some of the features, and $f_{B\times D}$ and  are matrices of features with size B×D representing the ground truth feature values and the predicted outputs, respectively [18].

The purpose of Equation 1 is to calculate the distance between the ground truth feature and the predicted feature. In each iteration, the binary mask $S_{b,j}$ is sampled independently from Bernoulli distribution, and the mask can only have a value of 0 or 1. During this process involving the masked value and its reconstruction, models are believed to learn implicit relationships between features.

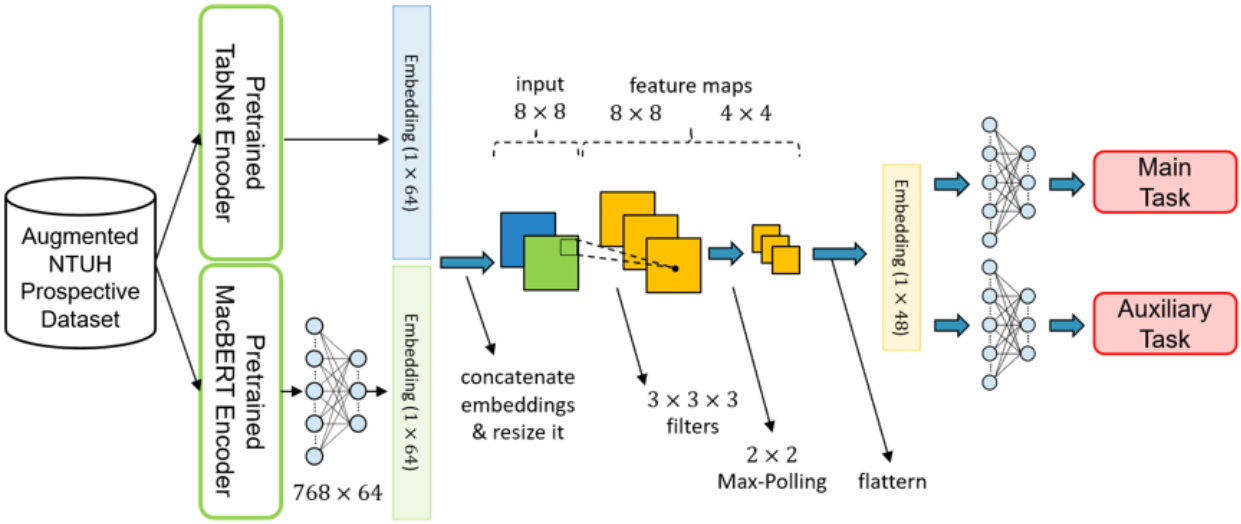
Pretraining of the MacBERT Encoder

BERT is a well-known language model that can be used to transform a word into a representation and understand the meaning behind the sentence. In addition, it performs consistently better than other language models (eg, ELMo [12] and GPT [14]) and also performs well in many different tasks. However, although the BERT model can be easily fine-tuned with an additional output layer to achieve outstanding performance, the pretraining process of the model is designed for general purposes. In this study, to better understand our medical data, we repretrained MacBERT (Chinese version of BERT) by applying MLM again. We extracted the text information from the NTUH prospective data set and then used the information to accomplish further training of MLM. All the settings of the training process followed the original set in MacBERT. By further training with MLM, the fine-tuned MacBERT could enrich its knowledge in specific domains.

Overall Model Architecture

The typical model architecture of each subsystem is shown in Figure 6. After the pretrained encoders are ready, the encoder weights are copied to the fine-tuning stage encoders. The typical model architecture can be divided into 4 main parts: input, encoders, classifiers, and output. First, in the input part, there are 2 data types, namely, structural data and free-text data. Since the prospective data set has only a limited amount of data, we sent it to the augmentation algorithm to obtain synthesized data and added them to the original data set. Second, the structural data and the free-text data are sent to the pretrained TabNet encoder and the pretrained MacBERT encoder, respectively. Afterward, to obtain a comprehensive representation of the data, 2 embeddings coming out from the pretrained TabNet encoder and pretrained MacBERT encoder are concatenated together. Third, the concatenated embeddings are passed through classifiers for output prediction.

Figure 6. Typical model architecture in the fine-tuning stage. MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.



Input

We used the augmented NTUH prospective data set in the fine-tuning stage. The data set contains 2 data types. The first is structural data, including patient demographics, triage information, and vital sign information. The second is free-text data, including patient chief complaints, nursing staff judgment descriptions, and transformed information from the structural data (Multimedia Appendix 4). However, since MacBERT is a Chinese BERT model, which is trained on simplified Chinese, we translated our text data from traditional Chinese to simplified Chinese to achieve better performance.

Encoders

As shown in Figure 6, since there were 2 types of data to be processed, we used the TabNet encoder and MacBERT encoder to extract feature information from structural data and free-text data, respectively. We then transformed these information pieces into high-dimensional embeddings for further training.

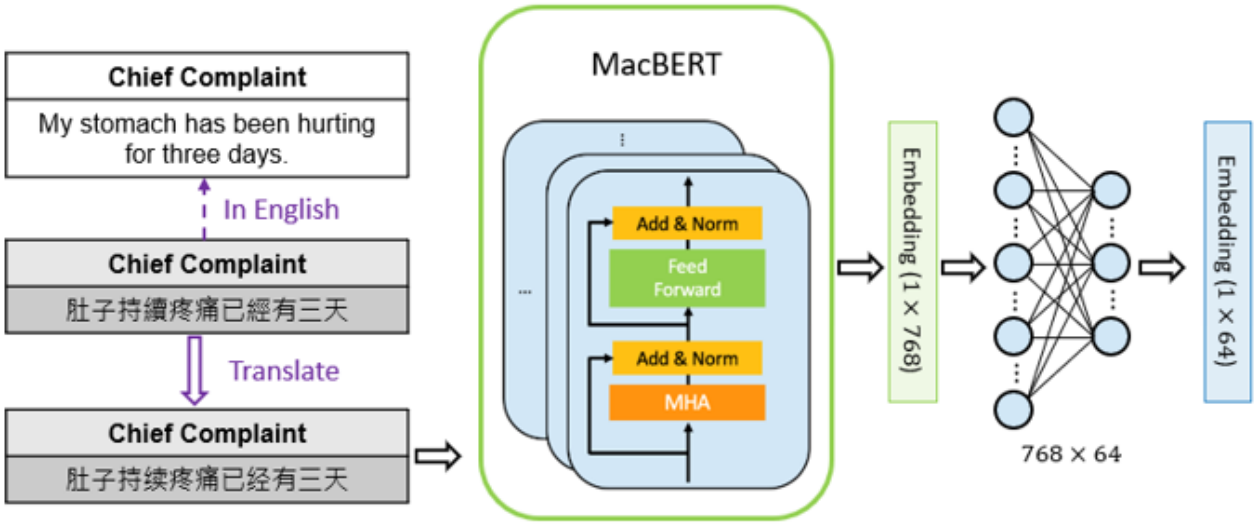
Pretrained Vital Sign Encoder

We used the pretrained TabNet encoder as our vital sign encoder. In the pretraining stage, we obtained some basic information of these medical data from the NTUH retrospective data set. As a result, to achieve better starting, the pretrained weights were directly deployed into our vital sign encoder. We stacked up 10 decision steps to build our vital sign encoder, and the dimensions of both the input and output were set to 64. A 1×64 vector was the final context vector.

Pretrained Language Model Encoder

As chief complaints are manually recorded by nurses and most of them are written in traditional Chinese, it is better to find a language model that has been trained on a Chinese corpus and can handle Chinese text well. MacBERT is an improved BERT model with novel MLM as a correction pretraining task, which mitigates the discrepancy between pretraining and fine-tuning. Moreover, it has been trained on simplified Chinese corpora, which is more suitable for our work. As a result, we decided to adopt MacBERT from Hugging Face as the chief complaint text encoder in our proposed model, instead of the original BERT model. On the other hand, we observed that the text in our data set might contain different languages, including English and Chinese. Therefore, to make MacBERT applicable to our case, we translated the text into a uniform language, namely, simplified Chinese, before sending it into MacBERT. However, since we wanted the contributions from the vital sign encoder and the MacBERT encoder to be comparable, a fully connected layer was placed after the output vector from MacBERT to decrease the vector dimension from 1×768 to 1×64. The entire process explaining how we handled the text data is shown in Figure 7.

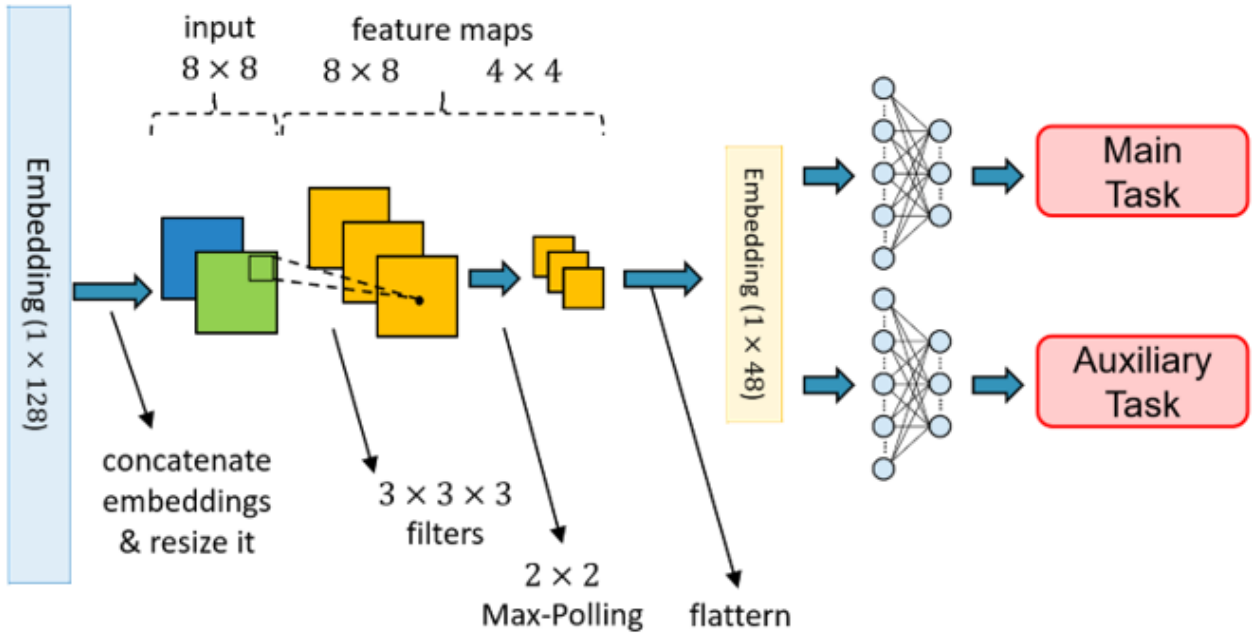
Figure 7. The entire process of handling text data. MacBERT: Chinese version of bidirectional encoder representations from transformers.



Classifiers

All the inputs were encoded into high-dimensional embeddings by the encoders mentioned in the previous stage. It is believed that both embeddings have different facets of information; therefore, instead of adding these vectors together, we concatenated these 2 vectors to obtain richer patient information before sending them into the classifiers. Moreover, in our study, we adopted the multi-task learning architecture to learn shared representation and avoid overfitting problems. As a result, there were 2 classifiers for predicting different targets, where each classifier had a 1-layer convolutional neural network and a 2-layer multi-layer perceptron. The details of the process are shown in Figure 8.

Figure 8. Components of the classifiers.



Output

In contrast to most single-output machine learning methods, our proposed model has a multi-task model architecture. Multi-task learning is a type of machine learning method by which the multi-output outcome can be learned simultaneously in a shared model. In addition to the data efficiency advantages, such an approach can reduce overfitting by leveraging auxiliary information and allowing fast learning. Since target prediction loss will update the encoders, the encoders can avoid being overfitted and learn more general knowledge. As there were 3 medical outcomes in our system, we designed 3 models with slight differences to handle different tasks. The details of these 3 models are shown in Figures 9 to 11.

Figure 9. The model architecture of triage level prediction. FL: focal loss; MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.

$$Loss_{total} = FL_{triage} + 0.5 \times FL_{hos} + 10^{-3} \times Loss_{sparse}$$

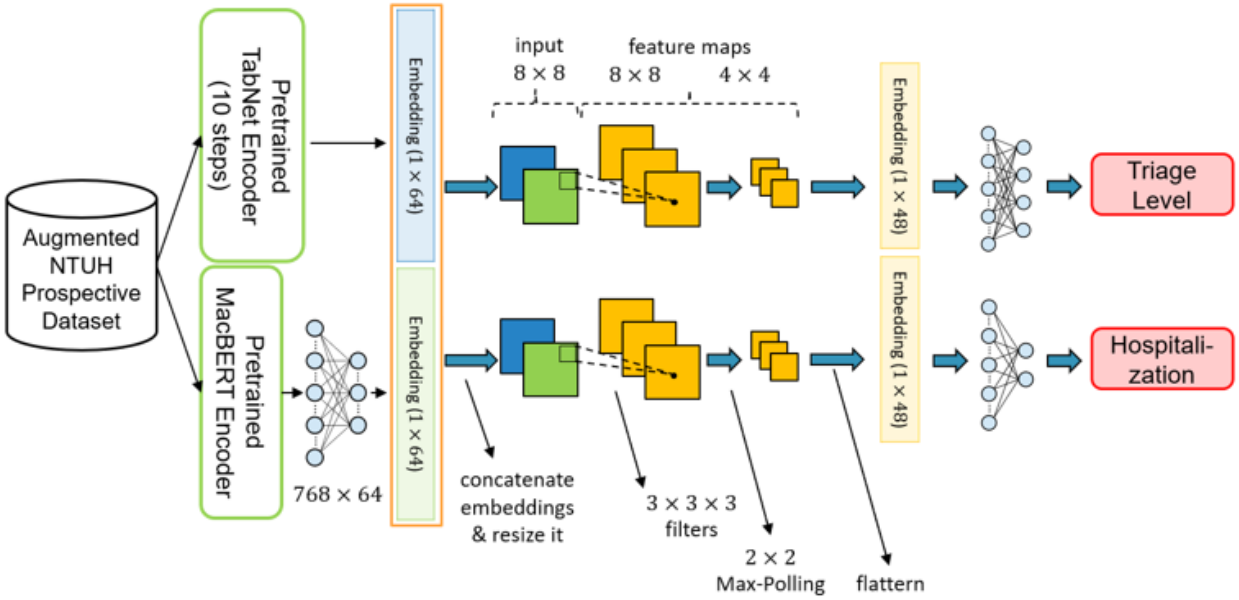


Figure 10. The model architecture of hospitalization prediction. FL: focal loss; MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.

$$Loss_{total} = FL_{hos} + 0.3 \times FL_{triage} + 10^{-3} \times Loss_{sparse}$$

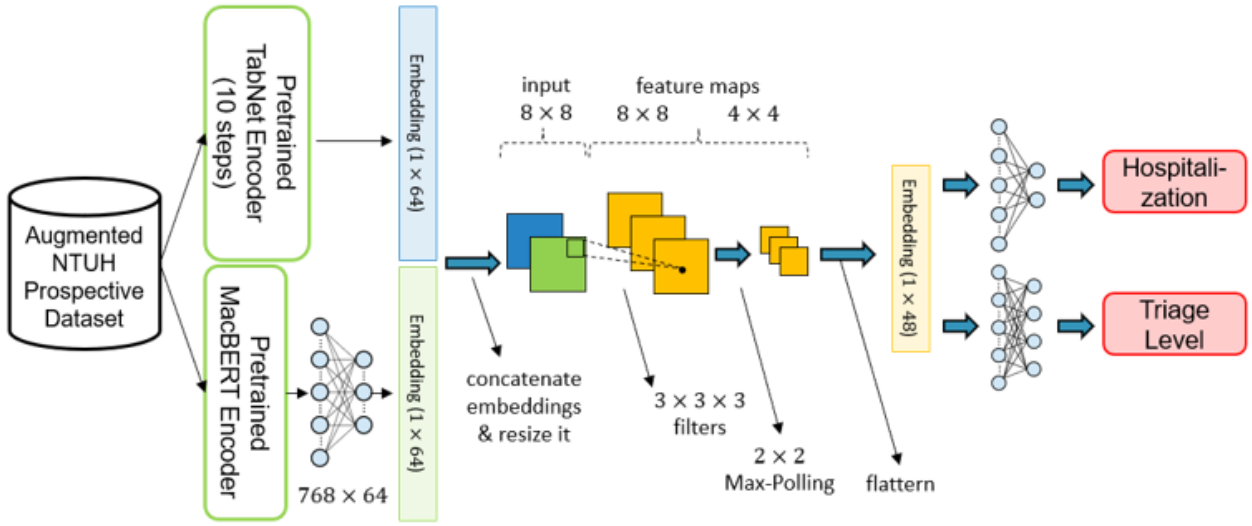
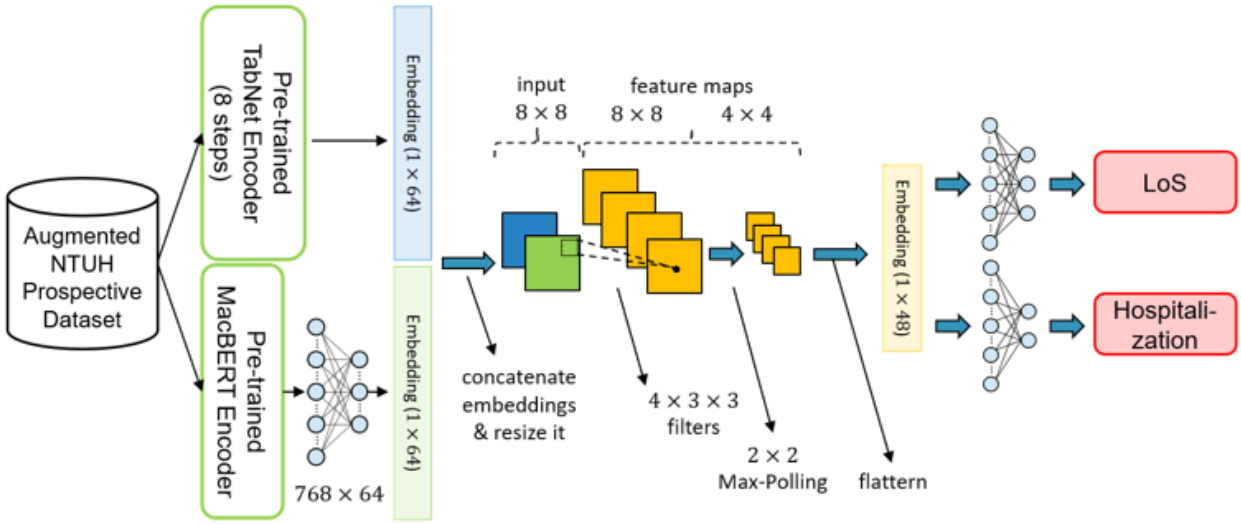


Figure 11. The model architecture of length of stay (LoS) prediction. FL: focal loss; MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.

$$Loss_{total} = FL_{LoS} + 0.3 \times FL_{hos} + 10^{-3} \times Loss_{sparse}$$



Loss Function

Total loss combines focal loss and sparse entropy loss as follows:

where λ_1 is a hyperparameter for determining the learning direction of the model via controlling the balance between the main task and related task, and λ_{sparse} is a hyperparameter for controlling the sparsity of the TabNet encoder, where a greater parameter is associated with a greater effect of the tabular data on the entire model, and the TabNet encoder tends to select 1 feature in 1 decision step.

In order to assess the performance of the model, the focal loss function was utilized by comparing the ground truth label with the probability distributions over network predictions, which has been shown as follows:

where $\hat{y}$ is the model prediction, y is the ground truth value, superscript i refers to sample i, y_k is 0 or 1 (indicating whether a class label is the correct classification among K classes),  denotes the confidence score of class k, and γ is a hyperparameter that is set to 2 in our study.

TabNet uses sparse entropy loss (first proposed in [23]) to provide a favorable inductive bias for data sets where most features are redundant. The sparse entropy loss can not only help the model to select salient features from all attributes of the sample, but also fasten the training process. The equation is as follows:

where N_{steps} denotes how many decision steps are stacked up in the model, B is the batch size, D is the total number of features, M represents the mask, $M_{b,j}$ [i] refers to the mask at the i^{th} step with batch sample b and feature j, and ε is a small number to maintain numerical stability.

Results

Experimental Setup

A series of experiments were conducted to validate the effectiveness of our design. The details of our system environment are presented below. We conducted our experiments on the Ubuntu 20.04 operating system with PyTorch 1.7.1 and Python 3.9.7, and all training procedures were performed on a computer with a Nvidia RTX 3090 graphics card, an Intel Core i7-1070K processor, and 32 GB of RAM.

Training Settings

The Adam optimizer with an initial learning rate of 0.01 was used in our experiments, and it was adjusted by the "ReduceLROnPlateau" scheduler with the patient value set as 15. Meanwhile, if the loss did not improve for 50 epochs, an early stop action was taken.

All experiments were carefully conducted in the following steps: (1) The data set was divided into 3 parts (training set, validation set, and testing set in the ratio of 8:1:1); (2) The training set was used to generate synthesized data to make up the gap between classes, and the synthesized data were added into the original training data set; (3) Our design was evaluated by taking the average test performance for 10 trials, as the division of the data set might have varied effects on the experiment results.

Evaluation Metrics

Since our data set was obviously imbalanced, the accuracy performance cannot represent the effectiveness of our system. As a result, in our experiment, the evaluation metrics included

precision, recall, and F1-score. Precision measures the rate of ground truth classes that are predicted correctly. Recall measures the portion of each class of our prediction that is actually that class. Finally, F1-score represents the harmonic mean between precision and recall. Their formulas are as follows:

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Data Characteristics

Our study included 2 data sets. One data set was the NTUH retrospective data set, which contains a collection of the past EHRs of 268,716 visits from 2009 to 2015, and the other data set was the NTUH prospective data set, which contains data collected with patient consent in the NTUH ED from May 26, 2020, to February 21, 2022, and includes 901 ED patient records after removal of unreasonable and missing data. Table 2 summarizes the data characteristics of vital sign information in these 2 data sets. Despite similar average values across all fields in the 2 data sets, on performing statistical tests using *P*-values, we found that there was a significant difference between the 2 data sets. However, we believe that using data with the same data collection background but different distributions can still effectively improve the robustness and generalization ability of the model. By pretraining on diverse data, the model can learn more general representations, leading to improvements in the final predictions.

On the other hand, the distributions for different tasks are shown in Multimedia Appendix 5. It is worth mentioning that the distribution gap of the triage level between the retrospective data set and prospective data set was greater than the distribution gaps for hospitalization and length of stay. This is because hospitalization and length of stay are based on facts, and in contrast to the triage level in the retrospective data set, the triage level in the prospective data set comes from physician diagnosis. As it is believed that the doctor's triage level can assign patient acuity more accurately, we used it as our golden label for predicting the triage level. Another reason for the distribution gap could be the difficulty in collecting data from more severe patients.

Table 2. Patient characteristics in the National Taiwan University Hospital retrospective and prospective data sets.

Variable	NTUH[a] retrospective data set	NTUH prospective data set
Age (years), mean (SD)	49.1 (19.98)	52.4 (18.98)
Sex, n (%)		
Female	141,783 (52.8)	450 (50.1)
Male	126,933 (47.2)	450 (49.9)
Arrival time, n (%)		
7 AM to 3 PM	10,2256 (42.8)	518 (57.4)
3 PM to 11 PM	11,4970 (38.0)	289 (32.1)
11 PM to 7 AM	5,1490 (19.2)	94 (10.5)
Systolic blood pressure (mmHg), mean (SD)	136.3 (26.79)	132.4 (24.78)
Diastolic blood pressure (mmHg), mean (SD)	80.8 (15.22)	79.8 (13.91)
Pulse (beats/min), mean (SD)	88.8 (18.74)	89.5 (18.74)
Oxygen saturation (%), mean (SD)	97.0 (3.09)	97.7 (1.69)
Respiration (breaths/min), mean (SD)	18.2 (2.16)	18.8 (2.04)
Body temperature (°C), mean (SD)	37.0 (0.82)	36.7 (0.65)
Pain index (scale), n		
0	134,292	357
1-3	9,554	368
4-6	60,526	140
7-10	64,344	36

[a]NTUH: National Taiwan University Hospital.

Experimental Results

We compared our model's performance regarding triage level, hospitalization, and length of stay against the performance of other machine learning methods. As the data of only 901 ED visits were finally included in our study, it was a challenge to obtain a robust model with great capability to identify critical patients.

Unlike other work on triage level prediction, since we endeavored to fix the bias of traditional rule-based system triage, such as the ESI and TTAS, we used the diagnosis results provided by the physician as our golden label. As shown in Table 3, it is worth noting that our triage model achieved a nearly 30% improvement in 4 metrics, including accuracy, precision, recall, and F1-score, when compared to the results obtained from other models. These outstanding results show the promising potential of our proposed model.

As shown in Table 4, we can observe that our hospitalization model achieved the highest performance in 3 metrics, including precision, recall, and F1-score. Although the support vector machine (SVM) model achieved an accuracy of 91.2%, it may tend to predict the majority (discharge) owing to the low precision and recall values. From the previous discussion, it can be seen that our model is the most discriminative model.

Additionally, our proposed model outperformed other models. Although the study design and data set in our study are different from those in other studies, it is worth indicating that with the help of retrospective data pretraining, the model can learn more than with only the use of prospective data. Our proposed model achieved promising results, with 3%-6% improvement in accuracy (Table 5).

As shown in Table 6, although most of the models achieved an accuracy of higher than 70%, their performances on other metrics revealed that these models tend to predict the majority class. Nevertheless, except for accuracy, our length of stay model outperformed other machine learning methods in the other 3 metrics, indicating the capability of our length of stay model for discrimination.

Table 3. Performance comparison between our model and other machine learning methods in the “triage level” task.

Method	Accuracy	Precision	Recall	F1-score
TabNet [18]	0.425	0.436	0.410	0.423
NODE[a] [19]	0.472	0.324	0.328	0.324
Random forest [24]	0.354	0.506	0.300	0.376
XGBoost[b] [15]	0.351	0.394	0.308	0.345
SVM[c] [25]	0.340	0.581	0.268	0.367
Our model	0.633[d]	0.686[d]	0.633[d]	0.658[d]

[a]NODE: neural oblivious decision ensembles.

[b]XGBoost: extreme gradient boosting.

[c]SVM: support vector machine.

[d]Highest value.

Table 4. Performance comparison between our model and other machine learning methods in the “hospitalization” task.

Methods	Accuracy	Precision	Recall	F1-score
TabNet [18]	0.791	0.701	0.702	0.701
NODE[a] [19]	0.752	0.622	0.689	0.653
Random forest [24]	0.821	0.765	0.674	0.717
XGBoost[b] [15]	0.829	0.651	0.679	0.655
SVM[c] [25]	0.912[d]	0.456	0.500	0.477
Our model	0.822	0.811[d]	0.823[d]	0.817[d]

[a]NODE: neural oblivious decision ensembles.

[b]XGBoost: extreme gradient boosting.

[c]SVM: support vector machine.

[d]Highest value.

Table 5. Performance comparison between our model and the models in other related studies in the "hospitalization" task.

Study	Data set	Study type	Accuracy	Precision	Recall	F1-score
Study by Raita et al [26]	NHAMCS[a]	Retrospective	—[b]	—	0.750	—
Study by Yao et al [27]	NHAMCS	Retrospective	0.775	0.820[c]	0.790	0.804
Study by Leung et al [28]	NTUH[d]	Prospective	0.805	0.806	0.790	0.798
Our study	NTUH	Prospective	0.822[c]	0.811	0.823[c]	0.817[c]

[a]NHAMCS: National Hospital Ambulatory Medical Care Survey.

[b]Not reported.

[c]Highest value.

[d]NTUH: National Taiwan University Hospital.

Table 6. Performance comparison between our model and other machine learning methods in the "length of stay" task.

Methods	Accuracy	Precision	Recall	F1-score
TabNet [18]	0.683	0.654	0.665	0.659
NODE[a] [19]	0.721	0.616	0.589	0.602
Random forest [24]	0.754	0.606	0.444	0.512
XGBoost[b] [15]	0.744	0.523	0.446	0.481
SVM[c] [25]	0.791[d]	0.263	0.333	0.294
Our model	0.713	0.786[d]	0.713[d]	0.747[d]

[a]NODE: neural oblivious decision ensembles.

[b]XGBoost: extreme gradient boosting.

[c]SVM: support vector machine.

[d]Highest value.

Ablation Studies

Effectiveness of Multimodality

Experiments were conducted to demonstrate the superior performance of our proposed model. Since our model comprised the TabNet encoder and the language model encoder, we designed an experiment to show that the performance of a model leveraging both vital sign information and text information is better than that of a model using only 1 information modality. Table 7 shows that the proposed model achieved the best performance when both modalities were used. The results suggest that both structural and text data contribute to model prediction. The greater performance of the model using only tabular data than that using only text data could be attributed to the advantage of pretraining, as the vital sign encoder was pretrained with a large volume of retrospective data.

Table 7. The effectiveness of different modalities in the "triage level" task.

Methods	Accuracy	Precision	Recall	F1-score
Only tabular data	0.575	0.613	0.568	0.589
Only text data	0.439	0.119	0.250	0.162
Our method (tabular data + text data)	0.633[a]	0.686[a]	0.633[a]	0.658[a]

[a]Highest value.

Effectiveness of Multitask Training and Data Augmentation

Multitask learning experiments confirmed that the approach does offer advantages like improving data efficiency, reducing overfitting through shared representations, and allowing fast learning by leveraging auxiliary information. However, in order to obtain a more robust feature extractor, in a general setting, the targets in the multitask learning model should be related. As a result, in the experiments, we selected triage level prediction and hospitalization as our 2 outputs. It is believed that a patient assigned to level 1 or 2 should have a higher probability of admission to the hospital after being discharged from the ED. Moreover, since data distribution in triage labels is unbalanced, we attempted to narrow the distribution gap by using the method of data augmentation. Table 8 shows that both multitask learning and augmentation contributed to better performance.

Table 8. The effectiveness of different architectures in the "triage level" task.

Methods	Accuracy	Precision	Recall	F1-score
Multitask	0.500	0.369	0.500	0.425
Single task + augmentation	0.583	0.600	0.582	0.591
Single task	0.458	0.506	0.455	0.479
Our method (multitask + augmentation[a])	0.633[b]	0.686[b]	0.633[b]	0.658[b]

[a]The method of data augmentation used in our proposed model is described in the "Data Augmentation" subsection.

[b]Highest value.

Effectiveness of Different Language Models

Experiments were conducted to evaluate the performance between different language models (Table 9). In our original data set, the chief complaint was written in traditional Chinese. However, no language model has been trained on traditional Chinese. Hence, to solve this problem, we first translated the text features into different languages before sending them to the respective language models. The results showed that the model using MacBERT as the language encoder was better than models using other approaches.

Table 9. The effectiveness of different language models in the "triage level" task.

Methods	Data language	Accuracy	Precision	Recall	F1-score
Multilingual BERT[a]	Simplified Chinese	0.500	0.369	0.500	0.425
Multilingual BERT	English	0.583	0.600	0.582	0.591
BERT	English	0.458	0.506	0.455	0.479
Our method (MacBERT[b])	Simplified Chinese	0.633[c]	0.686[c]	0.633[c]	0.658[c]

[a]BERT: bidirectional encoder representations from transformers.

[b]MacBERT: Chinese version of BERT.

[c]Highest value.

Effectiveness of Different Fusion Methods

Experiments were conducted to demonstrate the superior performance of our proposed model. As our model directly concatenated the decreased embedding from the language model and the embedding from the vital sign encoder, we designed an experiment to show that it is necessary to make contributions for the text data and structural data to be comparable, and direct concatenation fusion can preserve more information than addition fusion. In Table 10, the first experiment involves the model adding 2 embeddings (text and vital sign embeddings) together with a learnable scale value to balance the gap between the text and vital sign embeddings, and the second experiment involves directly using the embedding from the language model instead of passing another fully connected network to decrease its dimension. The results suggest that making 2 embeddings to be comparable and using a direct concatenation fusion method can contribute to better performance.

Table 10. The effectiveness of different fusion methods in the "triage level" task.

Methods	Accuracy	Precision	Recall	F1-score
Experiment 1 (addition fusion)	0.548	0.580	0.547	0.563
Experiment 2 (no concatenation fusion)	0.583	0.634	0.583	0.607
Our method	0.633[a]	0.686[a]	0.633[a]	0.658[a]

[a]Highest value.

Interpretability

Although machine learning models can provide remarkably good prediction results, models need to provide explanations of the results that humans can understand easily. In our proposed model, for structural features, the attentive transformer from TabNet generated the mask to mask out different features in each decision step and observed how these features affect the model performance. As a final step, the attentive transformer calculated the importance of features by adding up the mask values of each step. On the other hand, BertViz [29] is an interactive tool that can visualize attention in transformer language models such as BERT. By acquiring attention scores from transformer layers in language models, BertViz can point out important words that contribute to the predicted result.

Multimedia Appendix 6 provides an inference example from the field test, and Multimedia Appendix 7 provides the prediction results of the inference sample for hospitalization. In this example, the patient shows acute change during the triage process, extremely high systolic and diastolic blood pressure, and an unusual Glasgow Coma Scale (GCS) score. As shown

in Multimedia Appendix 7 our system recommended admission of the inferenced patient, and the patient was actually admitted to the hospital. Our system not only successfully provided the correct suggestion to the nursing staff, but also indicated that acute change, systolic blood pressure, diastolic blood pressure, GCS-E, and GCS-M have important effects on the prediction result. As for text analysis, we used the concept from BertViz to extract attention scores for each token from the language model and visualize these attention scores. Although the language model had a hierarchy of linguistic signals from phrase to semantic features, it is believed that the deeper layer of the language model holds more information of the whole sentence [30]. Hence, we extracted the attention score from the ninth layer of the language model for further visualization (Multimedia Appendix 8).

System Application

Triage aims to prioritize patients in the ED and ration care toward those patients who need immediate care. However, recently, owing to the rising number of elderly patients and the high volume of low-acuity ED visits under waiting, patients tend to wait for very long to see the physician. This situation can cause several severe clinical outcomes such as increased mortality rates.

With the advancement in technology and popular application of computers nowadays, we wonder whether machine learning methods can help to mitigate the overcrowding problem in the ED. Therefore, we developed a triage system based on our proposed model and adopted it in the NTUH ED to provide stable and reasonable clinical AI suggestions to nursing staff. For application in the real world, we should take the running time of the system into account. The entire running time of each part is shown in Multimedia Appendix 9. The system takes no more than 10 seconds to make clinical predictions.

Before the system is officially launched, we planned a field test to ensure that the system can achieve promising performance in the real world. Finally, we included almost 6500 ED patients in our analysis from September 30, 2022, to December 30, 2022. The distributions of hospitalization and length of stay between these patients were quite different as compared to the NTUH prospective data set (Multimedia Appendix 10 and Multimedia Appendix 11). Especially for length of stay, patients who stayed in the ED for over 24 hours were much less in this data set than in both NTUH data sets (Multimedia Appendix 11). Moreover, since our golden triage level depended on the physician's diagnosis, it was challenging to label all patients in the field test; however, we evaluated our system in another way, which will be discussed later. The distribution gap between both NTUH data sets and the field test is presented in Multimedia Appendix 12.

As shown in Multimedia Appendix 13 and Multimedia Appendix 14, there was a slight performance gap between the experiments on the earlier mentioned data sets and the real-world data. However, from the results of the confusion matrix, it can be seen that in the case of "patients actually discharged," 2085 out of 2539 (82.1%) discharged patients were accurately predicted and were recommended to be discharged by the system. On the other hand, in the case of "patients actually admitted," 194 out of 316 (61.4%) patients were accurately predicted and were recommended to stay in the hospital.

As mentioned previously, for length of stay, there was a large distribution gap between our field test data set and the NTUH prospective data set. Multimedia Appendix 15 and Multimedia Appendix 16 show that the system cannot perform as good as it does in local experiments. However, from the results of the confusion matrix, we can observe that the system has a better capability of discriminating patients who stay for less than 6 hours, and the system tends to underestimate patients who stay in the ED for 6 to 24 hours.

Finally, Multimedia Appendix 17 and Multimedia Appendix 18 show that although the newly collected data did not have the golden triage level labels provided by the doctors, the distribution of the triage level indicated that the model predicted a fairly even distribution, while the system triage still mainly predicted level 3.

Discussion

Limitations

Although our proposed model showed good preliminary results compared to the results of other machine learning methods, it still has a long way to go. For instance, despite our model's ability to incorporate various language models, it may not perform well for languages where specific language models are not available in the training data set. Second, as we need to translate the text into a uniform language initially and the sentence in the data is not always complete, a better translator and some postprocessing techniques are needed to alleviate the problems. Additionally, as retrospective data lack a label in triage level prediction, expansion of the data set for training the model should help the model to learn a wider range of patterns and should enhance model performance. Moreover, since our proposed model can allow efficient learning of image or text encoders in the presence of multimodality along with tabular data, further work can add images or speech information into our model to help it achieve better performance.

Conclusion

Emergency services are an essential aspect of the health care system in hospitals, and the demand for these services has increased exponentially in recent years. Although Taiwan has established a standard process of assigning patients to different emergency levels, there is insufficient capacity to ensure precise assignment. Most patients are over-triaged or under-triaged, which can waste limited medical resources or have severe consequences such as patient mortality.

In this study, we aimed to design a deep learning prediction system that can prioritize patients and assign patients to appropriate triage levels. To obtain rich information from patients, our proposed model not only uses vital sign information, but also leverages text information.

Our system included a well-pretrained vital sign encoder and a repretrained MacBERT encoder. Additionally, by using the multitask learning and data augmentation method, we successfully obtained promising results for triage level

prediction, hospitalization prediction, and length of stay prediction. For triage level prediction, there were nearly 30% improvements in 4 metrics compared with other machine learning methods, including accuracy, precision, recall, and F1-score. Different modalities and model architectures have also been studied for ablation effectiveness. Moreover, our proposed model also provides clinicians with interpretability to understand the reasons behind the model predictions.

In conclusion, our system improved the prediction of 3 different medical outcomes when compared with other machine learning methods. With the pretrained vital sign encoder and repretrained MLM MacBERT encoder, our multimodality model can provide a deeper insight into the characteristics of EHRs. Additionally, by providing interpretability, we believe that the proposed system can assist nursing staff and physicians in taking appropriate medical decisions.

Acknowledgments

This research was supported by the Joint Research Center for AI Technology and All Vista Healthcare under the Ministry of Science and Technology of Taiwan (grants 111-2223-E-002-008 and 111-2634-E-002-021), and by the Center for Artificial Intelligence and Advanced Robotics, National Taiwan University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison between related studies and our proposed work.
[PNG File , 338 KB - medinform_v12i1e48862_app1.png]

Multimedia Appendix 2
An example of uncleaned data.
[PNG File , 54 KB - medinform_v12i1e48862_app2.png]

Multimedia Appendix 3
An example of the number change in the data using the synthetic minority oversampling technique (SMOTE) algorithm and Tomek Links algorithm.
[PNG File , 87 KB - medinform_v12i1e48862_app3.png]

Multimedia Appendix 4
The data construction of the input (left is the structural data; right is the free-text data).
[PNG File , 603 KB - medinform_v12i1e48862_app4.png]

Multimedia Appendix 5
Distribution of each task.
[PNG File , 286 KB - medinform_v12i1e48862_app5.png]

Multimedia Appendix 6
An inference example from the field test.
[PNG File , 279 KB - medinform_v12i1e48862_app6.png]

Multimedia Appendix 7
Prediction result and feature importance of the inferenced patient for hospitalization from the field test (structural data).
[PNG File , 185 KB - medinform_v12i1e48862_app7.png]

Multimedia Appendix 8
Text visualization of the inference patient for hospitalization from the field test.
[PNG File , 520 KB - medinform_v12i1e48862_app8.png]

Multimedia Appendix 9
The running time of each part in the system.
[PNG File , 85 KB - medinform_v12i1e48862_app9.png]

Multimedia Appendix 10
The distribution gap between both National Taiwan University Hospital data sets and the field test for hospitalization.
[PNG File , 42 KB - medinform_v12i1e48862_app10.png]

Multimedia Appendix 11
The distribution gap between both National Taiwan University Hospital data sets and the field test for length of stay.
[PNG File , 58 KB - medinform_v12i1e48862_app11.png]

Multimedia Appendix 12
The distribution gap between both National Taiwan University Hospital data sets and the field test.
[PNG File , 44 KB - medinform_v12i1e48862_app12.png]

Multimedia Appendix 13
The performance of hospitalization in the field test.
[PNG File , 28 KB - medinform_v12i1e48862_app13.png]

Multimedia Appendix 14
The performance (truth) of hospitalization in the field test.
[PNG File , 24 KB - medinform_v12i1e48862_app14.png]

Multimedia Appendix 15
The performance of length of stay in the field test.
[PNG File , 29 KB - medinform_v12i1e48862_app15.png]

Multimedia Appendix 16
The performance (truth) of length of stay in the field test.
[PNG File , 45 KB - medinform_v12i1e48862_app16.png]

Multimedia Appendix 17
The prediction of triage in the field test.
[PNG File , 51 KB - medinform_v12i1e48862_app17.png]

Multimedia Appendix 18
Graph showing the prediction of triage in the field test.
[PNG File , 31 KB - medinform_v12i1e48862_app18.png]

References

1. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS One 2018 Jul 20;13(7):e0201016 [FREE Full text] [doi: 10.1371/journal.pone.0201016] [Medline: 30028888]
2. Guttmann A, Schull MJ, Vermeulen MJ, Stukel TA. Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. BMJ 2011 Jun 01;342(jun01 1):d2983-d2983 [FREE Full text] [doi: 10.1136/bmj.d2983] [Medline: 21632665]
3. Bullard MJ, Unger B, Spence J, Grafstein E, CTAS National Working Group. Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) adult guidelines. CJEM 2008 Mar 21;10(2):136-151. [doi: 10.1017/s1481803500009854] [Medline: 18371252]
4. Ng C, Yen Z, Tsai JC, Chen LC, Lin SJ, Sang YY, TTAS national working group. Validation of the Taiwan triage and acuity scale: a new computerised five-level triage system. Emerg Med J 2011 Dec 12;28(12):1026-1031. [doi: 10.1136/emj.2010.094185] [Medline: 21076055]
5. Tanabe P, Travers D, Gilboy N, Rosenau A, Sierzega G, Rupp V, et al. Refining Emergency Severity Index Triage Criteria. Acad Emergency Med 2005 Jun;12(6):497-501. [doi: 10.1111/j.1553-2712.2005.tb00888.x]
6. Choi SW, Ko T, Hong KJ, Kim KH. Machine Learning-Based Prediction of Korean Triage and Acuity Scale Level in Emergency Department Patients. Healthc Inform Res 2019 Oct;25(4):305-312 [FREE Full text] [doi: 10.4258/hir.2019.25.4.305] [Medline: 31777674]
7. Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Acad Emerg Med 2004 Jan 08;11(1):59-65 [FREE Full text] [doi: 10.1197/j.aem.2003.06.013] [Medline: 14709429]

8. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Acad Emerg Med 2000 Mar 28;7(3):236-242 [FREE Full text] [doi: 10.1111/j.1553-2712.2000.tb01066.x] [Medline: 10730830]
9. Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Acad Emerg Med 2001 Feb 28;8(2):170-176 [FREE Full text] [doi: 10.1111/j.1553-2712.2001.tb01283.x] [Medline: 11157294]
10. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar. [doi: 10.3115/v1/D14-1162]
11. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv. 2013. URL: https://arxiv.org/abs/1301.3781 [accessed 2024-02-19]
12. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv. 2018. URL: https://arxiv.org/abs/1802.05365 [accessed 2024-02-19]
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2024-02-19]
14. Radford A, Narasimhan K, Salimans T, Sutskever I. mproving Language Understanding by Generative Pre-Training. OpenAI. 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2024-02-19]
15. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
16. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018 Presented at: 32nd International Conference on Neural Information Processing Systems; December 3-8, 2018; Montréal, Canada. [doi: 10.5555/3327757.3327770]
17. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach CA. [doi: 10.5555/3294996.3295074]
18. Arik S, Pfister T. TabNet: Attentive Interpretable Tabular Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2021 Presented at: AAAI Conference on Artificial Intelligence; February 2-9, 2021; Virtual p. 6679-6687. [doi: 10.1609/aaai.v35i8.16826]
19. Popov S, Morozov S, Babenko A. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. arXiv. 2019. URL: https://arxiv.org/abs/1909.06312 [accessed 2024-02-19]
20. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. arXiv. 2020. URL: https://arxiv.org/abs/2012.06678 [accessed 2024-02-19]
21. Ivanov O, Wolf L, Brecher D, Lewis E, Masek K, Montgomery K, et al. Improving ED Emergency Severity Index Acuity Assignment Using Machine Learning and Clinical Natural Language Processing. J Emerg Nurs 2021 Mar;47(2):265-278.e7 [FREE Full text] [doi: 10.1016/j.jen.2020.11.001] [Medline: 33358394]
22. Liu Y, Gao J, Liu J, Walline JH, Liu X, Zhang T, et al. Development and validation of a practical machine-learning triage algorithm for the detection of patients in need of critical care in the emergency department. Sci Rep 2021 Dec 15;11(1):24044 [FREE Full text] [doi: 10.1038/s41598-021-03104-2] [Medline: 34911945]
23. Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. 2004 Presented at: 17th International Conference on Neural Information Processing Systems; December 1, 2004; Vancouver, British Columbia, Canada. [doi: 10.5555/2976040.2976107]
24. Breiman L. Random Forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
25. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications 1998;13(4):18-28. [doi: 10.1109/5254.708428]
26. Raita Y, Goto T, Faridi M, Brown D, Camargo C, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019 Feb 22;23(1):64 [FREE Full text] [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
27. Yao LH, Leung KC, Hong JH, Tsai CL, Fu LC. A System for Predicting Hospital Admission at Emergency Department Based on Electronic Health Record Using Convolution Neural Network. 2020 Presented at: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC); October 11-14, 2020; Toronto, ON. [doi: 10.1109/SMC42975.2020.9282952]
28. Leung KC, Lin YT, Hong DY, Tsai CL, Huang CH, Fu LC. A Novel Interpretable Deep-Learning-Based System for Triage Prediction in the Emergency Department: A Prospective Study. 2021 Presented at: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC); October 17-20, 2021; Melbourne, Australia.
29. Vig J. BertViz: A tool for visualizing multihead self-attention in the BERT model. 2019 Presented at: ICLR 2019 Debugging Machine Learning Models Workshop; May 2019; New Orleans, LA.

30. Jawahar G, Sagot B, Seddah D. What Does BERT Learn about the Structure of Language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy. [doi: 10.18653/v1/P19-1356]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
CatBoost: category boosting
C-NLP: clinical natural language processing
DT: decision tree
ED: emergency department
EHR: electronic health record
ELMo: embeddings from language models
ESI: Emergency Severity Index
GCS: Glasgow Coma Scale
GPT: generative pretrained transformer
MacBERT: Chinese version of bidirectional encoder representations from transformers
MLM: mask language modeling
NLP: natural language processing
NTUH: National Taiwan University Hospital
SMOTE: synthetic minority oversampling technique
TTAS: Taiwan Triage Acuity Scale
XGBoost: extreme gradient boosting

Edited by A Benis; submitted 10.05.23; peer-reviewed by U Sinha, A Garcia Abejas, D Hu; comments to author 26.09.23; revised version received 20.11.23; accepted 05.01.24; published 01.04.24.

Please cite as:
Lin YT, Deng YX, Tsai CL, Huang CH, Fu LC
Interpretable Deep Learning System for Identifying Critical Patients Through the Prediction of Triage Level, Hospitalization, and Length of Stay: Prospective Study
JMIR Med Inform 2024;12:e48862
URL: https://medinform.jmir.org/2024/1/e48862
doi:10.2196/48862
PMID:38557661

Original Paper

Development of a Cohort Analytics Tool for Monitoring Progression Patterns in Cardiovascular Diseases: Advanced Stochastic Modeling Approach

Arindam Brahma[1*], PhD; Samir Chatterjee[2*], PhD; Kala Seal[1*], PhD; Ben Fitzpatrick[3*], PhD; Youyou Tao[1*], PhD

[1]Department of Information Systems and Business Analytics, College of Business, Loyola Marymount University, Los Angeles, CA, United States
[2]School of Information Systems and Technology, Claremont Graduate University, Claremont, CA, United States
[3]Department of Mathematics, Seaver College of Science and Engineering, Loyola Marymount University, Los Angeles, CA, United States
[*]all authors contributed equally

Corresponding Author:
Arindam Brahma, PhD
Department of Information Systems and Business Analytics
College of Business
Loyola Marymount University
1 LMU Drive
Los Angeles, CA, 90045
United States
Phone: 1 9493021030
Email: arin.brahma@lmu.edu

Abstract

Background: The World Health Organization (WHO) reported that cardiovascular diseases (CVDs) are the leading cause of death worldwide. CVDs are chronic, with complex progression patterns involving episodes of comorbidities and multimorbidities. When dealing with chronic diseases, physicians often adopt a “watchful waiting” strategy, and actions are postponed until information is available. Population-level transition probabilities and progression patterns can be revealed by applying time-variant stochastic modeling methods to longitudinal patient data from cohort studies. Inputs from CVD practitioners indicate that tools to generate and visualize cohort transition patterns have many impactful clinical applications. The resultant computational model can be embedded in digital decision support tools for clinicians. However, to date, no study has attempted to accomplish this for CVDs.

Objective: This study aims to apply advanced stochastic modeling methods to uncover the transition probabilities and progression patterns from longitudinal episodic data of patient cohorts with CVD and thereafter use the computational model to build a digital clinical cohort analytics artifact demonstrating the actionability of such models.

Methods: Our data were sourced from 9 epidemiological cohort studies by the National Heart Lung and Blood Institute and comprised chronological records of 1274 patients associated with 4839 CVD episodes across 16 years. We then used the continuous-time Markov chain method to develop our model, which offers a robust approach to time-variant transitions between disease states in chronic diseases.

Results: Our study presents time-variant transition probabilities of CVD state changes, revealing patterns of CVD progression against time. We found that the transition from myocardial infarction (MI) to stroke has the fastest transition rate (mean transition time 3, SD 0 days, because only 1 patient had a MI-to-stroke transition in the dataset), and the transition from MI to angina is the slowest (mean transition time 1457, SD 1449 days). Congestive heart failure is the most probable first episode (371/840, 44.2%), followed by stroke (216/840, 25.7%). The resultant artifact is actionable as it can act as an eHealth cohort analytics tool, helping physicians gain insights into treatment and intervention strategies. Through expert panel interviews and surveys, we found 9 application use cases of our model.

Conclusions: Past research does not provide actionable cohort-level decision support tools based on a comprehensive, 10-state, continuous-time Markov chain model to unveil complex CVD progression patterns from real-world patient data and support clinical decision-making. This paper aims to address this crucial limitation. Our stochastic model–embedded artifact can help clinicians in efficient disease monitoring and intervention decisions, guided by objective data-driven insights from real patient

data. Furthermore, the proposed model can unveil progression patterns of any chronic disease of interest by inputting only 3 data elements: a synthetic patient identifier, episode name, and episode time in days from a baseline date.

(JMIR Med Inform 2024;12:e59392) doi:10.2196/59392

KEYWORDS

healthcare analytics; eHealth; disease monitoring; cardiovascular disease; disease progression model; myocardial; stroke; decision support; continuous-time Markov chain model; stochastic model; stochastic; Markov; cardiology; cardiovascular; heart; monitoring; progression

Introduction

Chronic conditions are defined as conditions that last for 1 year or more, require continuous medical care, or limit the ability to perform daily activities. They pose significant health challenges and financial burdens in the United States and worldwide [1,2]. In the United States, 6 in 10 adults have at least 1 chronic condition, and 4 in 10 have 2 or more [1]. Worldwide, nearly one-third of adults worldwide have with multiple chronic conditions [2].

Among these chronic conditions, cardiovascular diseases (CVDs), defined by the World Health Organization (WHO) as a group of disorders of the heart and blood vessels, are a significant public health concern worldwide. The WHO reported that CVDs are the leading cause of death worldwide, with an estimated 17.9 million deaths from CVDs in 2019, accounting for 32% of all deaths worldwide [3]. The American Heart Association also reported that in 2020, there were 928,713 CVD-related deaths in the United States, making it the leading cause of death in the United States [4].

CVDs are characteristically chronic, with a pattern of progression through many stages and episodic instances over time. Moreover, they are often associated with various comorbidities such as congestive heart failure (CHF), myocardial infarctions (MI), coronary heart disease, angina, stroke, and other complications [4]. For managing complex chronic conditions such as CVDs with long progression cycles, a detailed understanding of the progression pattern of disease states over time is essential [5,6]. This knowledge not only facilitates clinical decision-making but also enables hospitals to allocate their resources better [5]. Such models developed using representative population data can enable physicians to compare a patient's progress with the patterns in the base population model to evaluate whether an intervention strategy reduces the transition probabilities between states of interest [7].

When dealing with chronic diseases, physicians often adopt a "watchful waiting" strategy and actions are postponed until information from an evolving clinical scenario is available [8]. However, data-driven clinical decision support tools with the ability to generate transition probabilities and progression paths can allow the development of effective intervention strategies at a cohort level, leading to better treatment outcomes. For this purpose, researchers have proposed and developed various quantitative disease progression models based on mathematical functions to understand the progression patterns of complex chronic diseases [6]. Quantitative disease progression models can be applied to track and describe the changes in disease progression over time and enable physicians to continually monitor and tailor treatment strategies and interventions [5,6]. Such models have significantly contributed to managing chronic progressive diseases such as Parkinson disease [5] and can be a critical precursor to policy development in cancer control [9]. Work on quantitative disease progression modeling has been found addressing various conditions such as Alzheimer disease and glaucoma [10,11], chronic kidney disease [12], abdominal aortic aneurysm [13], multiple sclerosis [14], and cardiovascular disorders, such as hypertrophic cardiomyopathy [15].

Among these quantitative disease progression models, stochastic models such as Markov models are widely applied to analyze disease processes [16-18]. Soper et al [19], for example, studied the dynamic progression of COVID-19 during the course of hospitalization using a continuous-time hidden Markov model. A Markov process can be constructed as a discrete-time Markov Chain (DTMC) model when the observations of the events are captured at a fixed recurring interval of time. However, DTMC's approach to time quantification may not be suitable for diseases requiring frequent monitoring over short periods and observation over extended spans ranging from years to decades [20]. In contrast, events can be modeled as a continuous-time Markov chain (CTMC) when the recurring periods of observations are not fixed. CTMC models offer a realistic approach, supporting state transitions at any instant in a continuous time scale [21,22]. This motivated our study to adopt CTMC as the stochastic modeling method to unveil time-variant progression patterns of CVDs. There are other stochastic modeling approaches in the disease domain. We have provided a comparative analysis of the CTMC approach with 5 other approaches in Multimedia Appendix 1.

To better understand the extent of work in CTMC applications in disease progression modeling, we conducted a literature search (period: 2000-2023) and analyzed the studies that applied CTMC—an advanced stochastic modeling method—to model disease progression. The details of the paper search process and the results are described in Multimedia Appendix 2. We found only 7 papers that used the CTMC approach to model progressions of various diseases. The diseases studied in these papers are chronic or complex, requiring long-term observation and management, such as fibrosis, myelodysplastic syndromes, diabetic foot complications, Alzheimer disease, and chronic kidney disease. Each of these papers addresses a gap in understanding disease progression. For example, Meyer et al [23] aim to estimate progression time in fibrosis stages; Nicora et al [24] used simulation to generate longitudinal event data from cross-sectional patient data and build a CTMC model to

arrive at the transition probabilities representing various stages of myelodysplastic syndrome progression. Begun et al [25] note the lack of knowledge about diabetic foot progression dynamics.

This analysis indicates that although researchers clearly recognize the need and clinical benefits of time-variant stochastic models to understand the progression of chronic diseases, only a few papers have applied the CTMC approach to model disease progression, and notably, none applied CTMC to CVDs. Most of these papers are limited in scope, focused on the disease progression encompassing a limited number of states of a single disease, and do not provide any framework for application to other chronic diseases. Furthermore, within the scope of CVDs, to the best of our knowledge, there is no paper providing actionable cohort-level analytical tools based on a comprehensive, 10-state, CTMC model to unveil complex progression patterns from real-world patient data and support clinical decision-making. Thus, our paper aims to address this crucial limitation in extant health informatics literature. Specifically, our proposed CTMC model aims not only to offer tools for clinicians to make informed intervention and treatment decisions for patients with CVDs, based on objective data-driven insights from real patient data, but also be adaptable for studying the progression of other chronic and complex diseases that require monitoring over time.

Novelties and contributions of this paper include the following:

- Application of advanced stochastic modeling, CTMC, to real-life patient cohort data can uncover new knowledge about CVD progression patterns and transition probabilities.
- Physicians can potentially use the digital data visualization system as an eHealth cohort analytics tool in CVD management, and we have found 9 impactful application use cases externally validated by a panel of 7 cardiologists.
- The proposed model can unveil progression patterns of any chronic disease of interest by inputting only 3 data elements: a synthetic patient identifier, episode name, and episode time in days from a baseline date. This would allow future researchers to generate and study disease progression patterns for other chronic diseases.
- The CVD transition probabilities can help health care administrators calculate the anticipated patient mix at different CVD states for a future planning period. This can facilitate predictive resource planning, improved patient care, and cost savings.
- Results are reproducible and extendable as the data, code, and development framework are shared with the audience via a web-based repository.

Methods

Data

The data applied in this research were from a multicenter cohort study implemented by the National Heart Lung and Blood Institute (NHLBI), collected from 9 epidemiological studies (Sleep Heart Health Study) on heart and respiratory diseases comprising 5804 patient records and 4839 CVD episodes associated with 1274 patients [26]. These longitudinal data were collected in 3 cycles across 16 years, with the first collection in 1995 and 2 subsequent collections between 1995 and 2003. CVD events, including death, were tracked until 2011. The inclusion criteria of the cohort members were aged 40 years or older, with no history of sleep apnea treatment, no tracheostomy, and no current oxygen therapy.

Table 1 provides a snapshot of occurrences of various CVD episodes or states for 2 patients (patients IDs 200453 and 201195) randomly selected from the NHLBI data used in this research. CVD “states” in our model represent specific CVD events or episodes in the patient’s history. For this reason, in our model, we define CVD states as “episodic” states or events.

It is noteworthy in the above table that multiple CVD events occurred on the same day. These are examples of comorbidity and multimorbidity in real life for patients with CVD. In the Markov process, 1 patient can be only in 1 state at a time. Hence, for the same patient, every instance of comorbidity and multimorbidity event occurring simultaneously is defined as a single Markov state. Many patients can be in any of these Markov states for a patient population with CVD, but 1 patient can only be in 1 state at any given time. Based on this Markov principle and the time of occurrence of patients’ CVD episodes in our dataset, we began with 14 unique CVD Markov states as found in our data. However, 4 states were dropped as their occurrences were negligible compared to the others. This led us to a final set of 10 unique CVD states for further development of our model. The details of the record counts in the dataset for all 14 states are presented in Multimedia Appendix 3.

Table 1. Snapshot of cardiovascular disease episodes of 2 patients from the National Heart Lung and Blood Institute dataset.

Patient ID and episode name	Episode time (days from baseline)
200453	
Myocardial infarction	572
Congestive heart failure	2064
Congestive heart failure	2562
Myocardial infarction	2562
Congestive heart failure	2593
201195	
Congestive heart failure	1343
Congestive heart failure	1426
Angina	3086
Congestive heart failure	3086
Myocardial infarction	3086
Congestive heart failure	3143
Congestive heart failure	3183

Modeling Assumptions and Approximations

The modeling assumptions and approximations can be generalized into 2 categories—Markovian assumptions and CVD data–related assumptions. They are addressed in detail below.

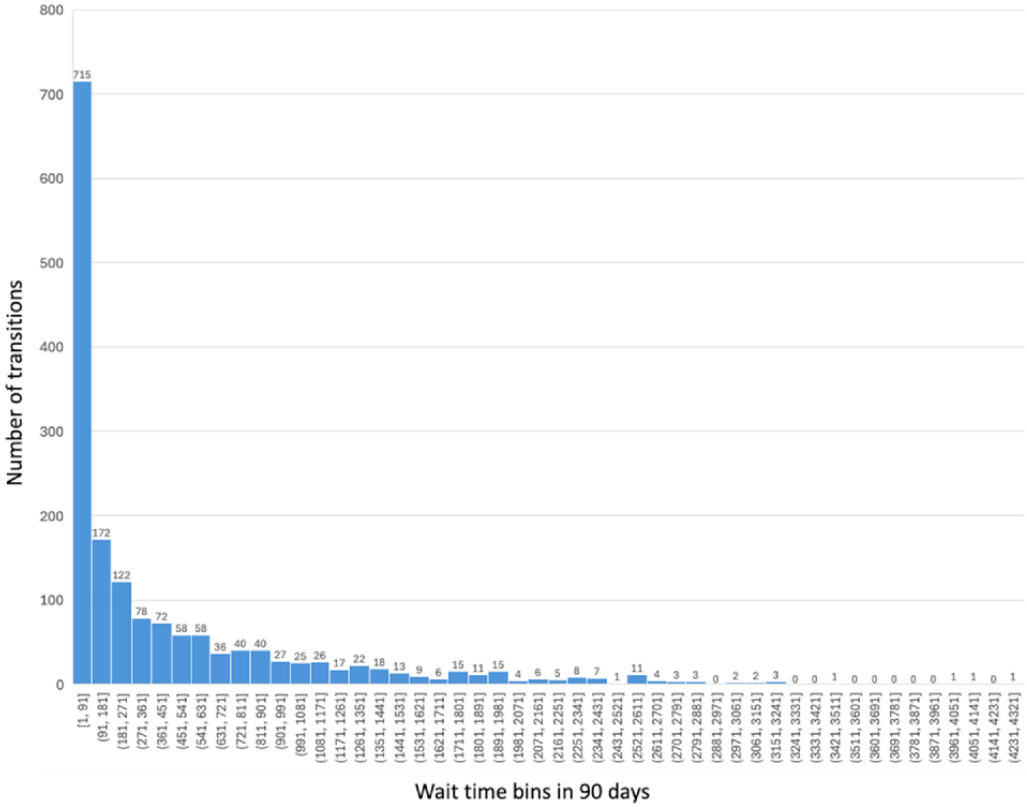
1. Finite number of Markovian states: There is a limited or finite number of possible states. States are collectively exhaustive and mutually exclusive. In CVDs, the same patient cannot be in 2 Markov states simultaneously. In our dataset of actual patients with CVD, we have found multimorbidity situations where a patient can have multiple events simultaneously. To comply with this Markovian rule, we have deliberately treated such multimorbidity events as single and unique Markov states (eg, simultaneous occurrences of angina and MI is defined as a separate unique state designated as “ANMI”).
2. Memory-less property of Markovian states: Markov states do not retain the memory of previous cycles or information from previous states leading to the future state. For example, in our model, we assume that all patients in the MI state, at the same time, have the same probability of transitioning to the angina state, irrespective of their previous history or path of reaching the previous MI state.
3. No immediate transition to the same state: Our dataset has patient instances where a patient reported the same CVD episode (eg, stroke) occurring consecutively with a time gap. In such cases, we preprocessed the data to combine them into 1 continuous state, as due to the nature of CVDs, it might not be accurate to assume that the patient had a complete remission during the interim period. For example, if a patient reported a stroke episode in day number 100 and reported a consecutive stroke again in day number 110, we assume that the patient was in the stroke state for the interim period of 10 days as well. In other words, the patient was continuously in the stroke state from 100th day to 110th day and this is counted as a single continuous disease state with longer wait time during the data preprocessing. However, if after a stroke the patient had an angina episode had a stroke again, our model state diagram would show a return to a stroke state again after the angina episode.
4. Exponential distribution approximation of wait times: Several studies and reviews support the use of exponential distribution approximations in time-to-event modeling for CVDs. For example, Sullivan [27] highlights that the exponential distribution is a common choice for modeling the time to cardiovascular events. In the application of the CTMC process in disease progression modeling for CVDs, holding time or wait time in a given state is approximated to be exponentially distributed.

Stochastic Modeling Theoretical Underpinnings

A stochastic process is a collection of random variables indexed by a variable *t*, usually representing time. A Markov Chain or Markov process is a stochastic model representing time-dependent cyclic processes. The primary advantage of a Markov process is the ability to describe, in a mathematically convenient form, the time-dependent transitions between health states [28].

As discussed in the previous section, in a disease progression process, the random time variable *t*—“wait time” or “hold time”—is known to follow an exponential distribution. As a validation, we computed our dataset’s “wait time” frequency distribution and found it to have an approximate exponential frequency distribution pattern (Figure 1), although literature indicates that it is not strictly followed in the medical field, but necessary for modeling disease progressions successfully [29].

Figure 1. Wait time–frequency distribution in the research data set.



Operational Behavior of CTMC

The operational behavior of a process modeled after CTMC can be described in the following steps, where a patient:

1. Stays X_i units of time in state i, where X_i is the random variable with an exponential distribution. The unit of time was measured in days in our study.
2. Jumps to a random next state j in a single step with probability $P(i,j)$
3. The behavior repeats across subsequent jumps.

Mathematically, CTMC can be defined by a tuple of 2 matrices: [S, R], where S is a matrix of s number of countable CTMC states and R is a transition rate matrix of ($s \times s$) size. The value of R(i,j) equals the rate at which a patient moves from state i to state j in 1 step. This R matrix is also known as the generator matrix (denoted by Q in later sections), which we will explain in the following subsection.

Generator Matrix—Q

The generator matrix, a fundamental matrix in CTMC calculations, is known by many names, such as infinitesimal generator matrix, transition rate matrix, intensity matrix, Q-matrix, etc. The generator matrix represents the rate (eg, number of transitions per day) at which transitions happen between states in 1 step. We denote it as Q-matrix with its elements as $q_{i,j}$, where

For diagonal matrix elements, $q(i,i)$ is the negative sum of the rest of the row so that the row sum of each row is equal to zero. There are several methods of calculating the generator matrix. For this study, we applied the Maximum Likelihood Estimator method, which Metzner et al [30] discussed in their paper named "Generator Estimation of Markov Jump Processes." In this method, for each state i of a finite number of states s, the generator matrix is computed as follows:

1. Calculate: $n_{i,j}$ = total number of transitions between state i to j for $i \neq j$
2. Calculate: $r_{i,j}$ = total wait time or hold time at state i before transitioning to j for $i \neq j$
3. Calculate: 
4. Place the negative of the row sum of nondiagonal positions as the diagonal entry so that the row sum of each row is zero. An infinitesimal generator matrix generates a continuous-time Markov process if and only if all off-diagonal entries are nonnegative and the sum over each row equals zero [30].

For example, if the current disease state is MI, as per our data, there are 31 cases ($n_{i,j}$) where patients have transitioned from MI to CHF state in 1 step. During the hold time or wait time computations, we found that these patients, involving 31 MI-CHF transitions, waited a total of 16,633 days ($r_{i,j}$) in the MI state before jumping to the CHF state. From these data points, $q_{i,j}$ (i=MI and j=CHF) can be calculated as 31 ÷ 16,633, which is 0.001864. This Q-matrix value is used to calculate the CTMC transition probability from MI to CHF at time t in the next step of the process.

CTMC Transition Probabilities

Transition probabilities from state *i* to *j* after waiting time *t* at state *i* are obtained by using the generator matrix, *Q*, computed earlier as $P_{i,j}(t)=e^{Qt}$, where the exponent of *e* is calculated by multiplying each element of *Q* matrix (or $q_{i,j}$) by the value of time *t*. Continuing the computation example from the previous section, if a patient is at a current state MI for 90 days, the probability of the patient transitioning to state CHF at the current time (*t*=90 days) can be calculated using the above exponential function. The calculated transition probability value from the $P_{i,j}(90)$ matrix (*i*=MI and *j*=CHF) will be 4.53% (Table 2).

Table 2. Continuous-time Markov chain (CTMC) transition probability matrix from state *i* to state *j* at a time of 90 days.

Beginning state	Immediate next state (%)										
	MI[a]	Stroke	Death	CHF[b]	CHMI[c]	CHST[d]	Angina	CHANMI[e]	ANMI[f]	CHAN[g]	Total
MI	0.63	40.18	19.60	4.53	1.15	5.47	20.23	1.74	1.27	5.21	100
Stroke	0.64	40.60	18.18	4.51	1.17	5.63	20.85	1.78	1.29	5.36	100
Death	0	0	100	0	0	0	0	0	0	0	100
CHF	0.67	22.75	13.84	33.27	1.75	3.14	14.51	1.96	1.55	6.55	100
CHMI	0.69	33.54	18.33	5.31	1.16	7.93	25.19	1.72	1.25	4.87	100
CHST	0.53	23.90	43.58	7.05	0.67	11.35	8.72	0.93	0.74	2.52	100
Angina	0.76	23.91	19.58	5.64	1.12	11.43	30.49	1.62	1.17	4.29	100
CHANMI	0.69	32.67	17.17	6.50	1.54	6.28	23.52	2.37	1.33	7.94	100
ANMI	0.64	39.05	18.56	4.85	1.29	5.14	19.94	1.88	1.34	7.32	100
CHAN	0.72	26.71	16.45	6.11	1.99	4.57	19.69	2.37	1.51	19.88	100

[a]MI: myocardial infarction.

[b]CHF: congestive heart failure.

[c]CHMI: congestive heart failure and myocardial infarction.

[d]CHST: congestive heart failure and stroke.

[e]CHANMI: congestive heart failure, angina, and myocardial infarction.

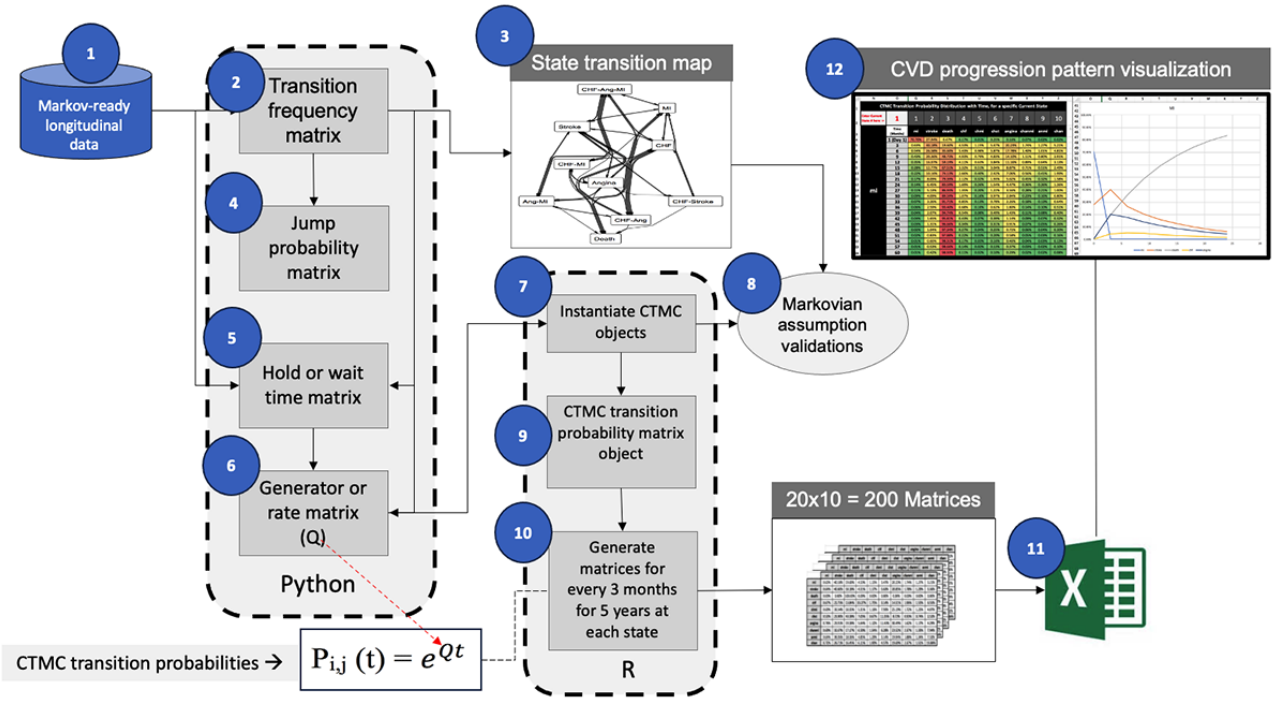
[f]ANMI: angina and myocardial infarction.

[g]CHAN: congestive heart failure and angina.

Model Development Technical Process

At this point, we want to highlight that, unlike the machine learning approach where the event probabilities are derived by training machine learning algorithms with historical data (supervised training), our approach uses a Markov Model, CTMC, which is a mathematical stochastic method to determine event probabilities and does not involve any sorts of training mechanism. Figure 2 illustrates our method used to generate a CTMC-based disease progression model for the CVDs. We have used all publicly available, open-source software tools and libraries for server-side module development (Python, Python Software Foundation, and R, R Foundation) and the widely used Microsoft Excel for front-end visualization artifacts. The *Results* section discusses further details.

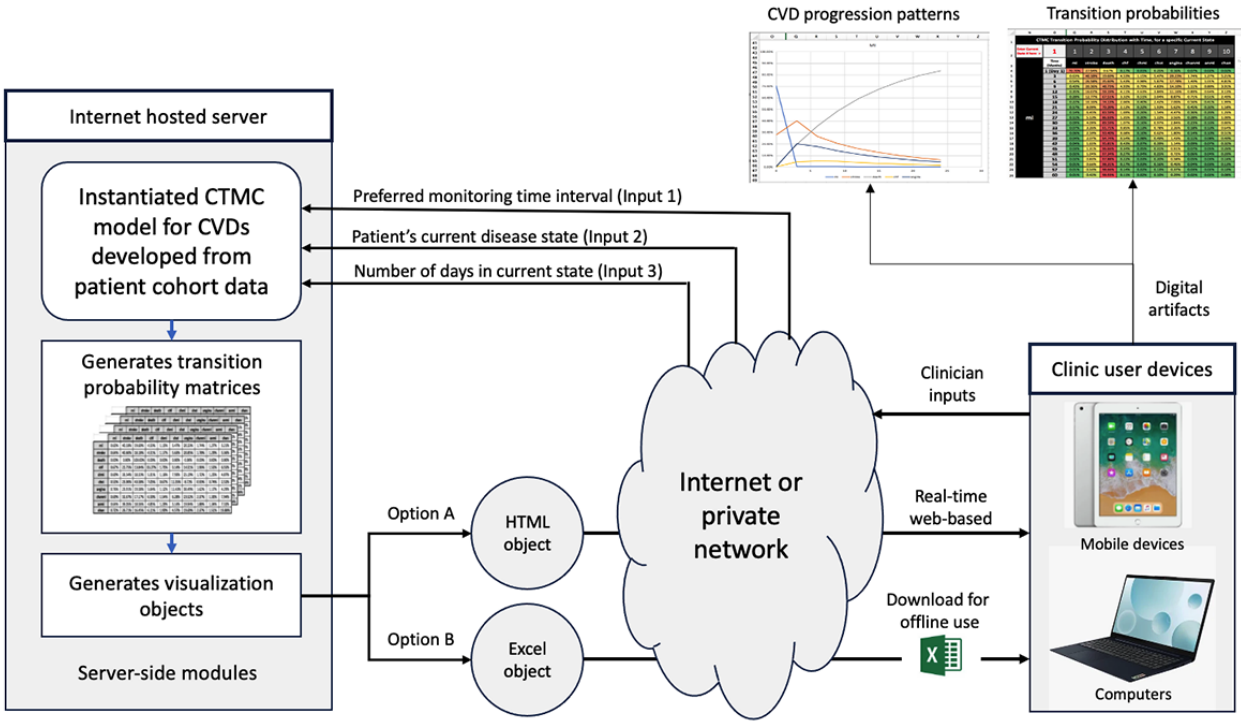
Figure 2. Development process for disease progression pattern generation and digital visualization. the circled numbers indicate the sequence of development steps and outputs. CTMC: continuous-time Markov chain; CVD: cardiovascular disease.



Deployment and User Interaction

Our developed system can be deployed as an eHealth solution for CVD clinics. In Figure 3, we present the high-level architecture of the proposed deployment configuration. Our proposed deployment design provides end users with the flexibility to use the system in real-time with an active internet or network connection using browsers, as well as the option of downloading the outputs as Excel artifacts and using them offline. We have provided the data, code, and Excel artifact (option B) via the web-based Mendeley repository [31] for reproducibility and reusability.

Figure 3. Proposed architecture for deployment of our system as an internet-based eHealth application for CVD clinicians with configurable monitoring time intervals. The architecture allows for synchronous web-based use (option A) as well as offline use via prior Excel (Microsoft Corp) artifact download (option B). CTMC: continuous-time Markov chain; CVD: cardiovascular disease.



Ethical Considerations

As discussed in this paper, cohort-level disease progression analytic tools are useful in many strategic clinical decisions. However, these are data-driven methods. Hence, ethical considerations applicable for data-driven clinical decision-making systems are important. Such considerations might include, but are not limited to, ensuring accuracy, bias-removal, fairness and equity, patient autonomy and consent, transparency and accountability, and privacy and data protection.

A Data Access and Use Agreement (the "DAUA") is executed between The Brigham and Women's Hospital, Inc., through its Division of Sleep and Circadian Disorders ("BWH") and Arin Brahma ("Data User") to facilitate access to and use of the de-identified sleep study and related covariate data originating from past NHLBI-funded research studies (the "Data"), by third-party researchers to conduct sleep research in accordance with NHLBI and BWH policies, procedures, and to the extent permitted by its Institutional Review Board (IRB) and institutional policies. The Data agreement can be made available by the authors upon request when deemed appropriate. The author has also received approval for the use of the Data for Cardiovascular Disease (CVD) research from the Institutional Review Board of Claremont Graduate University, CA, USA (Protocol ID is 3351; 01/11/2019).

Results

The results comprise the computational outputs from the method described in the previous section. They include the CVD Markov state model, various probability matrices, and the progression pattern visualization results. These are explained in detail below.

CVD Progression State Model

We generated the Markov state transition graph (see Figure 4) based on the jump frequency matrix (Table 3) information. The direction of the arrows indicates the direction of disease transition, and the thickness indicates the proportion of the patients transitioning between the states.

Figure 4 visually illustrates the chronic nature of CVDs involving various morbidity, comorbidity, and multimorbidity states. It can be observed that the state death only has incoming arrows. This property makes death an "absorbing state" in a Markov model.

Figure 4. Markov state diagram (generated from jump frequency matrix). Ang: angina; CHF: congestive heart failure; MI: myocardial infarction.

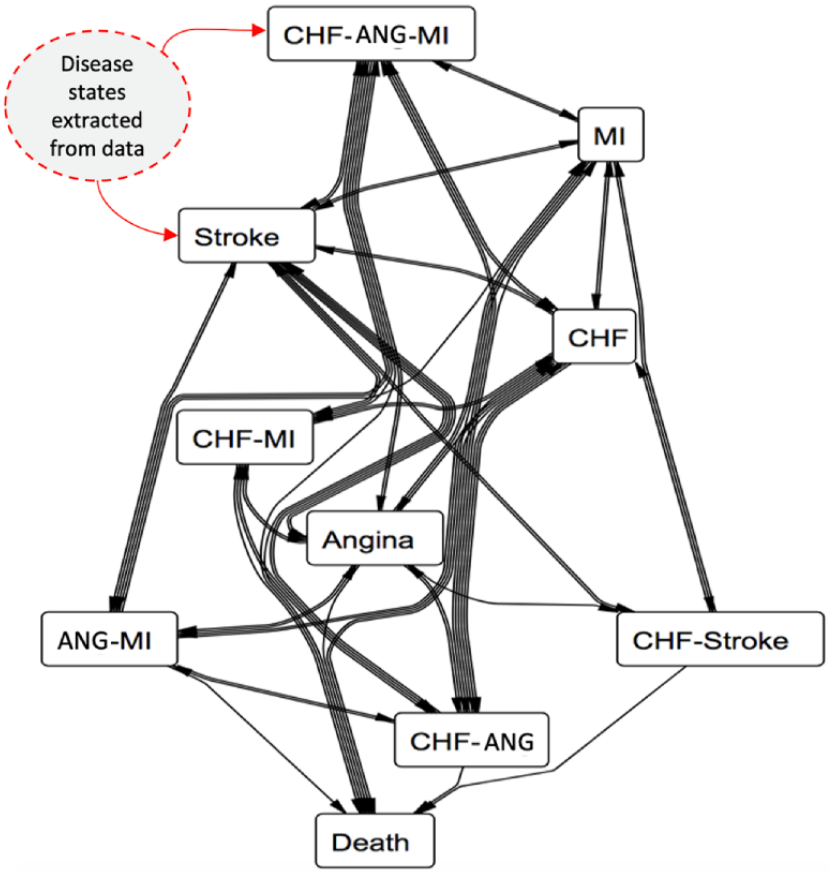


Table 3. Jump frequency matrix.

Beginning state	Immediate next state										
	MI[a]	Stroke	Death	CHF[b]	CHMI[c]	CHST[d]	Angina	CHANMI[e]	ANMI[f]	CHAN[g]	Total
MI	0	1	44	31	0	2	7	1	0	0	86
Stroke	10	0	74	18	2	9	3	6	7	2	131
Death	0	0	0	0	0	0	0	0	0	0	0
CHF	14	25	101	0	3	0	11	13	9	11	187
CHMI	0	1	8	6	0	0	2	0	0	1	18
CHST	1	6	3	4	0	0	0	0	0	0	14
Angina	9	11	12	39	1	1	0	7	7	14	101
CHANMI	7	4	10	19	1	0	13	0	4	4	62
ANMI	3	4	4	15	0	0	16	6	0	3	51
CHAN	1	2	6	19	2	0	8	5	1	0	44
Total	45	54	262	151	9	12	60	38	28	35	—[a]

[a]MI: myocardial infarction.

[b]CHF: congestive heart failure.

[c]CHMI: congestive heart failure and myocardial infarction.

[d]CHST: congestive heart failure and stroke.

[e]CHANMI: congestive heart failure, angina, and myocardial infarction.

[f]ANMI: angina and myocardial infarction.

[g]CHAN: congestive heart failure and angina.

[h]Not applicable.

CVD State Jump Probabilities

The jump probability matrix is computed from the jump frequency matrix in Table 3 and displayed in Table 4. This matrix provides the probability of a patient jumping from state *i* to state *j*, without accounting for the effect of wait time or hold time at the current state on the transition probability. The cells with a transition probability of 0% indicate that no patient transitioned between those states.

Table 4. Jump probability matrix.

Beginning state	Immediate next state (%)										
	MI[a]	Stroke	Death	CHF[b]	CHMI[c]	CHST[d]	Angina	CHANMI[e]	ANMI[f]	CHAN[g]	Total
MI	0	1.2	51.2	36	0	2.3	8.1	1.2	0	0	100
Stroke	7.6	0	56.5	13.7	1.5	6.9	2.3	4.6	5.3	1.5	100
Death	0	0	100	0	0	0	0	0	0	0	100
CHF	7.5	13.4	54	0	1.6	0	5.9	7	4.8	5.9	100
CHMI	0	5.6	44.4	33.3	0	0	11.1	0	0	5.6	100
CHST	7.1	42.9	21.4	28.6	0	0	0	0	0	0	100
Angina	8.9	10.9	11.9	38.6	1	1	0	6.9	6.9	13.9	100
CHANMI	11.3	6.5	16.1	30.6	1.6	0	21	0	6.5	6.5	100
ANMI	5.9	7.8	7.8	29.4	0	0	31.4	11.8	0	5.9	100
CHAN	2.3	4.5	13.6	43.2	4.5	0	18.2	11.4	2.3	0	100

[a]MI: myocardial infarction.

[b]CHF: congestive heart failure.

[c]CHMI: congestive heart failure and myocardial infarction.

[d]CHST: congestive heart failure and stroke.

[e]CHANMI: congestive heart failure, angina, and myocardial infarction.

[f]ANMI: angina and myocardial infarction.

[g]CHAN: congestive heart failure and angina.

CVD Continuous Time Transition Probabilities

Transition probabilities from state i to j after waiting time t at state i are obtained by using the generator matrix (Q) computed earlier as $P_{i,j}(t)=e^{Qt}$, where the exponent of e is calculated by multiplying each element of Q matrix (or $q_{i,j}$) by the value of time t. The generator matrix (Q) essentially represents the transition rates concerning time (eg, number of transitions per day) and is calculated before this step using the jump frequency and transition wait times.

If a patient is at a current state MI for 90 days, the probability of the patient transitioning to state CHF at the current time (t=90 days) can be calculated using the above exponential function. The calculated transition probability value from the $P_{i,j}(90)$ matrix (i=MI and j=CHF) will be 4.53% (Table 2).

This is the final set of matrices of the CTMC model that leads to the digital data visualization of the temporal progression pattern of CVDs. In CTMC, the transition probability from state i to state j depends on the current value of time, as the transition probability changes depending on how long a patient is at the initial state i. For every such value of t (time), a different CTMC transition probability matrix exists for all state i to state j transitions.

CVD Progression Pattern (Including Digital Data Visualization)

The CTMC transition probability matrix described above captures the transition probabilities at a specific recurring time interval scale of 3 months (90 days), as we assumed that, in general, CVD practitioners would review their patients' progress every 3 months. However, the model allows the generation of the above matrix in any granularity (continuous) time scale, such as by days, weeks, or months. Our model autogenerated the CTMC probability matrices for every current state at a 3-month incremental progression scale for 5 years. This led to twenty 10-states by 10-states transition probability matrices. We developed a data aggregation and visualization method that automatically combines and organizes the probabilities from 20 matrices for each disease on a 5-year running time scale. This transformation presents each interstate transition probability matrix (Figures 5 and 6 left) in a "heat map" style with a value-based color scale (green is lowest, yellow is medium, and red is severe). Furthermore, this also automatically generates a time series trend graph (Figures 5 and 6 right) of transition probabilities that reveal the disease progression patterns graphically over 5 years when the current disease state of a patient is inputted by a clinician through the digital interface. Using this system, a CVD practitioner can visualize the temporal pattern and various "what if" scenarios to assist in decision-making regarding treatment or intervention strategies. For example, a CVD practitioner might be interested in knowing: if a patient had MI 3 months back (means current time t=3 months), what would be the probability of the patient transitioning to stroke or angina? From the following CTMC matrix generated for t=3 months, or 90 days by our model (Table 2), one can observe that if the current state of the patient is MI (rows), the probability of having a stroke is 40.18% (highest) and angina 20.23% (next highest). Hence, the CVD practitioner might recommend treatment or lifestyle interventions so that the risk of stroke or angina can be reduced.

Figure 5. Digital data visualization: 3-month interval transition probabilities (left) and graphical progression pattern (right). Current state entered is 1 (MI)—best viewed in color. ANMI: angina and myocardial infarction; CHAN: congestive heart failure and angina; CHANMI: congestive heart failure, angina, and myocardial infarction; CHF: congestive heart failure; CHMI: congestive heart failure and myocardial infarction; CHST: congestive heart failure and stroke; CTMC: continuous-time Markov chain; MI: myocardial infarction.

CTMC transition probability distribution with time, for a specific current state

Enter current state # here ->	1	1	2	3	4	5	6	7	8	9	10
	Time (Months)	MI (%)	stroke (%)	death (%)	CHF (%)	CHMI (%)	CHST (%)	angina (%)	CHANMI (%)	ANMI (%)	CHAN (%)
MI	0	70.70	27.94	0.67	0.17	0.01	0.25	0.16	0.07	0.02	0.02
	3	0.63	40.18	19.60	4.53	1.15	5.47	20.23	1.74	1.27	5.21
	6	0.54	26.58	35.60	5.43	0.98	5.87	17.78	1.40	1.01	4.81
	9	0.43	20.36	48.73	4.93	0.79	4.83	14.10	1.11	0.80	3.91
	12	0.35	16.07	59.19	4.11	0.63	3.84	11.16	0.89	0.64	3.13
	15	0.28	12.77	67.51	3.32	0.51	3.04	8.87	0.71	0.51	2.49
	18	0.22	10.16	74.13	2.66	0.40	2.42	7.06	0.56	0.41	1.99
	21	0.17	8.09	79.39	2.12	0.32	1.93	5.62	0.45	0.32	1.58
	24	0.14	6.45	83.59	1.69	0.26	1.54	4.47	0.36	0.26	1.26
	27	0.11	5.13	86.93	1.35	0.20	1.22	3.56	0.28	0.21	1.00
	30	0.09	4.09	89.59	1.07	0.16	0.97	2.84	0.23	0.16	0.80
	33	0.07	3.26	91.71	0.85	0.13	0.78	2.26	0.18	0.13	0.64
	36	0.06	2.59	93.40	0.68	0.10	0.62	1.80	0.14	0.10	0.51
	39	0.04	2.07	94.74	0.54	0.08	0.49	1.43	0.11	0.08	0.40
	42	0.04	1.65	95.81	0.43	0.07	0.39	1.14	0.09	0.07	0.32
	45	0.03	1.31	96.66	0.34	0.05	0.31	0.91	0.07	0.05	0.26
	48	0.02	1.04	97.34	0.27	0.04	0.25	0.72	0.06	0.04	0.20
	51	0.02	0.83	97.88	0.22	0.03	0.20	0.58	0.05	0.03	0.16
	54	0.01	0.66	98.31	0.17	0.03	0.16	0.46	0.04	0.03	0.13
	57	0.01	0.53	98.66	0.14	0.02	0.13	0.37	0.03	0.02	0.10
	60	0.01	0.42	98.93	0.11	0.02	0.10	0.29	0.02	0.02	0.08

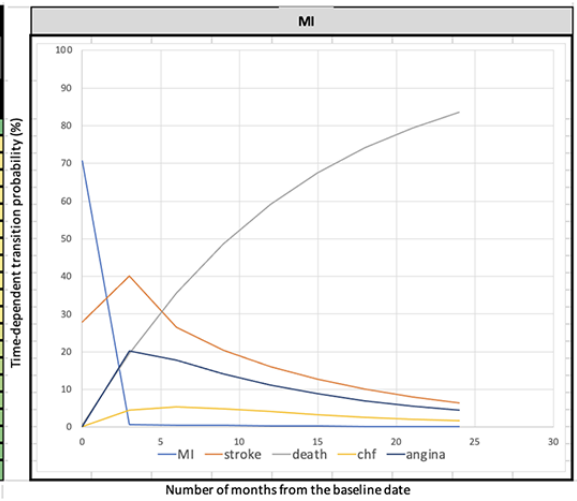


Figure 6. Digital data visualization: 3-month interval transition probabilities (left) and graphical progression pattern (right). Current state entered is 2 (stroke)—best viewed in color. ANMI: angina and myocardial infarction; CHAN: congestive heart failure and angina; CHANMI: congestive heart failure, angina, and myocardial infarction; CHF: congestive heart failure; CHMI: congestive heart failure and myocardial infarction; CHST: congestive heart failure and stroke; CTMC: continuous-time Markov chain; MI: myocardial infarction.

CTMC transition probability distribution with time, for a specific current state

Enter current state # here ->	2	1	2	3	4	5	6	7	8	9	10
	Time (Months)	MI (%)	stroke (%)	death (%)	CHF (%)	CHMI (%)	CHST (%)	angina (%)	CHANMI (%)	ANMI (%)	CHAN (%)
stroke	0	0.06	98.46	0.22	0.08	0.08	0.06	0.66	0.11	0.14	0.14
	3	0.64	40.60	18.18	4.51	1.17	5.63	20.85	1.78	1.29	5.36
	6	0.55	27.00	34.48	5.50	1.00	6.00	18.12	1.42	1.03	4.90
	9	0.44	20.71	47.85	5.01	0.81	4.92	14.34	1.13	0.82	3.98
	12	0.35	16.35	58.49	4.18	0.65	3.90	11.35	0.90	0.65	3.18
	15	0.28	12.99	66.95	3.38	0.51	3.10	9.02	0.72	0.52	2.54
	18	0.22	10.34	73.68	2.70	0.41	2.46	7.18	0.57	0.41	2.02
	21	0.18	8.23	79.04	2.16	0.33	1.96	5.71	0.46	0.33	1.61
	24	0.14	6.56	83.31	1.72	0.26	1.56	4.55	0.36	0.26	1.28
	27	0.11	5.22	86.70	1.37	0.21	1.24	3.62	0.29	0.21	1.02
	30	0.09	4.16	89.41	1.09	0.17	0.99	2.89	0.23	0.17	0.81
	33	0.07	3.31	91.57	0.87	0.13	0.79	2.30	0.18	0.13	0.65
	36	0.06	2.64	93.28	0.69	0.11	0.63	1.83	0.15	0.11	0.52
	39	0.05	2.10	94.65	0.55	0.08	0.50	1.46	0.12	0.08	0.41
	42	0.04	1.67	95.74	0.44	0.07	0.40	1.16	0.09	0.07	0.33
	45	0.03	1.33	96.61	0.35	0.05	0.32	0.93	0.07	0.05	0.26
	48	0.02	1.06	97.30	0.28	0.04	0.25	0.74	0.06	0.04	0.21
	51	0.02	0.85	97.85	0.22	0.03	0.20	0.59	0.05	0.03	0.17
	54	0.01	0.67	98.29	0.18	0.03	0.16	0.47	0.04	0.03	0.13
	57	0.01	0.54	98.63	0.14	0.02	0.13	0.37	0.03	0.02	0.11
	60	0.01	0.43	98.91	0.11	0.02	0.10	0.30	0.02	0.02	0.08

Analysis of CVD Progression Pattern

We explain our observations using an illustrative example (Figure 5—left and right), where the current CVD state of a hypothetical patient (representing population CVD progression pattern) is MI at *t*=1 (day 1 of month 1), the probability of staying in that state drops rapidly with time. In contrast, the probability of stroke, death, and CHF keeps increasing. Among these states, the probability of stroke starts with a higher probability of 28% at *t*=month 1. Continuing the analysis from Figure 5, the probability of death and angina starts at near zero but increases at a higher rate. At *t*=month 3, the probability of stroke, angina, death, and CHF peak at 40%, 20%, 20%, and 5%, respectively. At this point, if the patient transitions to stroke (40%), there is a 36% probability of death at *t*=month 6, compared to an 18% probability of angina and a 6% probability of CHF. If the patient escapes death and continues to be in the stroke state, then the probability of death crosses the 50% mark at *t*=month 9, and the death probability increases rapidly from then onwards. In 21 months, the death probability crosses the 80% mark, signifying a high probability of mortality.

Although the above findings meet the core objectives of this study, we performed further analysis of the dataset, which contributed additional interesting findings on CVD patterns. We use only 4 states—CHF, stroke, MI, and angina—for this analysis as they collectively account for 88.2% (231/262) of all CVD-related deaths (101/262, 38.6%; 74/262, 28.2%; 44/262, 16.8%; and 12/262, 4.6%, respectively). The details of this analysis are presented in Multimedia Appendix 4, and a summary of the key findings from this analysis is presented.

1. Transition rates and mean transition times: The fastest transition rate is from MI to stroke (mean transition time 3, SD 0 days, because only 1 patient had myocardial infarction to stroke transition in the dataset), and the slowest is from MI to angina (mean transition time 1457, SD 1449 days).

2. First CVD episodes: CHF is the most probable first episode (371/840, 44.2%), followed by stroke (216/840, 25.7%), angina (140/840, 16.7%), and MI (113/840, 13.5%).
3. Most dominant CVD episodes immediately after an episode: CHF is the most dominant episode after an occurrence of angina, whereas after CHF, stroke, and MI, death is the most dominant episode.

Model Validation

As practiced commonly in validating empirical models, such as machine learning predictive models, one would ideally hold some testing data out to compare to a model built on training data. Fitzpatrick [31], in his paper "Issues in Reproducible Simulation Research," states that the problem context of probabilistic models can often make this type of testing difficult. He discussed this aspect of stochastic model validations and laid a guideline for researchers working with such models. In the paper, Fitzpatrick refers to a multistage validation model from North and Macal [32] that includes validation of requirements, data, face, process, theory, and output. We used this model for internal evaluations during our model development cycles, as described in detail in Multimedia Appendix 5.

Model Artifacts, Reproducibility, and Generalizability

One of the significant outputs of this research is the CTMC model embedded visualization artifact that CVD clinicians can use as a cohort-level decision-support tool for several applications and use cases as described in detail in Multimedia Appendix 6. The user interface of this 10-state model artifact has an input field (pointed out by the text "Enter Current State# here " in Figures 5 and 6), where a clinician can be informed of the transition probability values and graphical progression patterns by entering the current CVD state of a cohort under investigation. This Excel artifact is self-contained, ready to use, and can be operated by having just the Microsoft Excel application on any computer. The artifact can be downloaded from the code repository [29].

Multimedia Appendix 7 provides more detailed information and instructions for future researchers interested in reusing these research methods, codes, and outputs to investigate the progression patterns of other chronic diseases or CVDs from different population datasets.

Discussion

Principal Findings

This research applies a well-established stochastic modeling—CTMC models on real-life patient cohort data to uncover previously unknown knowledge about CVD transition probabilities and complex progression patterns. The choice of CTMC overcame DTMC's limitations on time quantification for disease progression modeling. There are other types of Markov models applied for stochastic modeling in the disease domain. The choice of appropriate modeling method is often driven by the model's application goals (eg, patient-specific decisions vs cohort-level decisions). We have provided a detailed comparison of our choice of CTMC with 6 other methods in Multimedia Appendix 1. For complex chronic diseases such as CVDs, the availability of population-level transition probability matrices of all 10 disease states and respective progression patterns close a crucial gap in the literature and opens the door for future research on the disease progression of CVDs.

We find that the transition from MI to stroke has the fastest transition rate (mean transition time 3, SD 0 days, because only 1 patient had myocardial infarction to stroke transition in the dataset), and MI to angina is the slowest (mean transition time 1457, SD 1449 days). CHF is the most probable first episode (371/840, 44.2%), followed by stroke (216/840, 25.7%). This result reflects the epidemiological characteristics of our study cohort. However, if a dataset from a different cohort is inputted in our model, the patterns might differ.

External Validations of Clinical Applications and Usefulness

The research also has many practical implications. We organized an expert panel of 7 practicing cardiologists and conducted a combination of presentations, open-ended interviews, and structured web-based surveys to validate the applications and usefulness of our model. We have discussed the methods, results, and analysis of this external validation process in detail in Multimedia Appendix 6. Here, we are presenting 9 clinical applications and use cases that were evident from the interviews and survey.

1. This tool can help cardiologic clinics understand the effectiveness of their intervention strategies related to quality metrics, such as blood pressure thresholds for interventions.
2. Comparing the clinic-level transition rates with population-level rates built into this tool, clinics can identify a specific cohort that has significantly higher rates for certain transitions (eg, angina-to-death) and then back-identify the patients for analysis of variables leading to such outcomes.
3. Clinics can use the transition probabilities generated by this tool based on their patient cohort data to compute the patient state mix for a future period. This can allow predictive resource planning, improved patient care, and cost savings.
4. This system can be populated with CVD episode data of various clinic cohorts and subcohorts and their patterns (frequency, transition rates, probabilities, etc) can be compared. This can reveal the trend differences concerning various cohort control features such as demographics, education, nutrition profile, treatment alternatives, etc.
5. Such intercohort benchmarking can delineate best practices concerning cost-effectiveness, manpower economy, and fund allocations (eg, Affordable Care Act incentives) when the candidate cohorts are selected from disparate clinics.
6. The utility can be further extended to understand epidemiological trends of different geographical and population segments, leading to valuable inputs for the health care policymakers and administrators enabling them to target specific regions based on pattern differences.
7. This tool can also help comparison of CVD characteristics at a national level, for example, some countries might have more stroke compared to others having more CHF. This

can lead to necessary preventive measures and national health policies.

8. It can help identify temporal changes in CVD trends in patients because of shifts in major treatment paradigms (eg, prestatin vs statin period pattern changes)
9. This tool can be potentially integrated with clinic electronic health record systems to continually monitor temporal pattern shifts over time and send out automated notifications and reports based on preset quality metrics thresholds.

Study of Impact of Various Risk and Control Factors on Epidemiological Trends

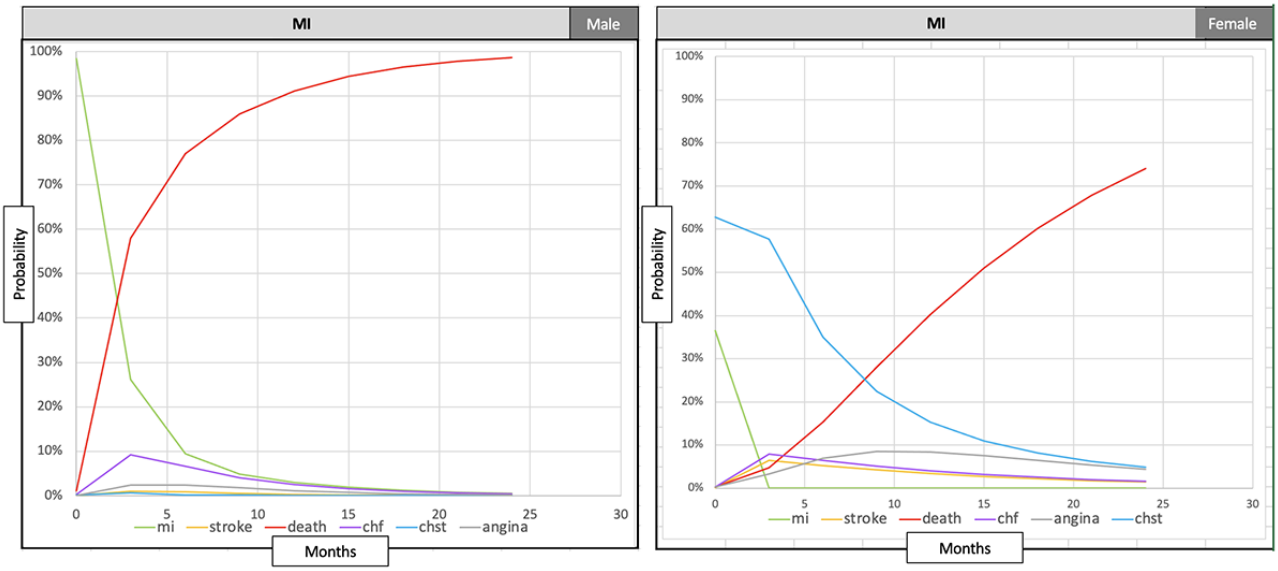
The clinical applications suggested by the expert panel point to the tool's general ability to reveal epidemiological trend differences across various subcategories within a given cohort or across different cohorts. Such subcategories might include various population risk factor groups (age, gender, smoking behavior, education levels, etc), geographical attributes (urban, rural, coastal, inland, state, country, etc), treatment policies (quality metrics), interventions (medication or procedures), and so on.

To demonstrate this key capability of our tool, we chose a population risk factor—"gender"—as an example and used the tool to visualize the CVD progression trend differences between male and female patients from our research cohort. To accomplish this, we partitioned the input dataset based on the "gender" subcategory. The data partitioning based on "gender" resulted in 278 male patients involved in 471 CVD episodes and 355 female patients involved in 592 CVD episodes. We then applied our research tool to generate 2 separate outputs corresponding to each of the genders. The visualization of the results (Figure 7) reveals distinct and interesting differences in the progression patterns between male and female patients within the cohort. In the following paragraph, we briefly outline some of the significant trend differences revealed between male and female patients when the current CVD state is "MI."

First, we observe that the transition rate to state "death" (in red color) is significantly slower for the female patients compared to the male patients. The progression curve for females is noticeably flatter compared to their male counterparts. Thus, the female patients take almost 15 months to cross the probability threshold of 50%, whereas the male patients cross that threshold in just about 2.5 months. Another key observable difference is revealed in the progression pattern of the multimorbidity state "CHST" (congestive heart failure and stroke in blue color). The graph shows that female patients have more than a 50% probability of transitioning to "CHST" immediately following their current "MI" state over the next 4 months. In contrast, male patients have negligible probabilities of transitioning to "CHST" following the current "MI" state. Also, the probabilities of transitioning out of the current "MI" state (in green color) to any other state have distinctly different patterns between the male and female patient groups. Such trend comparisons between subgroups or different cohorts might trigger further study and analysis by the clinicians leading to improved treatment strategies, quality metrics, or interventions.

In the *Deployment and User Interaction* section, we have also provided a proposed deployment architecture that will allow health care providers to implement the solution as an eHealth application over the internet. Furthermore, our cohort-level clinical decision support system can be used for other chronic diseases to unveil their progression patterns, requiring only 3 data elements as inputs: a synthetic patient identifier, disease episode name, and episode time in days from a baseline date. This approach can encourage and facilitate further studies on disease progression patterns of other chronic diseases, not limited to CVDs by future researchers.

Figure 7. CVD transition probability pattern differences between male and female patients when the patients' current state is "MI." CHF: congestive heart failure; CHST: congestive heart failure and stroke; CTMC: continuous-time Markov chain; MI: myocardial infarction.



Limitations

Our study has a few limitations. As per the Markovian assumptions, the process's behavior after any cycle is solely based on its state within that cycle, indicating that it does not retain the memory of previous cycles. This assumption about the memory-less state is tenuous in medicine since a patient's

history may give the physician important information about their condition in the present and future. Hence, this is an approximation in our model. It is noted from the literature that although the Markovian assumption is necessary to model prognosis with a finite number of states, it is not followed strictly in medical problems [33]. Next, the data used for this research comprised CVD episode records of 1274 patients. The generalizability of this model will significantly improve with a larger number of diverse patients with CVD data. The CTMC approach has its limitations as well for patient-specific decision support applications. Since a patient is likely to be associated with a specific medical history, medication, interventions, demographic risk factors, etc, each of the Markov states must be unique for all such combinations. This makes the design and computation of the CTMC model very complex, less interpretable, and computationally heavy. Other Markov methods such as the Markov decision process can model patient-specific sequential treatment decision-making processes more efficiently. These limitations also lead to future opportunities for research. Since the framework and the tool developed in this research are disease agnostic, reproducibility-tested, and the reproducible package has been publicly shared, future researchers can use this tool to populate with episodic data of any chronic disease of interest and apply them to similar clinical use cases and applications. Additionally, to improve the generalizability of its application for CVDs, data from various other cohorts can be modeled and validated by simply reusing our model with new datasets.

Conclusion

Stochastic disease progression models developed from fully observed real patient cohort data to compare a patient's CVD progression with a population pattern can provide better intervention-based decision-making capabilities to physicians. However, such models do not currently exist. This research uses CTMC methods to develop a disease progression model from the population data of 1274 patients associated with 4839 CVD episodes across 16 years. This study unveiled distinct CVD progression patterns and characteristics from the fully observed longitudinal data of patient cohorts. The results are actionable with the code and data framework shared with the audience. The model also serves as an eHealth decision support tool with digital visualization of progression patterns and opens the door for many practical applications, including proactive resource planning at hospitals. Our study results are reproducible and extendable to other chronic diseases. Despite certain limitations, this research contributes significantly to the literature and possible practical clinical applications.

Acknowledgments

We would like to thank Luanda Grazette, MD, Vishal Gupta, MD, George Ponce, MD, Jayanta Datta, MD, Wesam Alhejily, MD, Shilanjan Roy, MD, Satyajit Bose, MD, and Michael Neely, MD for providing us with valuable insights and feedback on application use cases and usefulness of our model. We also thank graduate research assistant Sebastian Van Hemert for his contributions to the literature search, testing the model's reproducibility, and developing code documentation for reproducibility.

Data Availability

The master dataset was obtained from a publicly available source—Sleep Heart Health Study by the National Heart Lung and Blood Institute (NHLBI) [26]. The preprocessed version of the dataset used for the continuous-time Markov chain (CTMC) model development is available at the publicly accessible web-based Mendeley repository [29].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Markov model types applied in the disease domain.
[DOCX File , 20 KB - medinform_v12i1e59392_app1.docx]

Multimedia Appendix 2
Literature search process and results.
[DOCX File , 22 KB - medinform_v12i1e59392_app2.docx]

Multimedia Appendix 3
CVD states and their occurrences. CVD: cardiovascular disease.
[DOCX File , 16 KB - medinform_v12i1e59392_app3.docx]

Multimedia Appendix 4
Additional analysis of CVD patterns. CVD: cardiovascular diseases.
[DOCX File , 70 KB - medinform_v12i1e59392_app4.docx]

Multimedia Appendix 5
Model validation.
[DOCX File , 21 KB - medinform_v12i1e59392_app5.docx]

Multimedia Appendix 6
Cardiologist expert panel interviews and web-based survey.
[DOCX File , 400 KB - medinform_v12i1e59392_app6.docx]

Multimedia Appendix 7
Model reproducibility information.
[DOCX File , 215 KB - medinform_v12i1e59392_app7.docx]

References

1. About chronic diseases. National Center for Chronic Disease Prevention and Health Promotion. 2022. URL: https://www.cdc.gov/chronic-disease/about/?CDC_AAref_Val=https://www.cdc.gov/chronicdisease/about/index.htm [accessed 2024-08-21]
2. Hajat C, Stein E. The global burden of multiple chronic conditions: a narrative review. Prev Med Rep 2018;12:284-293 [FREE Full text] [doi: 10.1016/j.pmedr.2018.10.008] [Medline: 30406006]
3. Cardiovascular diseases (CVDs). World Health Organization. 2021. URL: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [accessed 2024-08-21]
4. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. Circulation 2023;147(8):e93-e621. [doi: 10.1161/CIR.0000000000001123] [Medline: 36695182]
5. Venuto CS, Potter NB, Dorsey ER, Kieburtz K. A review of disease progression models of Parkinson's disease and applications in clinical trials. Mov Disord 2016;31(7):947-956 [FREE Full text] [doi: 10.1002/mds.26644] [Medline: 27226141]
6. Cook SF, Bies RR. Disease progression modeling: key concepts and recent developments. Curr Pharmacol Rep 2016;2(5):221-230 [FREE Full text] [doi: 10.1007/s40495-016-0066-x] [Medline: 28936389]
7. Kazmi S, Kambhampati C, Cleland JG, Cuthbert J, Kazmi KS, Pellicori P, et al. Dynamic risk stratification using Markov chain modelling in patients with chronic heart failure. ESC Heart Fail 2022;9(5):3009-3018 [FREE Full text] [doi: 10.1002/ehf2.14028] [Medline: 35736536]
8. Magni P, Quaglini S, Marchetti M, Barosi G. Deciding when to intervene: a Markov decision process approach. Int J Med Inf 2000;60(3):237-253 [FREE Full text] [doi: 10.1016/s1386-5056(00)00099-x]
9. Gulati R, Inoue L, Katcher J, Hazelton W, Etzioni R. Calibrating disease progression models using population data: a critical precursor to policy development in cancer control. Biostatistics 2010;11(4):707-719 [FREE Full text] [doi: 10.1093/biostatistics/kxq036] [Medline: 20530126]
10. Zhou J, Liu J, Narayan VA, Ye J. Modeling disease progression via fused sparse group lasso. 2012 Presented at: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 12-16, 2012; Beijing, China. [doi: 10.1145/2339530.2339702]
11. Sukkar R, Katz E, Zhang Y, Raunig D, Wyman B. Disease progression modeling using Hidden Markov models. Annu Int Conf IEEE Eng Med Biol Soc 2012;2012:2845-2848. [doi: 10.1109/EMBC.2012.6346556] [Medline: 23366517]
12. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, et al. A predictive model for progression of chronic kidney disease to kidney failure. JAMA 2011;305(15):1553-1559. [doi: 10.1001/jama.2011.451] [Medline: 21482743]
13. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multistate Markov models for disease progression with classification error. Statistician 2003;52((Part 2)):193-209. [doi: 10.1111/1467-9884.00351]
14. Ziemssen T, Piani-Meier D, Bennett B, Johnson C, Tinsley K, Trigg A, et al. A physician-completed digital tool for evaluating disease progression (Multiple Sclerosis Progression Discussion Tool): validation study. J Med Internet Res 2020;22(2):e16932 [FREE Full text] [doi: 10.2196/16932] [Medline: 32049062]
15. Pičulin M, Smole T, Žunkovič B, Kokalj E, Robnik-Šikonja M, Kukar M, et al. Disease progression of hypertrophic cardiomyopathy: modeling using machine learning. JMIR Med Inform 2022;10(2):e30483 [FREE Full text] [doi: 10.2196/30483] [Medline: 35107432]
16. Hovinen E, Kekki M, Kuikka S. A theory to the stochastic dynamic model building for chronic progressive disease processes with an application to chronic gastritis. J Theor Biol 1976;57(1):131-152. [doi: 10.1016/s0022-5193(76)80009-4] [Medline: 957649]
17. Niño-Torres D, Ríos-Gutiérrez A, Arunachalam V, Ohajunwa C, Seshaiyer P. Stochastic modeling, analysis, and simulation of the COVID-19 pandemic with explicit behavioral changes in Bogotá: a case study. Infect Dis Model 2022;7(1):199-211 [FREE Full text] [doi: 10.1016/j.idm.2021.12.008] [Medline: 35005324]

18. He S, Tang S, Rong L. A discrete stochastic model of the COVID-19 outbreak: forecast and control. Math Biosci Eng 2020;17(4):2792-2804 [FREE Full text] [doi: 10.3934/mbe.2020153] [Medline: 32987496]
19. Soper BC, Cadena J, Nguyen S, Chan KHR, Kiszka P, Womack L, et al. Dynamic modeling of hospitalized COVID-19 patients reveals disease state-dependent risk factors. J Am Med Inform Assoc 2022;29(5):864-872 [FREE Full text] [doi: 10.1093/jamia/ocac012] [Medline: 35137149]
20. Leiva-Murillo JM, Rodríguez AA, Baca-García E. Visualization and prediction of disease interactions with continuous-time Hidden Markov models. 2011 Presented at: NIPS 2011 Workshop on Personalized Medicine; December 12-17, 2011; Granada, Spain URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=7ac2ecf7b26afacdc813d3fe39c1ddfb4d019bfa/>
21. Begun A, Icks A, Waldeyer R, Landwehr S, Koch M, Giani G. Identification of a multistate continuous-time nonhomogeneous Markov chain model for patients with decreased renal function. Med Decis Making 2013;33(2):298-306. [doi: 10.1177/0272989X12466731] [Medline: 23275452]
22. Zadeh A, Jeyaraj A. A multistate modeling approach for organizational cybersecurity exploration and exploitation. Decis Support Syst 2022;162:113849 [FREE Full text] [doi: 10.1016/j.dss.2022.113849]
23. Meyer LF, Musante CJ, Allen R. A continuous-time Markov chain model of fibrosis progression in NAFLD and NASH. Front Med (Lausanne) 2023;10:1130890 [FREE Full text] [doi: 10.3389/fmed.2023.1130890] [Medline: 37324150]
24. Nicora G, Moretti F, Sauta E, Della Porta M, Malcovati L, Cazzola M, et al. A continuous-time Markov model approach for modeling myelodysplastic syndromes progression from cross-sectional data. J Biomed Inform 2020;104:103398 [FREE Full text] [doi: 10.1016/j.jbi.2020.103398] [Medline: 32113003]
25. Begun A, Morbach S, Rümenapf G, Icks A. Study of disease progression and relevant risk factors in diabetic foot patients using a multistate continuous-time Markov chain model. PLoS One 2016;11(1):e0147533 [FREE Full text] [doi: 10.1371/journal.pone.0147533] [Medline: 26814723]
26. Sleep heart health study. National Sleep Research Resource. URL: https://sleepdata.org/datasets/shhs [accessed 2024-08-21]
27. Sullivan L. Survival analysis. Boston University School of Public Health. 2024. URL: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival_print.html [accessed 2024-12-06]
28. Cher DJ, Lenert LA. Rapid approximation of confidence intervals for Markov process decision models: applications in decision support systems. J Am Med Inform Assoc 1997;4(4):301-312 [FREE Full text] [doi: 10.1136/jamia.1997.0040301] [Medline: 9223036]
29. Brahma A, Chatterjee S, Tao Y, Seal KC, Fitzpatrick B. Markov CTMC CVD progression framework compressed data and code for JMIR publication. Mendeley Data. 2024. URL: https://data.mendeley.com/datasets/jjx685vj6k/2 [accessed 2024-08-24]
30. Metzner P, Dittmer E, Jahnke T, Schütte C. Generator estimation of Markov jump processes. J Comput Phys 2007;227(1):353-375 [FREE Full text] [doi: 10.1016/j.jcp.2007.07.032]
31. Fitzpatrick B. Issues in reproducible simulation research. Bull Math Biol 2019;81(1):1-6 [FREE Full text] [doi: 10.1007/s11538-018-0496-1] [Medline: 30191471]
32. North M, Macal C. Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation. USA: Oxford University Press; 2007:313.
33. Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. Med Decis Making 1993;13(4):322-338. [doi: 10.1177/0272989X9301300409] [Medline: 8246705]

Abbreviations

ANMI: angina and myocardial infarction
CHF: congestive heart failure
CHST: congestive heart failure and stroke
CTMC: continuous-time Markov chain
CVD: cardiovascular disease
DTMC: discrete-time Markov chain
MI: myocardial infarction
NHLBI: National Heart Lung and Blood Institute
WHO: World Health Organization

Edited by C Lovis; submitted 10.04.24; peer-reviewed by B Wang, X Liu, L Meyer, GK Gupta, H Younes; comments to author 07.05.24; revised version received 13.06.24; accepted 17.08.24; published 24.09.24.

Please cite as:
Brahma A, Chatterjee S, Seal K, Fitzpatrick B, Tao Y
Development of a Cohort Analytics Tool for Monitoring Progression Patterns in Cardiovascular Diseases: Advanced Stochastic Modeling Approach
JMIR Med Inform 2024;12:e59392
URL: https://medinform.jmir.org/2024/1/e59392
doi:10.2196/59392
PMID:39316426

Original Paper

Toward Better Semantic Interoperability of Data Element Repositories in Medicine: Analysis Study

Zhengyong Hu[1*], MM; Anran Wang[1*], MS; Yifan Duan[1*], MM; Jiayin Zhou[1], MM; Wanfei Hu[1], MM; Sizhu Wu[1], PhD

Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China
[*]these authors contributed equally

Corresponding Author:
Sizhu Wu, PhD
Institute of Medical Information/Medical Library
Chinese Academy of Medical Sciences & Peking Union Medical College
No. 3 Yabao Road, Chaoyang District
Beijing, 100020
China
Phone: 86 52328760
Email: Wu.sizhu@imicams.ac.cn

Abstract

Background: Data element repositories facilitate high-quality medical data sharing by standardizing data and enhancing semantic interoperability. However, the application of repositories is confined to specific projects and institutions.

Objective: This study aims to explore potential issues and promote broader application of data element repositories within the medical field by evaluating and analyzing typical repositories.

Methods: Following the inclusion of 5 data element repositories through a literature review, a novel analysis framework consisting of 7 dimensions and 36 secondary indicators was constructed and used for evaluation and analysis.

Results: The study's results delineate the unique characteristics of different repositories and uncover specific issues in their construction. These issues include the absence of data reuse protocols and insufficient information regarding the application scenarios and efficacy of data elements. The repositories fully comply with only 45% (9/20) of the subprinciples for Findable and Reusable in the FAIR principle, while achieving a 90% (19/20 subprinciples) compliance rate for Accessible and 67% (10/15 subprinciples) for Interoperable.

Conclusions: The recommendations proposed in this study address the issues to improve the construction and application of repositories, offering valuable insights to data managers, computer experts, and other pertinent stakeholders.

(JMIR Med Inform 2024;12:e60293) doi:10.2196/60293

KEYWORDS

data element repository; FAIR; ISO/IEC 11179; metadata; semantic interoperability

Introduction

Background

The sharing of medical data can enhance the efficiency of medical research, bolster transparency within the field of medicine, and respond to the stringent demands for research reproducibility and data openness [1]. Nonetheless, medical data present challenges due to their high complexity in semantics and heterogeneity, and they lack standards and uniform specifications at the level of fields and value domains. For instance, the numeric value "18" could represent the age at which a patient started smoking in one study, while in another, it might signify the total number of years a person has been smoking. This issue of semantic ambiguity in the data renders it challenging for other researchers to comprehend and use the information. It impedes the integration, comparison, and joint analysis of different data sets [2], thereby obstructing data sharing.

Metadata, essentially data about data, offers a solution to address such issues. Metadata can describe data, providing researchers with a comprehensive overview to aid understanding and application. Furthermore, it supports more precise retrieval and traceability. When data are accurately associated with metadata (such as "18" being linked to an individual's total years of

smoking), its semantics become much more straightforward. Metadata has already found applications in various fields, including molecular biology [3,4] and clinical medicine [5,6]. Guidelines for data management and sharing, such as the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, also provide specifications for metadata to ensure that data are Findable, Accessible, Interoperable, and Reusable [7]. However, researchers often find that creating and annotating metadata are time-consuming and prone to errors [8]. This makes it challenging to ensure metadata quality and increases metadata heterogeneity across research. Hence, using standardized metadata for data collection to achieve semantic consistency from the inception of the data life cycle is essential to maximize semantic interoperability across multiple data sources.

Data elements (DEs) are vital components of metadata, representing indivisible data units within a given context. The underlying framework of DEs can furnish rich metadata information, including unique identifiers, definitions, and value domains, among other attributes. The DE repository represents a platform structured in accordance with a standardized framework dedicated to the construction, storage, administration, and dissemination of DEs. Within this repository, DEs adhere to rigorous standardization, with their conceptual aspects, value ranges, and related attributes systematically linked to controlled vocabularies and other terminological systems. A DE repository facilitates the unified management and maintenance of internal metadata, ensuring semantic consistency and reducing the cost associated with redundant design efforts for project-specific metadata. By fostering the reuse and sharing of standardized DEs, barriers to data integration are diminished, thus propelling applications such as cross-institutional and cross-study meta-analyses in the realm of medical data [9]. This, in turn, unlocks the value of medical data.

Currently, the prevailing international standards for DEs and repository construction are set by the ISO/IEC (International Organization for Standardization/International Electrotechnical Commission) 11179 standard. The ISO/IEC 11179 standard establishes a conceptual model for DEs and their repositories, while also providing regulations for activities such as DE registration and management. Many DE repositories have been constructed in the medical field based on the ISO/IEC 11179 standard. However, the broader application of DE repositories has not yet been achieved, often limited to specific projects or internal use within particular institutions [10]. As the central platform for storing, managing, and sharing DEs and metadata, the degree of completeness in its construction directly influences the practical usage of DEs. Current research tends to focus more on the specific technical aspects and standards for constructing DE repositories. Simultaneously, there is a discernible deficiency in evaluating and analyzing typical repositories in the medical domain.

Literature Review

Medical DE

Data elements, defined and standardized by the ISO/IEC 11179 standard, constitute the minors units for collecting, processing, and disseminating data [11]. The definition of DEs should ideally encompass 3 aspects—research questions, data acquisition, and data storage—to reflect the life cycle of a repository best [12]. DEs play a pivotal role in standardizing clinical data collection, enhancing data quality, facilitating secondary analysis and applications [13,14], and serving as a base for systems based on artificial intelligence (AI) [15].

Currently, the development of Des primarily relies on multidomain expert consensus and collaboration, often achieved through iterative Delphi methods for discussion, identification, and refinement of relevant DEs [16]. This approach ensures the professionalism of DEs within specific domains but demands considerable time and personnel involvement. National Institute of Neurological Disorders and Stroke (NINDS) categorizes the development of common data elements (CDEs) into 4 phases: discovery, internal validation, external validation, and distribution [17]. Numerous domains or projects have undergone multiple iterations of DE development, such as Stroke V2.0 CDE [18]. More granular domain-specific DEs have been developed or reached consensus, spanning therapeutic methods [19], examinations [20], and others. With the continuous expansion of DEs, Kim et al [21] proposed a comprehensive representation of real-world clinical semantics by defining semantic relationships and constraints between DEs.

The application and evaluation of DEs have indeed garnered considerable attention. For instance, Fitzgerald et al [22] analyzed the seizure burden using clinical data in childhood epilepsies collected from CDE-based forms within the electronic medical record. Evaluation studies encompass DE quality [23] and the effectiveness of data collection. Chen et al [24] assessed the data collection effectiveness of DEs in real-world scenarios, while Ryan et al [25] separately evaluated data capture rates for DEs in in-person and virtual visits scenarios.

Recently, several studies have sought to advance the application of AI technologies throughout the life cycle of DEs. Natural language processing can assist in extracting specific DEs from clinical documents [26-28]. Renner et al [29] explored the use of artificial neural networks to semiautomatically map DE models to the BRIDG model, thereby reducing the burden of manual mapping by experts. In addition, DEs play a role in collecting high-quality data to aid in training machine learning algorithms, further expanding their applications in the health information domain [30]. Littlefield et al [31], based on data collected through DEs, compared the performance of major machine learning algorithms with traditional statistical methods.

DE Repository

DE repositories serve as platforms for storing and managing DEs, facilitating standardization, and promoting the integration and sharing of medical data through both top-down and bottom-up approaches [32]. The bottom-up approach relies on users creating and maintaining their DEs. Hegselmann et al [33] have expanded upon this model by extracting real-world DEs from medical documents and standardizing them, thereby promoting the reusability of DEs. The DE repository can standardize metadata across various studies and institutions, facilitating data integration. Mallya et al [34] coordinated variables in 4 research endeavors through the effective usage of the DE repository.

Another crucial function of the DE repository is to ensure internal semantic consistency, thereby enhancing the semantic interoperability of DEs. One perspective suggests that the maintenance and updating of terms should be separated from the repository's operational tasks [35]. Schladetzky et al [36] developed the Mettertron system to enhance the linkage between the DE repository and the terminology system, simplifying terminology maintenance services. Meanwhile, mapping the repository model to the Web Ontology Language (OWL) ontology model can expand its semantic applications. Yuan and Li [37] constructed a semantic relation metamodel for the repository and defined mapping rules to the ontology model.

Recent research has also been conducted on data quality assessment based on the DE repository. For instance, Juárez et al [38] attempted to validate local data repositories by the central DE repository of networks, thereby providing a comparative method for assessing data quality across different sites. Kapsner et al [39] centralized the maintenance of data quality checks by associating data quality assessment tools with DE definitions in the DE repository.

Related Works

Current research lacks a comprehensive evaluation and analysis of multiple typical medical DE repositories. Ulrich et al [40] referenced information about specific metadata repositories in evaluating the application of the metadata exchange language QL4MDR. Hegselmann et al [33] also provided a brief overview of repository practices based on the ISO/IEC 11179 standard in his study on Pragmatic MDR. Nonetheless, both studies stopped short of providing a detailed evaluation or analysis and did not endeavor to suggest an analytical framework or standard.

Sasse et al [41] conducted a literature review on semantic annotation services for biomedical metadata. Through the review, they identified 10 supporting tools and conducted a detailed comparison based on 7 criteria. While their comparative dimensions are unidimensional and more aligned with tools rather than repositories, the variables in their semantic services provide a reference for the semantic dimensions in constructing the analytical framework for this study.

Stoehr et al [42] assessed the portal usability of the CoMetaR repository. They divided the web page into different modules and used the Think Aloud method along with a usability scale, conducting a combined quantitative and qualitative evaluation. While their method of module-based usability assessment provides insights for constructing usability evaluation dimensions in this study, it is worth noting that their focus is on optimizing the web page's interaction and does not compare it with the web pages of other repositories. Reichenpfader et al [43] similarly assessed the usability of the Portal of Medical Data Models (MDM-Portal) repository by analyzing the users' experience with the web page through various tests. The dimensions they analyzed also provide insights for the usability evaluation in this study.

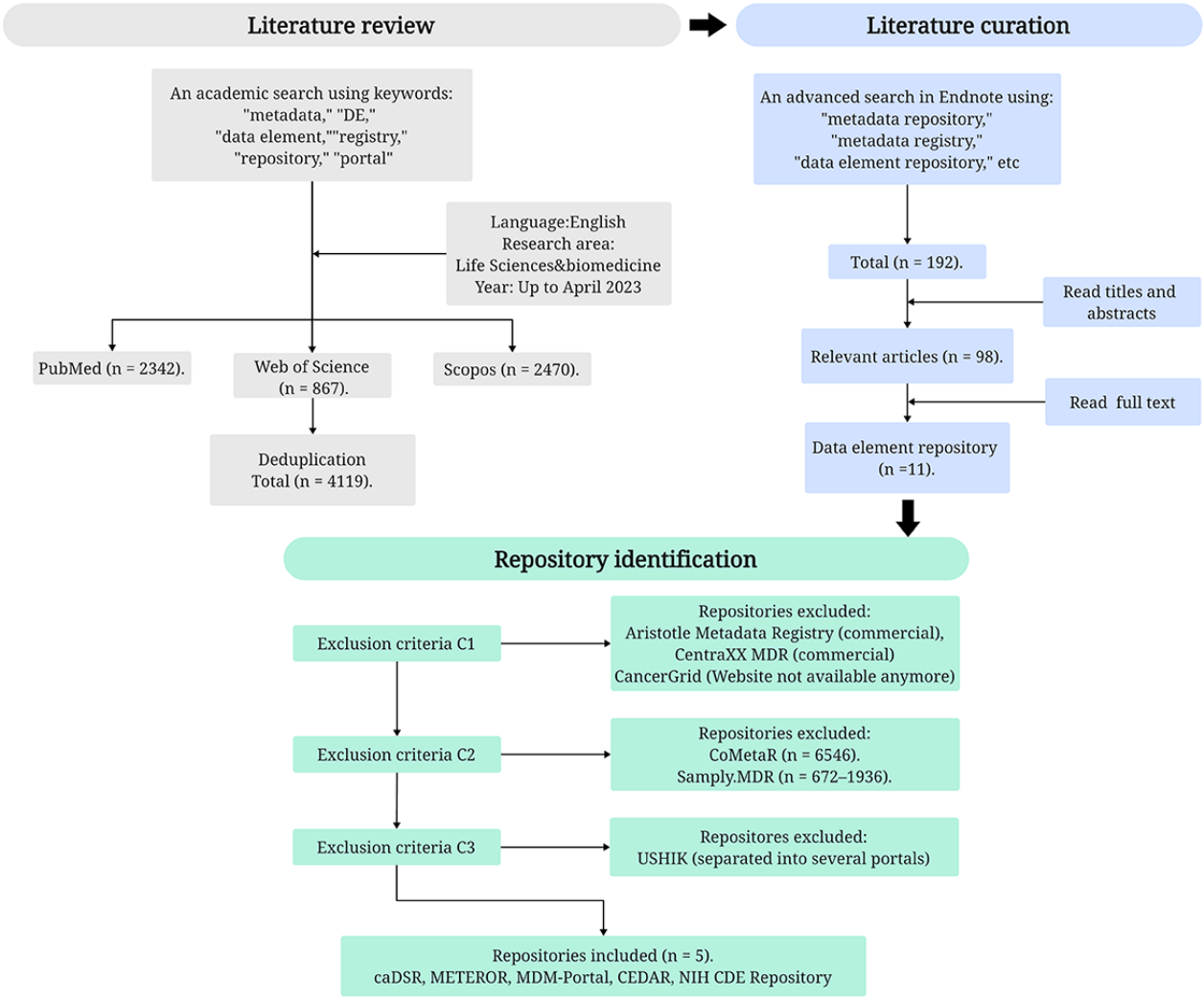
Objectives

The primary objective of this study is to explore potential issues and promote the broader application of DE repositories within the medical field by evaluating and analyzing typical repositories. Furthermore, we also endeavor to address the gap in the existing literature concerning the lack of evaluation of DE repositories, offering an overview of the typical DE repository construction in the medical field.

Methods

The method used in this study for screening medical DE repositories involves three distinct steps: (1) literature review, (2) literature curation, and (3) repositories identification (Figure 1).

Figure 1. Flowchart of screening the data element repository for this study through literature review. The upper left gray part is the first step: searching literature from various databases. The upper right blue part is the second step: obtaining the data element repository through further screening and reading of literature. The green part below is the third step: obtaining the data element repository for this study according to the 3 inclusion and exclusion criteria: C1, C2, and C3. caDSR: Cancer Data Standards Registry and Repository; CDE: Common Data Element; CEDAR: Center for Expanded Data Annotation and Retrieval; DE: Data Element; MDM-Portal: Portal of Medical Data Model; METEOR: Metadata Online Registry; NIH: National Institutes of Health.



Literature Review

This study conducted literature searches on PubMed, Web of Science, and Scopus. The searches were performed using a combination of keywords such as “metadata,” “data element,” and “DE,” combined with “repository,” “registry,” “platform,” and “portal.” The language was restricted to English, and the research area was focused on life sciences or biomedicine. Up to April 2023, a total of 4119 papers were retrieved.

Repository Curation

The retrieved literature was imported into Endnote, and an advanced search was conducted explicitly targeting titles or abstracts containing terms such as “metadata repository,” “metadata registry,” and “data element repository.” After this secondary screening, a total of 192 papers were obtained. After reviewing the titles and abstracts of these papers, 98 papers related to DE repositories were identified and subsequently read in full. In the end, 11 DE repositories (shown in Table 1) within the medical field were gathered. The information and data related to DE repositories were primarily collected from three sources: (1) the portals of various repositories, (2) relevant literature, and (3) project archives up to April 2023.

Table 1. Eleven data elements repositories retrieved from literature (repository URLs in references).

Data element repositories	Country
Samply.MDR [44]	Germany
MDM.Portal [45]	Germany
CoMetaR [46]	Germany
CentraXX MDR [47]	Germany
CancerGrid (2005-2010) [48]	United Kingdom
METEOR (METeOR) [49]	Australia
Aristotle Metadata Registry [50]	Australia
caDSR [51]	United States
USUIK [52]	United States
NIH CDE Repository [53]	United States
CEDAR [54]	United States

Repository Identification

To facilitate a more effective comparison, we established inclusion and exclusion criteria for screening the 11 repositories. The specific inclusion and exclusion criteria and the process are as follows:

1. C1: DE repositories should be open-access public platforms, meeting noncommercial or managed by nonprofit organizations (such as universities or research institutions).
2. C2: The repository's metadata or DE resources should comprise more than 20,000 records.
3. C3: We required the repository to have a well-established, independent portal to support access.

Five DE repositories were ultimately included (Table 2): Cancer Data Standards Registry and Repository (caDSR) [55], NIH (National Institutes of Health) CDE Repository, MDM-Portal [2], Metadata online registry (METEOR), and Center for Expanded Data Annotation and Retrieval (CEDAR) [56].

Table 2. Basic information of the 5 data element repositories included in this study.

Repositories	Country	First release year	Hosted by
caDSR[a]	America	2003	National Cancer Institute
NIH[b] CDE[c] Repository	America	2015	National Library of Medicine
METEOR[d]	Australia	2022	Australian Institute of Health and Welfare
MDM[e]	Germany	2012	Heidelberg University Hospital
CEDAR[f]	America	2014	Stanford University

[a]caDSR: Cancer Data Standards Registry and Repository.

[b]NIH: National Institutes of Health.

[c]CDE: Common Data Element.

[d]METEOR: Metadata Online Registry.

[e]MDM: Medical Data Model.

[f]CEDAR: Center for Expanded Data Annotation and Retrieval.

Analysis Framework

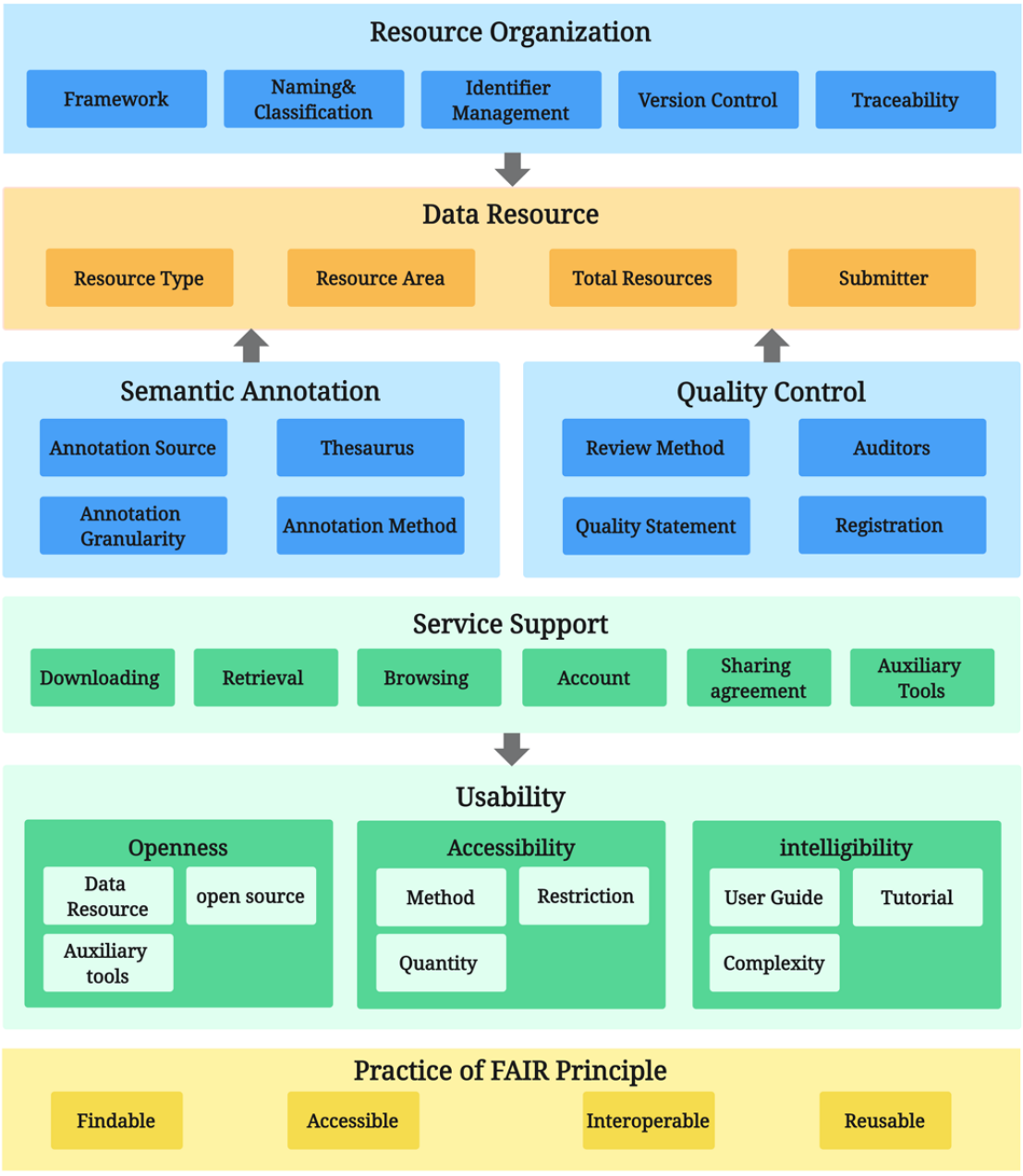
We aimed to comprehensively analyze the repositories, encompassing multiple dimensions, including technology, management, and services. To achieve this, we developed a comprehensive analysis framework consisting of 7 dimensions and 36 secondary indicators (Figure 1). The 7 dimensions include the following:

1. *Data resources*: providing an overview of the repository's data resources, including data volume, data types, data sources, coverage, and domains.
2. *Resource organization*: focusing on how metadata or DE resources are effectively organized and managed throughout their life cycle, including underlying frameworks, traceability, and version control.
3. *Quality control*: analyzing how the platform ensures the quality of stored data.
4. *Semantic annotation*: assessing how the repository achieves internal semantic consistency to enhance semantic interoperability.
5. *Service support*: examining the services offered to users by the repository, including basic services, such as retrieval and download, and advanced features such as analysis tools.

6. *Usability*: evaluating the platform's openness, accessibility, and intelligibility, including the availability of support documents and training materials.
7. *Practice of FAIR principles*: finally, analyzing the repository's adherence to the FAIR principles as a supplementary assessment.

Data resources and services dimensions are primarily determined by repository and portal characteristics, while resource organization, quality control, and semantics leverage insights from relevant literature and the ISO/IEC 11179 standard. Practice of FAIR adheres to the FAIR principle and its 15 subprinciples. Furthermore, 4 experts in data management, data warehousing construction, and data standardization participated in consultations to refine the analysis framework. Their input informed the division, naming, and selection of secondary indicators for the dimensions. The analysis framework was further refined based on expert suggestions primarily through (1) revising the name of Semantic Annotations and Service Support dimensions; (2) dividing the Usability dimension into 3 distinct modules: openness, accessibility, and intelligibility; and (3) adding more granular secondary indicators, such as source link and historical versions encoding, to enhance the depth of analysis (Figure 2). For a detailed description of the indicators included in this analytical framework, see Multimedia Appendix 1.

Figure 2. The analysis framework constructed in this study. The 7 light-colored parts represent the 7 analysis dimensions, and the dark rectangle in the middle is the specific indicator of each dimension. FAIR: Findable, Accessible, Interoperable, and Reusable.



Results

Data Resource

A comparison of the data resources of the 5 repositories was conducted (Table 3). The data resources of all 5 repositories are comprehensive, but each has its emphasis on specific subdomains. For example, caDSR focuses on cancer-related DEs, METEOR emphasizes health and welfare, while others encompass DEs from various biomedical research domains. The types of resources in the repositories include elements and forms. However, the names for DEs are not yet standardized across the repositories and may consist of terms such as CDE, DE, and Field, among others. Resources in caDSR, METEOR, and NIH CDE Repository are sourced from government and institutional research projects and are released through a top-down approach. In contrast, the other platforms rely on contributions from individual users, following a bottom-up data source model. The latter category tends to have a larger volume of resources, with MDM, for instance, cataloging the most extensive collection of DEs, totaling up to 500,000 elements and more than 20,000 forms.

Table 3. Analysis results of the 5 data element repositories in the data resource dimension.

Repositories	Area	Type	Total amount	Submitter
caDSR[a]	Cancer research, etc	DEs[b]	71,743 DEs	NIH[c] research institutes and programs
NIH CDE[d] Repository	Biomedical field	DEs and forms[e]	20,970 DEs; 1704 forms	NIH research institutes and programs
METEOR[f]	Health and welfare	DEs	21,180 DEs	Australian health department or research institution
MDM[g]	Clinical trials, special diseases, etc	DEs and forms	500,000 DEs; 24,810 forms	Individual user or project submissions
CEDAR[h]	Biomedical field	DEs and forms	120,829 DEs; 2000 forms	Individual user or project submissions

[a]caDSR: Cancer Data Standards Registry and Repository.

[b]DEs: Data Elements.

[c]NIH: National Institutes of Health.

[d]CDE: Common Data Elements.

[e]Forms: forms composed of data elements (eg, case report forms, questionnaires).

[f]METEOR: Metadata Online Registry.

[g]MDM: Medical Data Model.

[h]CEDAR: Center for Expanded Data Annotation and Retrieval.

Resource Organization

Regarding repository frameworks, all repositories except for MDM are constructed based on the ISO/IEC 11179 standard (Table 4). The DEs in these repositories are built upon the conceptual model of DEs and value domains outlined in the ISO/IEC 11179 standard. METEOR has extended this framework by introducing a top-level category called “data set specification” (DSS). This category is used to group specific DEs. For example, the “Diabetes (Clinical) DSS” in METEOR contains DEs related to standardized data collection for patients with diabetes. In MDM, which uses a custom DE framework, the attributes of DEs are relatively concise. They typically include only the DE description, data type, concept, and value domain information.

All repositories have assigned unique identifiers to their resources, although the granularity of the assignment varies. In the case of MDM, the smallest unit assigned an internal identifier is a form, and unique identifiers are not provided for individual DEs. On the other hand, other repositories assign unique identifiers at the level of individual DEs. Furthermore, the encoding of unique identifiers is standardized only within caDSR and METEOR. Simultaneously, some other repositories have inconsistent encoding methods for resources at the same hierarchical level, or they directly reference the source identifiers.

Regarding external provenance, most platforms can provide basic provenance information for DEs, such as the data submitter or the source institution. Among them, MDM provides the most detailed information, including the owner or institution of the DE, source links, and partial contact information. Regarding internal referencing and provenance, METEOR demonstrates the most comprehensive practice. It can support granularity to value domains, object classes, and properties. DEs in METEOR are listed with links to the attributes they reference and from which elements they are derived. Corresponding attributes such as value domains and object classes also provide links to all DEs that reference them. This allows for bidirectional provenance between elements and attributes.

Table 4. Analysis results of the 5 data element repositories in the resource organization dimension, with further explanation of identifiers, traceability, and version control indicators.

Repositories	Framework	Naming specification	Classification scheme	Identifier		Traceability				Version control	
				Name	Encoding	Submitter/source information	Source link	Source identifier	Internal citation link	Historical versions accessible	Version encoding
caDSR[a]	ISO/IEC 11179	Yes	Yes	Public ID	7 digits	Yes	No	No	No	No	Yes
NIH[b] CDE[c] Repository	ISO/IEC 11179	No	Yes	Identifiers	N/A[d]	Yes	No	Yes	Yes	Yes	No
METEOR[e]	ISO/IEC 11179	Yes	Yes	Identifiers	6 digits	Yes	Yes	No	Yes	Yes	No
MDM[f]	N/A	No	Yes	Public ID	N/A	Yes	Yes	No	No	Yes	Yes
CEDAR[g]	ISO/IEC 11179	No	No	Identifiers	N/A	N/A	No	Yes	No	No	No

[a]caDSR: Cancer Data Standards Registry and Repository.

[b]NIH: National Institutes of Health.

[c]CDE: Common Data Elements.

[d]N/A: not applicable.

[e]METEOR: Metadata Online Registry.

[f]MDM: Medical Data Model.

[g]CEDAR: Center for Expanded Data Annotation and Retrieval.

The version number formats for DEs in most repositories lack uniformity; in some cases, no version numbers are provided. In addition, some repositories do not allow access to historical versions of DEs, making them inaccessible for viewing. MDM has better version control practices in place. Historical versions of DEs are accessible and come with a standardized version number format. The version number includes a detailed editing data and information about the editor (eg, “4/6/22-Smith”). This allows users to navigate and browse historical versions using the version number as a reference.

Quality Control

DE quality control is primarily achieved through the audit process during registration. Currently, the audit process relies mainly on manual review, and it is evident shown in Table 5 that all 3 top-down repositories have established governance committees to conduct quality control audits. The audit process includes reviewing the basic attributes of elements (such as concepts and value domains), mapping or references between elements and controlled vocabularies, and the domain-specific expertise of elements. This audit process helps ensure the quality and authority of the published DEs, ensuring that their structural attributes are correct and appropriately specialized within their respective domains. However, it can be resource-intensive and time-consuming, requiring the involvement of experts. The bottom-up repositories MDM, on the other hand, cannot implement this process in the same way. Instead, it relies on repository administrators to conduct quality control audits. While this method can ensure only the basic structural integrity of data elements, its higher review efficiency makes it more suitable for bottom-up repositories handling large volumes of data element submissions.

A complete and well-defined registration workflow is a crucial part of quality control of DEs. MDM and CEDAR have not provided an entire registration process, while other repositories offer information on the registration workflow for DEs within the platform. They also assign identifiers for different registration statuses. METEOR and caDSR have more comprehensive registration statuses, with a finer-grained classification. In addition, only the NIH CDE Repository provides quality identifiers for DEs and includes only 1-level identifier (NIH-Endorsed). Other repositories do not appear to have detailed quality scoring or rating information. Only the NIH CDE Repository provides quality indicators for DEs, including a single-level indicator (NIH-Endorsed). Conversely, MDM relies on users to rate DEs, and other repositories do not seem to have detailed quality ratings or grading content.

Table 5. Analysis results of the 5 data element repositories in the quality control dimension, demonstrating the actions of each data element repository in data element quality control.

Repositories	Review method	Auditors	Quality mark	Registration			
				Registration workflow	Status identifier	Status type	Quality control records/documents
caDSR[a]	Manual review	Committee experts	No	Yes	Full life cycle	10	No
NIH[b] CDE[c] Repository	Manual review	Committee experts	NIH-Endorsed CDE	Yes	Full life cycle	2	No
METEOR[d]	Manual review	committee experts	No	Yes	Full life cycle	9	Yes
MDM[e]	Manual review	Portal administrator	No	N/A[f]	No	N/A	Partially provided
CEDAR[g]	N/A	N/A	No	N/A	No	N/A	No

[a]caDSR: Cancer Data Standards Registry and Repository.

[b]NIH: National Institutes of Health.

[c]CDE: Common Data Element.

[d]METEOR: Metadata Online Registry.

[e]MDM: Medical Data Model.

[f]N/A: not applicable.

[g]CEDAR: Center for Expanded Data Annotation and Retrieval.

Semantic Annotation

The repositories achieve semantic annotation by standardizing the mapping of DEs to terminology systems, ensuring internal semantic consistency (Table 6). The primary terminology systems used by these repositories include Unified Medical Language System (UMLS), Logical Observation Identifiers Names and Codes (LOINC), and Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), with others such as National Cancer Institute Thesaurus (NCIT) and National Center for Biomedical Ontology (NCBO) also being used. METEOR has developed its internal glossary and achieves semantic annotation through metadata items called “Glossary Items (GIs).” GIs share the same DE framework as other elements but store the definition of a term. Other DEs can achieve semantic annotation by referencing the corresponding GI associated with a specific term. Creating and referencing internal glossaries effectively harnesses the advantages of the ISO/IEC 11179 DE framework. GIs essentially facilitate clustering according to the DE framework, including object class, property, value domain, and more. DEs belonging to the same object class can be associated with the terminology item by referencing it. For instance, by querying the GI item “person,” you can observe all DEs that reference this term as their object class. This clustering enhances the interrelatedness of DEs at the conceptual level. However, the shortcomings of internal glossaries are also evident. If DEs need to be used across institutions, there is a need for remapping terminology, or semantic inconsistencies may persist. Regarding semantic interoperability, internal glossaries are less effective referencing internationally recognized terminology repositories.

Table 6. Analysis results of the 5 data element repositories in the semantic annotation dimension, presenting the measures taken by each data element repository to semantically standardize data elements.

Repositories	Annotation source	Mapping vocabulary	Granularity	Annotation method	Annotation content
caDSR[a]	Controlled vocabulary	NCIT[b]	DE[c] concept and permissible value	Manual mapping	Terms and links
NIH[d] CDE[e] Repository	Controlled vocabulary	NCIT, UMLS[f], etc	DE concept and permissible value	Manual mapping	Terms and coding
METEOR[g]	Self-built vocabulary	Self-built vocabulary	DE concept	Manual mapping	Terms and links
MDM[h]	Controlled vocabulary	UMLS, LOINC,[i] and SNOMED CT[j]	DE concept and description	Automatic mapping	Terms and coding
CEDAR[k]	Controlled vocabulary	NCBO[l]	DE concept and permissible value	Manual mapping	Terms and links

[a]caDSR: Cancer Data Standards Registry and Repository.

[b]NCIT: National Cancer Institute Thesaurus.

[c]DE: Data Element.

[d]NIH: National Institutes of Health.

[e]CDE: Common Data Element.

[f]UMLS: Unified Medical Language System.

[g]METEOR: Metadata Online Registry.

[h]MDM: Medical Data Model.

[i]LOINC: Logical Observation Identifiers Names and Codes.

[j]SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms.

[k]CEDAR: Center for Expanded Data Annotation and Retrieval.

[l]NCBO: National Center for Biomedical Ontology.

Service Support

A robust retrieval system can enhance the discoverability of data resources within repositories. Each of the 5 repositories possesses unique search capabilities, for instance, caDSR and NIH CDE Repository allow users to search by the names of NIH-affiliated institutions (Table 7). METEOR and MDM allow users to construct search queries using Boolean operators and keywords. Furthermore, these platforms also differ in their secondary filtering criteria, with caDSR and METEOR supporting additional filters such as submitting organization, registration status, and registering organization, among others.

The repositories offer personalized services to users, including features such as personal favorites in NIH CDE Repository and METEOR, enabling users to collect elements of interest and record and browse their own created or edited metadata. In CEDAR, DEs are organized in a folder structure, facilitating the categorization and management of metadata.

With regard to data element download and export services, all repositories except CEDAR offer support for multiple export formats. MDM supports export in 18 formats, including Comma Separated Values (CSVs) and Operational Data Model (ODM), but it is limited to exporting data by form and allows only 50 downloads per week. METEOR provides Word and PDF export formats with lower levels of structure, which can impact interoperability. caDSR and NIH CDE Repository allow DEs and forms to be exported in various structured document formats such as EXCEL, XML, and JSON, providing a relatively comprehensive download service. On the other hand, CEDAR offers only JSON source code for elements without direct download capabilities. Although it has a REST API interface, it may not be as convenient for nonbatch exports.

All 4 platforms except CEDAR provide web-based metadata comparison tools, but they differ in the dimensions they support for comparison. MDM and METEOR can perform horizontal comparisons for all information of 2 DEs, while caDSR supports comparisons for multiple DEs. NIH CDE Repository offers vertical comparisons, allowing users to compare DEs with their historical versions. In addition to the comparison tools, MDM also provides a rich set of auxiliary tools, including ODMedit (for creating ODM format DEs and forms) [57], CDEGenerator (for visualizing concept frequencies in forms) [58], OpenEDC (for web-based data collection using forms), and more. MDM offers a more significant number of tools and functionality than other repositories.

Table 7. Analysis results of the 5 data element repositories in the service dimension, mainly presenting the various services provided by each data element repository on its portal website to help users better use the repository and data elements.

Repositories	Features of retrieval	Results secondary screening	Register account	Account service	Sharing agreement	Download service	Download granularity	Export format	Comparison tool	Other tools
caDSR[a]	Abbreviation of the institute's name, identifier, etc	Registration status, submitter, etc	N/A[b]	N/A	N/A	Unlimited downloads, batch download	DE and form	EXCEL, XML, and JSON	Compare 2 or more DEs[c]	Form creation
NIH[d] CDE[e] Repository	Abbreviation of the institute name, identifier, etc	Data type, submitter, etc	UTS[f] account	Personal favorites, browsing history, etc	N/A	Unlimited downloads, batch download	DE and form	EXCEL, XML, JSON	Compare different versions of DE	Not support
METEOR	Keywords, identifier, Boolean operators, etc	Registration organization, data type, etc	Internal account	Personal favorites and settings, browsing history, etc	N/A	Unlimited downloads, batch download	DE	Word, PDF	Compare 2 DEs	DE creation
MDM[g]	Keywords, Boolean operators, wild card character, etc	Keywords, research field	Internal account	Personal favorites, browsing history, etc	Four version CC 4.0 licenses	50 forms per week	Form	CSV, EXCEL, SQL, and other 18 formats	Compare 2 or more DEs	Web-based date capture, visualization, visual analysis tools, etc
CEDAR[h]	Keywords, terminology, etc	Data type, version, etc	Internal account	API[i] keys, personal folder	N/A	Not support	Not support	JSON code	Not support	DE and form creation

[a]caDSR: Cancer Data Standards Registry and Repository.

[b]N/A: not applicable.

[c]DEs: Data Elements.

[d]NIH: National Institutes of Health.

[e]CDE: Common Data Element.

[f]UTS: UMLS Terminology Services.

[g]MDM: Medical Data Model.

[h]CEDAR: Center for Expanded Data Annotation and Retrieval.

[i]API: Application Programming Interface.

Usability

We analyze the usability of the repositories from 3 perspectives: openness, accessibility, and intelligibility. Openness focuses on the extent to which the repository's resources and services are available for browsing and use. Among the 5 repositories in the study, access is typically restricted by requiring user accounts. Regarding data resources, caDSR, METEOR, and MDM provide unrestricted browsing access, including both forms and DEs. However, the NIH CDE Repository restricts viewing some semantic annotation content. Regarding services, MDM and CEDAR restrict auxiliary tools to logged-in users, including web-based creation and submission of DEs, among other features. In contrast, the DE creation and registration tools in the other 3 top-down repositories are not open to regular users. CEDAR requires registration for access to all services and resources, but it provides source code and technical documentation on GitHub. In summary, caDSR and METEOR exhibit the highest level of openness regarding resources and tools (Table 8).

Accessibility considers the types of accessible resources, the methods of accessibility, and the extent to which resources are accessible. There are primarily 2 ways to access repository resources: web downloads and application programming interface (API) interfaces. CEDAR does not provide web downloads and offers only JSON source code and an API interface. MDM requires user login for downloading forms and performing batch downloads, with a limit of 50 forms per week. In contrast, caDSR and NIH CDE Repository allow free downloads and batch exports of DEs without the need for login, making them relatively more accessible regarding resource availability.

Intelligibility focuses on the availability of supplementary information provided by the repositories and the complexity of constructing DEs. First, all 5 repositories offer user guide

documents on their portals, which introduce basic information and operations. In addition, CEDAR and caDSR have Archive and Wiki web pages to provide further information and support. The repositories also pay attention to teaching concepts related to DEs. Since not all users have a computer-related background, all 4 platforms except MDM provide introductions or tutorials on metadata, DEs, and the ISO/IEC 11179 standard.

In addition, most repositories lack descriptions and visual representations of their data resources' coverage areas and quantities. On its portal page, MDM provides visualizations of its DEs categorized by proportion, which can help users understand the resources within the repository. Compared with the complexity of DEs across the 5 platforms, MDM benefits from its self-built framework, resulting in relatively more straightforward and more concise DEs with better comprehensibility. In contrast, other platforms build their DEs based on the ISO/IEC 11179 standard and often expand or subdivide the framework, increasing the amount of information and complexity, which can affect comprehensibility.

Table 8. Analysis results of the 5 data element repositories in the usability dimension, mainly focusing on openness, accessibility, and usability, and comprehensively evaluating the usability of each data element repository.

Repositories	Openness					Accessibility				Intelligibility		
	Open access	Restriction	Create and submit	Open source	Auxiliary tool	Method	Limitation	Batch download	Quantity limitation	User guide	DE tutorial	DE complexity
caDSR[a]	DEs[b]	No	No	No	Yes	Download and API[c]	No	Yes	No	Document	Yes	High
NIH[d] CDE[e] Repository	DEs and forms	Partial DEs	No	No	Yes	Download and API	No	Yes	No	Documents	Yes	High
METEOR[f]	DEs	No	No	No	Yes	Download	No	Require log-in	No	Document	Yes	Middle
MDM[g]	DEs and forms	No	Yes	No	Partially require log-in	Download	No	Require log-in	50 forms per week	Video	No	Low
CEDAR[h]	DEs and forms	All resources	Yes	Yes	Require log-in	JSON code and API	Require log-in	No	No	Video and document	Yes	Middle

[a]caDSR: Cancer Data Standards Registry and Repository.

[b]DEs: Data Elements.

[c]API: Application Programming Interface.

[d]NIH: National Institutes of Health.

[e]CDE: Common Data Elements.

[f]METEOR: Metadata Online Registry.

[g]MDM: Medical Data Model.

[h]CEDAR: Center for Expanded Data Annotation and Retrieval.

Practice of FAIR Principles

Finally, this study supplemented the analysis by evaluating the extent to which the 5 repositories comply with the FAIR principles. The level of compliance was categorized into 4 groups: complies completely, complies entirely, fails to comply, and unclear. The detailed content of each principle in FAIR can be found in Multimedia Appendix 2.

In Figure 3, a horizontal tally was conducted, with each of the 4 subprinciples of FAIR considered separately. The proportions of different levels of compliance to the subprinciples were calculated individually. For instance, considering the findable subprinciple, which comprises 4 principles (F1-F4), there are 20 cells. The proportions of "complies completely," "complies partly," "fails to comply," and "unclear" were then calculated for these 20 cells. The same process was applied to the remaining 3 subprinciples. Based on this step, Figure 4A was generated, depicting the overall adherence of repositories to each subprinciple. Figure 4B, calculated using the same method on a column basis, illustrates each repository's implementation of the FAIR principles.

Figure 3. Visualization of FAIR (Findable, Accessible, Interoperable, and Reusable) practices in 5 repositories, with practices divided into 4 levels: complies completely, complies partly, fails to comply, and unclear. Detailed subprinciples of FAIR are shown in Multimedia Appendix 2. caDSR: Cancer Data Standards Registry and Repository; CDE: Common Data Element; CEDAR: Center for Expanded Data Annotation and Retrieval; MDM: medical data model; METEOR: metadata online registry; NIH: National Institutes of Health.

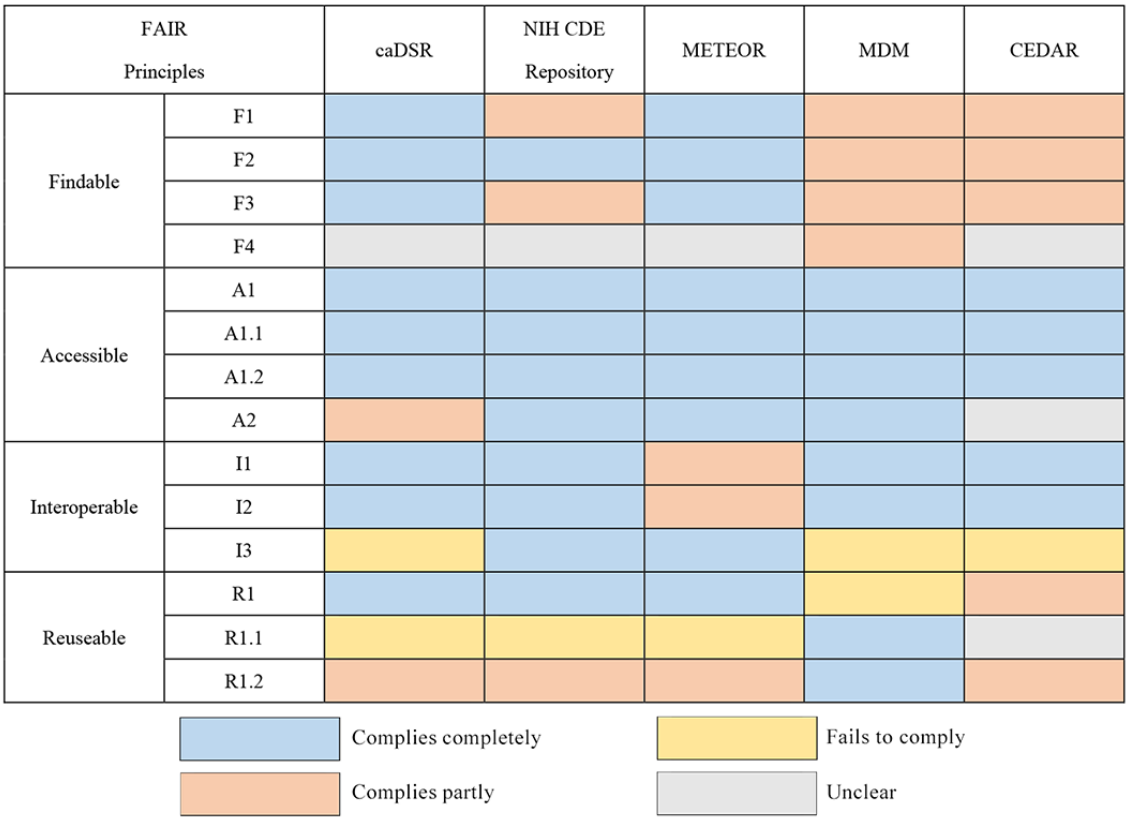
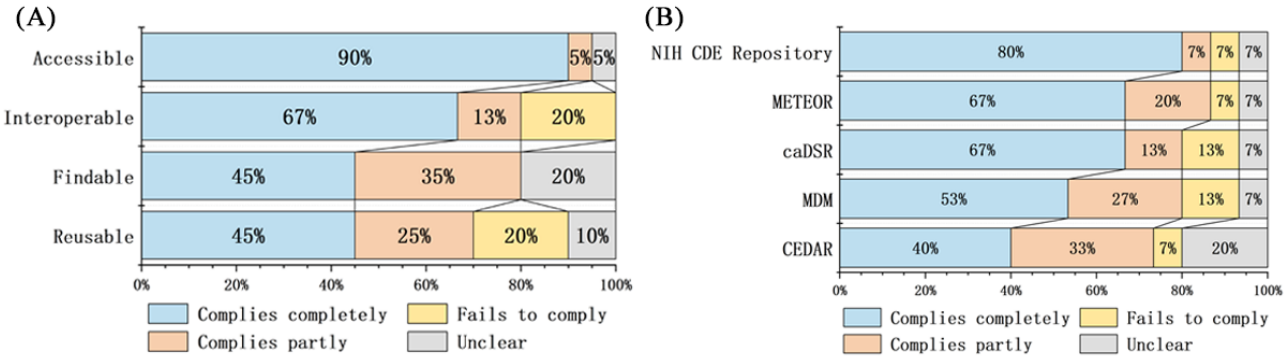


Figure 4. Statistics on the practice of FAIR (Findable, Accessible, Interoperable, and Reusable) principles based on Figure 3. (A) The figure starts from the perspective of FAIR principles and horizontally counts the practice of the 4 subprinciples in Figure 3. (B) The figure starts from the perspective of data element repositories and vertically counts the practice of FAIR principles in each repository in Figure 3. caDSR: Cancer Data Standards Registry and Repository; CDE: Common Data Element; CEDAR: Center for Expanded Data Annotation and Retrieval; MDM: medical data model; METEOR: metadata online registry; NIH: National Institutes of Health.



In comparison, among the 5 repositories, the NIH CDE Repository demonstrates the highest level of compliance with the FAIR principles, while CEDAR falls behind the other 4 platforms. When examining the 4 subprinciples of FAIR, overall, Accessibility is relatively well practiced and at the same time, Findability and Reusability have lower percentages of full compliance, indicating subpar adherence to these principles' aspects.

In the "Findable" principle, F4 "(Meta)data are registered or indexed in a searchable resource" is crucial for ensuring the discoverability of data resources on the web. Most of the 5 repositories analyzed in this study did not fully practice this aspect, impacting their web-based resources' discoverability. Only MDM has been registered in the international academic domain of registry and indexing for repositories, such as re3data and FAIR sharing.

In terms of the "Specific (meta)data are referred to by their identifier" subprinciple of interoperability, 3 repositories did not fully comply. This is mainly due to a lack of rich cross-referencing between DEs. METEOR had the best compliance with this subprinciple, as it provides comprehensive reference information for DEs on their detail pages.

Regarding the practice of "Reusability," the issues with the repositories primarily focus on data usage licenses and source information. Most repositories lack clear data usage licenses, which hinders data sharing. In addition, source information is often limited, with most repositories providing only submitter and time stamp information. Details about how the data were created and whether they had been previously published are typically not provided, impacting reusability.

Discussion

Principal Findings

The results of the analysis provide us with an overview of 5 DE repositories. The 2 approaches in repositories, namely, the top-down and bottom-up approaches, bring about differences and distinct characteristics regarding resources, semantics, and quality control. The community-driven, bottom-up approach where users submit resources, as seen in MDM and CEDAR, results in a richer pool of resources. This implies that repositories of this type need to implement more automation in various activities, including automated verification and terminology mapping. On the other hand, the top-down approach is the opposite of community-driven models. It relies on collaboration among experts from various domains. Expert committees are involved in designing, creating, and reviewing DEs in all 3 repositories following this approach. DEs following this approach have higher quality and authority, with a finer granularity in semantic annotations. However, consideration should still be given to their applicability outside the specific institution or research context. For example, DEs provided by repositories such as caDSR and NIH CDE Repository may be tailored to particular NIH-affiliated institutions and research scenarios. Conversely, community-driven DEs have a broader source base, potentially better reflecting real-world research situations. Their cross-study applicability might be more extensive.

Balancing the complexity and usability of DEs and repository metamodels is crucial. The data model structures built upon the ISO/IEC 11179 standard can be complex, and clinical researchers may not easily understand their underlying frameworks. It is essential to strike a balance to ensure that the repository remains user-friendly and accessible to its intended audience. Simplifying the framework, however, can complicate the organization of the repository. This may reduce the available information, which can negatively impact activities such as DE deduplication, establishing relationships, and hinder the development of advanced applications such as intelligent recommendations. While the self-built model of MDM is simple and user-friendly, it can organize resources only at the level of forms, lacking granularity down to the level of individual DEs. In contrast, repositories such as caDSR, built on the ISO/IEC 11179 standard, require more investment in learning and usage, but they offer more comprehensive and detailed management and organization capabilities.

Standardize Data Sharing

Promoting data sharing does not necessarily mean unrestricted sharing. DE sharing also requires clear agreements and statements. Among the 5 repositories in this study, only MDM provides 4 different versions of the CC-4.0 license as options for form resources, which offers clarity in licensing for these resources. The other 4 repositories have not provided such information on their platforms, and their affiliated institutions' data policies regarding the applicability to resources within these repositories are also somewhat unclear. Overall, these repositories seem to focus less on data sharing and reuse.

In the rapidly evolving landscape of open science, many mature examples of data-sharing strategies can serve as valuable references [59,60]. DEs are a form of data, and designing their sharing strategies can benefit from looking at the practices of other data-sharing platforms. We recommend that repositories clearly define protocols for sharing and reusing DEs in their portals. Furthermore, they should offer granularity down to the level of individual DEs, allowing resource submitters to choose specific sharing agreements. This approach can prevent unrestricted sharing and ensure greater control over DE access and usage.

The Interconnected Ecosystem of Repositories

While DE repositories facilitate the integration of DEs across institutions and projects, the gaps between DE repositories should not become new barriers to integration. In this study, the 5 repositories analyzed do not support direct sharing and exchange of resources among each other. Instead, resources must be exported and then recreated in the target repository. However, the exported formats may not be highly structured, and there is no support for importing these files for quick creation in another repository.

Despite most repositories being built based on the ISO/IEC 11179 standard, there is still a lack of interoperability and data exchange between these repositories. These limitations suggest establishing a comprehensive interconnected ecosystem for DE repositories. Both top-down and bottom-up approaches can complement each other in achieving this goal, thereby avoiding redundant construction and facilitating domain-specific developments. This can ultimately lead to more efficient and collaborative medical research efforts.

To build the interconnected ecosystem of repositories, our recommendations are as follows:

- Choose standardized repository frameworks (such as the ISO/IEC 11179 standard) and terminology systems (eg, UMLS) to avoid the need for secondary mapping of underlying frameworks or semantics.
- Enhance the export of DEs to provide more structured documents, such as CSV and JSON.
- Develop DE creation features that offer rapid import services, supporting content creation from structured documents.
- Consider developing a unified interface, like QL4MDR [40].

Enrich Information About DEs

A significant portion of DEs in DE repositories remains at the level of satisfying basic framework information. That is, they provide fundamental semantic information but lack application-oriented details. This includes contextual

information such as applicable scenarios, background details, and application outcomes. In this scenario, DEs are isolated fragments scattered throughout the repository, providing users minimal application support. Users are left uncertain whether a DE adheres to a particular standard or belongs to a specific data set, making it challenging to select accurate DEs and organize them into the required format. The repository also falls short in delivering advanced services such as intelligent recommendations.

Therefore, this study suggests that DE repositories should enrich the application information of DEs to support their practical use. We categorize application information into two aspects: (1) Application scenarios and background details: specifying the scenarios for which DEs are applicable, whether generic or specialized, and the standards or data sets from which they originate. Such contextual information can assist the repository in better associating and organizing relevant DEs. (2) Performance-related information: this can include statistics on the number of applications of a DE and user ratings, feedback, and other relevant details.

Furthermore, we recommend that the repositories consider using ontology resources to provide standardized terminology. Mature ontology repositories and tool kits, such as NCBO BioPortal [61] and Ontology Lookup Service [62], offer a wealth of ontology resources and support the download and localization of various ontology resources or their invocation through APIs. By using methods such as precise matching and semantic similarity calculation, DEs can be mapped to ontology terms, thereby standardizing DE concepts, value domains, and so on. This can provide specific term annotations for DEs and further enrich the available information.

Focus on Sensitive Data Protection

The existing repository contains DEs that collect sensitive information such as ID numbers, addresses, and phone numbers. However, these elements lack specific classification or identification to indicate that these DEs are used to collect sensitive data and may need to be deidentified or deleted. While the repository does not contain original research data, this remains a crucial issue for subsequent DE usage. We propose that the repository should align with the Health Insurance Portability and Accountability Act [63] or other relevant regulations and map the repository's DEs with the protected personal health information. The repository should create classifications and identifiers for privacy-related DEs. This will serve as a reminder to users about the sensitivity of such data and promote standardized usage practices.

Addressing the balance between FAIR data-sharing principles and privacy protection, we emphasize that FAIR promotes secure, compliant, and interoperable data sharing, not unrestricted dissemination. It advocates for data classification and the application of tailored sharing environments. Privacy data can be deidentified or directly removed during the aggregation phase. In subsequent sharing and reuse processes, while adhering to FAIR principles, we should establish a secure usage environment and sharing guidelines for the data. This includes data classification and grading, implementing differential sharing protocols, and using privacy-enhancing technologies such as privacy computing and federated learning to control data accessibility. This approach ensures effective data sharing and reuse under the FAIR principles while upholding privacy protection.

Implications

Theoretical Implications

In contrast to existing research that mainly concentrates developing specific technical aspects of DE repository construction, this study compares 5 typical DE repositories within the medical field and systematically evaluates and analyzes them. Furthermore, this study introduces a novel analysis framework consisting of 7 dimensions and 36 secondary indicators, based on the ISO/IEC 11179 standard and integrated with the FAIR principles. While this study focuses on the analysis of 5 DE repositories, we are confident that the proposed framework holds broad applicability to a wide range of repositories in the medical field. First, the 5 repositories included in this study have good representativeness, and their functions basically cover small repositories such as samply.MDR and CoMetaR. Therefore, the dimensions and indicators constructed by referring to these repositories can better cover general DE repositories and have more detailed content to be mined. Second, the ISO/IEC 11179 standard is an internationally used standard for the construction of DE repositories, and the FAIR principle is also a widely recognized data management and sharing guideline. Therefore, the dimensions and indicators constructed based on these 2 documents also have good applicability. Finally, in the process of constructing the analysis framework, we invited experts in data management and standardization to discuss and suggest the analysis framework. Simultaneously, the ISO/IEC 11179 standard provides specific definitions for the concept model of DEs and standardizes related management activities. Integrating of these 2 components in the analytical framework serves as the foundation for potential future research endeavors, allowing for further refinement of relevant standards and theories related to DE repositories.

Practical Implications

The practical significance of this study lies in its potential to drive the construction of DE repositories, facilitating a more robust implementation of the FAIR principles during the construction and management processes. This, in turn, contributes to a more substantial role in the data-driven advancement of medicine. For DE repository administrators, this study's findings assist them in understanding the repository's strengths and limitations, offering the necessary information for further improvements to the repository.

In addition, the integrated information on DE repositories from this research may hold practical implications for individuals involved in medical informatics research. For clinical research data managers, this information can assist them in gaining a better understanding of DE repositories. They can use this knowledge to make informed choices regarding suitable repositories and reuse DEs, reducing redundant design work in the clinical research process. For computer experts developing medical information systems, this research encompasses resource organization and management information from

multiple repositories, along with service design offered by web apps. This can reference the top-level structure of DE repositories within their respective institutes.

Conclusions and Limitations

Medical DEs enhance data quality, foster data reuse, and maximize the value of data in the era of health big data. They also form the foundational basis for AI-based medical systems. This study, using a constructed multidimensional analytical framework, evaluates and analyzes the current state of construction of typical medical DE repositories. It summarizes the characteristics of different repositories and provides recommendations based on identified issues. This study's findings can promote the broader application of DE repositories, ensuring that DEs and repositories better serve clinical and medical research needs. Furthermore, this research can have applications in medical knowledge organization, and semantic representation, thus contributing to the development of AI technologies in medicine.

This study also has some limitations and areas for future improvement. First, the study had limited inclusion of databases, focusing solely on comprehensive, noncommercial DE repositories, all in the English language. Smaller or domain-specific repositories may have been overlooked. Furthermore, the data primarily came from repository websites and literature, with little attention given to other sources such as social media accounts. This approach might have missed some of the latest updates or changes. Therefore, future research will consider expanding the scope to include more repositories for analysis, relaxing constraints related to quantity and language. In addition, efforts will be made to enhance the generality of the analysis framework and develop a practical model for DE repositories.

Acknowledgments

This work was supported by the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences Program (grant 2021-I2M-1-057).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete overview of the analysis framework.
[DOCX File , 24 KB - medinform_v12i1e60293_app1.docx]

Multimedia Appendix 2
Details of the FAIR principle.
[DOCX File , 16 KB - medinform_v12i1e60293_app2.docx]

References

1. Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, CASIMIR Rome Meeting participants. Post-publication sharing of data and tools. Nature 2009 Sep 10;461(7261):171-173 [FREE Full text] [doi: 10.1038/461171a] [Medline: 19741686]
2. Dugas M, Hegselmann S, Riepenhausen S, Neuhaus P, Greulich L, Meidt A, et al. Compatible data models at design stage of medical information systems: leveraging related data elements from the MDM portal. Stud Health Technol Inform 2019 Aug 21;264:113-117. [doi: 10.3233/SHTI190194] [Medline: 31437896]
3. Courtot M, Cherubin L, Faulconbridge A, Vaughan D, Green M, Richardson D, et al. BioSamples database: an updated sample metadata hub. Nucleic Acids Res 2019 Jan 08;47(D1):D1172-D1178 [FREE Full text] [doi: 10.1093/nar/gky1061] [Medline: 30407529]
4. Vempati U, Chung C, Mader C, Koleti A, Datar N, Vidović D, et al. Metadata standard and data exchange specifications to describe, model, and integrate complex and diverse high-throughput screening data from the Library of Integrated Network-based Cellular Signatures (LINCS). J Biomol Screen 2014 Jun;19(5):803-816 [FREE Full text] [doi: 10.1177/1087057114522514] [Medline: 24518066]
5. Pacheco AGC, Krohling RA. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. IEEE J Biomed Health Inform 2021 Sep;25(9):3554-3563. [doi: 10.1109/JBHI.2021.3062002] [Medline: 33635800]
6. Olar A, Biricz A, Bedőházi Z, Sulyok B, Pollner P, Csabai I. Automated prediction of COVID-19 severity upon admission by chest X-ray images and clinical metadata aiming at accuracy and explainability. Sci Rep 2023 Mar 14;13(1):4226 [FREE Full text] [doi: 10.1038/s41598-023-30505-2] [Medline: 36918593]
7. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]

8. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data sharing by scientists: practices and perceptions. PLoS One 2011;6(6):e21101 [FREE Full text] [doi: 10.1371/journal.pone.0021101] [Medline: 21738610]
9. Stellmach C, Hopff SM, Jaenisch T, Nunes de Miranda SM, Rinaldi E, NAPKON, LEOSS, ORCHESTRA,ReCoDID Working Groups. Creation of standardized common data elements for diagnostic tests in infectious disease studies: semantic and syntactic mapping. J Med Internet Res 2024 Jun 10;26:e50049 [FREE Full text] [doi: 10.2196/50049] [Medline: 38857066]
10. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. J Biomed Inform 2020 Jul;107:103421 [FREE Full text] [doi: 10.1016/j.jbi.2020.103421] [Medline: 32407878]
11. Pahuja G. Comparative study of metadata standards and metadata repositories. 2011 Presented at: 2ND International Conference on Methods and Models in Science and Technology; 2011 November 19-20; Jaipur, India. [doi: 10.1063/1.3669934]
12. Stausberg J, Harkener S, Burgmer M, Engel C, Finger R, Heinz C, et al. Metadata definition in registries: what is a data element? Stud Health Technol Inform 2022 May 25;294:174-178. [doi: 10.3233/SHTI220432] [Medline: 35612051]
13. Berenspöhler S, Minnerup J, Dugas M, Varghese J. Common data elements for meaningful stroke documentation in routine care and clinical research: retrospective data analysis. JMIR Med Inform 2021 Oct 12;9(10):e27396 [FREE Full text] [doi: 10.2196/27396] [Medline: 34636733]
14. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials 2016 Dec;13(6):671-676 [FREE Full text] [doi: 10.1177/1740774516653238] [Medline: 27311638]
15. Zare S, Meidani Z, Ouhadian M, Akbari H, Zand F, Fakharian E, et al. Identification of data elements for blood gas analysis dataset: a base for developing registries and artificial intelligence-based systems. BMC Health Serv Res 2022 Mar 08;22(1):317 [FREE Full text] [doi: 10.1186/s12913-022-07706-y] [Medline: 35260155]
16. Hirji SA, Salenger R, Boyle EM, Williams J, Reddy VS, Grant MC, et al. Expert consensus of data elements for collection for enhanced recovery after cardiac surgery. World J Surg 2021 Apr;45(4):917-925. [doi: 10.1007/s00268-021-05964-1] [Medline: 33521878]
17. Grinnon ST, Miller K, Marler JR, Lu Y, Stout A, Odenkirchen J, et al. National Institute of Neurological Disorders and Stroke Common Data Element Project - approach and methods. Clin Trials 2012 Jun;9(3):322-329 [FREE Full text] [doi: 10.1177/1740774512438980] [Medline: 22371630]
18. Gay K, Collie D, Sheikh M, Saver J, Warach S, Wright C, et al. National Institute of Neurological Disorders and Stroke Common Data Elements: Stroke Version 2.0 Recommendations. Stroke 2021 Mar;52(Suppl_1):52. [doi: 10.1161/str.52.suppl_1.p192]
19. Vemulapalli S, Simonato M, Ben Yehuda O, Wu C, Feldman T, Popma JJ, et al. Minimum core data elements for transcatheter mitral therapies: scientific statement by PASSION CV, HVC, and TVTR. JACC Cardiovasc Interv 2023 Jun 26;16(12):1437-1447 [FREE Full text] [doi: 10.1016/j.jcin.2023.03.034] [Medline: 37380225]
20. Boesch RP, de Alarcon A, Piccione J, Prager J, Rosen R, Sidell DR, Aerodigestive Research Collaborative. Consensus on triple endoscopy data elements preparatory to development of an aerodigestive registry. Laryngoscope 2022 Nov;132(11):2251-2258. [doi: 10.1002/lary.30038] [Medline: 35122443]
21. Kim HH, Park YR, Lee S, Kim JH. Composite CDE: modeling composite relationships between common data elements for representing complex clinical data. BMC Med Inform Decis Mak 2020 Jul 03;20(1):147 [FREE Full text] [doi: 10.1186/s12911-020-01168-0] [Medline: 32620117]
22. Fitzgerald MP, Kaufman MC, Massey SL, Fridinger S, Prelack M, Ellis C, CHOP Pediatric Epilepsy Program Collaborative, et al. Assessing seizure burden in pediatric epilepsy using an electronic medical record-based tool through a common data element approach. Epilepsia 2021 Jul;62(7):1617-1628 [FREE Full text] [doi: 10.1111/epi.16934] [Medline: 34075580]
23. Vest JR, Adler-Milstein J, Gottlieb LM, Bian J, Campion TR, Cohen GR, et al. Assessment of structured data elements for social risk factors. Am J Manag Care 2022 Jan 01;28(1):e14-e23 [FREE Full text] [doi: 10.37765/ajmc.2022.88816] [Medline: 35049262]
24. Chen EK, Edelen MO, McMullen T, Ahluwalia SC, Dalton SE, Paddock S, et al. Developing standardized patient assessment data elements for Medicare post-acute care assessments. J Am Geriatr Soc 2022 Apr;70(4):981-990. [doi: 10.1111/jgs.17648] [Medline: 35235210]
25. Ryan ME, Warmin A, Binstadt BA, Correll CK, Hause E, Hobday P, Pediatric Rheumatology Care, Outcomes Improvement Network. Capturing critical data elements in Juvenile Idiopathic Arthritis: initiatives to improve data capture. Pediatr Rheumatol Online J 2022 Sep 29;20(1):83 [FREE Full text] [doi: 10.1186/s12969-022-00745-z] [Medline: 36175929]
26. Wyles CC, Fu S, Odum SL, Rowe T, Habet NA, Berry DJ, et al. External validation of natural language processing algorithms to extract common data elements in THA operative notes. J Arthroplasty 2023 Oct;38(10):2081-2084. [doi: 10.1016/j.arth.2022.10.031] [Medline: 36280160]
27. Fu S, Wyles CC, Osmon DR, Carvour ML, Sagheb E, Ramazanian T, et al. Automated detection of periprosthetic joint infections and data elements using natural language processing. J Arthroplasty 2021 Feb;36(2):688-692 [FREE Full text] [doi: 10.1016/j.arth.2020.07.076] [Medline: 32854996]

28. Han P, Fu S, Kolis J, Hughes R, Hallstrom BR, Carvour M, et al. Multicenter validation of natural language processing algorithms for the detection of common data elements in operative notes for total hip arthroplasty: algorithm development and validation. JMIR Med Inform 2022 Aug 31;10(8):e38155 [FREE Full text] [doi: 10.2196/38155] [Medline: 36044253]
29. Renner R, Li S, Huang Y, van der Zijp-Tan AC, Tan S, Li D, et al. Using an artificial neural network to map cancer common data elements to the biomedical research integrated domain group model in a semi-automated manner. BMC Med Inform Decis Mak 2019 Dec 23;19(Suppl 7):276 [FREE Full text] [doi: 10.1186/s12911-019-0979-5] [Medline: 31865899]
30. Rajamohan AG, Patel V, Sheikh-Bahaei N, Liu CJ, Go JL, Kim PE, et al. Common data elements in head and neck radiology reporting. Neuroimaging Clin N Am 2020 Aug;30(3):379-391. [doi: 10.1016/j.nic.2020.05.002] [Medline: 32600638]
31. Littlefield A, Cooke J, Bagge C, Glenn C, Kleiman E, Jacobucci R, et al. Machine learning to classify suicidal thoughts and behaviors: implementation within the common data elements used by the military suicide research consortium. Clinical Psychological Science 2021 Mar 15;9(3):467-481. [doi: 10.1177/2167702620961067]
32. Stausberg J, Löbe M, Verplancke P, Drepper J, Herre H, Löffler M. Foundations of a metadata repository for databases of registers and trials. Stud Health Technol Inform 2009;150:409-413. [doi: 10.3233/978-1-60750-044-5-409] [Medline: 19745342]
33. Hegselmann S, Storck M, Gessner S, Neuhaus P, Varghese J, Bruland P, et al. Pragmatic MDR: a metadata repository with bottom-up standardization of medical metadata through reuse. BMC Med Inform Decis Mak 2021 May 17;21(1):160 [FREE Full text] [doi: 10.1186/s12911-021-01524-8] [Medline: 34001121]
34. Mallya P, Stevens LM, Zhao J, Hong C, Henao R, Economou-Zavlanos N, et al. Facilitating harmonization of variables in Framingham, MESA, ARIC, and REGARDS studies through a metadata repository. Circ Cardiovasc Qual Outcomes 2023 Nov;16(11):e009938. [doi: 10.1161/CIRCOUTCOMES.123.009938] [Medline: 37850400]
35. Wiedekopf J, Ulrich H, Drenkhahn C, Kock-Schoppenhauer A, Ingenerf J. TermiCron - Bridging the Gap Between FHIR Terminology Servers and Metadata Repositories. Stud Health Technol Inform 2022 Jun 06;290:71-75. [doi: 10.3233/SHTI220034] [Medline: 35672973]
36. Schladetzky J, Kock-Schoppenhauer A, Drenkhahn C, Ingenerf J, Wiedekopf J. Mettertron - bridging metadata repositories and terminology servers. Stud Health Technol Inform 2023 Sep 12;307:243-248. [doi: 10.3233/SHTI230721] [Medline: 37697859]
37. Yuan J, Li H. Research on standardization of semantic relation and ontology representation based on MDR. 2022 Presented at: 2022 IEEE 8th International Conference on Computer and Communications (ICCC); 2022 December 09-12; Chengdu, China p. 1490-1494. [doi: 10.1109/iccc56324.2022.10065685]
38. Juárez D, Schmidt EE, Stahl-Toyota S, Ückert F, Lablans M. A generic method and implementation to evaluate and improve data quality in distributed research networks. Methods Inf Med 2019 Sep;58(2-03):86-93 [FREE Full text] [doi: 10.1055/s-0039-1693685] [Medline: 31514209]
39. Kapsner LA, Mang JM, Mate S, Seuchter SA, Vengadeswaran A, Bathelt F, et al. Linking a consortium-wide data quality assessment tool with the MIRACUM Metadata Repository. Appl Clin Inform 2021 Aug;12(4):826-835 [FREE Full text] [doi: 10.1055/s-0041-1733847] [Medline: 34433217]
40. Ulrich H, Kern J, Tas D, Kock-Schoppenhauer AK, Ückert F, Ingenerf J, et al. QLMDR: a GraphQL query language for ISO 11179-based metadata repositories. BMC Med Inform Decis Mak 2019 Mar 18;19(1):45 [FREE Full text] [doi: 10.1186/s12911-019-0794-z] [Medline: 30885183]
41. Sasse J, Darms J, Fluck J. Semantic metadata annotation services in the biomedical domain—a literature review. Applied Sciences 2022 Jan 13;12(2):796. [doi: 10.3390/app12020796]
42. Stöhr MR, Günther A, Majeed RW. The Collaborative Metadata Repository (CoMetaR) web app: quantitative and qualitative usability evaluation. JMIR Med Inform 2021 Nov 29;9(11):e30308 [FREE Full text] [doi: 10.2196/30308] [Medline: 34847059]
43. Reichenpfader D, Glauser R, Dugas M, Denecke K. Assessing and improving the usability of the medical data models portal. Stud Health Technol Inform 2020 Jun 23;271:199-206. [doi: 10.3233/SHTI200097] [Medline: 32578564]
44. Kadioglu D, Breil B, Knell C, Lablans M, Mate S, Schlue D, et al. Samply.MDR—a metadata repository and its application in various research networks. Stud Health Technol Inform 2018;253:50-54 [FREE Full text] [doi: 10.3233/978-1-61499-896-9-50] [Medline: 30147039]
45. MDM.Portal. URL: https://medical-data-models.org/ [accessed 2023-04-07]
46. CoMetaR. URL: https://data.dzl.de/cometar/web/ [accessed 2023-04-11]
47. CentraXX MDR. URL: https://www.toolpool-gesundheitsforschung.de/produkte/centraxx [accessed 2023-04-12]
48. CancerGrid (2005-2010). URL: https://www.cs.ox.ac.uk/projects/cancergrid/ [accessed 2023-04-12]
49. Metadata Online Registry. URL: https://meteor.aihw.gov.au/content/181414 [accessed 2023-04-16]
50. Aristotle Metadata Registry. URL: https://www.aristotlemetadata.com/ [accessed 2023-04-21]
51. caDSR. URL: https://cadsr.cancer.gov/onedata/Home.jsp [accessed 2023-04-22]
52. United States Health Information Knowledgebase (USHIK). URL: https://www.ahrq.gov/data/ushik.html [accessed 2023-04-27]
53. NIH Common Data Elements (CDE) Repository. URL: https://cde.nlm.nih.gov/ [accessed 2023-04-12]
54. CEDAR. URL: https://metadatacenter.org/ [accessed 2023-04-10]

55. Nadkarni PM, Brandt CA. The Common Data Elements for cancer research: remarks on functions and structure. Methods Inf Med 2006;45(6):594-601 [FREE Full text] [doi: 10.1055/s-0038-1634121] [Medline: 17149500]
56. O'Connor MJ, Warzel DB, Martínez-Romero M, Hardi J, Willrett D, Egyedi AL, et al. Unleashing the value of common data elements through the CEDAR workbench. AMIA Annu Symp Proc 2019;2019:681-690 [FREE Full text] [Medline: 32308863]
57. Dugas M, Meidt A, Neuhaus P, Storck M, Varghese J. ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository. BMC Med Res Methodol 2016 Jun 01;16:65 [FREE Full text] [doi: 10.1186/s12874-016-0164-9] [Medline: 27245222]
58. Varghese J, Fujarski M, Hegselmann S, Neuhaus P, Dugas M. CDEGenerator: an online platform to learn from existing data models to build model registries. Clin Epidemiol 2018;10:961-970 [FREE Full text] [doi: 10.2147/CLEP.S170075] [Medline: 30127646]
59. Waithira N, Mutinda B, Cheah PY. Data management and sharing policy: the first step towards promoting data sharing. BMC Med 2019 Apr 17;17(1):80 [FREE Full text] [doi: 10.1186/s12916-019-1315-8] [Medline: 30992010]
60. Paltoo DN, Rodriguez LL, Feolo M, Gillanders E, Ramos EM, Rutter JL, National Institutes of Health Genomic Data Sharing Governance Committees. Data use under the NIH GWAS data sharing policy and future directions. Nat Genet 2014 Sep;46(9):934-938 [FREE Full text] [doi: 10.1038/ng.3062] [Medline: 25162809]
61. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009 Jul;37(Web Server issue):W170-W173 [FREE Full text] [doi: 10.1093/nar/gkp440] [Medline: 19483092]
62. Côté RG, Jones P, Apweiler R, Hermjakob H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics 2006 Feb 28;7:97 [FREE Full text] [doi: 10.1186/1471-2105-7-97] [Medline: 16507094]
63. Atchinson B, Fox DM. The politics of the Health Insurance Portability and Accountability Act. Health Aff (Millwood) 1997;16(3):146-150. [doi: 10.1377/hlthaff.16.3.146] [Medline: 9141331]

Abbreviations

AI: artificial intelligence
API: application programming interface
caDSR: Cancer Data Standards Registry and Repository
CDE: Common Data Element
CEDAR: Center for Expanded Data Annotation and Retrieval
CSV: Comma Separated Values
DE: data element
DSS: data set specification
FAIR: Findable, Accessible, Interoperable, and Reusable
GI: Glossary Item
ISO/IEC: International Organization for Standardization/International Electrotechnical Commission
LOINC: Logical Observation Identifiers Names and Codes
MDM-Portal: Portal of Medical Data Models
METEOR: Metadata Online Registry
NCBO: National Center for Biomedical Ontology
NCIT: National Cancer Institute Thesaurus
NIH: National Institutes of Health
NINDS: National Institute of Neurological Disorders and Stroke.
ODM: Operational Data Model
OWL: Web Ontology Language
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 11.05.24; peer-reviewed by C Gaudet-Blavignac, AJ Ponsero; comments to author 16.06.24; revised version received 07.07.24; accepted 21.07.24; published 30.09.24.

Please cite as:
Hu Z, Wang A, Duan Y, Zhou J, Hu W, Wu S
Toward Better Semantic Interoperability of Data Element Repositories in Medicine: Analysis Study
JMIR Med Inform 2024;12:e60293
URL: https://medinform.jmir.org/2024/1/e60293
doi:10.2196/60293
PMID:39348178

The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation

Yong Nam Gwon[1,*], MD; Jae Heon Kim[1,*], MD, PhD; Hyun Soo Chung[2], MD; Eun Jee Jung[2], MD; Joey Chun[1,3], MD; Serin Lee[1,4], MD; Sung Ryul Shim[5,6], MPH, PhD

1

2

3

4

5

6

*these authors contributed equally

Corresponding Author:
Sung Ryul Shim, MPH, PhD

Abstract

Background: A large language model is a type of artificial intelligence (AI) model that opens up great possibilities for health care practice, research, and education, although scholars have emphasized the need to proactively address the issue of unvalidated and inaccurate information regarding its use. One of the best-known large language models is ChatGPT (OpenAI). It is believed to be of great help to medical research, as it facilitates more efficient data set analysis, code generation, and literature review, allowing researchers to focus on experimental design as well as drug discovery and development.

Objective: This study aims to explore the potential of ChatGPT as a real-time literature search tool for systematic reviews and clinical decision support systems, to enhance their efficiency and accuracy in health care settings.

Methods: The search results of a published systematic review by human experts on the treatment of Peyronie disease were selected as a benchmark, and the literature search formula of the study was applied to ChatGPT and Microsoft Bing AI as a comparison to human researchers. Peyronie disease typically presents with discomfort, curvature, or deformity of the penis in association with palpable plaques and erectile dysfunction. To evaluate the quality of individual studies derived from AI answers, we created a structured rating system based on bibliographic information related to the publications. We classified its answers into 4 grades if the title existed: A, B, C, and F. No grade was given for a fake title or no answer.

Results: From ChatGPT, 7 (0.5%) out of 1287 identified studies were directly relevant, whereas Bing AI resulted in 19 (40%) relevant studies out of 48, compared to the human benchmark of 24 studies. In the qualitative evaluation, ChatGPT had 7 grade A, 18 grade B, 167 grade C, and 211 grade F studies, and Bing AI had 19 grade A and 28 grade C studies.

Conclusions: This is the first study to compare AI and conventional human systematic review methods as a real-time literature collection tool for evidence-based medicine. The results suggest that the use of ChatGPT as a tool for real-time evidence generation is not yet accurate and feasible. Therefore, researchers should be cautious about using such AI. The limitations of this study using the generative pre-trained transformer model are that the search for research topics was not diverse and that it did not prevent the hallucination of generative AI. However, this study will serve as a standard for future studies by providing an index to verify the reliability and consistency of generative AI from a user's point of view. If the reliability and consistency of AI literature search services are verified, then the use of these technologies will help medical research greatly.

(JMIR Med Inform 2024;12:e51187) doi:10.2196/51187

KEYWORDS

artificial intelligence; search engine; systematic review; evidence-based medicine; ChatGPT; language model; education; tool; clinical decision support system; decision support; support; treatment

Introduction

The global artificial intelligence (AI) health care market size was estimated to be at US $15.1 billion in 2022 and is expected to surpass approximately US $187.95 billion by 2030, growing at an annualized rate of 37% during the forecast period from 2022 to 2030 [1]. In particular, innovative applications of medical AI are expected to increase in response to medical demand, which will explode in 2030 [2,3].

A large language model (LLM) is a type of AI model that opens up great possibilities for health care practice, research, and education, although scholars have emphasized the need to proactively address the issue of unvalidated and inaccurate information regarding its use [4,5]. One of the best-known LLMs is ChatGPT (OpenAI). It was launched in November 2022. Similar to other LLMs, ChatGPT is trained on huge text data sets in numerous languages, allowing it to respond to text input with humanlike responses [4]. Developed by the San Francisco–based AI research laboratory OpenAI, ChatGPT is based on a generative pre-trained transformer (GPT) architecture. It is considered an advanced form of a chatbot, an umbrella term for a program that uses a text-based interface to understand and generate responses. The key difference between a chatbot and ChatGPT is that a chatbot is usually programmed with a limited number of responses, whereas ChatGPT can produce personalized responses according to the conversation [4,6].

Sallam's [5] systematic review (SR) sought to identify the benefits and current concerns regarding ChatGPT. That review advises that health care research could benefit from ChatGPT, since it could be used to facilitate more efficient data set analysis, code generation, and literature reviews, thus allowing researchers to concentrate on experimental design as well as drug discovery and development. The author also suggests that ChatGPT could be used to improve research equity and versatility in addition to its ability to improve scientific writing. Health care practice could also benefit from ChatGPT in multiple ways, including enabling improved health literacy and delivery of more personalized medical care, improved documentation, workflow streamlining, and cost savings. Health care education could also use ChatGPT to provide more personalized learning with a particular focus on problem-solving and critical thinking skills [5]. However, the same review also lays out the current concerns, including copyright issues, incorrect citations, and increased risk of plagiarism, as well as inaccurate content, risk of excessive information leading to an infodemic on a particular topic, and cybersecurity issues [5].

A key question regarding the use of ChatGPT is if it can use evidence to identify premedical content. Evidence-based medicine (EBM) provides the highest level of evidence in medical treatment by integrating clinician experience, patient value, and best-available scientific information to guide decision-making on clinical management [7]. The principle of EBM means that the most appropriate treatment plan for patients should be devised based on the latest empirical research evidence. However, the scientific information identified by ChatGPT is not yet validated in terms of safety or accuracy according to Sallam [5], who further suggests that neither doctors nor patients should rely on it at this stage. In contrast, another study by Zhou et al [8] found that answers provided by ChatGPT were generally based on the latest verified scientific evidence, that is, the advice given followed high-quality treatment protocols and adhered to guidelines from experts.

In medicine, a clinical decision support system (CDSS) uses real-time evidence to support clinical decision-making. This is a fundamental tool in EBM, which uses SRs based on a systematic, scientific search of a particular subject. If ChatGPT becomes a CDSS, it is fundamental to determine whether it is capable of performing a systematic search based on real-time generation of evidence in the medical field. Therefore, this study will be the first to determine whether ChatGPT can search papers for an SR. In particular, this study aims to present a standard for medical research using generative AI search technology in the future by providing indicators for the reliability and consistency of generative AI searches from a user's perspective.

Methods

Ethical Considerations

As per 45 CFR §46.102(f), the activities performed herein were considered exempt from institutional review board approval due to the data being publicly available. Informed consent was not obtained, since this study used previously published deidentified information that was available to the general public. This study used publicly available data from PubMed, Embase, and Cochrane Library and did not include human participant research.

Setting the Benchmark

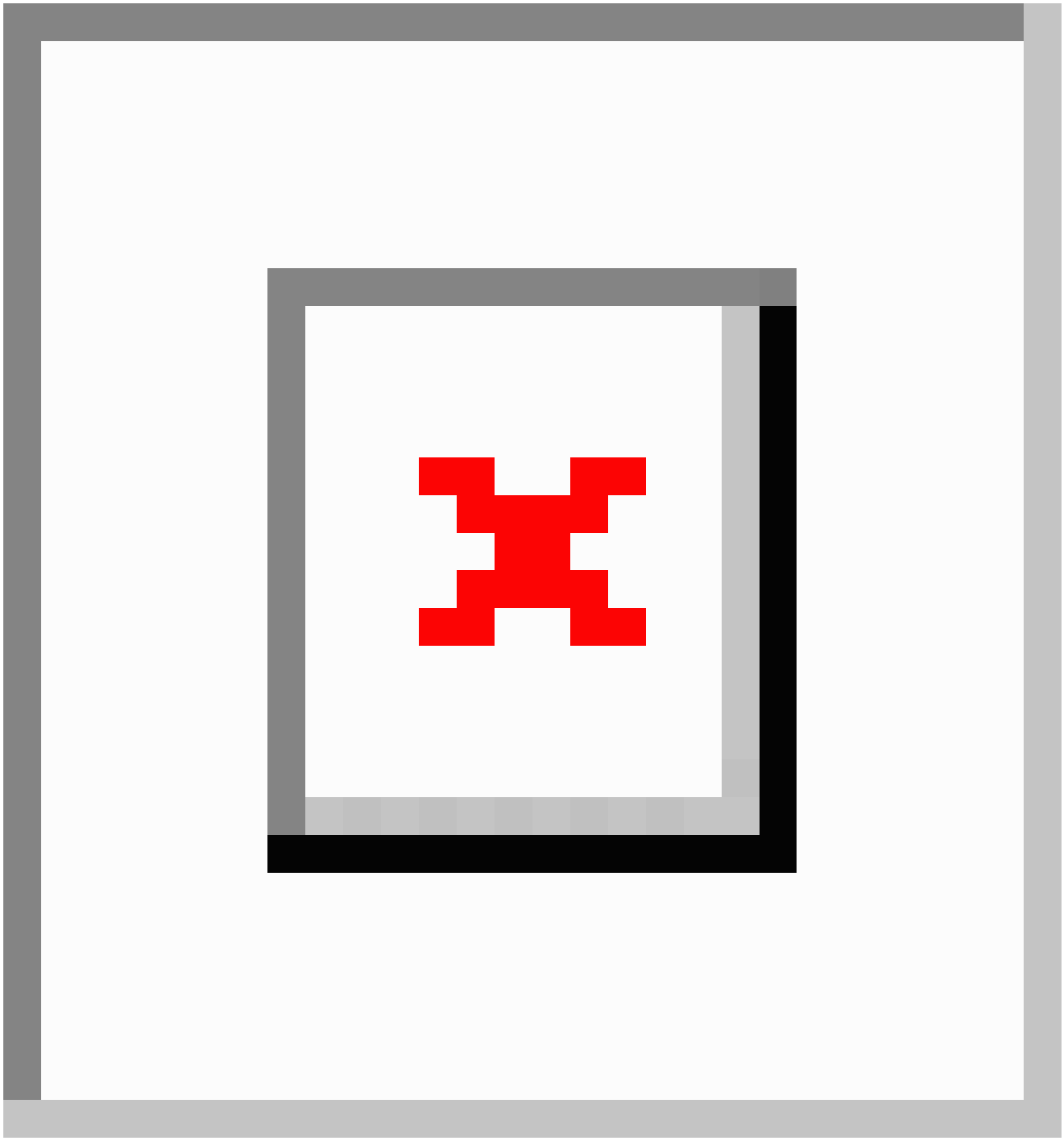
To determine whether ChatGPT, currently the most representative LLM, is capable of systematic searches, we set an SR that was performed by human experts as a benchmark and checked how many studies were finally included in the benchmark were presented by ChatGPT. We chose Lee et al [9] as the benchmark for the following reasons. First, Lee et al [9] performed an SR and meta-analysis about the medical treatment for Peyronie disease (PD) with human experts. PD typically presents with discomfort, curvature, or deformity of the penis in association with palpable plaques and erectile dysfunction [10]. Second, it was easy to compare the results of ChatGPT and the benchmark, because we had full information about the interim process and results of the study. Third, a sufficient amount of studies has been published about the medical treatment for PD, but there is still no consensus answer. So, we expected to assess the sole ability of ChatGPT as a systematic search tool with sufficient data while avoiding any possible pretrained bias. Lastly, with the topic of Lee et al [9], we could build questions that start broad and become more specific and add some conditions that could test ChatGPT's comprehension about scientific research. For example, questions could not only be built broadly by asking about "medical treatment for Peyronie's disease" but also specifically by asking about "oral therapy for Peyronie's disease" or "colchicine for Peyronie's disease." Because Lee et al [9] only contained randomized controlled trials (RCTs), we could add a condition

to the questions to restrict the study type to RCTs, which could be useful to assess the comprehension of ChatGPT.

Systematic Search Formula of Benchmark

Lee et al [9] used the following search query in PubMed and Cochrane Library: *("penile induration"[MeSH Terms] OR "Peyronie's disease"[Title/Abstract]) AND "male"[MeSH Terms] AND "randomized controlled trial"[Publication Type]*, and the following query in Embase: *('Peyronie disease'/exp OR 'Peyronie's diseas':ab,ti) AND 'male'/exp AND 'randomized controlled trial'/de*. After the systematic search, a total of 217 records were identified. Studies were excluded for the following reasons: not RCTs, not perfectly fit to the topic, not enough sample size or outcome, and not written in English. Finally, 24 RCTs were included in the SR, with only 1 RCT published in 2022 (Figure 1) [9]. The characteristics of all studies included in Lee et al [9] are summarized in Section S1 in Multimedia Appendix 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for Lee et al [9]. RCT: randomized controlled trial.



Methodology of Systematic Search for ChatGPT

Based on the search formula used in Lee et al [9], a simple mandatory prompt in the form of a question was created, starting with comprehensive questions and gradually asking more specific questions (Textbox 1). For example, questions could be built as "Could you show RCTs of colchicine for Peyronie's disease in PubMed?" with the treatment and database changed

under the same format. In addition to mandatory questions, we added questions about treatment additionally provided by ChatGPT during the conversation. Considering the possibility that ChatGPT might respond differently depending on the interaction, we arranged questions into 2 logical flows, focusing on database and treatment, respectively (Figure 2 and Figure S1 in Multimedia Appendix 1). We asked about search results from 4 databases: PubMed [11], Google (Google Scholar) [12], Cochrane Library [13], and ClinicalTrials.gov [14]. PubMed is a leading biomedical database offering access to peer-reviewed articles. Google Scholar provides a wide-ranging index of scholarly literature, including medical studies. Cochrane Library specializes in high-quality evidence through SRs and clinical trials. ClinicalTrials.gov, managed by the National Library of Medicine, serves as a comprehensive repository for clinical study information globally. These databases collectively serve researchers by providing access to diverse and credible sources, facilitating literature reviews and evidence synthesis, and informing EBM in the medical field. They play crucial roles in advancing medical knowledge, supporting informed decision-making, and ultimately improving patient care outcomes [11-14]. These 4 databases were easy to access and contained most of the accessible studies. Each question was repeated at least twice. We extracted the answers and evaluated the quality of information based on the title, author, journal, and publication year (Sections S2-S5 Multimedia Appendix 1).

Textbox 1. Mandatory question prompts.

Basic format of questions

- "Could you show RCTs of (A) for Peyronie's disease in (B)?"

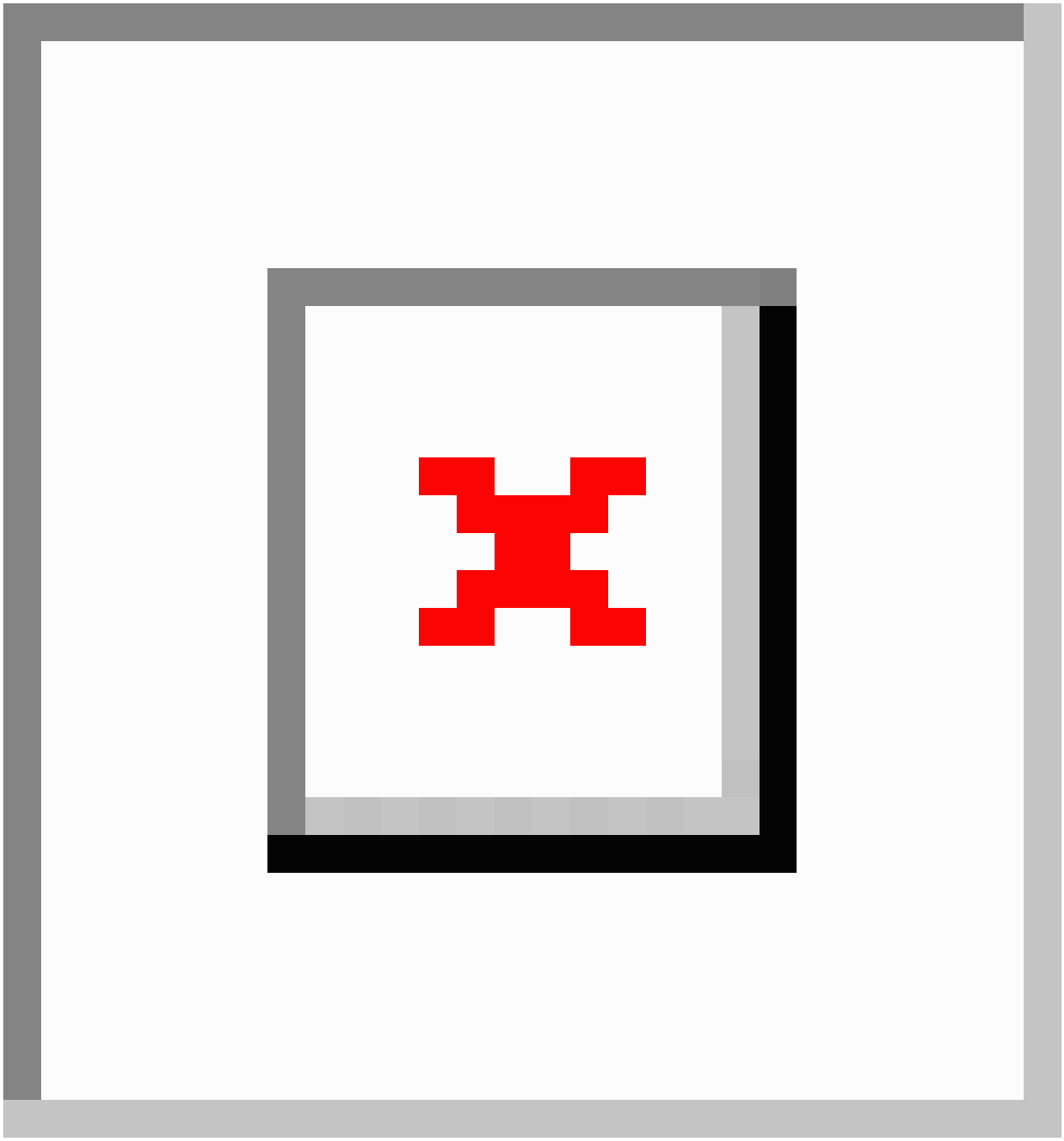
(A) Treatment category and specific treatment

- **Oral therapy**
  - Vitamin E, colchicine, L-carnitine, potassium aminobenzoate, tamoxifen, pentoxifylline, tadalafil, L-arginine, and sildenafil
- **Intralesional therapy**
  - Verapamil, interferon-a2B, collagenase *Clostridium histolyticum*, transdermal electromotive administration, hyaluronidase, triamcinolone, mitomycin C, super-oxide dismutase, and 5-fluorouracil
- **Mechanical therapy**
  - Extracorporeal shockwave therapy, iontophoresis, traction therapy, vacuum, penile massage, and exercise shockwave therapy
- **Topical therapy**
  - 5-Alpha-reductase inhibitors, superficial heat, diclofenac gel, collagenase *Clostridium histolyticum* gel, verapamil gel, potassium aminobenzoate gel, and propionyl-L-carnitine gel

(B) Database

- PubMed
- Google (Google Scholar)
- Cochrane Library
- ClinicalTrials.gov

Figure 2. Logical flow and results focusing on database for ChatGPT. RCT: randomized controlled trial.



We used the GPT-3.5 version of ChatGPT, which was pretrained with data before 2021, for the systematic search and evaluated how many RCTs that were included in Lee et al [9] were present in the search results from ChatGPT. To assess the reliability of ChatGPT's answers, we also evaluated whether the studies presented actually existed. ChatGPT's response style and the amount of information might vary from answer to answer. Thus, we evaluated the accuracy of the responses by prioritizing a match by (1) title; (2) author, journal, and publication year; and (3) other items.

To obtain higher-quality responses, it is important to structure the prompts using refined language that is well understood by the LLM [15-17]. In this study, we performed the following fine-tuning to clearly convey the most important content or information. We first defined roles and provided context and input data before asking complete questions to get responses, and we used specific and varied examples to help the model narrow its focus and produce more accurate results [18,19]. During the prompt engineering, the treatment category, specific treatment, and target databases were structured in order, and the order was changed in the detailed elements to induce consistent answers. Details of this are presented in Multimedia Appendix 1.

Quality Assessment of Answers

To evaluate the quality of individual studies derived from AI answers, we created a structured rating system based on bibliographic information related to the publications (Table 1).

We classified its answers into 4 grades if the title existed: A, B, C, and F. No grade was given for a fake title or no answer.

Table . Grade table based on bibliographic information.

Grade	Title actually exists	PICOS[a]	Essential information				Accessory information				Definition of grade
			Title	Author	Journal	Publication year	Issue number	Page number	DOI	PMID	
A	Yes	✓[b]	✓	✓	✓	✓	✓	✓	✓	✓	All bibliographic information matched
B	Yes	✓	✓	✓	✓	✓	Any X[c]	Any X	Any X	Any X	PICOS and essential information matched, but not accessory information
C	Yes	X[d]	✓	✓	✓	✓	N/A[e]	N/A	N/A	N/A	Essential information matched, but not PICOS
F	Yes	N/A	✓	Any X	Any X	Any X	N/A	N/A	N/A	N/A	Title matched, but not other essential information

[a]PICOS: population, intervention, comparison, outcome, and study design (research questions).

[b]Matched.

[c]Any mismatch in essential information or accessory information.

[d]Mismatch.

[e]N/A: not assessed.

A grade of "A" was given to an answer that was appropriate for the question and perfectly consistent with the actual study. For example, for the question "Could you show all RCTs of medical treatment for Peyronie's disease in Google?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey F. Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. J Urol. 1993; 149(1):56-58." This answer was perfectly consistent with actual study because the authors, title, journal, publication year, issue number, and even page number matched, and it was appropriate because it was an RCT, was about medical treatment for PD, and could be found in Google.

A grade of "B" was given to an answer that was appropriate for the question and the title, author, journal, and publication year matched with the actual study. For example, for the question "Could you show RCTs of Intralesional therapy for Peyronie's disease in Cochrane Library?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey F. Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. J Urol. 1993 Feb; 149(2):56-8. doi: 10.1016/s0022-5347(17)36148-1. PMID: 8426418." This answer was appropriate because it was an RCT, was about intralesional therapy for PD, and could be found in Cochrane Library. However, this answer was consistent with the title, authors, journal, and publication year of actual study but inconsistent with the accessory information for publication month, issue number, page number, DOI, and PMID. So, it was able to specify the presented study with essential information, but the accessory information was not consistent with the actual study.

A grade of "C" was given to an answer that was inappropriate for the question but the title, author, journal, and publication year matched with the actual study. For example, for the question "Could you show RCTs of Pentoxifylline for Peyronie's disease in PubMed?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey F. Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. J Urol. 1993; 149(1):56-8. doi: 10.1016/s0022-5347(17)36091-7. PMID: 8417228." This answer was consistent with the title, authors, journal, and publication year of the actual study, but it was inappropriate because it was not about the use of pentoxifylline for PD.

A grade of "F" was given to an answer that was inconsistent with the author, journal, or publication year of the actual study that matched the title, making it impossible to be specified. Thus, it was considered as a fake study that was generated. For

example, for the question "Could you show RCTs of collagenase *Clostridium histolyticum* for Peyronie's disease in PubMed?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey FJ, & Collagenase Study Group. (2012). Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. The Journal of urology, 187(3), 948-953." This answer was consistent with the title of the actual study but inconsistent with the authors, publication year, and so on.

Searching Strategy for Bing AI

To compare with ChatGPT, we performed the same process with Bing AI [20], also known as "New Bing," an AI chatbot developed by Microsoft and released in 2023. Since Bing AI functions based on the huge AI model "Prometheus" that includes OpenAI's GPT-4 with web searching capabilities, it is expected to give more accurate answers than the GPT-3.5 version of ChatGPT. We performed the conversation with the "Precise" tone. Because Bing AI limited the number of questions per session to 20, we did not arrange questions into 2 logical flows (Section S6 in Multimedia Appendix 1). We compared the number of studies included in the benchmark [9] and provided by Bing AI. We also evaluated the reliability of answers with the same method described above or using links of websites presented by Bing AI (Figure S2 and Section S7 in Multimedia Appendix 1).

Results

Systematic Search Results via ChatGPT

A total of 639 questions were entered into ChatGPT, and 1287 studies were obtained (Table 2). The systematic search via ChatGPT was performed from April 17 to May 6, 2023. At the beginning of the conversation, we gave ChatGPT the role of a researcher conducting a systematic search who intended to perform a meta-analysis for more appropriate answers. At first, we tried to build question format by using the word "find," such as "Could you find RCTs of medical treatment for Peyronie's disease?" However, ChatGPT did not present studies and only suggested how to find RCTs in a database, such as PubMed. Therefore, we changed the word "find" to "show," and ChatGPT presented lists of RCTs. For comprehensive questions, ChatGPT did not give an answer, saying that it did not have the capability to show a list of RCTs as an AI language model. However, when questions were gradually specified, it created answers (Sections S2 and S4 in Multimedia Appendix 1).

Table . Quality assessment of answers from ChatGPT and Bing AI[a].

Searcher, setting, and question level		Grade, n				Studies, n
		A	B	C	F	
ChatGPT						
Database setting						
	Comprehensive question	1	0	3	5	56
	Category-specific question	1	1	8	18	124
	Treatment-specific question	4	7	67	87	545
	Total	6	8	78	110	725
Treatment setting						
	Comprehensive question	0	0	0	1	27
	Category-specific question	0	0	4	8	61
	Treatment-specific question	1	10	85	92	474
	Total	1	10	89	101	562
Total		7	18	167	211	1287
Bing AI						
Comprehensive question		0	0	1	0	1
Category-specific question		0	0	7	0	7
Treatment-specific question		19	0	20	0	40
Total		19	0	28	0	48
Human[b]		24	0	0	0	24

[a]AI: artificial intelligence.

[b]From Lee et al [9].

Of the 1287 studies provided by ChatGPT, only 7 (0.5%) studies were perfectly eligible and 18 (1.4%) studies could be considered suitable under the assumption that they were real studies if only the title, author, journal, and publication year matched (Table 2). Among these, only 1 study was perfectly consistent with studies finally included in Lee et al [9], and 4 studies were matched under the assumption (Sections S1, S3, and S5 in Multimedia Appendix 1).

Specifically, systematic search via ChatGPT was performed in 2 logical flow schemes, database setting and treatment setting (Figure 2 and Figure S1 in Multimedia Appendix 1). With the logical flow by database setting, among the 725 obtained studies, 6 (0.8%) and 8 (1.1%) studies were classified as grade A and grade B, respectively (Table 1). Of these, 1 grade A study and 1 grade B study were included in Lee et al [5]. With the logical flow by treatment setting, among the 562 obtained studies, 1 (0.2%) study was classified as grade A and 10 (1.8%) studies were classified as grade B. Of these, 3 grade B studies were included in the benchmark [9] (Table 2).

It was common for answers to be changed. There were many cases where answers contradicted themselves. In addition, there were cases where the answer was “no capability” or “no RCT found” at first, but when another question was asked and the previous question was asked again, an answer was given. ChatGPT showed a tendency to create articles by rotating some format and words. Titles presented were so plausible that it was almost impossible to identify fake articles until an actual search was conducted. The presented authors were also real people. Titles often contained highly specific numbers, devices, or brand names that were real. There were some cases where it was possible to infer which articles ChatGPT mimicked in the fake answers (Sections S3 and S5 in Multimedia Appendix 1). Considering these characteristics, when generating sentences, ChatGPT seemed to list words with a high probability of appearing among pretrained data rather than presenting accurate facts or understanding questions.

In conclusion, of the 1287 studies presented by ChatGPT, only 1 (0.08%) RCT matched the 24 RCTs of the benchmark [9].

Systematic Search Results via Bing AI

For Bing AI, a total of 223 questions were asked and 48 studies were presented. Among the 48 obtained studies, 19 (40%) studies were classified as grade A. There were no grade B

studies (Table 2). Because Bing AI always gave references with links to the websites, all studies presented by Bing AI existed. However, it also provided wrong answers about the study type, especially as it listed reviews as RCTs. Of the 28 studies with grade C, 27 (96%) were not RCTs and 1 (4%) was about a different treatment. Only 1 study had no grade because of a fake title; it presented a study registered in PubMed while pretending that it was the result of a search in ClinicalTrials.gov. However, the study was not in ClinicalTrials.gov (Section S7 in Multimedia Appendix 1).

Bing AI had more accurate answers than ChatGPT since it provides actual website references. However, it also showed a tendency to give more answers to more specific questions, similar to ChatGPT. For example, with a comprehensive question, Bing AI said "I am not able to access or search specific databases." However, with more specific questions, it found studies or answered "I couldn't find any RCTs' without mention about accessibility." In most cases, Bing AI either failed to find studies or listed too few studies to be used as a systematic searching tool.

In conclusion, of the 48 studies presented by Bing AI, 2 (4%) RCTs matched the 24 RCTs of the benchmark [9].

Discussion

Principal Findings

This paper's researchers sought to determine whether ChatGPT could conduct a real-time systematic search for EBM. For the first time, researchers compared the performance of ChatGPT with classic systematic searching as well as the Microsoft Bing AI search engine. Although Zhou et al [8] suggested that ChatGPT answered qualitative questions based on recent evidence, this study found that ChatGPT's results were not based on a systematic search (which is the basis for an SR), meaning that they could not be used for real-time CDSS in their current state.

With recent controversy regarding the risks and benefits of advanced AI technologies [21-24], ChatGPT has received mixed responses from the scientific community and academia. Although many scholars agree that ChatGPT can increase the efficiency and accuracy of the output in writing and conversational tasks [25], others suggest that the data sets used in ChatGPT's training might lead to possible bias, which not only limits its capabilities but also leads to the phenomenon of hallucination—apparently scientifically plausible yet factually inaccurate information [24]. Caution around the use of LLMs should also bear in mind security concerns, including the potential of cyberattacks that deliberately spread misinformation [25].

When applying the plug-in method in this study, especially when using PubMed Research [26], the process worked smoothly and there was not a single case of hallucination of fake research (by providing information along with a link), regardless of the designation of a specific database engine. Among the responses, 21 RCTs were included in the final SR, and out of a total of 24, all RCTs except 3 were provided. This is a very encouraging result. However, there is no plug-in that allows access to other databases yet, and if the conversation is long, the response speed is very slow. Furthermore, although it is a paid service, it only provides a total of 100 papers, so if more than 100 RCTs are searched, the user must manually search all papers. Ultimately, it is not intended for conducting an efficient and systematic search, as additional time and effort are required. If a more efficient plug-in is developed, this could play a promising part in systematic searches.

Although Sallam's [5] SR suggests that academic and scientific writing as well as health care practice, research, and education could benefit from the use of ChatGPT, this study found that ChatGPT could not search scientific articles properly, with a 0.08% (1/1287) of probability of the desired paper being presented. In the case of Bing AI using GPT-4, this study showed that Bing AI could search scientific articles with a much higher accuracy than ChatGPT. However, the probability was only 4% (2/48). It was still an insufficient probability for performing systematic research. Moreover, fake answers generated by ChatGPT, known as hallucinations, caused researchers to spend extra time and effort by checking the accuracy of the answers. A typical problem with generative AI is that it creates hallucinations. However, this is difficult to completely remove due to the principle of generative AI. Therefore, if it cannot be prevented from the pretraining of the model, efforts to increase reliability and consistency in the use of generative AI in medical care by checking the accuracy from the user's point of view are required, as shown in this study. Unlike ChatGPT, Bing AI did not generate fake studies. However, the total number of studies presented was too small. Very few studies have focused on the scientific searching accuracy of ChatGPT. Although this paper found many articles about the use of ChatGPT in the medical field, the majority concerned the role of ChatGPT as an author. Although the latter might accelerate writing efficiency, it also confirms the previously mentioned issues of transparency and plagiarism.

Wang et al [27] have recently investigated whether ChatGPT could be used to generate effective Boolean queries for an SR literature search. The authors suggest that ChatGPT should be considered a "valuable tool" for researchers conducting SRs, especially for time-constrained rapid reviews where trading off higher precision for lower recall is generally acceptable. They cite its ability to follow complex instructions and generate high-precision queries. Nonetheless, it should be noted that building a Boolean query is not a complex process. However, selecting the most appropriate articles for an SR is critical, which might be a more useful subject to examine in relation to the use of ChatGPT. Moreover, although Aydın and Karaarslan [28] have indicated that ChatGPT shows promise in generating a literature review, the iThenticate plagiarism tool found significant matches in paraphrased elements.

In scientific research, the most time-consuming and challenging task can be the process of filtering out unnecessary papers on the one hand and identifying those that are needed on the other hand. This difficult yet critical task can be daunting. It discourages many researchers from participating in scientific research. If AI could replace this process, it will be easier to collect and analyze data from the selected papers. Recently, commercial literature search services using generative AI models

have emerged. Representative examples include Covidence [29], Consensus [30], and Elicit [31]. The technical details of these commercial AI literature search services are unknown, but they are based on LLMs using GPT. Therefore, these search services are not only insufficient to verify hallucinations but also lack information in the search target databases. Even if there may be mistakes, the researcher should aim for completeness, and unverified methods should be avoided. Although this study did not use a commercial literature search service, it manually searched the target databases one by one. If the reliability and consistency of AI literature search services are verified, the use of these technologies will help medical research greatly

This study suggests that ChatGPT still has limitations in academic search, despite the recent assertion from Zhou et al [8] about its potential in searching for academic evidence. Moreover, although ChatGPT can search and identify guidance in open-access guidelines, its results are brief and fragmentary, often with just 1 or 2 sentences that lack relevant details about the guidelines.

Arguably, more concern should be placed on the potential use of ChatGPT in a CDSS than its role in education or writing draft papers. On the one hand, if AI such as ChatGPT is used within a patient-physician relationship, this is unlikely to affect liability since the advice is filtered through professionals' judgment and inaccurate advice generated by AI is no different from erroneous or harmful information disseminated by a professional. However, ChatGPT lacks sufficient accuracy and speed to be used in this manner. On the other hand, ChatGPT could also be used to give direct-to-consumer advice, which is largely unregulated since asking AI directly for medical advice or emotional support acts outside the established patient-physician relationship [32]. Since there is a risk of patient knowing inaccurate information, the medical establishment should seek to educate patients and guardians about the risk of inaccurate information in this regard.

Academic interest in ChatGPT to date has mainly focused on potential benefits including research efficiency and education, drawbacks related to ethical issues such as plagiarism and the risk of bias, as well as security issues including data privacy. However, in terms of providing medical information and acting as a CDSS, the use of ChatGPT is currently less certain because its academic search capability is potentially inaccurate, which is a fundamental issue that must be addressed.

The limitation of this study is that it did not address various research topics, because only 1 research topic was searched when collecting target literature. In addition, due to the time difference between the start of the study and the review and evaluation period, the latest technology could not be fully applied because it could become an outdated technology in a field of study where technology advances rapidly, such as generative AI. For example, there have already been significant technological advances since new AI models such as ChatGPT Turbo (4.0) were released between the time we started this study and the current revised time point.

This paper thus suggests that the use of AI as a tool for generating real-time evidence for a CDSS is a dream that has not yet become a reality. The starting point of evidence generation is a systematic search and ChatGPT is unsuccessful even for this initial purpose. Furthermore, its potential use in providing advice directly to patients in a direct-to-consumer form is concerning, since ChatGPT could provide inaccurate medical information that is not evidence based and can result in harm. For the proper use of generative AI in medical care in the future, it is suggested that a feedback model that evaluates accuracy according to experts' perspective, as done in this study, and then reflects it back into an LLM is necessary.

Conclusion

This is the first study to compare AI and conventional human SR methods as a real-time literature collection tool for EBM. The results suggest that the use of ChatGPT as a tool for real-time evidence generation is not yet accurate and feasible. Therefore, researchers should be cautious about using such AI. The limitations of this study using the GPT model are that the search for research topics was not diverse and that it did not prevent the hallucinations of generative AI. However, this study will serve as a standard for future studies by providing an index to verify the reliability and consistency of generative AI from a user's point of view. If the reliability and consistency of AI literature search services are verified, the use of these technologies will help medical research greatly.

Acknowledgments

This work was supported by the Soonchunhyang University Research Fund. This body had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Authors' Contributions

SRS had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. YNG, HSC, EJJ, JC, SL, and SRS contributed to the analysis and interpretation of data. YNG, HSC, SRS, and JHK contributed to the drafting of the manuscript. SRS and JHK contributed to critical revision of the manuscript for important intellectual content. YNG and SRS contributed to statistical analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional logical flow diagrams, characteristics of studies included in Lee et al [9], ChatGPT and Microsoft Bing transcripts, and grade classification for answers.
[DOCX File, 2209 KB - medinform_v12i1e51187_app1.docx]

References

1. Artificial intelligence (AI) in healthcare market (by component: software, hardware, services; by application: virtual assistants, diagnosis, robot assisted surgery, clinical trials, wearable, others; by technology: machine learning, natural language processing, context-aware computing, computer vision; by end user) - global industry analysis, size, share, growth, trends, regional outlook, and forecast 2022-2030. Precedence Research. 2023 Feb. URL: https://www.precedenceresearch.com/artificial-intelligence-in-healthcare-market [accessed 2024-03-31]
2. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
3. Zahlan A, Ranjan RP, Hayes D. Artificial intelligence innovation in healthcare: literature review, exploratory analysis, and future research. Technol Soc 2023 Aug;74:102321. [doi: 10.1016/j.techsoc.2023.102321]
4. Models. OpenAI. URL: https://platform.openai.com/docs/models/gpt-3-5 [accessed 2023-06-14]
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
6. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]
7. Evidence-Based Medicine Working Group. Evidence-based medicine. a new approach to teaching the practice of medicine. JAMA 1992 Nov 4;268(17):2420-2425. [doi: 10.1001/jama.1992.03490170092032] [Medline: 1404801]
8. Zhou Z, Wang X, Li X, Liao L. Is ChatGPT an evidence-based doctor? Eur Urol 2023 Sep;84(3):355-356. [doi: 10.1016/j.eururo.2023.03.037] [Medline: 37061445]
9. Lee HY, Pyun JH, Shim SR, Kim JH. Medical treatment for Peyronie's disease: systematic review and network Bayesian meta-analysis. World J Mens Health 2024 Jan;42(1):133. [doi: 10.5534/wjmh.230016]
10. Chung E, Ralph D, Kagioglu A, et al. Evidence-based management guidelines on Peyronie's disease. J Sex Med 2016 Jun;13(6):905-923. [doi: 10.1016/j.jsxm.2016.04.062] [Medline: 27215686]
11. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/about/ [accessed 2023-06-14]
12. Google Scholar. URL: https://scholar.google.com/ [accessed 2023-06-14]
13. Cochrane Library. URL: https://www.cochranelibrary.com/ [accessed 2023-06-14]
14. ClinicalTrials.gov. URL: https://classic.clinicaltrials.gov/ [accessed 2023-06-14]
15. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv. Preprint posted online on Nov 28, 2023. [doi: 10.48550/arXiv.2311.16452]
16. Ziegler A, Berryman J. A developer's guide to prompt engineering and LLMs. GitHub Blog. 2023 Jul 17. URL: https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/ [accessed 2023-07-17]
17. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/blog/chatgpt [accessed 2023-10-16]
18. Reid R. How to write an effective GPT-3 or GPT-4 prompt. Zapier. 2023 Aug 3. URL: https://zapier.com/blog/gpt-prompt/ [accessed 2023-10-14]
19. Prompt engineering for generative AI. Google. 2023 Aug 8. URL: https://developers.google.com/machine-learning/resources/prompt-eng?hl=en [accessed 2024-04-23]
20. Bing. URL: https://www.bing.com/ [accessed 2024-04-30]
21. de Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023 Apr 25;11:1166120. [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
22. Howard J. Artificial intelligence: implications for the future of work. Am J Ind Med 2019 Nov;62(11):917-926. [doi: 10.1002/ajim.23037] [Medline: 31436850]
23. Tai MCT. The impact of artificial intelligence on human society and bioethics. Tzu Chi Med J 2020 Aug 14;32(4):339-343. [doi: 10.4103/tcmj.tcmj_71_20] [Medline: 33163378]
24. Wogu IAP, Olu-Owolabi FE, Assibong PA, et al. Artificial intelligence, alienation and ontological problems of other minds: a critical investigation into the future of man and machines. Presented at: 2017 International Conference on Computing Networking and Informatics (ICCNI); Oct 29 to 31, 2017:; Lagos, Nigeria p. 1-10. [doi: 10.1109/ICCNI.2017.8123792]
25. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Frontiers in Computing and Intelligent Systems 2023 Jan 5;2(2):81-83. [doi: 10.54097/fcis.v2i2.4465]
26. PubMed Research. whatplugin.ai. URL: https://www.whatplugin.ai/plugins/pubmed-research [accessed 2024-04-30]
27. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? arXiv. Preprint posted online on Feb 9, 2023. [doi: 10.48550/arXiv.2302.03495]
28. Aydın Ö, Karaarslan E. OpenAI ChatGPT generated literature review: digital twin in healthcare. In: Aydın Ö, editor. Emerging Computer Technologies 2: İzmir Akademi Dernegi; 2022:22-31. [doi: 10.2139/ssrn.4308687]

29. Covidence. URL: https://www.covidence.org/ [accessed 2024-04-24]
30. Consensus. URL: https://consensus.app/ [accessed 2024-04-24]
31. Elicit. URL: https://elicit.com/ [accessed 2024-04-24]
32. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023 Apr 25;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
EBM: evidence-based medicine
GPT: generative pre-trained transformer
LLM: large language model
PD: Peyronie disease
RCT: randomized controlled trial
SR: systematic review

Edited by A Castonguay; submitted 24.07.23; peer-reviewed by IG Jeong Jeong, J Noh, L Zhu, S Pandey, TG Rhee; revised version received 31.03.24; accepted 04.04.24; published 14.05.24.

Please cite as:
Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, Shim SR
The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation
JMIR Med Inform 2024;12:e51187
URL: https://medinform.jmir.org/2024/1/e51187
doi:10.2196/51187

Implementation of the World Health Organization Minimum Dataset for Emergency Medical Teams to Create Disaster Profiles for the Indonesian SATUSEHAT Platform Using Fast Healthcare Interoperability Resources: Development and Validation Study

Hiro Putra Faisal[1,2], MD; Masaharu Nakayama[1], MD, PhD

1

2

Corresponding Author:
Masaharu Nakayama, MD, PhD

Abstract

Background: The National Disaster Management Agency (*Badan Nasional Penanggulangan Bencana*) handles disaster management in Indonesia as a health cluster by collecting, storing, and reporting information on the state of survivors and their health from various sources during disasters. Data were collected on paper and transferred to Microsoft Excel spreadsheets. These activities are challenging because there are no standards for data collection. The World Health Organization (WHO) introduced a standard for health data collection during disasters for emergency medical teams (EMTs) in the form of a minimum dataset (MDS). Meanwhile, the Ministry of Health of Indonesia launched the SATUSEHAT platform to integrate all electronic medical records in Indonesia based on Fast Healthcare Interoperability Resources (FHIR).

Objective: This study aims to implement the WHO EMT MDS to create a disaster profile for the SATUSEHAT platform using FHIR.

Methods: We extracted variables from 2 EMT MDS medical records—the WHO and Association of Southeast Asian Nations (ASEAN) versions—and the daily reporting form. We then performed a mapping process to match these variables with the FHIR resources and analyzed the gaps between the variables and base resources. Next, we conducted profiling to see if there were any changes in the selected resources and created extensions to fill the gap using the Forge application. Subsequently, the profile was implemented using an open-source FHIR server.

Results: The total numbers of variables extracted from the WHO EMT MDS, ASEAN EMT MDS, and daily reporting forms were 30, 32, and 46, with the percentage of variables matching FHIR resources being 100% (30/30), 97% (31/32), and 85% (39/46), respectively. From the 40 resources available in the FHIR ID core, we used 10, 14, and 9 for the WHO EMT MDS, ASEAN EMT MDS, and daily reporting form, respectively. Based on the gap analysis, we found 4 variables in the daily reporting form that were not covered by the resources. Thus, we created extensions to address this gap.

Conclusions: We successfully created a disaster profile that can be used as a disaster case for the SATUSEHAT platform. This profile may standardize health data collection during disasters.

(JMIR Med Inform 2024;12:e59651) doi:10.2196/59651

KEYWORDS

WHO EMT MDS; FHIR; SATUSEHAT; disaster; implementation; development; validation; emergency medical team; disaster management; Indonesia; Fast Healthcare Interoperability Resources; resources; interoperability; electronic medical records; EMR; reporting; disaster profile; health data; health data collection; World Health Organization; EMT; WHO; MDS; minimum dataset

Introduction

In Indonesia, disaster management is conducted by the National Disaster Management Agency (*Badan Nasional Penanggulangan Bencana*) [1]. Specifically, health issues during disasters are mandated to health clusters, whose members consist of regional health services, rapid health assessment teams, and emergency medical teams (EMTs) from various institutions. The health cluster collects, records, and reports information on survivors and their health conditions during a disaster [2]. According to the World Health Organization (WHO), collecting patient data requires the use of a nationally accepted reporting form or an approved dataset that is reported periodically. This report must include copies for the patient [3]. The current standard form available in Indonesia is the Rapid Health Assessment Form established by the Ministry of Health (MoH). This form summarized only the number of survivors and their general situation [2]. Meanwhile, the EMTs from various

institutions recorded the survivors' health status using their forms [4]. This information was collected by *Badan Nasional Penanggulangan Bencana* during daily reporting meetings and transferred to Microsoft Excel spreadsheets [5]. This activity is time-consuming and often inaccurate [6]. Additionally, the lack of coordination due to the decreased number of officers affected by the disaster may have impacted data collection and information exchange [7]. Consequently, this affects the handling of survivors at disaster locations and directs them to nearby health facilities.

Several institutions have tried to develop applications to record medical data [8], construct a documentation form [9-11], and create the minimum dataset (MDS) needed during disasters [11,12]. However, none of these are available for general use. To address this problem, the WHO introduced an MDS for EMTs in 2017 for use during disaster events [12]. This form has been tested and used to assess several disasters worldwide [13]. In Southeast Asian countries, the EMT MDS was introduced through the Project for Strengthening the Association of Southeast Asian Nations (ASEAN) Regional Capacity on Disaster Health Management (ARCH Project) [14]. The ARCH Project aimed to strengthen disaster management for ASEAN members through collaboration with the Japan International Cooperation Agency. One of the main goals is to use the WHO EMT MDS standards in the ASEAN region.

Indonesia's MoH launched the SATUSEHAT platform based on Fast Healthcare Interoperability Resources (FHIR) [15]. This action aims to integrate and perform interoperability among health care facilities. Thus, the government has targeted all health care facilities to have electronic medical records by the end of 2023 and be interoperable using this platform [16].

FHIR is a standard for exchanging health care information through health information systems developed by the health care standards organization Health Level Seven International [17]. FHIR uses a representational state-transfer (REST) application programming interface (API), a common web service architecture that aims to make it easier for health care systems to share and access data, eventually improving the interoperability of health care information. The main features of FHIR are modularity, standardized resources, and interoperability, thus making it easy to work with specific data elements such as patient demographics, health conditions, and medications with different health care systems.

With this momentum, we conducted this research to map the WHO EMT MDS form to FHIR so that it can be integrated into the SATUSEHAT platform. This will help survivors of disasters to record, report, and refer to systems.

Methods

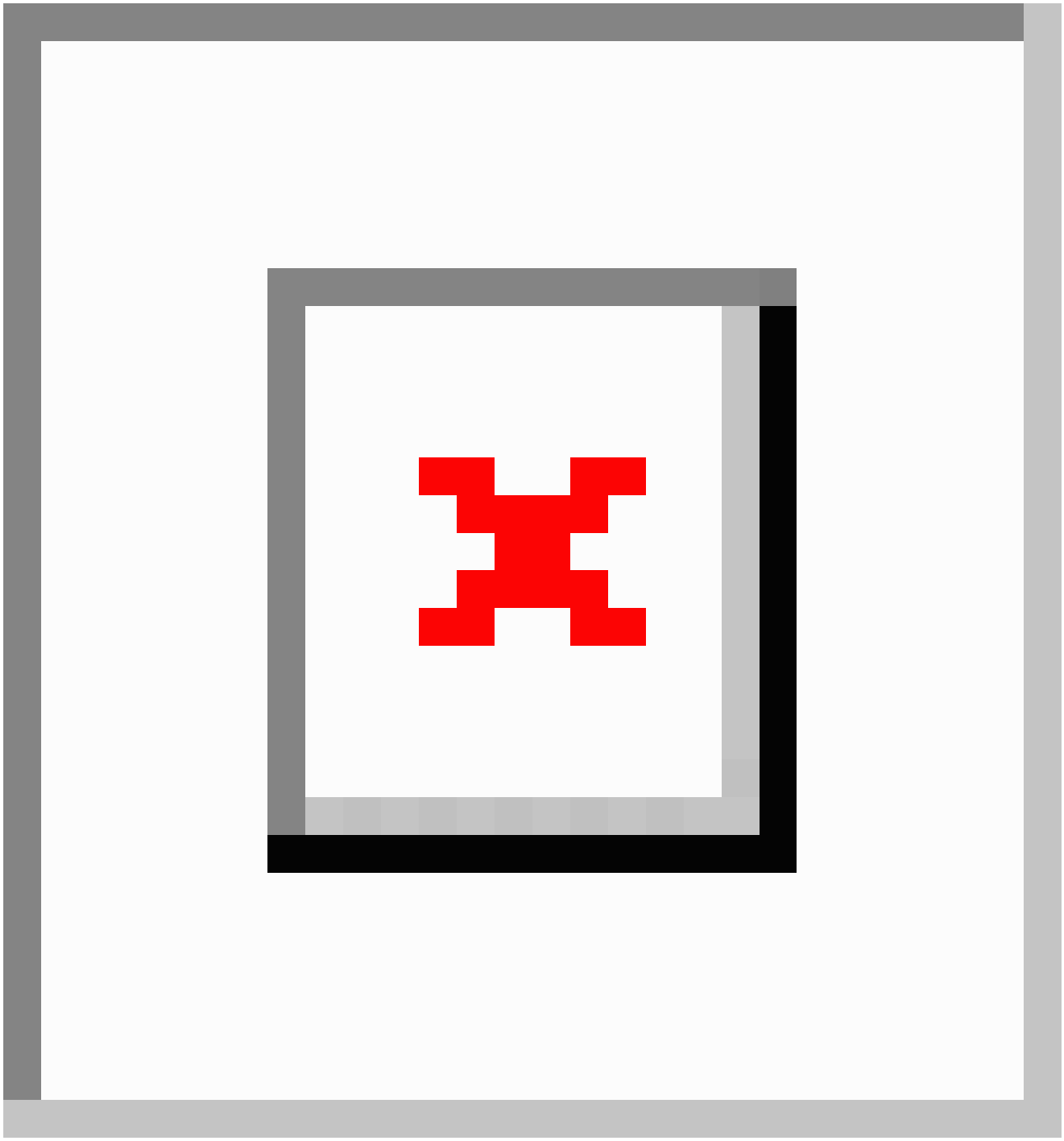
Extract Variables From the Medical Records

The EMT MDS consists of several forms that must be filled out during disasters. The set includes a medical records form combined with a tick-box section and the daily reporting form, both of which are sent to the EMT Coordination Cell (EMTCC). In the medical records section and tick-box form, EMT officers fill in patient data such as identity, medical history, vital signs, physical examination, and therapy management, which must be signed by a doctor. In this study, the tick box only helps to highlight the patient's condition, which is divided into age, sex, health events, procedures and outcomes, and context, without the patient's identity to maintain anonymity. Afterward, the form will be collected by the EMTCC to aggregate the data into the daily reporting form. The daily reporting form is part of the EMT MDS designed by the WHO to be used in the EMTCC office. The daily reporting form itself has already been tested during disasters in some countries [13,18,19]. It contains basic information regarding the team, the location to which the team was assigned, daily summaries of the facility, accumulation of data from the tick boxes as MDS statistics, and information regarding crucial needs and risks [20].

This study extracted data elements from the medical records form and daily reporting form. We obtained 2 variations of medical records from the WHO EMT MDS and the ASEAN EMT MDS, which are adaptations for ASEAN countries developed by the ARCH Project. These forms are illustrated in Figure 1.

We chose this medical record form for Indonesia since the Indonesian government, as a member of ASEAN states, has already been trained to use the EMT MDS during regional collaboration drills on the ARCH Project [21,22]. This project aims to strengthen disaster health management in ASEAN countries and meet the WHO EMT MDS as a standard operational procedure [23]. Thus, digitizing this set will be an appropriate choice.

Figure 1. Form of medical records taken from the (A) WHO EMT MDS, (B) ASEAN EMT MDS, and (C) daily reporting form. ASEAN: Association of Southeast Asian Nations; EMT: emergency medical team; MDS: minimum dataset; WHO: World Health Organization.



Mapping and Gap Analysis

Next, we manually mapped these data elements onto the base resources. Based on the results, we conducted a gap analysis between the data elements and FHIR resources. We divided the EMT MDS elements into the medical record and daily reporting form, including the tick-box section. Next, we examined whether each data element and the FHIR base resources matched.

Profiling and Validation

Profiling is the process of defining the profile created by setting the cardinality and creating an extension to the data elements that have definitions in the resource base. We used the Forge application since it is a common tool used in profiling research [24,25]. Next, we validated the profiles using the SIMPLIFIER.NET website [26].

Implementation

We implemented the profiles in the HAPI framework using an open-source FHIR server [27]. We used the Insomnia application, an open-source application that performs the API testing and development process, to perform the POST request to the server with body content using JSON [28].

Ethical Considerations

Since our research did not involve human subjects and no actual patient data were included in this study, ethics approval was not required.

Results

Mapping

We created a list and group of variables extracted from the WHO and ASEAN EMT MDS medical records. The data were divided into 2 categories: medical and daily reporting forms. Subsequently, the base FHIR was mapped. The list of variables and mapping process results from the WHO EMT MDS and ASEAN EMT MDS medical records forms are shown in Table 1. Meanwhile, Table 2 shows the variables in the daily reporting form mapped to the FHIR resources.

From the 40 resources available in the FHIR ID core, we selected 10, 14, and 9 resources for the WHO EMT MDS, ASEAN EMT MDS, and daily reporting form, respectively, as displayed in Table 3. The total numbers of variables from the WHO EMT MDS medical record form, the ASEAN EMT MDS medical record form, and the daily reporting form with variables that matched the FHIR resources are listed in Table 4.

Table . Mapping results for the medical forms.

Variable		WHO[a] EMT[b] MDS[c]	ASEAN[d] EMT MDS	FHIR[e] resources
Team name		N/A[f]	✓[g]	Organization
Site		N/A	✓	Location
Date		✓	✓	Encounter
ID		✓	✓	Patient
Name		✓	✓	Patient
Age		N/A	✗[h]	N/A
Nickname		✓	N/A	Patient
Present address		✓	✓	Patient
Triage category		N/A	✓	Observation
Hazards (if any)		N/A	✓	Observation
Breastfeeding		✓	✓	Observation
Arm circumference (<5 y)		✓	N/A	Observation
Vaccination		✓	✓	Immunization
Allergy		✓	✓	AllergyIntolerance
Past history		✓	✓	Condition
Medication		✓	✓	MedicationStatement
Chief complaints		✓	✓	Observation
Onset		N/A	✓	Condition
Trauma		N/A	✓	Observation
Vital signs				
	BT[i]	✓	✓	Observation
	PR[j]	✓	✓	Observation
	BP[k]	✓	✓	Observation
	RR[l]	✓	✓	Observation
	O_2 sat[m]	N/A	✓	Observation
	GCS[n]	N/A	✓	Observation
	Pain score	N/A	✓	Observation
Weight		✓	N/A	Observation
Height		✓	N/A	Observation
History of present illness		✓	N/A	Encounter and Condition
Physical examination		✓	✓	Observation
Diagnosis		✓	✓	Condition
Investigation		N/A	✓	Observation, Procedure, and ImagingStudy
Drug name or dose (management in ASEAN MDS)		✓	✓	Medication and Medication-Request
Procedure		N/A	✓	Procedure
Staff signature				
	Reception	✓	✓	Practitioner
	Doctor	✓	N/A	Practitioner
	MDS	✓	N/A	Practitioner

Variable		WHO[a] EMT[b] MDS[c]	ASEAN[d] EMT MDS	FHIR[e] resources
	Nurse	✓	N/A	Practitioner
	Drug	✓	N/A	Practitioner
	Examination	✓	N/A	Practitioner
	Data input	✓	✓	Practitioner
Memo		✓	✓	Observation

[a]WHO: World Health Organization.
[b]EMT: emergency medical team.
[c]MDS: minimum dataset.
[d]ASEAN: Association of Southeast Asian Nations.
[e]FHIR: Fast Healthcare Interoperability Resources.
[f]N/A: not applicable.
[g]✓: match.
[h]╳: not a match.
[i]BT: body temperature.
[j]PR: pulse rate.
[k]BP: blood pressure.
[l]RR: respiratory rate.
[m]O_2 sat: oxygen saturation.
[n]GCS: Glasgow Coma Scale.

Table . Mapping results in the daily reporting form with tick boxes.

Variables		Match or not a match	FHIR[a] resources
MDS[b] statistics			
Age		✗[c]	N/A[d]
Sex		✓[e]	Patient, Observation
Health events		✓	Condition
Procedure and Outcome			
	Procedure	✓	Procedure
	Outcome	✓	Condition, Observation, and ServiceRequest
Context			
	Relation	✗	N/A
	Protection	✓	Condition
Daily reporting form			
Team information			
	Organization name	✓	Organization
	Team name	✓	Organization
	Type 1 mobile	✓	Organization
	Type 1 fixed	✓	Organization
	Type 2	✓	Organization
	Type 3	✓	Organization
	Specialized cell	✓	Organization
	Contact person(s) name(s)	✓	Organization
	Phone number	✓	Organization
	Email	✓	Organization
	Estimated date departure	✗	N/A
	Date of activity	✓	Location
	Time of reporting	✓	Encounter
Location			
	State	✗	N/A
	City	✗	N/A
	Village	✗	N/A
	Facility name	✓	Location
	Geo-tag (latitude)	✓	Location
	Geo-tag (longitude)	✓	Location
Daily summary			
	Total number of new consultation	✓	Encounter
	New admission	✓	Encounter
	Live birth	✓	Patient and Encounter
	Total bed capacity	✓	Location
	Empty inpatient bed (non-ICU[f])	✓	Location
	Empty ICU	✓	Location
Needs and Risks			

Variables		Match or not a match	FHIR[a] resources
Immediate report			
	Unexpected death	✓	Communication
	Notifiable disease	✓	Communication
	Protection issues #	✓	Communication
	Critical incident to EMT[g] and/or community	✓	Communication
	Any other issue requiring immediate reporting	✓	Communication
Community risks			
	WASH[h]	✓	Communication
	Community or suspected over infectious disease	✓	Communication
	Environmental risk or exposure	✓	Communication
	Shelter or nonfood items	✓	Communication
	Food insecurity	✓	Communication
Operational constrains			
	Logistics or operational support	✓	Communication
	Supply	✓	Communication
	Human resources	✓	Communication
	Finance	✓	Communication
	Others	✓	Communication

[a]FHIR: Fast Healthcare Interoperability Resources.
[b]MDS: minimum dataset.
[c]✕: not a match.
[d]N/A: not applicable.
[e]✓: match.
[f]ICU: intensive care unit.
[g]EMT: emergency medical team.
[h]WASH: water, sanitation and hygiene.

Table . FHIR[a] usability.

FHIR resources	WHO[b] EMT[c] MDS[d]	ASEAN[e] EMT MDS	Daily reporting form
Organization	N/A[f]	✓[g]	✓
Location	N/A	✓	✓
Encounter	✓	✓	✓
Patient	✓	✓	✓
Observation	✓	✓	✓
Immunization	✓	✓	N/A
AllergyIntolerance	✓	✓	N/A
Condition	✓	✓	✓
MedicationStatement	✓	✓	N/A
MedicationRequest	✓	✓	N/A
Procedure	N/A	✓	✓
ServiceRequest	✓	✓	✓
Practitioner	✓	✓	N/A
ImagingStudy	N/A	✓	N/A
Communication	N/A	N/A	✓
Total, n	10	14	9

[a]FHIR: Fast Healthcare Interoperability Resources.
[b]WHO: World Health Organization.
[c]EMT: emergency medical team.
[d]MDS: minimum dataset.
[e]ASEAN: Association of Southeast Asian Nations.
[f]N/A: not applicable.
[g]✓: use.

Table . Calculation of EMT[a] MDS[b] variables mapped to FHIR[c] resources.

Form	Variables, n	Matching variables, n (%)
WHO[d] EMT MDS	30	30 (100)
ASEAN[e] EMT MDS	32	31 (97)
Daily reporting form	46	39 (85)

[a]EMT: emergency medical team.
[b]MDS: minimum dataset.
[c]FHIR: Fast Healthcare Interoperability Resources.
[d]WHO: World Health Organization.
[e]ASEAN: Association of Southeast Asian Nations.

Gap Analysis

From the gap analysis, we found several data elements that did not match the FHIR resources. Most data came from the daily reporting form. The list is presented in Table 5 . We created extensions for the age, relationship, and estimated date departure elements for this problem. Age did not contain a data element in the FHIR resources. Because age is related to patients, we created an extension of the patient resource to define it. The JSON file is shown in Figure 2.

Relation status is defined as whether the patient's condition is related to a disaster event. Thus, an extension of the condition resource is deemed appropriate to explain this condition. By definition, the estimated date of departure is when the team ends its service at specific locations. This allows the EMTCC to plan another EMT if necessary. Because this definition differs from the operational service hours in which the data element is available in the Location resource, we added an extension under the Organization resource to define the end period of the EMT service.

Sex in medical records uses a different value set from FHIR, divided into 3 groups: male, pregnant female, and nonpregnant female. Although sex is only divided into 2 groups, it adds pregnancy status to the female criteria. Since the pregnancy status already has a code in Logical Observation Identifiers Names and Codes (LOINC; "pregnancy status reported," code 11449 - 6), we bundled the Patient and Observation resources for this variable.

Table . Gap analysis results.

Data elements	Type	Form	Remarks	Possible FHIR[a] resources
Age	String	ASEAN[b] EMT[c] MDS[d] and daily reporting form	No exact definition in the data element of FHIR	Patient
State	Code	Daily reporting form	The data element available in ID core extension	Location
City	Code	Daily reporting form	The data element available in ID core extension	Location
Village	Code	Daily reporting form	The data element available in ID core extension	Location
Gender	Code	Daily reporting form	The definition implicate in 2 resources	Patient and Observation
Relation	Code	Daily reporting form	No exact definition in the data element of FHIR	Condition
Estimated date departure	Date	Daily reporting form	No exact definition in the data element of FHIR	Organization

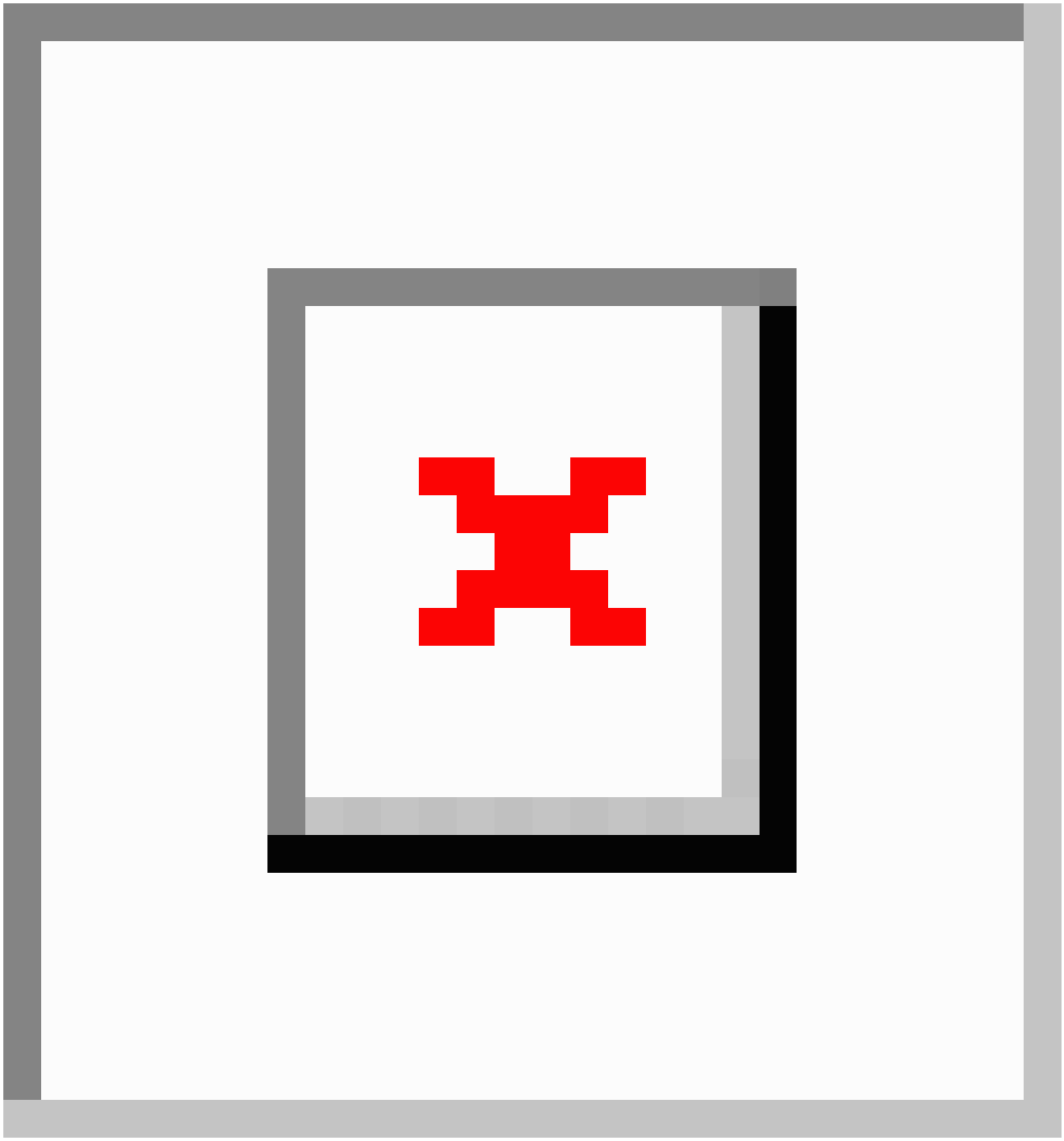
[a]FHIR: Fast Healthcare Interoperability Resources.

[b]ASEAN: Association of Southeast Asian Nations.

[c]EMT: emergency medical team.

[d]MDS: minimum dataset.

Figure 2. Extensions for (A) age and (B) relation data elements.



Validation and Implementation

We validated the resources for the FHIR ID core using SIMPLIFIER.NET. Example results are shown in Figure 3. The proposed disaster FHIR profile is implemented using the HAPI framework. We used the Insomnia application to perform requests on the FHIR REST server.

We created several mock datasets based on the SATUSEHAT public example [29]. We managed to have samples for each resource. However, we added extensions that are not defined in the SATUSEHAT platform. Afterward, we validated the disaster profile using the HAPI FHIR server. We created a JSON file for each resource to implement the POST protocol. The POST protocol validates the request body and stores the resources in a database. We successfully validated 15 resources used in this profile (an example message is shown in Figure 4).

Figure 3. Validation results on SIMPLIFIER.NET.

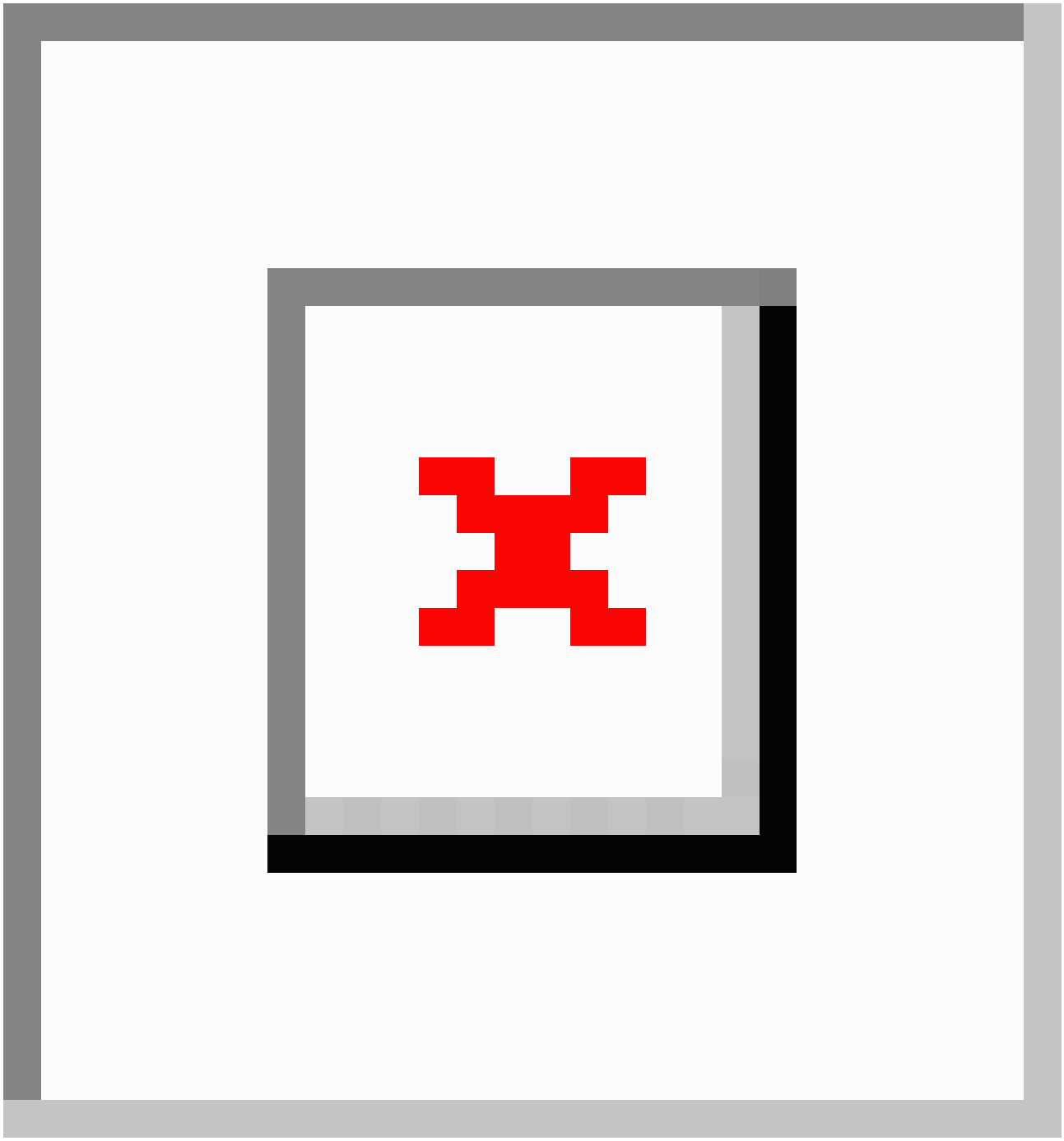
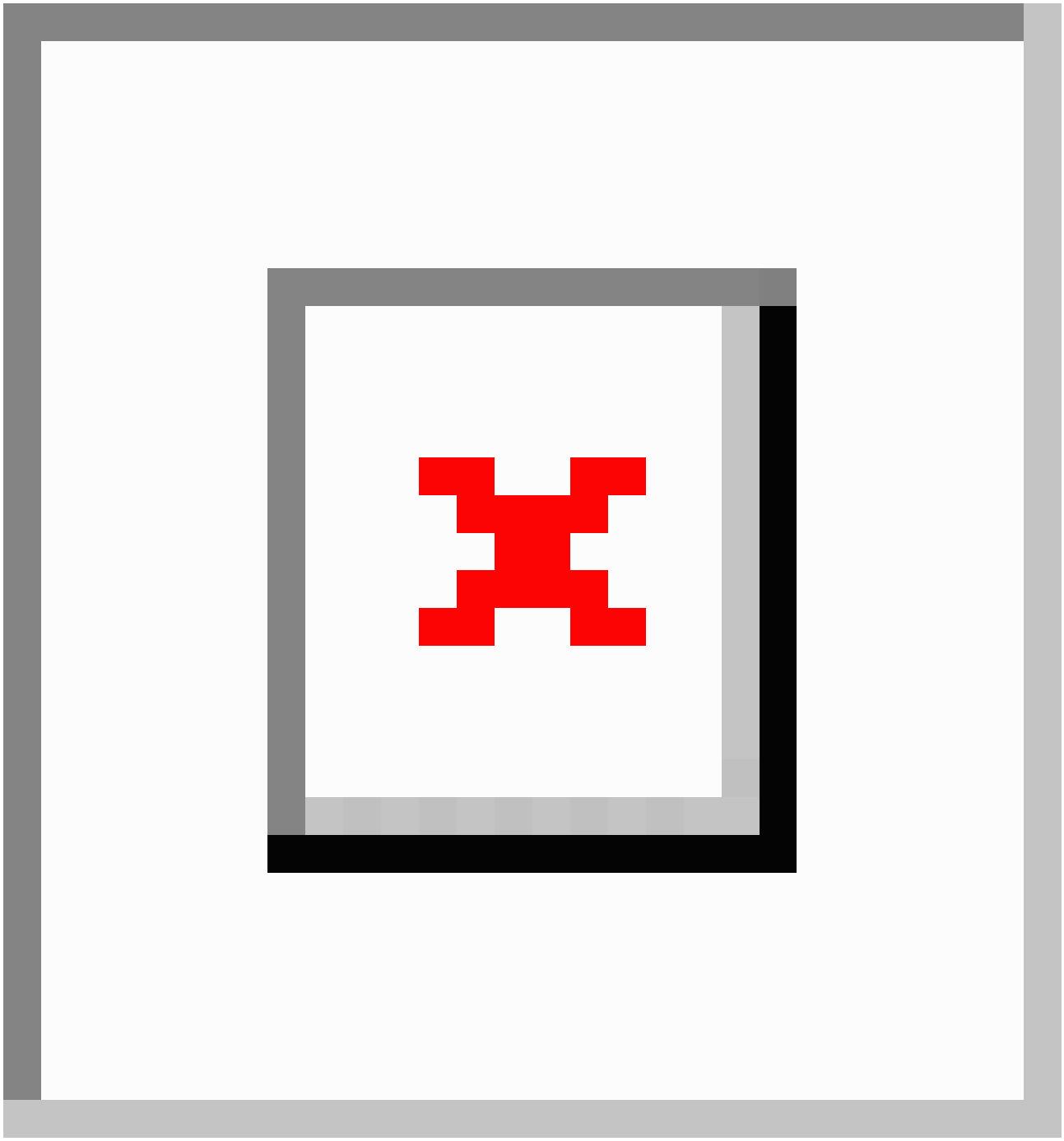


Figure 4. A JSON of the disaster profile was created based on HAPI. The profile mentioned (left) matches with the StructureDefinition resource (right).



Discussion

Principal Findings

Data collection plays a significant role during disasters. With major calamities requiring support from international aid workers, implementing a standard data format can greatly facilitate the collection of health-related information. We chose the WHO and ASEAN EMT MDS forms to map onto the FHIR profiles. Although the EMT MDS is a standard developed by the WHO for data collection during disasters, the technical implementation designs for disaster situations in which data are transferred manually using an Excel spreadsheet [18]. As daily reports from various sources may be used by the MoH or EMTCC to make decisions, shifting from paper-based documentation to a digital process may increase the efficiency and effectiveness of collecting daily reports from various sources. In this study, we successfully created and validated the profile to the FHIR ID core profile using base resources. A FHIR profile has generally been created for specific use cases; however, few studies have performed disaster-related profiling. Several studies have created profiles related to disasters, such as COVID-19 [30,31].

To the best of our knowledge, this is the first study that has developed FHIR profiles for natural disasters. Our profiles may standardize health data collection, resulting in real-time actionable insights during disaster scenarios. For example, the interoperability facilitated by FHIR allows instant data sharing between disparate health systems, ensuring that international

aid workers and medical facilities can access up-to-date patient records, treatment histories, and socioeconomic backgrounds. This seamless exchange of information is essential for coordinating effective and timely responses, significantly affecting survival rates and recovery outcomes.

Another advantage of using the FHIR format is that it optimizes the usability and quality of the health data. The EMT MDS form only records the number of disaster incidents. For example, if a patient visited twice in a day, 2 reports were generated [20]. Using FHIR, we can retrieve existing patient data from the database. The data entered during a disaster can be saved on the SATUSEHAT platform and reused when referring disaster survivors to nearby health care facilities [32]. By doing so, the quality and safety of patient care can be enhanced. These interoperability features have also enabled us to improve the efficiency of clinical data utilization during and after disasters. Ayaz et al [33] discussed the implementation of FHIR applications in their systematic literature review. Balch et al [34] defined the use of FHIR data standards in machine learning–enabled clinical information systems. Thus, using the data available on the SATUSEHAT platform would be helpful for researchers to perform studies related to disasters, such as identifying patterns of disease outbreaks and improving decision-making processes.

We mapped over 80% of the WHO and ASEAN EMT MDS medical record variables onto FHIR resources. Meanwhile, the variables that cannot be mapped to FHIR resource are mainly in the tick-box sections. Variables that did not have data elements, such as age and sex, were due to the characteristics of the EMT MDS form itself. The EMT MDS form was created to facilitate the collection of data and reports on the conditions occurring in the field [35]. This report form is an accumulation of the various daily report forms used by the EMTCC and MoH to provide care and treatment. Thus, age and sex were classified in a classification format to simplify grouping.

Limitations

This study faced a notable challenge in developing a comprehensive value set for EMT MDS tick-box variables. FHIR uses a standard value set to define medical terminology, such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) or LOINC codes. We could not define the value options for health events and procedure variables because the value options from these variables were a group of several conditions. For example, health events have a “major extremity injury” option, which by definition indicates any upper- and lower-extremity injury requiring hospitalization and/or spinal or general anesthesia [36]. However, SNOMED-CT did not have codes for this category; it did have codes to define each of the upper- and lower-extremity injuries (the detailed list is available in Multimedia Appendix 1). Some studies have reported similar findings when mapping the values to SNOMED-CT or LOINC codes [37,38]. A previous study also mentioned that SNOMED-CT requires disaster codification adjustments [39]. One solution that has been done is to submit new codes to the SNOMED-CT or LOINC committee. However, this process is often prolonged without a guarantee of acceptance of the proposed code [40]. An alternative, more immediate solution we considered was to create a new value set that aligns with the WHO EMT MDS standards [41]. This approach aims to bridge the current gap in codification and facilitate more accurate data representation and analysis for disaster health management.

Finally, because this study was based on the concept of the mapping process, we did not perform a functionality test using real data. We used a mock-up dataset to test the POST and GET functions on the FHIR server. In the future, we will conduct tests using data from the SATUSEHAT platform by opening up opportunities for collaboration with the Indonesian MoH. In addition, future work should focus on FHIR-based applications using disaster profiles for health data collection during disaster events.

Conclusions

This research aims to facilitate data collection using international standards, such as the WHO EMT MDS, and transfer it to the SATUSEHAT platform using FHIR as an interoperability standard. The proposed disaster profile was successfully implemented and validated using the FHIR server. This research will be beneficial for facilitating data collection using international standards during disasters and transferring it to a national platform for health care activity using FHIR as an interoperability standard.

Acknowledgments

The authors acknowledge the Indonesia Endowment Fund for Education (LPDP) under the Ministry of Finance, Republic of Indonesia, for its scholarship funding support. This study was also supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” grant JPJ012425; by Japan Society for the Promotion of Science (JSPS) KAKENHI grant 23K11890; and by Ministry of Health, Labour and Welfare of Japan (MHLW) Program grant JPMH20AC1007. We appreciate Professor Tatsuhiko Kubo’s permission to use the images for Figure 1.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table of minimum dataset mapping to Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT).
[XLSX File, 15 KB - medinform_v12i1e59651_app1.xlsx]

References

1. Presiden Republik Indonesia. Peraturan presiden republik Indonesia nomor 8 tahun 2008 [Article in Indonesian]. Badan Nasional Penanggulangan Bencana. 2008. URL: https://web.bnpb.go.id/ppid//theme/file/Perpres_08_2008.pdf [accessed 2023-08-08]
2. Peraturan menteri kesehatan republik indonesia nomor 75 tahun 2019 tentang penanggulangan krisis kesehatan [Article in Indonesian]. Ministry of Health of the Republic of Indonesia. 2019. URL: http://hukor.kemkes.go.id/uploads/produk_hukum/PMK_No__75_Th_2019_ttg_Penaggulangan_Krisis_Kesehatan.pdf [accessed 2023-08-08]
3. Classification and minimum standards for emergency medical teams. World Health Organization. 2021 Jun 18. URL: https://www.who.int/publications/i/item/9789240029330 [accessed 2023-08-08]
4. Pedoman nasional penanggulangan krisis kesehatan. National guidelines for health crisis management [Article in Indonesian]. Center for Health Crisis Ministry of Health of the Republic of Indonesia. 2023. URL: https://pusatkrisis.kemkes.go.id/download/flgog/files49279Final_Pedoman_Nasional_Penanggulangan_Krisis_Kesehatan.pdf [accessed 2023-08-08]
5. Petunjuk teknis pengumpulan data dan informasi bencana [Article in Indonesian]. Pusdatinmas BNPB. 2014. URL: https://perpustakaan.bnpb.go.id/bulian/index.php?p=show_detail&id=793 [accessed 2023-08-08]
6. Morton M, Levy JL. Challenges in disaster data collection during recent disasters. Prehosp Disaster Med 2011 Jun;26(3):196-201. [doi: 10.1017/S1049023X11006339] [Medline: 22107771]
7. Ayuningtyas D, Windiarti S, Hadi MS, Fasrini UU, Barinda S. Disaster preparedness and mitigation in Indonesia: a narrative review. Iran J Public Health 2021 Aug;50(8):1536-1546. [doi: 10.18502/ijph.v50i8.6799] [Medline: 34917524]
8. Mamlin BW, Shivers JE, Glober NK, Dick JJ. OpenMRS as an emergency EMR-how we used a global good to create an emergency EMR in a week. Int J Med Inform 2021 May;149:104433. [doi: 10.1016/j.ijmedinf.2021.104433] [Medline: 33752170]
9. Schnall AH, Wolkin AF, Noe R, et al. Evaluation of a standardized morbidity surveillance form for use during disasters caused by natural hazards. Prehosp Disaster Med 2011 Apr;26(2):90-98. [doi: 10.1017/S1049023X11000112] [Medline: 21888728]
10. Shinchi K, Ashida H. Proposal of a model for medical records for international disaster relief operations. Mil Med 2003 Feb;168(2):120-123. [doi: 10.1093/milmed/168.2.120] [Medline: 12636139]
11. Tavakoli N, Jahanbakhsh M, Fooladvand M. Developing health information documentation in disaster. Int J Health Syst Disaster Manage 2013;1(1):11-15. [doi: 10.4103/2347-9019.122426]
12. Benin-Goren O, Kubo T, Norton I. Emergency medical team working group for minimum data set. Prehosp Disaster Med 2017 Apr 20;32(S1):S96. [doi: 10.1017/S1049023X17002473]
13. Kubo T, Chimed-Ochir O, Cossa M, et al. First activation of the WHO emergency medical team minimum data set in the 2019 response to tropical cyclone Idai in Mozambique. Prehosp Disaster Med 2022 Dec;37(6):727-734. [doi: 10.1017/S1049023X22001406] [Medline: 36325992]
14. Yanasan A, Pongpamon N, Pattanarattanamole R, et al. ARCH project and the global initiatives of disaster health management. Prehosp Disaster Med 2022 Feb;37(S1):s11-s15. [doi: 10.1017/S1049023X22000048] [Medline: 35253637]
15. Office of Assistant to Deputy Cabinet Secretary for State Documents & Translation. Health ministry launches 'SatuSehat' platform. Cabinet Secretariat of the Republic of Indonesia. 2022 Jul 26. URL: https://setkab.go.id/en/health-ministry-launches-satusehat-platform/ [accessed 2023-08-08]
16. Triferna P. Electronic medical record implementation deadline at 2023-end. ANTARA Indonesian News Agency. 2022 Sep 9. URL: https://en.antaranews.com/news/248945/electronic-medical-record-implementation-deadline-at-2023-end [accessed 2023-08-08]
17. FHIR v4.0.1. HL7. URL: https://hl7.org/fhir/R4/index.html [accessed 2023-08-22]
18. Kubo T, Odgerel CO. Reviewing the implementation of the emergency medical team minimum data set. Prehosp Disaster Med 2023 May;38(S1):s35. [doi: 10.1017/S1049023X23001292]
19. Armitage R, Afonso AT. WHO emergency medical teams minimal data set in conflict-stricken Ukraine: comparative analysis of a new primary health care coding tool. Prehosp Disaster Med 2022 Nov 22;37(S2):s57. [doi: 10.1017/S1049023X2200156X]
20. WHO EMT MDS Working Group, Japan Disaster Relief EMT Initiative Corresponding Unit. Instruction for the EMT MDS daily report. EMT MDS Gateway. URL: https://www.dropbox.com/scl/fo/y15kzry7pbvd0mb37vxqo/AKe14pXHNbb2G8nBhBVx8K4?e=1&preview=06_MDS_Instruction220315A.pptx&rlkey=6xjh0plnc6afghp9jhf275hpm&st=fpmkl8lc&dl=0 [accessed 2023-12-01]
21. Wuthisuthimethawee P, Satthaphong S, Phongphuttha W, et al. How the ARCH project could contribute to strengthening ASEAN regional capacities on disaster health management (DHM). Prehosp Disaster Med 2022 Feb;37(S1):s30-s43. [doi: 10.1017/S1049023X22000061] [Medline: 35253635]
22. Silapunt P, Fernando F, Catampongan J, et al. How the ARCH project has contributed to the development of the ASEAN regional collaboration mechanism on disaster health management. Prehosp Disaster Med 2022 Feb;37(S1):s16-s29. [doi: 10.1017/S1049023X2200005X] [Medline: 35253638]
23. ASEAN Coordinating Centre for Humanitarian Assistance. SASOP - standard operating procedure for regional standby arrangements and coordination of joint disaster relief and emergency response operations. Association of Southeast Asian

Nations (ASEAN). 2017. URL: https://asean.org/wp-content/uploads/2018/02/59.-December-2017-SASOP-6th-Reprint.pdf [accessed 2024-08-14]
24. Nan J, Xu LQ. Designing interoperable health care services based on Fast Healthcare Interoperability Resources: literature review. JMIR Med Inform 2023 Aug 21;11:e44842. [doi: 10.2196/44842] [Medline: 37603388]
25. Xiao D, Song C, Nakamura N, Nakayama M. Development of an application concerning Fast Healthcare Interoperability Resources based on standardized structured medical information exchange version 2 data. Comput Methods Programs Biomed 2021 Sep;208:106232. [doi: 10.1016/j.cmpb.2021.106232] [Medline: 34174764]
26. Forge. SIMPLIFIER.NET. URL: https://simplifier.net/forge [accessed 2023-11-06]
27. HAPI FHIR - the open source FHIR API for Java. HAPI FHIR. URL: https://hapifhir.io/ [accessed 2023-11-06]
28. The collaborative API development platform - Insomnia. Insomnia. URL: https://insomnia.rest/ [accessed 2023-11-06]
29. SATUSEHAT public. Postman. URL: https://www.postman.com/satusehat/workspace/satusehat-public/overview [accessed 2023-09-04]
30. Lichtner G, Haese T, Brose S, et al. Interoperable, domain-specific extensions for the German Corona Consensus (GECCO) COVID-19 research data set using an interdisciplinary, consensus-based workflow: data set development study. JMIR Med Inform 2023 Jul 18;11:e45496. [doi: 10.2196/45496] [Medline: 37490312]
31. Ladas N, Franz S, Haarbrandt B, et al. OpenEHR-to-FHIR: converting openEHR compositions to Fast Healthcare Interoperability Resources (FHIR) for the German Corona Consensus dataset (GECCO). Stud Health Technol Inform 2022 Jan 14;289:485-486. [doi: 10.3233/SHTI210963] [Medline: 35062196]
32. Blueprint of digital health transformation strategy 2024. Ministry of Health of the Republic of Indonesia. 2021. URL: https://oss2.dto.kemkes.go.id/artikel-web-dto/ENG-Blueprint-for-Digital-Health-Transformation-Strategy-Indonesia%202024.pdf [accessed 2024-08-14]
33. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929. [doi: 10.2196/21929] [Medline: 34328424]
34. Balch JA, Ruppert MM, Loftus TJ, et al. Machine learning-enabled clinical information systems using Fast Healthcare Interoperability Resources data standards: scoping review. JMIR Med Inform 2023 Aug 24;11:e48297. [doi: 10.2196/48297] [Medline: 37646309]
35. Jafar AJN, Sergeant JC, Lecky F. What is the inter-rater agreement of injury classification using the WHO minimum data set for emergency medical teams? Emerg Med J 2020 Feb;37(2):58-64. [doi: 10.1136/emermed-2019-209012] [Medline: 31911417]
36. WHO EMT minimum data set. EMT MDS Gateway. URL: https://www.dropbox.com/scl/fi/s98dbxuzdhw19kntzg1tu/MDS-Ver1.0_-Supple1_Definition.pdf?rlkey=gmgxildgckfl3wha8bnz6st2d&e=1&st=frxwi90u&dl=0 [accessed 2023-10-04]
37. Odigie E, Lacson R, Raja A, et al. Fast Healthcare Interoperability Resources, Clinical Quality Language, and Systematized Nomenclature of Medicine-Clinical Terms in representing clinical evidence logic statements for the use of imaging procedures: descriptive study. JMIR Med Inform 2019 May 13;7(2):e13590. [doi: 10.2196/13590] [Medline: 31094359]
38. Tetzlaff L, Purohit AM, Spallek J, Holmberg C, Schrader T. Evaluating interoperability in German critical incident reporting systems. Stud Health Technol Inform 2023 Sep 12;307:249-257. [doi: 10.3233/SHTI230722] [Medline: 37697860]
39. Lokmic-Tomkins Z, Block LJ, Davies S, et al. Evaluating the representation of disaster hazards in SNOMED CT: gaps and opportunities. J Am Med Inform Assoc 2023 Oct 19;30(11):1762-1772. [doi: 10.1093/jamia/ocad153] [Medline: 37558235]
40. Matney SA, Heale B, Hasley S, et al. Lessons learned in creating interoperable Fast Healthcare Interoperability Resources profiles for large-scale public health programs. Appl Clin Inform 2019 Jan;10(1):87-95. [doi: 10.1055/s-0038-1677527] [Medline: 30727002]
41. Rinaldi E, Saas J, Thun S. Use of LOINC and SNOMED CT with FHIR for microbiology data. Stud Health Technol Inform 2021 May 24;278:156-162. [doi: 10.3233/SHTI210064] [Medline: 34042889]

Abbreviations

API: application programming interface
ARCH Project: Project for Strengthening the Association of Southeast Asian Nations Regional Capacity on Disaster Health Management
ASEAN: Association of Southeast Asian Nations
EMT: emergency medical team
EMTCC: Emergency Medical Team Coordination Cell
FHIR: Fast Healthcare Interoperability Resources
LOINC: Logical Observation Identifiers Names and Codes
MDS: minimum dataset
MoH: Ministry of Health
REST: representational state-transfer
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms

WHO: World Health Organization

Edited by C Lovis; submitted 21.04.24; peer-reviewed by J Cuthbertson, P Yusuf; revised version received 10.07.24; accepted 21.07.24; published 28.08.24.

Please cite as:
Faisal HP, Nakayama M
Implementation of the World Health Organization Minimum Dataset for Emergency Medical Teams to Create Disaster Profiles for the Indonesian SATUSEHAT Platform Using Fast Healthcare Interoperability Resources: Development and Validation Study
JMIR Med Inform 2024;12:e59651
URL: https://medinform.jmir.org/2024/1/e59651
doi:10.2196/59651

Enhancing the Functionalities of Personal Health Record Systems: Empirical Study Based on the HL7 Personal Health Record System Functional Model Release 1

Teng Cao, MPH; Zhi Chen, MPH; Masaharu Nakayama, MD, PhD

Department of Medical Informatics, Tohoku University Graduate School of Medicine, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi, Japan

Corresponding Author:
Masaharu Nakayama, MD, PhD

Abstract

Background: The increasing demand for personal health record (PHR) systems is driven by individuals' desire to actively manage their health care. However, the limited functionality of current PHR systems has affected users' willingness to adopt them, leading to lower-than-expected usage rates. The HL7 (Health Level Seven) PHR System Functional Model (PHR-S FM) was proposed to address this issue, outlining all possible functionalities in PHR systems. Although the PHR-S FM provides a comprehensive theoretical framework, its practical effectiveness and applicability have not been fully explored.

Objective: This study aimed to design and develop a tethered PHR prototype in accordance with the guidelines of the PHR-S FM. It sought to explore the feasibility of applying the PHR-S FM in PHR systems by comparing the prototype with the results of previous research.

Methods: The PHR-S FM profile was defined to meet broad clinical data management requirements based on previous research. We designed and developed a PHR prototype as a web application using the Fast Healthcare Interoperability Resources R4 (FHIR) and Logical Observation Identifiers Names and Codes (LOINC) coding system for interoperability and data consistency. We validated the prototype using the Synthea dataset, which provided realistic synthetic medical records. In addition, we compared the results produced by the prototype with those of previous studies to evaluate the feasibility and implementation of the PHR-S FM framework.

Results: The PHR prototype was developed based on the PHR-S FM profile. We verified its functionality by demonstrating its ability to synchronize data with the FHIR server, effectively managing and displaying various health data types. Validation using the Synthea dataset confirmed the prototype's accuracy, achieving 100% coverage across 1157 data items. A comparison with the findings of previous studies indicated the feasibility of implementing the PHR-S FM and highlighted areas for future research and improvements.

Conclusions: The results of this study offer valuable insights into the potential for practical application and broad adoption of the PHR-S FM in real-world health care settings.

(JMIR Med Inform 2024;12:e56735) doi:10.2196/56735

KEYWORDS

fast healthcare interoperability resources; logical observation identifiers names and codes; personal health record system functional model; personal health records

Introduction

Personal health records (PHRs) are beneficial tools in modern health care, as they allow patients to access clinical information and share it with medical staff in a secure and confidential environment [1]. An increasing number of countries, including the United States [2,3], Japan [4,5], European Union-member nations [6,7], South Korea [8,9], and Indonesia [10], have invested significantly in the development and promotion of PHR systems to enhance the health care experience by improving patients' ability to manage their health information. However, despite the acknowledged benefits of PHRs and substantial investments by various countries, the usage rates of PHR systems have fallen short of expectations [11-13]. In the United Kingdom, over 48.7% of individuals have never interacted with any web-based patient-information-management service [14]. Similarly, in the United States, data from the 2023 Health Information National Trends Survey revealed that more than 43% of people had not accessed their web-based medical records or patient portals even once in the past year [15].

The Personal Health Record System Functional Model Release 1 (PHR-S FM) [16] was developed as a standardized framework for managing personal health information and has been certified as ISO/HL7 (International Organization for Standardization/Health Level Seven International) 16527 [17]. This model outlines a range of possible functionalities in PHR

systems, aiming to provide a standardized framework for designing, developing, and evaluating them. The PHR-S FM includes three main sections: (1) personal health (PH), which manages individual health data, encompassing a wide range of functionalities, from medical history to ongoing health conditions; (2) supportive (S), which facilitates the administrative and financial aspects of health care, thereby enabling smoother patient-provider interactions and backend processes; and (3) information infrastructure (IN), which ensures information privacy, security, interoperability, and ease of use. These 3 main sections contain several subsections, each with various functions. The sections and subsections delineate broad functional domains, whereas the functions offer detailed specifications of the features required in PHRs, adhering to a defined parent-child relationship. Each function is characterized by a function ID and Name and is described by a Statement, with its numbering indicating the parent-child relationship between sections and subsections. For example, a function ID "PH.3.1" would be the parent of "PH.3.1.1."

The PHR-S FM allows researchers to select appropriate functionalities to create a functional profile, which defines a subset of functionalities, thereby facilitating the implementation of PHR. Although the PHR-S FM provides a comprehensive theoretical framework, its practical effectiveness and applicability have not been fully explored. Researchers must precisely specify the actionable functions when creating profiles. However, due to the potential limitations of individual cases, variations may exist in the chosen function list, even if the research objectives are the same.

Harahap et al [18] conducted an extensive systematic review to identify the fundamental functionalities and challenges of the current PHRs. They thoroughly analyzed the essential functionalities required for effective and user-centric PHR systems and created a comprehensive PHR-S FM functional profile containing a subset of functions. This profile encompasses health and administrative records, medication management, communication, appointment management, education, and self-health monitoring. It also considers the challenges faced in PHR implementation, such as interoperability, security and privacy, usability, and data quality. Their functional profile provides a holistic framework for designing and developing PHR systems, ensuring that the systems meet health care users' evolving needs, thereby enhancing the effective deployment and user adaptation of PHR systems.

Building on previous research, this study aimed to design and develop a tethered PHR system prototype to determine the feasibility of providing servers that meet the PHR-S FM function profile. Moreover, this study examined the efficacy of the PHR-S FM by comparing its contents to that of previous studies.

Methods

Personal Health Record System Functional Model

We conducted a comprehensive analysis of Harahap's PHR-S FM profile and dissected each function into its subfunctions to create actionable levels. For instance, we expanded PH.2.5 into PH.2.5.1–PH.2.5.11, PH.3.1 into PH.3.1.1 and PH.3.1.2, and IN.2 into IN.2.1–IN.2.3.

Based on this classification, we created a new PHR-S FM profile to meet the broad requirements of clinical data management and ensure practical feasibility. We prioritized and implemented functions within the PH module that manage widely used core clinical data [7,19], such as medication lists (PH.2.5.6) and test results (PH.2.5.3). We omitted certain functions, including surgical history (PH.2.5.7), family health history (PH.2.5.8), genetic information (PH.2.5.9), social history (PH.2.5.10), nutrition and diet information (PH.2.5.11), communication with home monitoring devices (PH.3.1.2), medication management (PH.3.4), health education (PH.4), and appointment scheduling (PH.6.3). Although these functions are theoretically important, their operational complexity and the substantial resources required for their analysis exceeded the practical scope of this study.

Furthermore, we integrated essential auxiliary functions from the IN and S modules to support the effective operation of the PH module. These included displaying health records, ensuring system interoperability, and managing user access control. However, we excluded functions such as present ad hoc views of the health record (IN.1.3), standards version control (IN.2.2), application integration (IN.2.3), interoperability protocols (IN.2.4), secure messaging (IN.3.10), and insurance management (S.2.1), as they provided limited support and could unnecessarily complicate the prototype.

Finally, we established the PHR-S FM profile for this study (Table 1). In addition, we extracted detailed descriptions of each function from the ISO/HL7 16527 [17] standard (see Multimedia Appendix 1 for more details).

Table . The Personal Health Record System Functional Model Release 1 profile function list.

Function list sections	ID #
PH[a]	PH.1.1, PH.1.2, PH.2.5.1, PH.2.5.2, PH2.5.3, PH2.5.4, PH2.5.5, PH2.5.6, PH3.1.1
S[b]	S.1.3, S.1.5
IN[c]	IN.2.1, IN.3.3, IN.4

[a]PH: Personal Health

[b]S: Supportive

[c]IN: Information Infrastructure

PHR Design and Development

Based on the PHR-S FM profile from this study, we designed the functionality of a PHR prototype. It encompasses 8 distinct functions in the PH section: user demographics (PH.1.2), diagnosis information (PH.2.5.1), medications (PH.2.5.2), imaging test reports and laboratory test reports (PH.2.5.3), allergy information (PH.2.5.4), immunization (PH.2.5.5), visiting records (PH.2.5.6), and vital signs and Patient-Generated Health Data (PGHD; PH.3.1.1) [20]. IN.2.1 explicitly emphasizes the importance of interoperability standards in supporting information sharing between PHR and other systems. We adopted the Fast Healthcare Interoperability Resources Release 4 (FHIR R4) standard [21] to facilitate effective data exchange. FHIR has been widely recognized to effectively overcome data-sharing difficulties between medical information systems and has already become the preferred standard for achieving interoperability [22-24]. Table 2 lists the PHR data categories and their corresponding FHIR R4 resources. This illustrates the 8 types of medical data incorporated into the PHR prototype, specific details captured for each data type, and corresponding FHIR R4 resources. This mapping is essential to ensure data consistency and facilitate interoperability. Moreover, to ensure data consistency and achieve semantic interoperability, we standardized the user-entered PGHD data using the Logical Observation Identifiers Names and Codes (LOINC) system [25] and Unified Code of Units of Measurement (UCUM) system [26] and stored it in the FHIR Observation resources (see Multimedia Appendix 2 for details).

This study used Firely R4 [27] as the FHIR server and successfully developed a PHR prototype based on the PHR-S FM. The PHR prototype was designed as a web application optimized for mobile phones. When users log in, the PHR prototype actively retrieves their medical records to ensure that they always have the latest data. Figure 1 shows the main interface with a clear layout that allows users to access essential functions quickly and intuitively. The prototype is divided into 4 main sections: User Demographics, PGHD, Encounter History, and Comprehensive Records (see Multimedia Appendix 3 for details on the implementation of the PHR prototype, functional demonstration, and verification of consistency with the PHR-S FM functional profile).

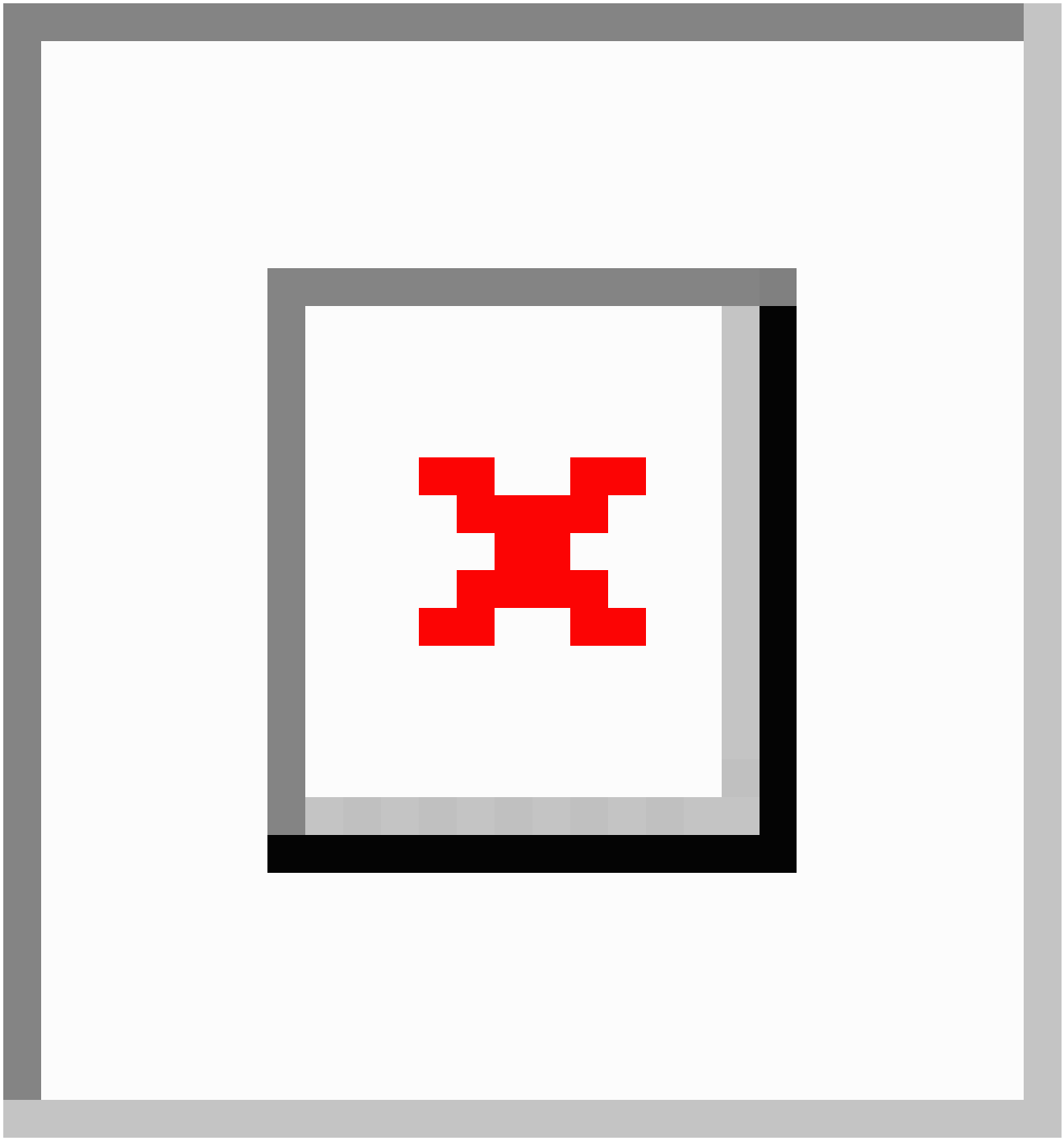
Table . Mapping table from PHR[a] prototype functions to FHIR[b] resources.

PHR-S FM ID#	PHR functions	Details	FHIR resources
PH.1.1, PH.1.2	User Demographics	Name, Gender, Birthday, Age, Phone Number, Address, Email.	Patient
PH.2.5.1	Diagnosis Information	Disease, Date, Doctor.	Condition, Practitioner
PH.2.5.2	Medications	Item Name, Dose/Unit, Way, Frequency, Start-time, End-time, Amount, Group, Property, Doctor.	Medication Request, Practitioner
PH.2.5.3	Test Results	Imaging Test reports: Item Name, Operate Time, PDF Report, Imaging Test Report, Critical Value.conclusion, Conclusion, Performer.	Diagnostic Report, Observation, ImagingStudy, Practitioner
PH.2.5.3	Test Results	Laboratory Test reports: Item Name, Operate Time, PDF Report, Item Detail, Value, Normal Range, Critical Value/ Detailed Test Names, Critical Value / value, Critical Value/ ReferenceRange, Critical Value / abnormal Flag, Performer.	DiagnosticReport, Observation, Practitioner
PH.2.5.4	Allergy Information	Allergen, Note Date, Severity, Reaction, Comment, Performer.	AllergyIntolerance, Practitioner
PH.2.5.5	Immunization	Vaccine Name, Datetime.	Immunization
PH.2.5.6	Visiting Records	Admission Time, Discharge Time, Visit Department, Hospital, Visit Type, Diagnosis, Doctor.	Encounter, Condition, Practitioner, Organization
PH.3.1.1	Observations and Care	Vital Signs: Datetime, Height, Weight, Temperature, Pulse Rate, Heart Rate, BMI Respiratory Rate, Blood Pressure, Pain Severity, Pediatric Head Occipital-Frontal Circumference Percentile, Body Mass Index (BMI) for Age, Pediatric Weight for Height, Head Occipital-Frontal Circumference, Doctor.	Observation, Practitioner, Organization
PH.3.1.1	Observations and Care	Patient-Generated Health Data (PGHD): Height, Weight, Temperature, Steps, BMI, Blood Pressure, Heart Rate, Respiration Rate, Smoking Habits.	Observation, Organization

[a]PHR: personal health record.

[b]FHIR: Fast Healthcare Interoperability Resources R4.

Figure 1. The main interface of the personal health record (PHR) prototype.



Validation

We used the Synthea dataset [28] to validate the effectiveness of the PHR prototype. Synthea was designed to generate realistic and synthetic medical records and has been extensively applied in multiple studies [29-32]. It extracts data from public datasets and statistical models, ensuring their authenticity and representativeness without compromising individual privacy. The dataset comprised comprehensive health records for over 1 million fictional patients in standardized formats, such as HL7 FHIR, C-CDA, and CSV. These records encompassed a wide range of medical information, including medication usage, allergies, medical history, and social health determinants, making them extremely valuable for developing, testing, and demonstrating PHR systems.

We randomly selected 5 patients from the Synthea dataset to validate the effectiveness of the PHR prototype. Their medical data were formatted in FHIR R4 and uploaded to the Firely server for testing. We analyzed FHIR resources and counted the number of FHIR resource types for each patient. The results were then compared with the types and quantities of data displayed in the PHR prototype. The evaluation method involved a quantitative analysis of the number of data entries available in the Synthea dataset against the number displayed in the PHR system, aiming to ascertain the prototype's effectiveness in data representation.

Ethical Considerations

As the Synthea dataset comprised solely synthetic data unrelated to any real individuals, considerations of legal and privacy issues were not required. Moreover, these data could be used for research without requiring patient consent or institutional review board approval.

Comparison With Prior Research

In our study, we conducted a comprehensive search of databases such as PubMed, IEEE Xplore, and ACM Digital Library using the keywords "PHR," "PHR-S FM," and "Personal Health Record System Functional Model." The search yielded 4 studies [33-36], but only 3 of these detailed the PHR-S FM functional profiles used as well as the implementation of PHR functionalities. Specifically, Katehakis et al [33] mentioned a PHR system for chronic care and home health, using the PHR-S FM and experiences from EU projects to ensure interoperability and meet health management needs. The project underscored the importance of adhering to standards like the HL7 PHR-S FM for better adoption of PHR systems. Saripalle et al [34] used HL7 FHIR to design an interoperable mobile PHR following PHR-S FM guidelines, achieving integration with OpenEMR and underscoring the benefits of a modular approach and standards-based API-driven data exchange. Chatterjee et al [35] developed a tethered PHR using FHIR resources and SNOMED-CT (Systematized Nomenclature of Medicine–Clinical Terms) terminology according to PHR-S FM guidelines, focusing on collecting PGHD from diverse sources for electronic health record integration.

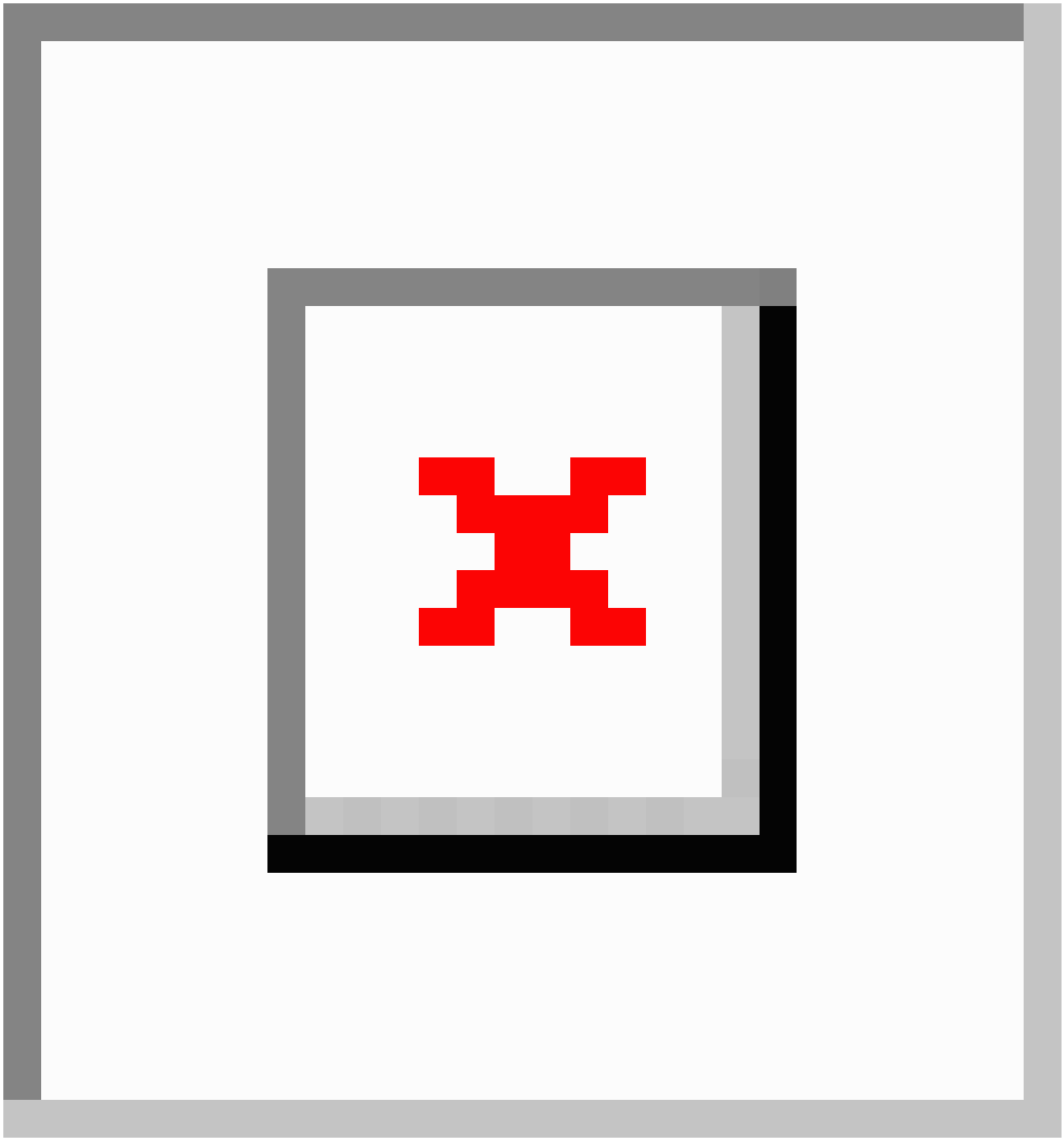
We analyzed and quantified the PHR-S FM functions in these 3 studies [33-35]. Using the PHR-S FM functional profile proposed by Harahap et al [18] as a benchmark, we conducted a comparative analysis of the results of these studies and of the present study to quantify the implementation of each functionality outlined in the profile. We conducted a survey on how many studies had implemented each function. The results were then subjected to a detailed discussion to ascertain the strengths and weaknesses of each PHR-S FM function. This approach aimed to evaluate the feasibility of the PHR-S FM framework and identify potential directions for future research.

Results

PHR Prototype Functionality

In this section, we verify and demonstrate the functionality of the PHR prototype (see Multimedia Appendix 4 for details). Figure 2 shows the sequence of events for communicating the FHIR resources between the PHR and FHIR servers. It outlines all the interactions between the PHR prototype and the Firely server based on FHIR resources. The PHR prototype enables users to create new PGHD, which undergo semantic conversion to ensure consistent terminology and units using LOINC and UCUM standards, then synchronizing them to the FHIR server as FHIR Observation resources. Concurrently, the PHR prototype demonstrates interoperability with the Firely server by using the RESTful API, efficiently retrieves various medical data, including allergies, immunization, vital signs, medications, diagnostic information, and laboratory and imaging test reports, while also uploading newly generated PGHD data back to the Firely server.

Figure 2. Sequence diagram of the synchronization of data between PHR and FHIR server.



Validation

We conducted a quantitative analysis of the PHR prototype using data from 5 patients extracted from the Synthea dataset (See Multimedia Appendix 5 for an example of verifying that the PHR correctly displayed data from Synthea). We identified 1157 data items across these patient profiles, with individual counts of 183, 154, 198, 327, and 295 (see Multimedia Appendix 6). The prototype accurately displayed all data entries contained within the patient profiles, achieving a coverage rate of 100%.

Entries were marked as not applicable if corresponding data points were absent in the patient profiles. The PHR prototype was designed to display 21 distinct FHIR resources and effectively presented data entries covering 20 out of 21 health data categories (95%), including user demographics, diagnosis information, medications, and laboratory test reports.

Specifically, for imaging test reports, no entries were recorded for any of the 5 patients. The PHR prototype could process and display such data, mirroring its performance with allergy information. Despite the absence of data in 4 out of the 5 patient profiles in the allergy category, the prototype successfully displayed all 7 entries for the 1 profile that included allergy data. This demonstrated the prototype's capability to fully render allergy information available in the dataset. This suggested that imaging test reports would be similarly displayed with complete accuracy if these data were available.

Comparison With the Findings of Previous Studies

This study used the PHR-S FM functional profile, as proposed by Harahap et al [18], as a foundational benchmark for conducting a comprehensive comparative analysis of 3 studies that used it. Table 3 presents a comparative analysis of our study with 3 other studies based on the PHR-S FM framework. In Table 3, "✓" was used to identify implemented functions, and "-" was used for unspecified functions. The results indicated that 5 functions were adopted and implemented in 4 studies, all of which originated from the PH module. Another 5 functions were implemented by 3 studies (3/4), 4 of which belonged to the PH module, and 1 was an interoperability function of the IN module. Furthermore, 2 functions were implemented in half of the studies (2/4), including the PH.2.5.10 (Manage Social History) function of the IN module. Moreover, 4 functions were implemented in 1 study (1/4). Of these, the "Manage Surgical History" function attracted attention, as it was the only function of the PH module implemented by only 1 study. In total, 13 functions were not implemented in any study.

Table . Comparative analysis of specific function completions in the PHR-S FM profile.

Function list sections ID #	Function name	Saripalle et al [34]	Chatterjee et al [35]	Katehakis et al [32]	This study	Function score
Personal Health (PH)						
PH.1.1	Identify and Maintain a PHR Account Holder Record	—	✓	✓	✓	3
PH.1.2	Manage PHR Account Holder Demographics	✓	✓	✓	✓	4
PH.2.5.1	Manage Problem Lists	✓	✓	✓	✓	4
PH.2.5.2	Manage Medication List	✓	—	✓	✓	3
PH.2.5.3	Manage Test Results	✓	✓	✓	✓	4
PH.2.5.4	Manage Allergy, Intolerance, and Adverse Reaction List	✓	✓	✓	✓	4
PH.2.5.5	Manage Immunization List	✓	—	✓	✓	3
PH.2.5.6	Manage Medical History	✓	✓	—	✓	3
PH.2.5.7	Manage Surgical History	—	—	✓	—	1
PH.2.5.8	Maintain Family History	—	—	—	—	0
PH.2.5.9	Manage Personal Genetic Information	—	—	—	—	0
PH.2.5.10	Manage Social History	✓	✓	—	—	2
PH.2.5.11	Nutrition and Diet Information	—	—	—	—	0
PH.3.1.1	Manage Personal Observations and Care	✓	✓	✓	✓	4
PH.3.1.2	Communication with Home Monitoring Devices	—	—	—	—	0
PH.3.4	Manage Medications	—	—	—	—	0
PH.4	Manage Health Education	—	—	—	—	0
PH.6.3	Communications between Provider and/or the PHR Account Holder's Representative	—	—	—	—	0
Supportive (S)						

Function list sections ID #	Function name	Saripalle et al [34]	Chatterjee et al [35]	Katehakis et al [32]	This study	Function score
S.1.3	Manage Health-Care Provider Information	—	—	—	✓	1
S.1.5	Manage Health-care Facility Information	—	—	—	✓	1
S.2.1	Capture and Read Health Insurance Account and Benefit Information	—	—	—	—	0
Information Infrastructure (IN)						
IN.1.3	Present Ad Hoc Views of the Health Record	—	—	—	—	0
IN.2.1	Interoperability Standards	✓	✓	-	✓	3
IN.2.2	Interoperability Standards Versioning and Maintenance	—	—	—	—	0
IN.2.3	Standards-Based Application Integration	—	—	—	—	0
IN.2.4	Interoperability Agreements	—	—	—	—	0
IN.3.3	Entity Access Control	—	✓	—	✓	2
IN.3.10	Secure Messaging	—	—	—	—	0
IN.4	Auditable Records	—	—	—	✓	1

Discussion

Principal Results

This study developed a PHR prototype guided by the PHR-S FM profile, which successfully managed medical data, including immunizations, allergies, vital signs, medications, diagnoses, and test results. The prototype used LOINC and FHIR R4 for semantic consistency and standardization of user-input data, ensuring data consistency and interoperability. The prototype was validated using the Synthea dataset and demonstrated 100% coverage in accurately displaying patient information. A comprehensive analysis of previous studies revealed the current implementation status of the PHR-S FM framework, which predominantly focuses on the PH module with limited emphasis on the IN and S modules. To the best of our knowledge, this was the first study to examine the feasibility of using PHR-S FM to develop and design PHRs.

Comparison With the Findings of Previous Studies

PHR-S FM Feasibility Analysis

The detailed analysis of the functional implementation revealed significant trends and differences (Table 3). The diversity in the extent of functionality implementation indicated varying levels of focus on specific features in the PHR-S FM framework across different studies.

Widely Implemented Functions

As shown in Table 3, 10 features, including PH.1.2 (Manage PHR Account Holder Demographics), PH.2.5.1 (Manage Problem Lists), and PH.2.5.3 (Manage Test Results), were implemented by at least 3 studies, of which 9 were attributed to PH modules. We found that almost all these functions aligned closely with the PHR core set of necessary criteria outlined in the ONC's Meaningful Use criteria [7]. The widespread adoption of these features may be related to their centrality in patient data management and essential compliance with healthcare standards [7,19].

IN.2.1 (Interoperability Standards), the only function not corresponding to the PH module, is a major theme in current

PHR research; however, with the advent of FHIR, interoperability can be achieved [37]. FHIR released the first version with normative content (FHIR R4) in 2019. Katehakis et al [33] denoted a missed opportunity for interoperability. This indicated that researchers should focus on keeping up to date with the latest technologies related to PHRs to expand and enhance their functionality. In addition, the PHR-S FM community should undertake more initiatives to promote FHIR adoption and improve the interoperability and use of PHR systems. In comparison to the integrated PHR by Katehakis et al [33], both Saripalle et al and Chatterjee et al [35] developed tethered PHRs using FHIR for interoperability. While Saripalle et al [34] integrated multiple health information, it lacked qualitative and quantitative evaluations. Chatterjee et al [35] focused on recording PGHDs, but whether the prototype could effectively retrieve and display medical data from institutions was unclear.

Limited Implementation Functions

Six features were implemented in at least one study. One function, PH.2.5.10 (Manage Social History), was partially implemented in this study, allowing smoking status to be recorded; however, it was not fully implemented due to the variety of data types involved and the complexity of data collection [38].

The PH.2.5.7 (Manage Surgical History) function was uniquely mentioned in Katehakis et al's study [33], which effectively applied the PHR-S FM framework in the context of real research projects in the European Union. This allowed researchers to bypass the complexities of implementing this detailed function. In contrast, the other 3 studies were at the prototype stage. However, other functions mentioned by Katehakis et al [33] were adopted by at least 2 other studies, highlighting the feasibility of the PHR-S FM framework across various research settings.

Functions Not Mentioned

Many functions were not mentioned because the implementation of certain features was not within the scope or the priority of a particular study. Moreover, collecting data for the PHR system was challenging, and some collected data were not useful. For instance, the intricacy of PH.2.5.9 (Manage Personal Genetic Information) and PH.2.5.11 (Nutrition and Diet Information) were infrequently used in PHRs [39]. The limited participation of the IN and S module functions may be attributed to the complexity of their implementation, such as IN.3.10 (Secure Messaging), which ensures the security of communication [40]. Although this function is significant in the PHR system, it faces technical challenges [11]. The S.2.1 (Capture and Read Health Insurance Account and Benefit Information) function was not implemented, likely due to the need for data from multiple sources and a wide variety of types, which rendered it costly and challenging to implement [41]. PHR conformed to the PHR-S FM functional profile, allowing researchers to effectively use it by tailoring the functional profile to suit their research needs. This flexibility emphasized the adaptability of PHR-S FM and demonstrated its use as a dynamic health care research field, despite challenges in achieving certain features.

Furthermore, this study uncovered a significant trend, with the S and IN modules used less frequently than the PH modules. The S module offers users administrative and financial support functionalities, and its implementation can enhance the effectiveness of health care services. Implementing the IN module ensures privacy, security, and interoperability, facilitating access and usability of PHR functions. Both modules are essential components of a comprehensive PHR system. Therefore, researchers should prioritize the S and IN modules when considering PH module functionalities, as they form the backbone that supports the optimal operation of PHRs.

Comparison With Studies Based on Other PHR Frameworks

To delve deeper into the development of PHR systems, this study conducted a comparative analysis with research based on the PHR-S FM and examined other significant studies. Lee et al [42], Li [43], and Song et al [44] presented further perspectives on PHR systems' development. Lee et al [42] developed a PHR prototype that organizes and visualizes personal health information according to a patient-centered journey map. However, their study neither achieved data interoperability nor enabled users to record PGHD data, making the application irrelevant for people without diseases. Li [43] proposed a service-oriented approach to integrating the electronic health record and PHR systems. However, the development and implementation of a service-oriented approach-based environment can be highly complex and expensive. Moreover, compared with the FHIR, they used the HL7 CDA to deliver messages, thereby increasing maintenance difficulty. Song et al [44] developed a patient summary application based on the international patient summary and FHIR R4 standards to mitigate information overload and reduce physicians' and nurses' workloads. Unlike traditional PHRs, which provide a comprehensive display of numerous information items, this application is limited in scope, displaying only the essential international patient summary standard information on a single screen. This design choice aims to simplify the user interface for clinicians but may restrict the availability of more detailed patient data. Thus, this study implemented more functions of the PHR-S FM and created an advanced PHR with more comprehensive functions that can better meet users' functional needs than those of previous studies.

Limitations

Although this study exhaustively explored the feasibility of PHR-S FM functionality, its scope was limited to 29 functions. As the PHR-S FM is a comprehensive framework designed to encompass the full range of functions that may be integrated into a PHR system, this study did not address its full potential. Future research could consider a broader set of features to more thoroughly evaluate the value of the PHR-S FM for practical applications.

In addition, the application developed in this study was a prototype and must be verified for use and evaluated for functionality in the future. Furthermore, the proposed prototype does not currently address privacy and security concerns. The platform SMART on FHIR [45] offers secure access control through the OAuth2 standard, ensuring that only authorized

users and applications can access a patient's PHR. Future research should delve deeper into the integration of this technology in PHRs. Finally, the data types included in the PHR in this study were limited and lacked information on appointments and surgical history. However, our modularly designed PHR allows for the addition of such content in the future without altering the foundational structure.

Conclusions

This study developed a PHR prototype based on the PHR-S FM profile and compared its results with the findings of previous studies to evaluate the framework's feasibility in the application of PHR. The findings demonstrated the us of the PHR-S FM as a valuable tool for PHR development and highlighted its potential to inform future technological enhancements in the PHR domain. The empirical data on function implementation offer a foundation for subsequent PHR system design. Moreover, this study provides actionable recommendations for applying the PHR-S FM in health care settings. The adaptability of the PHR-S FM framework is well suited to evolving health information management demands, suggesting avenues for continued research and optimization.

Acknowledgments

This study was supported by JST, the establishment of university fellowships toward the creation of science technology innovation (grant JPMJFS2102), by the Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" (grant JPJ012425), by JSPS KAKENHI (grant 23K11890), and by the MHLW Program (grant JPMH20AC1007).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The PHR-S FM (Personal Health Record System Functional Model) functional profile details.
[PDF File, 51 KB - medinform_v12i1e56735_app1.pdf]

Multimedia Appendix 2
Mapping table between the PGHD (Patient-Generated Health Data) and the LOINC (Logical Observation Identifiers Names and Codes) code systems.
[PDF File, 31 KB - medinform_v12i1e56735_app2.pdf]

Multimedia Appendix 3
The development of the personal health record application.
[PDF File, 161 KB - medinform_v12i1e56735_app3.pdf]

Multimedia Appendix 4
The details of the personal health record application.
[PDF File, 1891 KB - medinform_v12i1e56735_app4.pdf]

Multimedia Appendix 5
Verification of personal health record display for Synthea immunization data.
[PDF File, 373 KB - medinform_v12i1e56735_app5.pdf]

Multimedia Appendix 6
Comprehensive data display coverage of the personal health record prototype Across 5 patient profiles.
[PDF File, 101 KB - medinform_v12i1e56735_app6.pdf]

References

1. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126. [doi: 10.1197/jamia.M2025] [Medline: 16357345]
2. Woods SS, Schwartz E, Tuepker A, et al. Patient experiences with full electronic access to health records and clinical notes through the My HealtheVet Personal Health Record Pilot: qualitative study. J Med Internet Res 2013 Mar 27;15(3):e65. [doi: 10.2196/jmir.2356] [Medline: 23535584]
3. Ford EW, Hesse BW, Huerta TR. Personal health record use in the United States: forecasting future adoption levels. J Med Internet Res 2016 Mar 30;18(3):e73. [doi: 10.2196/jmir.4973] [Medline: 27030105]

4. Fukuda H, Murata F, Azuma S, et al. Development of a data platform for monitoring personal health records in Japan: the sustaining health by integrating next-generation ecosystems (SHINE) study. PLoS One 2023;18(2):e0281512. [doi: 10.1371/journal.pone.0281512] [Medline: 36787325]
5. Xiao D, Song C, Nakamura N, Nakayama M. Development of an application concerning fast healthcare interoperability resources based on standardized structured medical information exchange version 2 data. Comput Methods Programs Biomed 2021 Sep;208:106232. [doi: 10.1016/j.cmpb.2021.106232] [Medline: 34174764]
6. Determann D, Lambooij MS, Gyrd-Hansen D, et al. Personal health records in the Netherlands: potential user preferences quantified by a discrete choice experiment. J Am Med Inform Assoc 2017 May 1;24(3):529-536. [doi: 10.1093/jamia/ocw158] [Medline: 28011592]
7. Genitsaridi I, Kondylakis H, Koumakis L, Marias K, Tsiknakis M. Evaluation of personal health record systems through the lenses of EC research projects. Comput Biol Med 2015 Apr;59:175-185. [doi: 10.1016/j.compbiomed.2013.11.004] [Medline: 24315661]
8. Park HS, Kim KI, Chung HY, et al. A worker-centered personal health record app for workplace health promotion using national health care data sets: design and development study. JMIR Med Inform 2021 Aug 4;9(8):e29184. [doi: 10.2196/29184] [Medline: 34346894]
9. Jung SY, Kim JW, Hwang H, et al. Development of comprehensive personal health records integrating patient-generated health data directly from Samsung s-health and Apple health apps: retrospective cross-sectional observational study. JMIR Mhealth Uhealth 2019 May 28;7(5):e12691. [doi: 10.2196/12691] [Medline: 31140446]
10. Harahap NC, Handayani PW, Hidayanto AN. Integrated personal health record in Indonesia: design science research study. JMIR Med Inform 2023 Mar 14;11:e44784. [doi: 10.2196/44784] [Medline: 36917168]
11. Gagnon MP, Payne-Gagnon J, Breton E, et al. Adoption of electronic personal health records in Canada: perceptions of stakeholders. Int J Health Policy Manag 2016 Jul 1;5(7):425-433. [doi: 10.15171/ijhpm.2016.36] [Medline: 27694670]
12. Damen DJ, Schoonman GG, Maat B, Habibović M, Krahmer E, Pauws S. Patients managing their medical data in personal electronic health records: scoping review. J Med Internet Res 2022 Dec 27;24(12):e37783. [doi: 10.2196/37783] [Medline: 36574275]
13. Kahn JS, Aulakh V, Bosworth A. What it takes: characteristics of the ideal personal health record. Health Aff (Millwood) 2009;28(2):369-376. [doi: 10.1377/hlthaff.28.2.369] [Medline: 19275992]
14. Digital N. Patient Online Management Information (POMI). URL: https://app.powerbi.com/view?r=eyJfjoiMzM4YTkNWYtMzlNS00ZmExLWFkZDQtMmMzNDBkNzkxYTFmIiwidCI6IjM3YzM1NGIyLTg1YjAtNDdmNS1iMjIyLTA3YjQ4ZDc3NGVlMyJ9 [accessed 2024-03-14]
15. Strawley C, Richwine C. Individuals' access and use of patient portals and smartphone health apps.: ONC Data Brief; 2022. URL: https://www.healthit.gov/data/data-briefs/individuals-access-and-use-patient-portals-and-smartphone-health-apps-2022 [accessed 2024-03-14]
16. Health Level Seven International. Personal Health Record System Functional Model, Release 1 (PHR-S FM). 2016. URL: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=88 [accessed 2024-02-06]
17. ISO. HL7 personal health record system functional model, release 1 (PHRS FM). 2023. URL: https://www.iso.org/standard/57046.html [accessed 2024-10-04]
18. Harahap NC, Handayani PW, Hidayanto AN. Functionalities and issues in the implementation of personal health records: systematic review. J Med Internet Res 2021 Jul 21;23(7):e26236. [doi: 10.2196/26236] [Medline: 34287210]
19. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016 Mar 18;37(1):61-81. [doi: 10.1146/annurev-publhealth-032315-021353]
20. Hsueh PYS, Dey S, Das S, Wetter T. Making sense of patient-generated health data for interpretable patient-centered care: the transition from "more" to "better.". Stud Health Technol Inform 2017;245:113-117. [Medline: 29295063]
21. HL7 International. Fast Healthcare Interoperability Resources Release 4. 2019. URL: http://hl7.org/fhir/R4/ [accessed 2024-02-06]
22. Vorisek CN, Lehne M, Klopfenstein SAI, et al. Fast healthcare interoperability resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724. [doi: 10.2196/35724] [Medline: 35852842]
23. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - a review of the scientific literature. Stud Health Technol Inform 2019 Sep 3;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]
24. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The fast health interoperability resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021;9(7):e21929. [doi: 10.2196/21929]
25. HL7 International. HL7 FHIR release 4 observation resource. 2019. URL: https://hl7.org/fhir/R4/observation-vitalsigns.html [accessed 2024-02-06]
26. Institute R. The Unified Code of Units of Measurement (UCUM). URL: https://ucum.org/ [accessed 2024-04-10]
27. Firely Server. URL: https://server.fire.ly [accessed 2024-02-06]

28. Walonoski J, Kramer M, Nichols J, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2018 Mar 1;25(3):230-238. [doi: 10.1093/jamia/ocx079] [Medline: 29025144]
29. Phelan D, Gottlieb D, Mandel JC, et al. Beyond compliance with the 21st century cures act rule: a patient controlled electronic health information export application programming interface. J Am Med Inform Assoc 2024 Apr 3;31(4):901-909. [doi: 10.1093/jamia/ocae013] [Medline: 38287642]
30. Chen A, Chen DO. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. Sci Rep 2022 Oct 26;12(1):17917. [doi: 10.1038/s41598-022-23011-4] [Medline: 36289292]
31. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: pilot mixed-methods study. JMIR Form Res 2020 Jun 5;4(6):e16670. [doi: 10.2196/16670] [Medline: 32442148]
32. Katalinic M, Schenk M, Franke S, et al. Generation of a realistic synthetic laryngeal cancer cohort for AI applications. Cancers (Basel) 2024 Feb 1;16(3):639. [doi: 10.3390/cancers16030639] [Medline: 38339389]
33. Katehakis DG, Kondylakis H, Koumakis L, Kouroubali A, Marias K. Integrated care solutions for the citizen: personal health record functional models to support interoperability. Eur J Biomed Inform 2017;13(1):51-58. [doi: 10.24105/ejbi.2017.13.1.8]
34. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019 Jun;94:103188. [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
35. Chatterjee A, Pahari N, Prinz A. HL7 FHIR with SNOMED-CT to achieve semantic and structural interoperability in personal health data: a proof-of-concept study. Sensors (Basel) 2022 May 15;22(10):3756. [doi: 10.3390/s22103756] [Medline: 35632165]
36. Genitsaridi I, Kondylakis H, Koumakis L, Marias K, Tsiknakis M. Towards intelligent personal health record systems: review, criteria and extensions. Procedia Comput Sci 2013;21:327-334. [doi: 10.1016/j.procs.2013.09.043]
37. Tsafnat G, Dunscombe R, Gabriel D, Grieve G, Reich C. Converge or collide? Making sense of a plethora of open data standards in health care. J Med Internet Res 2024 Apr 9;26:e55779. [doi: 10.2196/55779] [Medline: 38593431]
38. Arsoniadis EG, Tambyraja R, Khairat S, et al. Characterizing patient-generated clinical data and associated implications for electronic health records. Stud Health Technol Inform 2015;216:158-162. [Medline: 26262030]
39. Lemke AA, Thompson J, Hulick PJ, et al. Primary care physician experiences utilizing a family health history tool with electronic health record-integrated clinical decision support: an implementation process assessment. J Community Genet 2020 Jul;11(3):339-350. [doi: 10.1007/s12687-020-00454-8] [Medline: 32020508]
40. Raghu TS, Frey K, Chang YH, Cheng MR, Freimund S, Patel A. Using secure messaging to update medications list in ambulatory care setting. Int J Med Inform 2015 Oct;84(10):754-762. [doi: 10.1016/j.ijmedinf.2015.06.003] [Medline: 26113460]
41. Barbarito F, Pinciroli F, Barone A, et al. Implementing the lifelong personal health record in a regionalised health information system: the case of Lombardy, Italy. Comput Biol Med 2015 Apr;59:164-174. [doi: 10.1016/j.compbiomed.2013.10.021] [Medline: 24238969]
42. Lee B, Lee J, Cho Y, et al. Visualisation of information using patient journey maps for a mobile health application. Appl Sci (Basel) 2023;13(10):6067. [doi: 10.3390/app13106067]
43. Li JQ. A service-oriented approach to interoperable and secure personal health record systems. Presented at: Proceedings of the 2017 11th IEEE Symposium on Service-Oriented System Engineering (SOSE); Apr 6-9, 2017; San Francisco, USA.
44. Song C, Nakayama M. Implementation of a patient summary web application according to the international patient summary and validation in common use cases in Japan. J Med Syst 2023 Sep 23;47(1):100. [doi: 10.1007/s10916-023-01993-6] [Medline: 37740823]
45. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908. [doi: 10.1093/jamia/ocv189] [Medline: 26911829]

Abbreviations

FHIR: Fast Healthcare Interoperability Resources
FM: Personal Health Record System Functional Model
HL7: Health Level Seven
HTML5: HyperText Markup Language 5
IN: information infrastructure
ISO/HL7: International Organization for Standardization/Health Level Seven International
LOINC: Logical Observation Identifiers Names and Codes
PGHD: Patient-Generated Health Data
PH: personal health
PHR: personal health record
S: supportive

SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms
UCUM: Unified Code of Units of Measurement

Edited by C Lovis; submitted 09.02.24; peer-reviewed by E Kimura, F Amar; revised version received 04.08.24; accepted 17.08.24; published 09.10.24.

Please cite as:
Cao T, Chen Z, Nakayama M
Enhancing the Functionalities of Personal Health Record Systems: Empirical Study Based on the HL7 Personal Health Record System Functional Model Release 1
JMIR Med Inform 2024;12:e56735
URL: https://medinform.jmir.org/2024/1/e56735
doi:10.2196/56735

Bridging Data Models in Health Care With a Novel Intermediate Query Format for Feasibility Queries: Mixed Methods Study

Lorenz Rosenau[1,*], MSc; Julian Gruendner[2,*], PhD; Alexander Kiel[3], BSc; Thomas Köhler[4,5,6]; Bastian Schaffer[2]; Raphael W Majeed[7], PhD

1

2

3

4

5

6

7

*these authors contributed equally

Corresponding Author:
Lorenz Rosenau, MSc

Abstract

Background: To advance research with clinical data, it is essential to make access to the available data as fast and easy as possible for researchers, which is especially challenging for data from different source systems within and across institutions. Over the years, many research repositories and data standards have been created. One of these is the Fast Healthcare Interoperability Resources (FHIR) standard, used by the German Medical Informatics Initiative (MII) to harmonize and standardize data across university hospitals in Germany. One of the first steps to make these data available is to allow researchers to create feasibility queries to determine the data availability for a specific research question. Given the heterogeneity of different query languages to access different data across and even within standards such as FHIR (eg, CQL and FHIR Search), creating an intermediate query syntax for feasibility queries reduces the complexity of query translation and improves interoperability across different research repositories and query languages.

Objective: This study describes the creation and implementation of an intermediate query syntax for feasibility queries and how it integrates into the federated German health research portal (Forschungsdatenportal Gesundheit) and the MII.

Methods: We analyzed the requirements for feasibility queries and the feasibility tools that are currently available in research repositories. Based on this analysis, we developed an intermediate query syntax that can be easily translated into different research repository–specific query languages.

Results: The resulting Clinical Cohort Definition Language (CCDL) for feasibility queries combines inclusion criteria in a conjunctive normal form and exclusion criteria in a disjunctive normal form, allowing for additional filters like time or numerical restrictions. The inclusion and exclusion results are combined via an expression to specify feasibility queries. We defined a JSON schema for the CCDL, generated an ontology, and demonstrated the use and translatability of the CCDL across multiple studies and real-world use cases.

Conclusions: We developed and evaluated a structured query syntax for feasibility queries and demonstrated its use in a real-world example as part of a research platform across 39 German university hospitals.

(JMIR Med Inform 2024;12:e58541) doi:10.2196/58541

KEYWORDS

feasibility; FHIR; CQL; eligibility criteria; clinical research; intermediate query format; healthcare interoperability; cohort definition; query; queries; interoperability; interoperable; informatics; portal; portals; implementation; develop; development; ontology; ontologies; JSON

Introduction

Background

In the rapidly evolving field of medical research, patient data have emerged as a critical resource. The vast amounts of data generated through clinical encounters, laboratory tests, imaging studies, and other patient interactions hold the potential to significantly advance our understanding of disease processes and treatment outcomes. Clinical Data Repositories (CDRs) are a valuable tool for storing, organizing, and retrieving this wealth of patient data. These repositories facilitate data storage in a structured and standardized manner, enabling researchers to query these data efficiently for various research purposes.

One key aspect of effectively using CDRs is the ability to perform feasibility queries. These queries allow researchers to assess the availability and adequacy of data for specific research questions before embarking on full-scale studies. Doing so can save considerable time and resources by identifying potential issues, such as insufficient sample size or a lack of necessary data elements.

Distributed Data Collections

The landscape of data repositories is not homogeneous. There are 2 primary approaches to data repository management: the classical single repository approach and the federated approach. Traditionally, these repositories have been centralized, pooling data from various sources into a single repository [1]. However, this classical approach has been challenged by the emergence of federated data repositories [1,2].

The classic single repository approach involves a centralized system where all data are stored and managed in one place. This solution offers the advantage of uniformity and ease of data management. It enables efficient data quality benchmarking at scale and the generation of derivatives, harmonized variables, and units of measure for comparable and consistent analytics [1]. However, it is often impractical or impossible to implement, especially when dealing with multiple institutions, each having its own schema for its clinical data repository.

On the other hand, the federated approach involves a network of repositories, each maintained by different institutions. These repositories operate independently but are interconnected for data sharing and collaboration. The data generally remain at the generating site, which offers the advantages of local curation by personnel deeply familiar with the data [1] and maintains data anonymity and security [2]. The data can then be analyzed using a federated approach or, if the correct patient consent is given, be transferred to a central data management unit for a specific analysis.

This approach respects individual institutions' autonomy and data governance policies, making it a more feasible option for multi-institutional collaborations [3-9] and can enhance the scope and depth of clinical research by enabling access to a broader range of data.

Despite the potential benefits of federated data repositories, performing feasibility queries across multiple CDRs presents significant challenges [10]. Each repository contains data originating from different source systems, leading to heterogeneity in data formats, terminologies, and quality. This heterogeneity can significantly complicate the process of data integration and harmonization, making it challenging to perform comprehensive and accurate feasibility queries [10].

Moreover, the federated nature of the system introduces additional complexities. Data privacy regulations and institutional policies may restrict the sharing and use of certain data, further complicating the query process. Additionally, the technical infrastructure required to support secure and efficient data exchange across multiple repositories can be challenging to implement and maintain.

Data Exchange Standards for Interoperability

In a federated network, the commitment to an interoperability standard becomes pivotal to tackling these challenges. Prominent examples include but are not limited to Fast Healthcare Interoperability Resources (FHIR) [11], OMOP CDM [12], i2b2 [13], and OpenEHR [14] share the commonality of being centered around the patients' medical history.

Agreeing on an interoperability standard only partially solves the challenge. While a health care data exchange standard facilitates the conversion of existing data into a common format at each hospital, a distributed feasibility query platform for the data is still missing.

Tools for Feasibility Queries

Besides the data integration standardization, interactive user interfaces enable researchers to design and submit feasibility queries. For this purpose, a multitude of tools for feasibility queries exist (eg, i2b2 [13,15], TriNetX [16], tranSMART [17], SampleLocator [18-20], Observational Health Data Science and Informatics [OHDSI] ATLAS [21], DZHK Feasibility Explorer [22]), each with its own data formats, standards, and query languages, including Structured Query Language (SQL), Clinical Quality Language (CQL), FHIR-Search, and Archetype Query Language (AQL). Consequently, querying across these different tools is difficult as there is no common query representation, and researchers must navigate these diverse tools and formats, particularly when dealing with cross-institutional data or distributed data storage within an institution.

Within the broader context of establishing a feasibility platform as part of the central German Portal for Health Data (FDPG), this research introduces a novel query syntax, serving as an intermediary between user interfaces and data repositories. This syntax is designed to be sufficiently flexible to ensure interoperability while maintaining simplicity. It focuses on the primary needs of a feasibility query, while allowing the syntax to be translated into repository-specific languages like FHIR-Search or CQL.

Our approach is grounded in the broader context of clinical research, where the reuse of eligibility criteria is common, whether in their original form or with modifications. These criteria are instrumental not just for feasibility studies but also for prescreening, data selection, extraction, and validation. Consequently, a need has emerged to decouple the representation of eligibility criteria from their implementation in specific

systems. A mechanism to express complex criteria and combinations thereof in a way that is both intuitive and adaptable to varying implementation needs is required.

In this study, we describe the development and application of the query syntax within the network of the Medical Informatics Initiative (MII), encompassing 39 German university hospitals, specifically, the FDPG feasibility platform and show how it achieves interoperability across different research platforms.

Methods

Requirement Analysis

In our pursuit to create an intermediate query syntax to express eligibility criteria, we performed a requirement analysis. Within it, we combine insight from feasibility queries and cohort selection, with the latter often manifesting as a query output in the form of cohort size rather than a list of discrete patient identifiers.

Our research reviewed existing tooling, namely i2b2, TriNetX, and OHDSI Atlas. We aimed to identify common functionalities and essential features across these tools. To obtain insight into the criteria's structure and complexity, we analyzed existing eligibility criteria from ClinicalTrials.gov [23] and incorporated the findings from Ross et al [24] and Gulden et al [25]. Moreover, we integrated insights from the usability study by Schüttler et al evaluating feasibility tools [26], conducted expert interviews and recursively synchronized the requirements within our project. This multifaceted analysis allowed us to infer a set of requirements crucial for developing our query syntax. These requirements were categorized into query expressiveness, interoperability, and accessibility.

Expressiveness Requirements

The query syntax should:

1. allow for the definition of inclusion and exclusion criteria
2. be expressed in Boolean logic.
3. allow the expression of exclusion criteria.
4. support at least patient as query subject (feasibility queries can be performed on different query subjects: find all patients with specific criteria, find all encounters with specific criteria, find all specimens with specific criteria).
5. use unique identifiers for criteria and concepts.
6. support the following filter on the criterion level:
    - existence of a criterion
    - numeric restriction
    - concept filter
    - time restrictions
    - attribute filters

Interoperability and Accessibility Requirements

The query syntax should:

1. provide an abstract (decoupling) layer between the user interface and the query execution.
2. have a low level of complexity and be easily translatable to different query languages.
3. be suitable for integration with the Health Level Seven International (HL7) FHIR standard used by the MII.
4. use a widely used data exchange format like JSON to ease parsing and generation
5. human readability or writability
6. ideally directly support the use of standard medical terminology (LOINC [Logical Observation Identifiers Names and Codes], SNOMED-CT [Systematized Nomenclature of Medicine–Clinical Terms], *ICD-10* [*International Statistical Classification of Diseases, Tenth Revision*], etc) to lower mapping efforts

Related Work and Existing Solutions

Analyzing the existing solutions, we found that none of the solutions met all the requirements. Most failed to have a formally defined low-complexity feasibility query syntax, and i2b2 was missing the direct relationship with the terminology on the syntax level. FHIR Search and the FHIR standard did not provide the ability to express a feasibility query in the required scope [25] at the time of our research. Other query languages that could have been candidates, like CQL or SQL, are complex or data model specific, making the translation between different data models and their representation, as well as the generation of the syntax by a user interface, challenging.

Evaluation

To evaluate the specification of the query syntax, we compared the final specification with our requirements and additionally demonstrated its applicability beyond the scope of FHIR by applying it to AQL.

We incorporated the solution into a large-scale real-world distributed feasibility query infrastructure, including a user interface, where it was integrated as the central intermediate query syntax. We further evaluated the applicability of the syntax to a wide range of clinical criteria and investigated its translatability, as well as how well it lends itself to creating a user interface for feasibility queries. Beyond the use in our projects based on German data sets and specifications, we also successfully applied the Clinical Cohort Definition Language (CCDL) to the international Synthea [27] data set.

Ethical Considerations

No ethics board decision is required as we are presenting a technical solution without working on patients' data.

Results

Based on the requirements of a team of experts, we created the "Clinical Cohort Definition Language," an intermediate query syntax for feasibility queries. The exchange format for the syntax was chosen to be JSON, which is currently widely used across the software community and is familiar to software developers from user interfaces, REST application programming interfaces (APIs), and query execution backends alike.

Criterion Types and Filters

The atomic component of CCDL is the criterion, serving as the foundational building block for inclusion or exclusion criteria. Each criterion is uniquely identified using a tuple of code system and code (which we named termCode) analogous to FHIR and OMOP-CDM (For conceptual equivalence between concepts

across medical terminologies, multiple termCodes can be provided, eg, the criterion for sleep apnea may be represented by the termCodes G47.3 from *ICD-10* and 73430006 from SNOMED-CT). Each termCode may have an additional "display" attribute, which serves purely as a visual representation to make the interpretation of a CCDL easier for humans. Within our CCDL, the criteria can occur as 1 of 4 different base types of criteria:

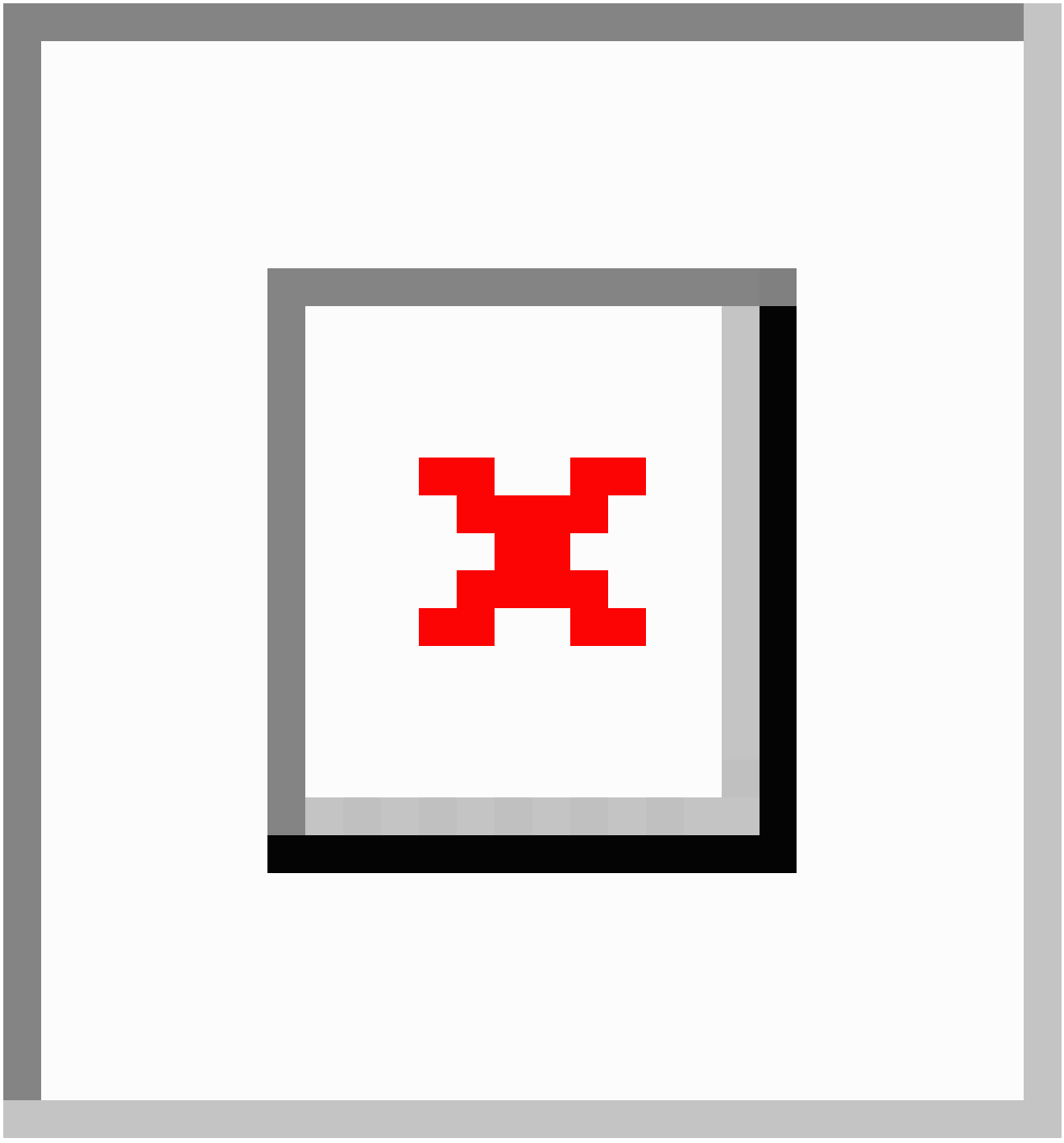
- Exist criteria with no additional filters (eg, conditions or a laboratory concept with no filter, like the existence of a hemoglobin value regardless of the value)
- Comparatively restricted numerical criteria (eg, hemoglobin laboratory value <12 g/dL)
- Range-restricted numerical criteria (eg, hemoglobin laboratory value between 10 and 12 g/dL)
- Value set restricted criteria (eg, gender=female or male)

Additionally, each criterion can be further restricted to a date range (eg, a Condition that occurred between January 1, 2024, and February 5, 2024), and it supports additional "attribute" filters, which can be added to each of the base types of criteria. The attribute filters support similar filters that identify the criterion types, ie, comparative numerical, comparative range, and value set restriction (eg, the body site=skin for a tissue specimen—see Multimedia Appendix 1).

The Explicit Logic Layer

The logic layer of the query aligns with existing solutions (i2b2/tranSMART/TriNetX) in representing the structured query as a combination of conjunctive normal form (CNF) and disjunctive normal form (DNF). Every criterion is embedded into the logic layer in a CNF for inclusion criteria and DNF for exclusion criteria (Figure 1). Inclusion and exclusion criteria are then logically combined via an AND NOT operator by subtracting the result of the exclusion criteria from the result of the inclusion criteria. Every feasibility query also receives a syntax version number and an additional description. The syntax version allows to distinguish the current version from future versions and changes, and the description allows the query to transport additional human-readable information about the query.

Figure 1. Structured query syntax top-level elements and logic layer. Certain criterion types will imply additional intrinsic logical relations. See ValueSet criteria and attribute filters and time restrictions.

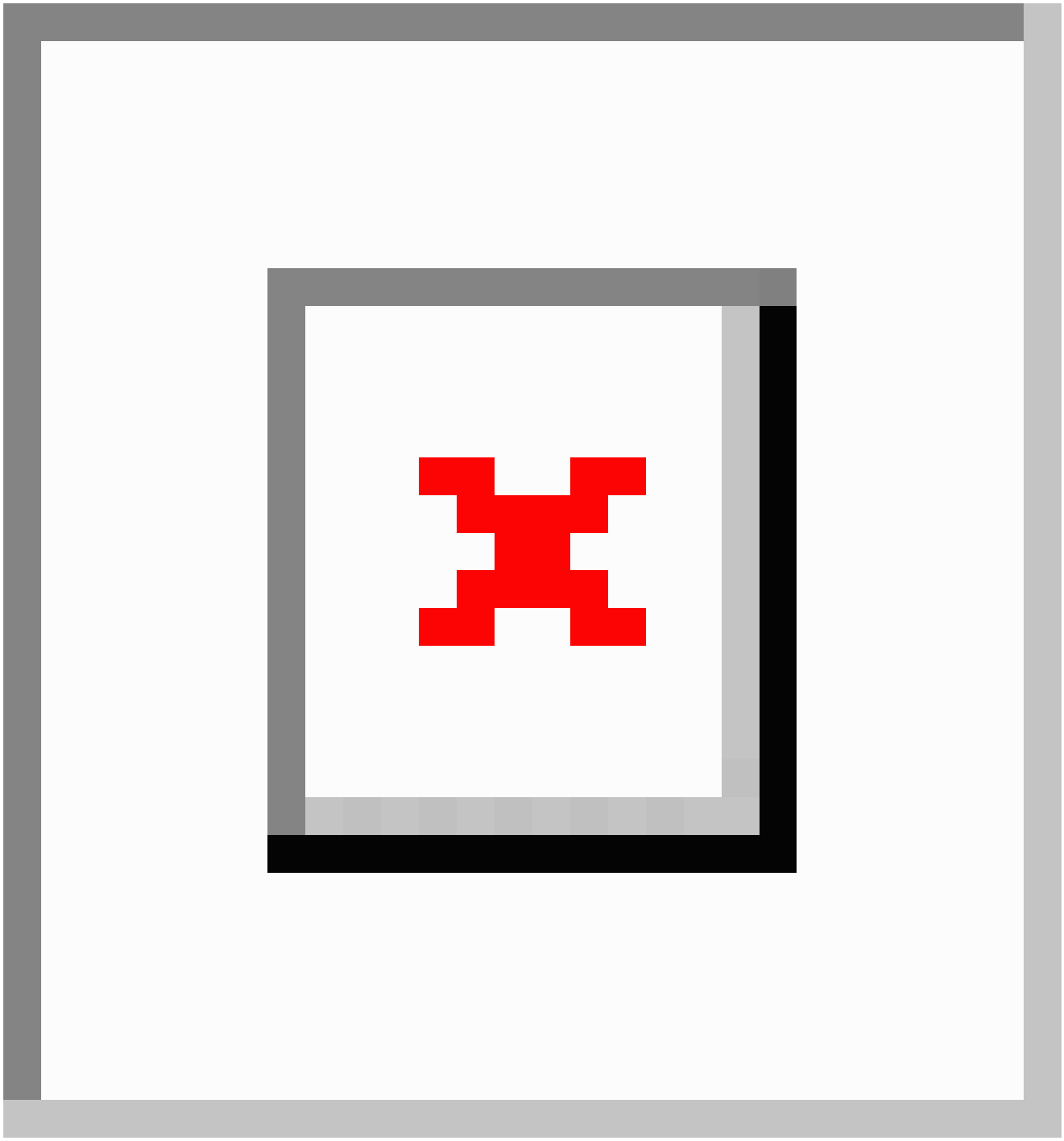


The Implicit Logic (Criteria Expansion)

Apart from the explicit logic layer across criteria, different types of criteria and their filters further impact the execution logic as follows.

ValueSet criteria (see Figure 2D) allow the selection of multiple values (concepts). In this case, the value selections are treated as OR choices. For example, gender = (male, female) expands to: (gender=male) OR (gender=female).

Figure 2. Different types of criteria definitions. (A) Simple conceptual criterion. (B) Numeric criterion with quantitative comparison. (C) Numeric criterion with range restriction. (D) ValueSet criterion.



Attribute filters for each criterion are additional filters that can be set for each criterion. All individual filters on a criterion are combined using AND. For example, a specimen of type “Tissue specimen” and body site “skin” only applies to specimens with the type of Tissue and the body site skin.

The same applies to time restrictions. In this example, the time restriction “between 2020-01-01 and 2021-01-01” will predictably be added using an AND conjunction of the type, body site, and time restriction.

Furthermore, there is an implicit OR expansion of criteria when the criterion-identifying code is a parent code of multiple child codes within a terminology hierarchy. For example, suppose a researcher adds the diagnosis of type 2 diabetes mellitus as a criterion (*ICD-10* code=E11). In that case, it can be expanded to search all subtypes of type 2 diabetes mellitus (E11, E11.3, E11.31, E11.30, E11.1, E11.11, E11.0, E11.01, E11.7, E11.75, E11.74, E11.73, E11.72, E11.4, E11.41, E11.40, E11.8, E11.81, E11.80, E11.2, E11.21, E11.20, E11.5, E11.51, E11.50, E11.6, E11.61, E11.60, E11.9, E11.91, E11.90) combining them using a logical OR operation).

Context-Dependent Criteria

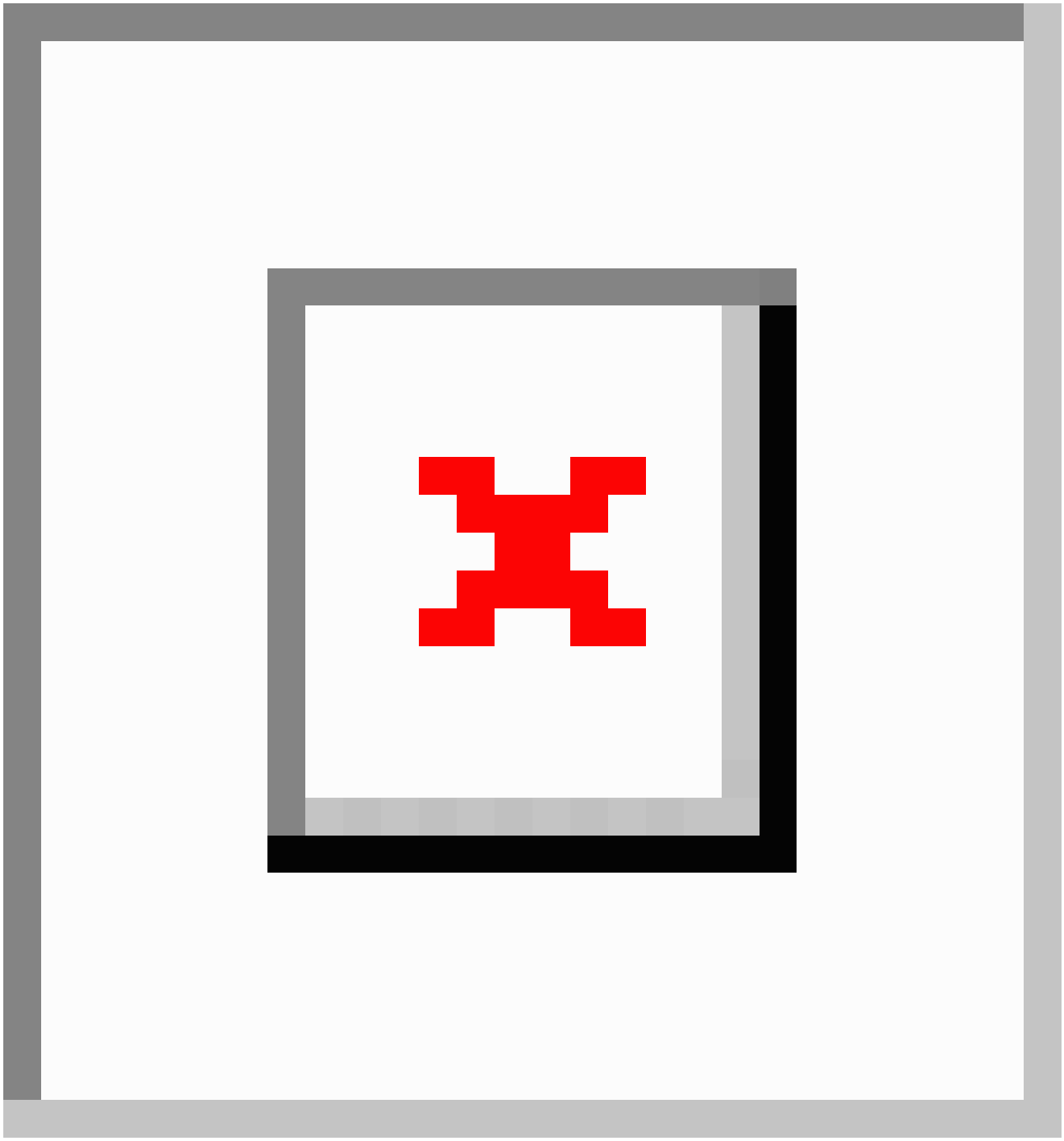
In some cases, a criterion cannot be uniquely defined by its term code within a terminology, making it impossible to map a criterion for execution. One example of this is the use of *ICD-10*

condition codes for causes of death, specimen-specific conditions, or the general condition of a patient.

In modern terminologies like SNOMED-CT, this can be resolved using postcoordination, where a combined code, which carries the context, is created. For example, 419620001|Death|:42752001|Due to|=22298006|Myocardial infarction| which, while in line with SNOMED Compositional Grammar [28], a template to express this is not currently part of the SNOMED-CT implementation.

The syntax we developed here allows for post-coordinated codes; however, we allow for an additional "context" attribute for some use cases where postcoordination is unsuitable. The context attribute is modeled after our termCode attribute and provides an extra term code to identify the context. Figure 3 provides an example for myocardial infarction as condition aand cause of death.

Figure 3. Myocardial infarction in 2 contexts (condition and cause of death).



Data Availability

As a technical solution to define the structure of the CCDL, we decided on the JSON Schema definition and made it publicly available [29]. The schema serves implementation guidance and validation purposes; the git repository also contains documentation examples, test data, and the capabilities to create matching test queries.

Requirement Verification

An analysis was performed based on the structure defined in the JSON schema to evaluate the developed intermediate query syntax. The following table (Table 1) presents the detailed results of this analysis:

This syntax efficiently meets a wide range of expressiveness and interoperability requirements, demonstrating capabilities in defining complex medical queries with standard terminologies and logical operators.

Table . CCDL[a] components and their purpose regarding the expressiveness requirements.

Component	Key properties	Purpose and function	Requirements met
inclusionCriteria	CNF[b] without negation	Conjunction of criteria with logical operators.	Expressive query formulation, boolean logic
exclusionCriteria	DNF[c] without negation	Allows negation of criteria for comprehensive exclusion.	Negation of criteria on a group level, Boolean logic
termCode	code, system, version, display	Identifies concepts using standard coding systems.	Standard medical terminology, uniqueness
criterion	context, termCodes, valueFilter, attibuteFilter, timeRestriction	Sets criteria with defined context, using term codes and filters.	Expressiveness of simple and complex eligibility criteria
timeRestriction	afterDate, beforeDate	Specifies time frame for criteria fulfillment.	Time restrictions
unit	code, display	Standardized unit definition, adhering to UCUM[d] units.	Use of standardized units
valueFilter	type (concept, quantity-comparator, etc)	Varied filtering types for flexible data querying.	Numeric restriction, concept restrictions
attributeFilters	type (concept, quantity-comparator, reference)	Mechanism for detailed filtering at the attribute level.	Detailed filtering, clinical relations

[a]CCDL: Clinical Cohort Definition Language.

[b]CNF: conjunctive normal form.

[c]DNF: disjunctive normal form.

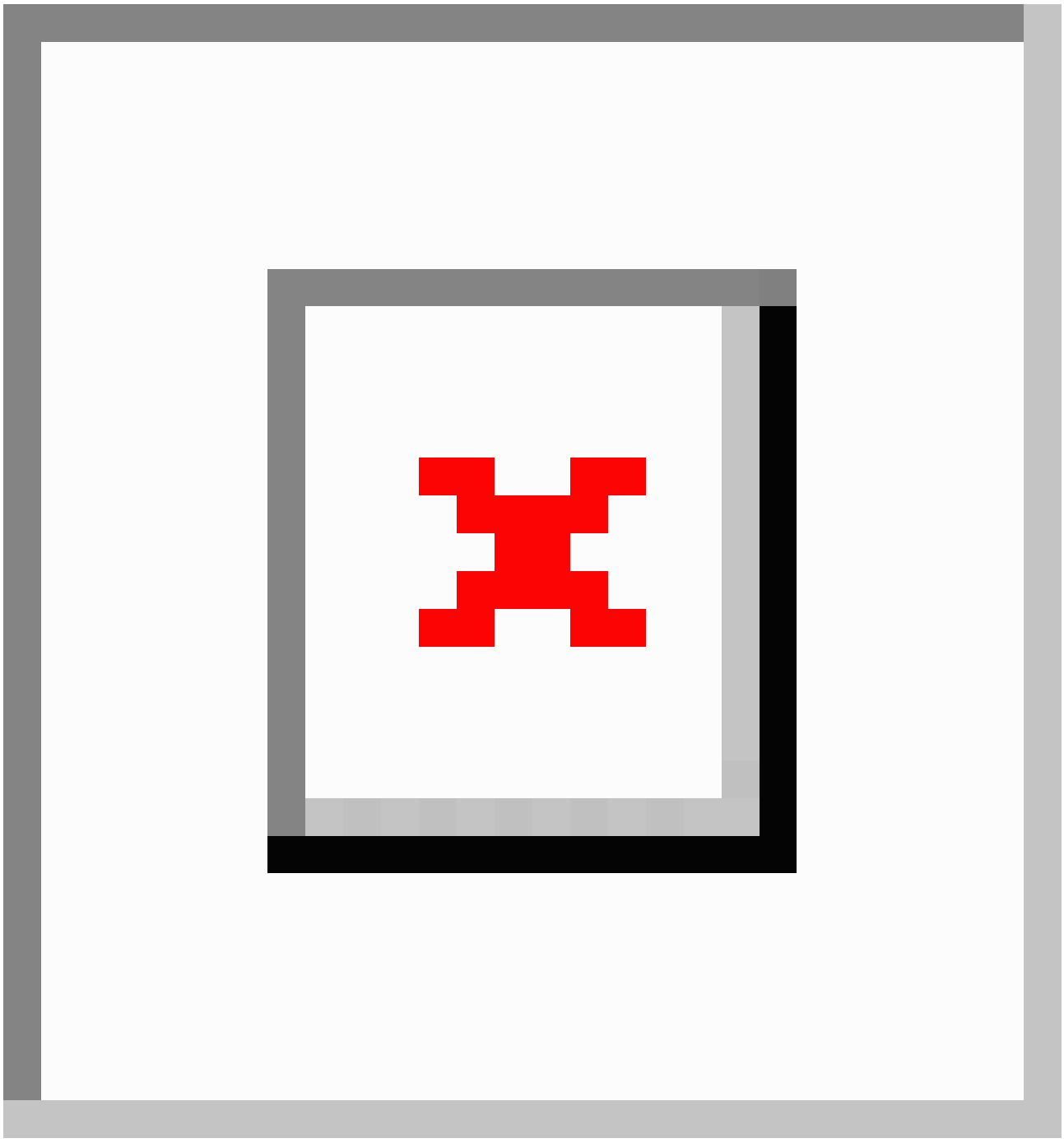
[d]UCUM: ___.

Evaluation and Use of the CCDL in Real-World Scenarios

We believe the potential of the CCDL extends beyond its application in the federated feasibility portal of the German Research Portal for Health. Nevertheless, the CCDL remains a crucial technical solution within the FDPG's feasibility portal.

We created a user interface for the feasibility queries in the FDPG (Figure 4), which generates the CCDL, demonstrating how it lends itself well to building feasibility query user interfaces [30,31]. The CCDL especially supports this as its design follows the typical way of feasibility query creation as seen in platforms such as the FDPG, i2b2, OMOP, and TrinetX. We evaluated the usability of the user interface across multiple projects [32,33] and embedded it in a German-wide distributed research infrastructure [10,34]. These evaluations highlighted the applicability of the CCDL to a feasibility query and that the usability issues found were not due to a lack of expressiveness of the CCDL. We further used the Synthea data set to test the CCDL against [35], demonstrated the ability of the CCDL to represent a wide range of criteria [36], and showed that it could be fully translated to FHIR Search [37], CQL [38], and AQL [39]. At the time of writing, almost 9000 CCDLs have been created and executed across Germany.

Figure 4. Example of a feasibility query in the central German Portal for Health Data (FDPG) feasibility portal to find patients with a leukocyte count within a normal range, with a malignant neoplasm of the brain, available tumor tissue specimen, and a CT scan after January 1, 2020, who did not take doxorubicin.



Discussion

Principal Findings

We presented an intermediate feasibility query syntax that separates concerns between the user interface and the execution of a feasibility query on different research repositories and their specific query languages. The syntax defined here fulfills all the interoperability and accessibility requirements while supporting a broad range of expressiveness requirements we identified by analyzing existing query tools. The solution is fully compatible with established medical terminology standards, notation of parameters, and restriction semantics.

The solution we describe here is compatible with the query logic established by i2b2 and, therefore, tranSMART and TriNetX. This means that tools like i2b2 or similar could be easily extended to produce our syntax.

The CCDL was further used as part of a larger infrastructure for feasibility queries in Germany and is currently used as the interface for feasibility queries within the German research portal for health, supporting feasibility queries across 39 university hospitals in Germany. We successfully created translation components for FHIR Search and CQL in the current implementation. Current research also indicates the adaptability for FHIR Pathling's aggregation API [40], and SQL. The criteria

content and the required reintroduction of data model—dependent information are obtained from an automatically generated search ontology [36].

Related Work

While the expression of eligibility criteria within a specific data model context is well established and adequately discussed in this work, research on a data-agnostic intermediate format for computable eligibility has been sparse in recent years.

Alper et al [41] closely align with our approach of representing eligibility criteria in a structured format, namely the FHIR EvidenceVariable, which currently does not directly support the representation of eligibility criteria but may be refined to do so. Presumably, an FHIR representation would provide a structure beyond the realm of the MII, which could add significant value and improve syntax interoperability. However, in the early stages, the challenges of adopting new solutions could have impeded the development presented here. Our ongoing communication with the HL7 working group, which focuses on Research Studies, gives us confidence that once a suitable FHIR Resource is established or adapted to meet the needs outlined in our publication, the established technical components could be efficiently modified to align with these changes. Parallels can be drawn to implementing structured eligibility criteria, as presented by Yuan et al [42] and Fang et al [43]. Their publications present a half-automated approach to generate feasibility queries based on free text study protocols from ClinicalTrials.gov [23]. Their system is built around the OHDSI data model and uses the concept IDs. After converting the free text criteria, they allow users to edit and download an intermediate representation in JSON format. Unfortunately, no clear implementation guidelines on the format are given by Yuan et al [42] and Fang et al [43]. However, recurring themes include differentiating inclusion and exclusion criteria and defining temporal constraints. To our knowledge, contrary to our approach, they do not allow for further restrictions beyond the value constraint on specific criteria.

Limitations

The separation of concerns, which the CCDL provides, also leads to the need for a mapping to identify the correct way of translating the CCDL information model to the local information model and terminology. The mapping allows the link between the specific data model and the criterion as identified in the CCDL to be created. One example of this is that for FHIR Search, the mapping for a condition criterion identified by a specific *ICD-10* code C50.0 would provide the information that the condition is found in the end point "/Condition" and the search parameter for the term code is "code" – Leading to the translated FHIR Search URL:

[fhir-base-url]/Condition?code=http://fhir.de/CodeSystem/bfarm/icd-10-gm|C50.0."

Further, additional information about the terminology is necessary to allow the selection of criteria within a terminology hierarchy, where the criterion resolves to multiple child criteria. Finally, this then requires the query executor and the CCDL-generating user interface to agree on criteria or terminology entries.

One common requirement currently not supported by the CCDL is temporal interdependencies between different criteria. Therefore, queries like a specific laboratory value within a certain period of diagnosis cannot be currently expressed using the CCDL.

We deliberately decided to delay the implementation of this extension as time dependencies significantly increase the complexity and performance requirements of any query execution.

The data model agnostic nature of the CCDL is inherently valuable. Its full potential—the capability to be used across different health care data models—requires more than technical translation. For cross-model query capability, the existence of the concepts in all target data models must be ensured.

Future Work

The CCDL described here provides a good base to make feasibility queries possible across various research repositories and close the gap between the different research repositories and their access. We have demonstrated the applicability of the CCDL to FHIR Search, CQL, and AQL; however, more repositories and other query languages, such as SQL on FHIR, OHDSI OMOP, or i2b2 might be added in the future. Further, one could imagine how separating the query syntax and execution would theoretically allow one to query different internationally distributed repositories such as FHIR, OMOP-CDM, and i2b2 simultaneously. Additionally, the CCDL is currently limited in how much it can express, and new capabilities will be added in the future. In this pursuit of making the CCDL more expressive, any extension must be weighed against the added complexity and overhead it introduces.

Conclusion

We presented a query syntax for medical feasibility queries, which creates an abstract layer between the user interface and the execution query language. We showed how it is flexible enough to be translated into different query languages and can be used to express various complex feasibility queries. The applicability of the query syntax was further demonstrated by embedding it into a large research project where it is used to query multiple millions of patients across 39 German university hospitals. The CCDL for feasibility queries will be extended in the future to allow more features, and we are currently working on a modified version for data selection and extraction.

Acknowledgments

The project was funded by the German Federal Ministry of Education and Research (BMBF) under the FDPG-PLUS Project (grants 01ZZ2309A, 01ZZ2309C, 01ZZ2309D, 01ZZ2309E, and 01ZZ2309F).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example of specimen criteria with an ICD-o-3 (International Classification of Diseases for Oncology, 3rd Edition) attribute indicating the location the specimen was taken from.
[PNG File, 62 KB - medinform_v12i1e58541_app1.png]

References

1. Pfaff ER, Girvin AT, Gabriel DL, et al. Synergies between centralized and federated approaches to data quality: a report from the national covid cohort collaborative. J Am Med Inform Assoc 2022 Mar 15;29(4):609-618. [doi: 10.1093/jamia/ocab217] [Medline: 34590684]
2. Prayitno, Shyu CR, Putra KT, et al. A systematic review of federated learning in the healthcare area: from the perspective of data properties and applications. Appl Sci (Basel) 2021 Nov 25;11(23):11191. [doi: 10.3390/app112311191]
3. Sebire NJ, Cake C, Morris AD. HDR UK supporting mobilising computable biomedical knowledge in the UK. BMJ Health Care Inform 2020 Jul;27(2):e100122. [doi: 10.1136/bmjhci-2019-100122] [Medline: 32723851]
4. Morrato EH, Lennox LA, Sendro ER, et al. Scale-up of the Accrual to Clinical Trials (ACT) network across the clinical and translational science award consortium: a mixed-methods evaluation of the first 18 months. J Clin Trans Sci 2020 Dec;4(6):515-528. [doi: 10.1017/cts.2020.505]
5. Litton JE. Launch of an infrastructure for health research: BBMRI-ERIC. Biopreserv Biobank 2018 Jun;16(3):233-241. [doi: 10.1089/bio.2018.0027] [Medline: 29781706]
6. AKTIN and SPoCK Research Group, Bienzeisler J, Triefenbach L, et al. A federated and distributed data management infrastructure to enable public health surveillance from intensive care unit data. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe JD, Pelayo S, et al, editors. Studies in Health Technology and Informatics: IOS Press; 2022. [doi: 10.3233/SHTI220507]
7. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-582. [doi: 10.1136/amiajnl-2014-002747] [Medline: 24821743]
8. Lawrence AK, Selter L, Frey U. SPHN - the Swiss personalized health network initiative. Stud Health Technol Inform 2020 Jun 16;270:1156-1160. [doi: 10.3233/SHTI200344] [Medline: 32570562]
9. Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med 2018 May;57(S 01):e50-e56. [doi: 10.3414/ME18-03-0003]
10. Gruendner J, Deppenwiese N, Folz M, et al. The architecture of a feasibility query portal for distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) patient data repositories: design and implementation study. JMIR Med Inform 2022 May 25;10(5):e36709. [doi: 10.2196/36709] [Medline: 35486893]
11. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR: Springer International Publishing; 2016. [doi: 10.1007/978-3-319-30370-3]
12. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. Ann Intern Med 2010 Nov 2;153(9):600-606. [doi: 10.7326/0003-4819-153-9-201011020-00010] [Medline: 21041580]
13. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130. [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
14. Kalra D, Beale T, Heard S. The openEHR foundation. Stud Health Technol Inform 2005;115:153-173. [Medline: 16160223]
15. Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009;16(5):624-630. [doi: 10.1197/jamia.M3191] [Medline: 19567788]
16. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. JCO Clin Cancer Inform 2018 Dec;2:1-10. [doi: 10.1200/CCI.17.00067] [Medline: 30652541]
17. Scheufele E, Aronzon D, Coopersmith R, et al. tranSMART: an open source knowledge management and high content data analytics platform. AMIA Jt Summits Transl Sci Proc 2014;2014:96-101. [Medline: 25717408]
18. Lablans M, Kadioglu D, Mate S, Leb I, Prokosch HU, Ückert F. Strategien zur Vernetzung von Biobanken: Klassifizierung verschiedener Ansätze zur Probensuche und Ausblick auf die Zukunft in der BBMRI-ERIC. Bundesgesundheitsbl 2016 Mar;59(3):373-378. [doi: 10.1007/s00103-015-2299-y]
19. Schüttler C, Prokosch HU, Hummel M, et al. The journey to establishing an IT-infrastructure within the German Biobank Alliance. PLoS ONE 2021;16(9):e0257632. [doi: 10.1371/journal.pone.0257632] [Medline: 34551019]
20. Schüttler C, Huth V, von Jagwitz-Biegnitz M, Lablans M, Prokosch HU, Griebel L. A federated online search tool for biospecimens (sample locator): usability study. J Med Internet Res 2020 Aug 18;22(8):e17739. [doi: 10.2196/17739] [Medline: 32663150]
21. ATLAS. GitHub. URL: https://github.com/OHDSI/Atlas/wiki/Home [accessed 2024-01-02]

22. Hoffmann J, Hanß S, Kraus M, et al. The DZHK research platform: maximisation of scientific value by enabling access to health data and biological samples collected in cardiovascular clinical studies. Clin Res Cardiol 2023 Jul;112(7):923-941. [doi: 10.1007/s00392-023-02177-5]
23. Home. : ClinicalTrials.gov URL: https://clinicaltrials.gov/ [accessed 2024-03-11]
24. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit Transl Bioinform 2010 Mar 1;2010:46-50. [Medline: 21347148]
25. Gulden C, Mate S, Prokosch HU, Kraus S. Investigating the capabilities of FHIR search for clinical trial phenotyping. In: German Medical Data Sciences: A Learning Healthcare System: IOS Press:2018. [doi: 10.3233/978-1-61499-896-9-3]
26. Schüttler C, Prokosch HU, Sedlmayr M, Sedlmayr B. Evaluation of three feasibility tools for identifying patient data and biospecimen availability: comparative usability study. JMIR Med Inform 2021 Jul 21;9(7):e25531. [doi: 10.2196/25531] [Medline: 34287211]
27. Walonoski J, Kramer M, Nichols J, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2018 Mar 1;25(3):230-238. [doi: 10.1093/jamia/ocx079]
28. Drenkhahn C, Ohlsen T, Wiedekopf J, Ingenerf J. WASP—A web application to support syntactically and semantically correct SNOMED CT postcoordination. Appl Sci (Basel) 2023 May 16;13(10):6114. [doi: 10.3390/app13106114]
29. Release v100 · medizininformatik-initiative/clinical-cohort-definition-language. URL: https://github.com/medizininformatik-initiative/clinical-cohort-definition-language/releases/tag/v1.0.0 [accessed 2024-03-18]
30. medizininformatik-initiative/feasibility-deploy. Medizininformatik-Initiative. 2024. URL: https://github.com/medizininformatik-initiative/feasibility-deploy [accessed 2024-06-16]
31. medizininformatik-initiative/feasibility-gui. Medizininformatik-Initiative. 2024. URL: https://github.com/medizininformatik-initiative/feasibility-gui [accessed 2024-06-16]
32. Sedlmayr B, Sedlmayr M, Kroll B, Prokosch HU, Gruendner J, Schüttler C. Improving covid-19 research of university hospitals in Germany: formative usability evaluation of the codex feasibility portal. Appl Clin Inform 2022 Mar;13(2):400-409. [doi: 10.1055/s-0042-1744549] [Medline: 35445386]
33. Schüttler C, Zerlik M, Gruendner J, et al. Empowering researchers to query medical data and biospecimens by ensuring appropriate usability of a feasibility tool: evaluation study. JMIR Hum Factors 2023;10:e43782. [doi: 10.2196/43782]
34. Prokosch HU, Gebhardt M, Gruendner J, et al. Towards a national portal for medical research data (FDPG): vision, status, and lessons learned. Stud Health Technol Inform 2023 May 18;302:307-311. [doi: 10.3233/SHTI230124] [Medline: 37203668]
35. flare/.github/integration-test at main. medizininformatik-initiative/flare.: GitHub URL: https://github.com/medizininformatik-initiative/flare/tree/main/.github/integration-test [accessed 2024-06-16]
36. Rosenau L, Majeed RW, Ingenerf J, et al. Generation of a Fast Healthcare Interoperability Resources (FHIR)-based ontology for federated feasibility queries in the context of COVID-19: feasibility study. JMIR Med Inform 2022 Apr 27;10(4):e35789. [doi: 10.2196/35789] [Medline: 35380548]
37. medizininformatik-initiative/flare: Feasibility Analysis Request Executor. URL: https://github.com/medizininformatik-initiative/flare [accessed 2024-01-04]
38. medizininformatik-initiative/sq2cql. URL: https://github.com/medizininformatik-initiative/sq2cql [accessed 2024-01-04]
39. Rosenau L, Ingenerf J. Structured queries to AQL: querying openEHR data leveraging a FHIR-based infrastructure for federated feasibility queries. In: MEDINFO 2023 — The Future Is Accessible: IOS Press; 2024:33-37. [doi: 10.3233/SHTI230922]
40. Grimes J, Szul P, Metke-Jimenez A, Lawley M, Loi K. Pathling: analytics on FHIR. J Biomed Semant 2022 Sep 8;13(1):23. [doi: 10.1186/s13326-022-00277-1]
41. Alper BS, Dehnbostel J, Shahin K, Ojha N, Khanna G, Tignanelli CJ. Striking a match between FHIR-based patient data and FHIR-based eligibility criteria. Learn Health Syst 2023 Oct;7(4):e10368. [doi: 10.1002/lrh2.10368] [Medline: 37860063]
42. Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc 2019 Apr 1;26(4):294-305. [doi: 10.1093/jamia/ocy178] [Medline: 30753493]
43. Fang Y, Idnay B, Sun Y, et al. Combining human and machine intelligence for clinical trial eligibility querying. J Am Med Inform Assoc 2022 Jun 14;29(7):1161-1171. [doi: 10.1093/jamia/ocac051] [Medline: 35426943]

Abbreviations

API: application programming interface
AQL: Archetype Query Language
CCDL: Clinical Cohort Definition Language
CDR: Clinical Data Repository
CNF: conjunctive normal form
CQL: Clinical Quality Language
DNF: disjunctive normal form

FDPG: central German Portal for Health Data
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven International
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
LOINC: Logical Observation Identifiers Names and Codes
MII: Medical Informatics Initiative
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms
SQL: Structured Query Language

Edited by C Lovis; submitted 18.03.24; peer-reviewed by B Alper, P Horki; revised version received 16.06.24; accepted 23.06.24; published 14.10.24.

Please cite as:
Rosenau L, Gruendner J, Kiel A, Köhler T, Schaffer B, Majeed RW
Bridging Data Models in Health Care With a Novel Intermediate Query Format for Feasibility Queries: Mixed Methods Study
JMIR Med Inform 2024;12:e58541
URL: https://medinform.jmir.org/2024/1/e58541
doi:10.2196/58541

Review

Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review

Yuan Peng[1], MSc; Franziska Bathelt[2], Dr Rer Nat; Richard Gebler[1], MSc; Robert Gött[3], Dipl.-Ing.; Andreas Heidenreich[4], Dipl.-Biol., Dr Rer Nat; Elisa Henke[1], MSc; Dennis Kadioglu[4,5], MSc; Stephan Lorenz[1], MSc; Abishaa Vengadeswaran[5], MSc; Martin Sedlmayr[1], Dr Rer Nat, Prof Dr

[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

[2]Thiem-Research GmbH, Cottbus, Germany

[3]Core Unit Datenintegrationszentrum, University Medicine Greifswald, Greifswald, Germany

[4]Department for Information and Communication Technology (DICT), Data Integration Center (DIC), Goethe University Frankfurt, University Hospital, Frankfurt am Main, Germany

[5]Institute for Medical Informatics, Goethe University Frankfurt, University Hospital Frankfurt, Frankfurt am Main, Germany

Corresponding Author:
Yuan Peng, MSc
Institute for Medical Informatics and Biometry
Carl Gustav Carus Faculty of Medicine
Technische Universität Dresden
Fetscherstraße 74
Dresden, 01307
Germany
Phone: 49 3514583648
Fax: 49 3514585738
Email: yuan.peng@tu-dresden.de

Abstract

Background: Multisite clinical studies are increasingly using real-world data to gain real-world evidence. However, due to the heterogeneity of source data, it is difficult to analyze such data in a unified way across clinics. Therefore, the implementation of Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT) processes for harmonizing local health data is necessary, in order to guarantee the data quality for research. However, the development of such processes is time-consuming and unsustainable. A promising way to ease this is the generalization of ETL/ELT processes.

Objective: In this work, we investigate existing possibilities for the development of generic ETL/ELT processes. Particularly, we focus on approaches with low development complexity by using descriptive metadata and structural metadata.

Methods: We conducted a literature review following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We used 4 publication databases (ie, PubMed, IEEE Explore, Web of Science, and Biomed Center) to search for relevant publications from 2012 to 2022. The PRISMA flow was then visualized using an R-based tool (Evidence Synthesis Hackathon). All relevant contents of the publications were extracted into a spreadsheet for further analysis and visualization.

Results: Regarding the PRISMA guidelines, we included 33 publications in this literature review. All included publications were categorized into 7 different focus groups (ie, medicine, data warehouse, big data, industry, geoinformatics, archaeology, and military). Based on the extracted data, ontology-based and rule-based approaches were the 2 most used approaches in different thematic categories. Different approaches and tools were chosen to achieve different purposes within the use cases.

Conclusions: Our literature review shows that using metadata-driven (MDD) approaches to develop an ETL/ELT process can serve different purposes in different thematic categories. The results show that it is promising to implement an ETL/ELT process by applying MDD approach to automate the data transformation from Fast Healthcare Interoperability Resources to Observational Medical Outcomes Partnership Common Data Model. However, the determining of an appropriate MDD approach and tool to implement such an ETL/ELT process remains a challenge. This is due to the lack of comprehensive insight into the characterizations of the MDD approaches presented in this study. Therefore, our next step is to evaluate the MDD approaches presented in this

study and to determine the most appropriate MDD approaches and the way to integrate them into the ETL/ELT process. This could verify the ability of using MDD approaches to generalize the ETL process for harmonizing medical data.

(JMIR Med Inform 2024;12:e52967) doi:10.2196/52967

KEYWORDS

ETL; ELT; Extract-Load-Transform; Extract-Transform-Load; interoperability; metadata-driven; medical domain; data harmonization

Introduction

Multisite clinical studies are increasingly using real-world data to gain real-world evidence, especially during the COVID-19 pandemic [1]. However, not all clinics use the same hospital information system, resulting in heterogeneity of data produced by different hospital information systems. These heterogeneous data are not semantically and syntactically interoperable. Therefore, it is difficult to analyze such data in a unified way across sites. For this, the heterogeneous data need to be harmonized and standardized, for example, by using a common data model (CDM) [2]. For example, the European Medical Agency [3] set up the DARWIN EU (Data Analysis and Real World Interrogation Network European Union) [4] to provide real-world evidence on use and adverse events of medicines across the European Union. DARWIN EU uses the Observational Medical Outcomes Partnership (OMOP) CDM [5] as the base model, which is provided by the Observational Health Data Sciences and Informatics [6] community. To participate in such networks, a transformation of local data is needed. A common approach is to develop an Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT) process. Both are used to harmonize heterogeneous data into the target systems. The only difference between them is the order of processing data. ETL transforms the data before loading them into the target systems, while ELT loads the data into the target systems first, and then transforms the data. Due to the different data formats and source systems, multiple ETL/ELT processes have to be implemented [7-10]. This work is time-consuming and hard to maintain [11].

Using a standard data exchange format can reduce the complexity of transforming heterogeneous data into CDMs. An example is the Fast Healthcare Interoperability Resources (FHIR) [12] format. FHIR is a communication standard and is provided by the Health Level 7 (HL7) [13]. In Germany, the Medical Informatics Initiative (MII) [14] provides a Core Data Set (CDS) [15] in FHIR format for enabling the interoperability of data across all university hospitals. Another German association “the National Association of Statutory Health Insurance Physicians” (KBV, German: Kassenärztliche Bundesvereinigung) [16] also provides a KBV CDS in FHIR format, which provides a stable foundation for the development of the medical information objects [17] (eg, immunization records and maternity records). Although both MII CDS and KBV CDS are based on the German HL7 Basis Profiles [18], the FHIR profiles defined in the 2 CDSs are not identical [19]. This is due to the different requirements of MII and KBV. For example, codes indicating departments within a clinic (eg, 0100 for internal medicine department) are defined in different value-sets and therefore use different coding systems. This also complicates the implementation and maintenance of ETL/ELT processes.

Furthermore, most countries try to standardize their electronic health records (EHR) data for research and to improve the interoperability of the data. Consequently, country-specific FHIR profiles are developed, for example, German HL7 Basis Profiles [18] and the US CDS [20]. Due to different languages (ie, German vs English), different structure definitions (eg, extensions and cardinality) and different coding systems (eg, system URL for International Classification of Diseases, 10, Revision: German Modification [21] vs system URL for International Classification of Diseases, 10, Clinical Modification [22]) used in the FHIR profiles, different ETL processes need to be implemented [8,23]. Although these are just a few examples, it is conceivable that with the expansion of supported use cases, the time required for implementing an ETL/ELT process increases massively, while the maintainability decreases. Therefore, the implementation of a generic ETL/ELT process for harmonizing local health data can guarantee the semantic and syntactic interoperability of research data across sites and countries.

Using metadata for the implementation of ETL/ELT processes is a promising approach, as stated by David Loshin [24]: “in order to organize data for analytical purposes, it will need to be extracted from the original source (source metadata), transformed into a representation that is consistent with the warehouse (target metadata) in a way that does not lose information due to differences in format and precision (structure metadata) and is aligned in a meaningful way (semantic metadata).” A very broad definition of metadata is “data about other data” [25]. Depending on the specific context of use, metadata can be classified into 3 types [26]:

- **Descriptive metadata**: the metadata is used for discovery and identification purposes, for example metadata for source and target data.
- **Structural metadata**: the metadata is used for managing data in information systems, for example, column names and table names in a database.
- **Administrative metadata**: the metadata exists within a database that provides additional information, for example, the name of a person, who has changed the data in a database.

Metadata can be represented by metadata languages (eg, Resource Description Framework and Notation3) [27]. Such languages are also called ontology languages. For enabling the interoperability of data from different source and target systems, rule languages (eg, Rule Markup Language and Semantic Web

Rule Language) can be used to define the transformation rules between them [27]. Therefore, the use of metadata is expected to improve the development and maintenance for transforming FHIR resources to OMOP CDM.

As a side note, we understand any (descriptive and structural) metadata-based approach used for developing ETL/ELT processes as metadata-driven (MDD) approach. This work focuses on providing an overview of the types of MDD approaches and their use in different thematic categories. The overview aims to identify a suitable MDD approach to enhance the data transformation from FHIR to OMOP CDM. This will be achieved by answering the following questions:

- Q1: What are the themes of application for MDD approaches?
- Q2: What types of MDD approaches exist in the literature?
- Q3: What are the reasons for the usage of MDD approaches?
- Q4: What tool was used to implement the MDD approach?

Methods

To answer our 4 research questions, we conducted a literature review. To ensure the transparency of the review process, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [28]. We used 4 publication databases (ie, PubMed, IEEE Explore, Web of Science, and Biomed Center) to search for relevant publications from 2012 to 2022 written in German or English (Textbox 1). The first search was performed on August 11, 2022, and the second one was on March 15, 2023, which in turn completed the search through December 31, 2022. The collected publications were loaded into the Zotero Citation Management program (Corporation for Digital Scholarship) [29] and the duplicates were manually removed. To better categorize the publications to be excluded, we defined 8 exclusion criteria (Textbox 2).

This review was a 2-fold process consisting of Title-Abstract-Screening (TAS) and full-text screening (FTS). Both screening processes used the same exclusion criteria listed in Textbox 2. The unique publications were divided into 2 groups based on their publication dates and uploaded to a research collaboration platform, Rayyan (Qatar Computing Research Institute and Cochrane Bahrain) [30], as 2 separate projects. Each publication group was assigned with 4 reviewers. The corresponding author reviewed all publications. The TAS was performed under the blind-modus, so that each reviewer could label the publication independently. The blind-modus was turned off after all publications were tagged and the conflicts were discussed and resolved. After that, all included publications were randomly divided into 2 groups and reloaded into Rayyan as a new project for FTS. Similar to TAS, 4 reviewers were assigned to each publication group and the corresponding author reviewed all publications. The FTS was also conducted under the blind-modus and followed the same review process as the TAS.

We extracted the content of all included publications based on the categories listed in Textbox 3. The extraction of publication content was done by the corresponding author and validated by 4 coauthors. The extracted content was stored in a spreadsheet for further analysis and visualization.

The result of the literature review was visualized using an R-based tool, which was developed based on PRISMA 2020 [31].

Textbox 1. Search string and publication databases.

Search string

PubMed

- ((meta data) OR (meta-data) OR (metadata) OR (ontology) OR (rules)) AND ((extract transform load) OR (ETL) OR (extract load transform) OR (ELT))

IEEE Explore

- ((“All Metadata”:metadata) OR (“All Metadata”:meta-data) OR (“All Metadata”:meta data) OR (“All Metadata”:ontology) OR (“All Metadata”:rules)) AND ((“All Metadata”:ETL) OR (“All Metadata”:extract transform load) OR (“All Metadata”:ELT) OR (“All Metadata”:extract load transform))

Web of Science

- (ALL=(metadata) OR ALL=(meta-data) OR ALL=(“meta data”) OR ALL=(ontology) OR ALL=(rules)) AND (ALL=(ETL) OR ALL=(“extract transform load”) OR ALL=(ELT) OR ALL=(“extract load transform”))

Biomed Center (BMC)

- (“meta data” OR meta-data OR metadata OR ontology OR rules) AND (“extract transform load” OR ETL OR “extract load transform” OR ELT)

Textbox 2. Labels and descriptions of exclusion criteria.

Wrong_abbreviation

- Publication does not contain Extract-Transform-Load (ETL) as “Extract-Transform-Load.”
- Publication does not contain Extract-Load-Transform (ELT) as “Extract-Load-Transform.”

Wrong_definition

- Publication does not use metadata in the context of “metadata of data in source or target.”
- Publication does not use rules in the context of “rules for data transformation.”

Only_etl_elt

- Publication describes only ETL/ELT.

Only_metadata

- Publication describes only metadata.

Wrong_focus

- Publication mentioned metadata and ETL/ELT, but the focus is not about data harmonization

Wrong_type

- Publication is not a conference paper or a journal publication

Foreign_language

- Publication is written in other languages than English and German

Wrong_content

- Publication does not mention ETL/ELT or metadata

Textbox 3. Categories for data extraction.

Theme

- The main theme of the work.

Metadata-driven method

- The used metadata-driven method in the work.

Metadata-driven method tool

- Tool which was used to conduct the metadata-driven method.

Purpose

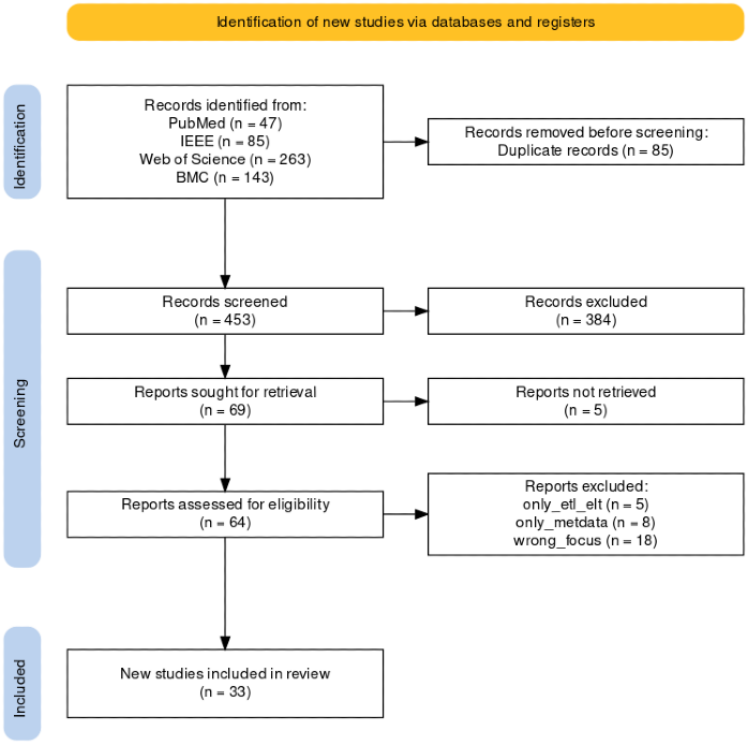
- The purpose of using the metadata-driven method.

Results

Literature Search

The literature search resulted in 538 publications. After removing 85 duplicates, 453 publications were screened during the TAS phase. By using the exclusion criteria defined in Textbox 2 and excluding the publications, which have no full-text, 64 publications were included for FTS. Finally, we included 33 publications in this work. The screening process and results are structured using the PRISMA flow diagram 2020 (Figure 1). A complete list of included publications is available in Multimedia Appendix 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. Generated using an R-based tool (reproduced from Haddaway et al [31], with permission from Neal R Haddaway).



Distribution of Publications

In order to gain an overview of the potential application focuses of MDDs (Q1) and thus an indication of where the approaches have proven beneficial, the focused theme of application was first evaluated. According to the extracted data, the focuses of all included publications are classified into 7 different categories, namely medicine (n=9) [10,32-39], data warehouse (n=13) [40-52], big data (n=4) [53-56], industry (n=4) [57-60], geoinformatics (n=1) [61], archaeology (n=1) [62], and military (n=1) [63]. This shows that data warehouse and medicine are the 2 categories that use the MDD approach the most.

MDD Approaches Used for Various Thematic Categories

Different types of MDD approaches were used across the thematic categories. To gain knowledge about the use of these types of MDD approaches in each category (Q2), the distribution of MDD approaches was investigated. Figure 2 shows the application of different types of MDD approaches in different thematic categories. The most frequently used type of MDD approach was ontology-based, where the ontology (using for example, resource description framework) of the source or target was applied in the ETL/ELT process. This approach was used in 6 categories, particularly in the categories of data warehouse [45-48,50,52] and medicine [10,32,35,37-39]. Another frequently used type of MDD approach was rule-based, which applied transformation rules generated based on the source and target to the ETL/ELT process. The rule-based approach was also widely used in the categories of data warehouse [40-43,49] and medicine [33,34,37,39]. All other MDD approaches besides the ontology-based and rule-based approaches were categorized as “other” (Table 1).

Figure 2. Metadata-driven approaches used in each thematic category.

Table 1. MDD[a] approaches that are categorized as "other."

MDD approach type and publication	Example
UML[b]-based	
Dhaouadi et al [46]	UML class diagram is used for modeling the transformation process
Graphic-based	
Dhaouadi et al [46]	BPMN[c] standard is used for modeling an ETL[d] process
Ad hoc formalisms-based	
Dhaouadi et al [46]	Entity Mapping Diagram is used for representing ETL tasks
MDA[e]-based	
Dhaouadi et al [46]	MDA is a multilayered framework with multiple submodules for separation of the specification of a functionality from its implementation
Message-based	
Novak et al [51]	"Normal message" contains information of mapping and transformation; "command message" configures the (execution) system
Template-based	
McCarthy et al [58]	A transformation template for each data source that manages the complex transformation process
Binding et al [62]	A template contains the mapping patterns which is then used for querying in database
Metadata-based[f]	
Ozyurt and Grethe [36]	Implementing a generic data transformation language to transform heterogeneous data from multiple sources to a common format
Tomingas et al [44]	Metadata of the source and target stored in a knowledge and metadata repository
Suleykin and Panfilov [60]	Metadata of the mapping path stored in a metadata management framework

[a]MDD: metadata-driven.

[b]UML: unified modeling language.

[c]BPMN: Business Process Model Notation.

[d]ETL: Extract-Transform-Load.

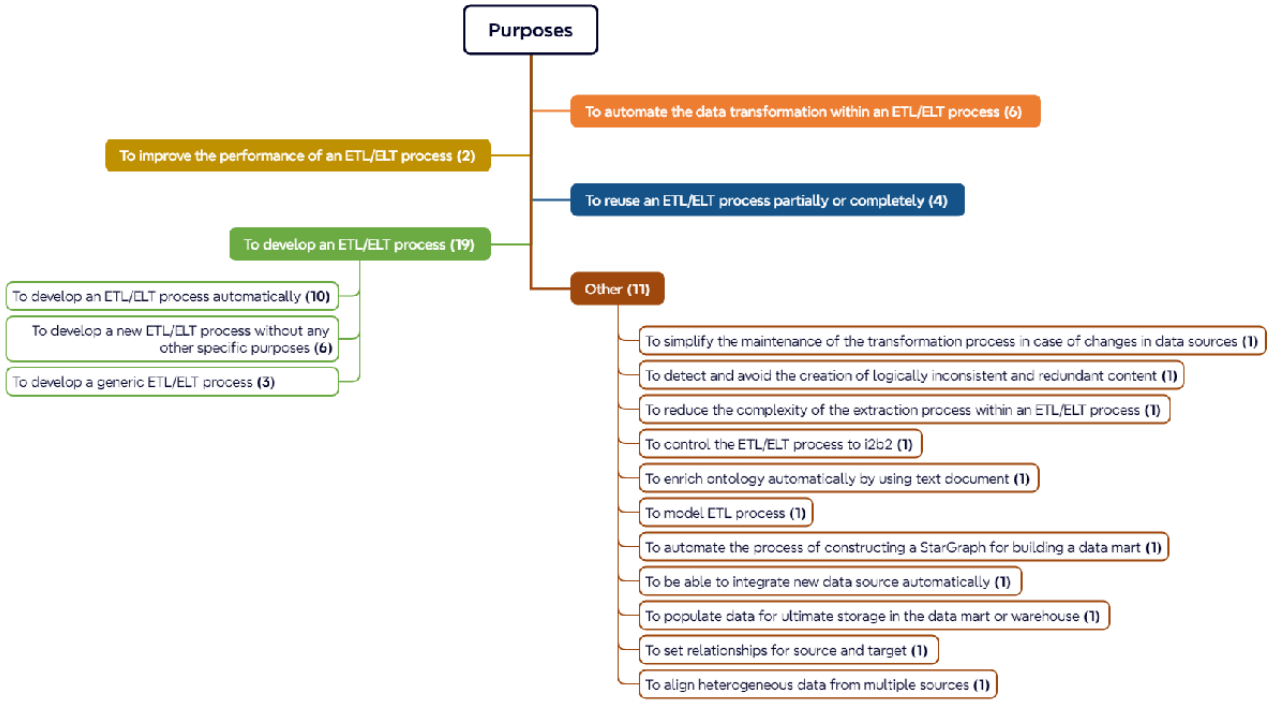
[e]MDA: Model Driven Architecture.

[f]Metadata-based approach: approach uses metadata without any specification.

Purposes of Using MDD Method for Data Harmonization

The purpose of using MDD approaches in each use case was then investigated to clarify the reasons why MDD approaches were used (Q3). Figure 3 shows different purposes of using MDD approaches in developing ETL/ELT processes based on the extracted data. The majority of publications describe the use of MDD approaches to develop an ETL/ELT process. This purpose can be divided into three detailed categories: (1) to automate the development of the ETL/ELT process [35,38,42,46,48-51,60], (2) to develop a generic ETL/ELT process [39,47,52], and (3) to develop a new ETL/ELT process without any further technical specifications [40,45,46,55,57,61]. Additionally, the transformation part of the ETL/ELT process could also be automated by applying an MDD approach [34,37,41,44,58,63]. For example, Chen and Zhao [41] described an MDD approach for the automatic generation of SQL scripts for data transformation. Moreover, using MDD approaches can also help to improve the performance of ETL/ELT processes [43,46] or to partially or fully reuse the ETL/ELT process [10,33,43,62]. Other goals (categorized as "Others" in Figure 3), such as simplifying the maintenance of the transformation process [37] and reducing the complexity of the extraction process [53], can also be realized by using MDD approaches in ETL/ELT processes.

Figure 3. Purposes of using MDD approaches in ETL/ELT process. ELT: Extract-Load-Transform; ETL: Extract-Transform-Load; i2b2: Informatics for Integrating Biology and the Bedside; MDD: metadata-driven.



Relationship Between Use Case and Used MDD Approach

As shown in the previous section, different MDD approaches were applied for different purposes. To further elucidate the reasons for choosing MDD approaches (Q3), the relationship between them was investigated. Table 2 lists the number of publications, which used a type of MDD approach to achieve a specific purpose. The ontology-based approach was used to achieve purposes (1) and (2), and (4)-(7). For example, Huang et al [63] created both local ontology (ontology based on the source data) and global ontology (ontology for the query processing) for the data transformation process, so that the data transformation from local ontology to global can be automated by applying ontology learning, ontology mapping, and ontology rules. Additionally, the ontology-based approach was also used to achieve other goals, such as controlling the ETL process to Informatics for Integrating Biology and the Bedside [32] and reducing the complexity of the extraction process [53]. Similar to the ontology-based approach, the rule-based approach was used to achieve the purposes of (1)-(3) and (5)-(7). Due to the reusability of the transformation rules, it was also possible to simplify the maintenance of the ETL/ELT process by applying rules in the process [37]. Other MDD approaches such as template-based [58,62], message-based [51], and metadata-based [41,44,48] were used to achieve the goals of (1)-(3) and (5)-(7). A metadata-based approach (eg, metadata management framework) can be used to develop the ETL tasks automatically [60]. The detailed information of Table 2 is available in the Multimedia Appendix 1.

Table 2. Relationships between purposes and MDD[a] approaches used.

Purposes		MDD approaches		
Number	Description	Ontology-based, n/N (%)	Rule-based, n/N (%)	Other, n/N (%)
(1)	To automate the data transformation within an ETL[b]/ELT[c] process	2/6 (33)	3/6 (50)	1/6 (17)
(2)	To reuse an ETL/ELT process (partially or completely)	1/4 (25)	2/4 (50)	1/4 (25)
(3)	To improve the performance of an ETL/ELT process	0/2 (0)	1/2 (50)	1/2 (50)
(4)	To develop a generic ETL/ELT process	3/3 (100)	0/3 (0)	0/3 (0)
(5)	To develop an ETL/ELT process automatically	5/9 (56)	2/9 (22)	2/9 (22)
(6)	To develop a new ETL/ELT process (without any other specific purposes)	4/6 (67)	1/6 (17)	1/6 (17)
(7)	Other	5/11 (45)	2/11 (18)	4/11 (36)

[a]MDD: metadata-driven.

[b]ETL: Extract-Transform-Load.

[c]ELT: Extract-Load-Transform.

Tools Used for Implementing MDD Approaches

Finally, we focused on the tools used to implemented MDD approaches (Q4). For achieving various purposes as shown in the previous section, different tools were used. As shown in Figure 4, each type of MDD approach can be implemented by using either an existing tool or a use case specific tool. Based on the included publications, the ontology-base approaches were mostly implemented using Protégé (Stanford Center for Biomedical Informatics Research) [64]. Protégé is an ontology editor, as well as OntoEdit (Institute AIFB, University of Karlsruhe and Ontoprise GmbH) [65]. The main reason for using an ontology editor is its ease of use and maintenance, as well as the various plug-ins. The use of case specific tools, such as ontology generator introduced by Kamil et al [45], generated ontologies based on the data definition language of the relational database. Both types of tools were used for creating and maintaining the ontology, which was then used to establish a generic mapping logic in the ETL/ELT process [32,50,52,54,55,61]. Another type of frequently used MDD approach is rule-based, which is used for phrasing and storing the transformation rules. The transformation rules can be stored in a mapping sheet [49], a CSV file [34], a YAML (YAML Ain't Markup Language) file [33] or a table within a database [43], which were implemented manually. Afterwards, the transformation rules could be used in the ETL/ELT process, for example, to enable the automatic transformation. Other types of MDD approaches can also be implemented by using existing tools (eg, knowledge and metadata repository [66]) or use case specific tools (eg, metadata repository [41] and metadata management framework [60]). For example, Ozyurt and Grethe [36] implemented a generic transformation language using the bioCADDIE Data Tag Suite (bioCADDIE Project) [67] (a metadata schema) to align heterogeneous data from multiple sources, which provided a basis for further analytic queries.

Figure 4. Tools used for developing the metadata-driven approach. MMF: metadata management framework; OWL: Web Ontology Language; YAML: YAML Ain't Markup Language.

Discussion

Principal Findings

Our literature review on the topic "metadata-driven ETL/ELT" includes all publications listed on PubMed, IEEE Explore, Web of Science, and Biomed Center on MDD ETL/ELT process from 2012 to 2022. In some context, the use of metadata is represented specifically using "ontology" or "rules." Therefore, we added "ontology" and "rules" into the search string to expand the search range.

With the review process presented, we were able to provide an overview of the thematic categories to which the MDD ETL/ELT processes were applied (Q1), the types of MDD approaches used in the ETL/ELT processes (Q2), the purposes of using MDD approaches (Q3), as well as the tools used to implement the MDD approaches (Q4).

Across all thematic categories, ontology-based and rule-based approaches are the most used approaches in the data warehouse and the medical thematic categories. In some cases, more than one MDD approach was used in the ETL/ELT process. For example, Del Carmen Legaz-García et al [39] used both ontology-based and rule-based approaches. Therefore, such publications were categorized as both MDD approach types.

Various tools can be used to implement MDD approaches. Unfortunately, we were not able to extract this information from all included publications. The reason for that is that some publications used proprietary or nontransferable approaches (eg, data-specific ontologies [39,62] and rules from Data Vault [DataVaultAlliance] [42]). Some other publications did not explicitly mention or describe the tools they used. Therefore, these publications were not included in the analysis of MDD tools used.

The results indicate that it is promising to implement a generic ETL/ELT process to transform different FHIR profiles to OMOP CDM automatically by applying MDD approaches. However, the results do not provide a trivial solution for this. For example, Huang et al [63] used an ontology-based approach to be able to automate the data transformation in an ETL/ELT process, while Ong et al [34] used a rule-based approach to achieve the same purpose. In some cases, more than one MDD approach were used as complements in order to accomplish the data transformation. For example, Pacaci et al [37] chose an ontology-based approach to automate the data transformation and a rule-based to simplify the maintenance of the transformation process in case of changes in data sources. By applying these 2 approaches in combination, the authors were able to transform EHR data from heterogeneous EHR systems into OMOP CDM. Therefore, determining an appropriate MDD approach and tool to implement a generic ETL/ELT process to transform FHIR to OMOP CDM automatically remains a challenge.

This work aimed to provide an overview of different types of MDD approaches and their tools. Consequently, this review lacks an analysis of detailing the specific traits of each MDD approach. This gap underscores the importance of providing a comprehensive insight into the characterizations of the MDD approaches presented in this study. This analysis will be conducted in the future to provide solid evidence for selecting the most suitable MDD approach and tool, or for considering using multiple MDD approaches in combination to implement the generic ETL/ELT process for transforming FHIR to OMOP CDM.

Conclusions

Our literature review shows that using MDD approaches to develop an ETL/ELT process can serve different purposes in different focus groups (ie, medicine, data warehouse, big data, industry, geoinformatics, archaeology, and military). The results show that it is promising to implement an ETL/ELT process by applying MDD approach for automating the data transformation from FHIR to OMOP CDM. However, the determination of an appropriate MDD approach and tool to implement such an ETL/ELT process remains a challenge. This is due to the lack of comprehensive insight into the characterizations of the MDD approaches presented in this study. Therefore, our next step is to evaluate the MDD approaches presented in this study and to determine the most appropriate MDD approaches and the way of integrating them into the MII CDS FHIR to OMOP CDM ETL process [8]. This could verify the ability of using MDD approaches to generalize the ETL process for harmonizing medical data [11].

Acknowledgments

This publication was partially funded by the German Federal Ministry of Education and Research (BMBF) Network of University Medicine 2.0: “NUM 2.0”, Grant No. 01KX2121, Project: NUM-Data integration center – NUM-DIZ. The Article Processing Charge was funded by the joint publication funds of the Technische Universität, Dresden, including the Carl Gustav Carus Faculty of Medicine, and the Sächsische Landesbibliothek—Staats- und Universitätsbibliothek, Dresden, as well as the Open Access Publication Funding of the Deutsche Forschungsgemeinschaft.

Authors' Contributions

All authors contributed substantially to this work. YP did the search string definition and publications for the review-process preparation. YP, FB, Robert G, AH, EH, DK, SL, and AV: screened the title and abstract. YP, FB, Richard G, Robert G, AH, EH, DK, SL, and AV screened the full text. YP did the data extraction. FB, DK, Robert G, and SL performed the data extraction validation. YP wrote the original draft. YP, FB, Richard G, Robert G, AH, EH, DK, SL, AV, and MS reviewed and edited the writing. MS handled the resources. All authors have read and agreed to the current version of the paper and take responsibility for the scientific integrity of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excel tables for extracted data from included publications.
[XLSX File (Microsoft Excel File), 462 KB - medinform_v12i1e52967_app1.xlsx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[DOCX File , 85 KB - medinform_v12i1e52967_app2.docx]

References

1. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022;22(1):287 [FREE Full text] [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
2. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
3. European Medicines Agency. URL: https://www.ema.europa.eu/en [accessed 2022-08-18]
4. Data Analysis and Real World Interrogation Network (DARWIN EU). 2021. URL: https://www.darwin-eu.org/ [accessed 2023-12-16]
5. The Book of OHDSI: Observational Health Data Sciences and Informatics. San Bernardino, CA: OHDSI; 2019. URL: https://ohdsi.github.io/TheBookOfOhdsi [accessed 2024-01-19]
6. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]

7. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the all of us research program: transforming i2b2 data into the OMOP common data model. PLoS One 2019;14(2):e0212463 [FREE Full text] [doi: 10.1371/journal.pone.0212463] [Medline: 30779778]
8. Peng Y, Henke E, Reinecke I, Zoch M, Sedlmayr M, Bathelt F. An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. Int J Med Inform 2023;169:104925 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104925] [Medline: 36395615]
9. Zoch M, Henke E, Reinecke I, Peng Y, Gebler R, Gruhl M, et al. Extract, transform and load German claim data to OMOP CDM—design and implications. Ger Medical Sci 2022:153 [FREE Full text] [doi: 10.3205/22gmds057]
10. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc 2016;23(5):909-915 [FREE Full text] [doi: 10.1093/jamia/ocv188] [Medline: 26911824]
11. Peng Y, Henke E, Sedlmayr M, Bathelt F. Towards ETL Processes to OMOP CDM Using Metadata and Modularization. Stud Health Technol Inform 2023;302:751-752 [FREE Full text] [doi: 10.3233/SHTI230256] [Medline: 37203486]
12. FHIR v4.0.1. HL7 International. URL: https://www.hl7.org/fhir/ [accessed 2022-04-05]
13. Kabachinski J. What is Health Level 7? Biomed Instrum Technol 2006;40(5):375-379 [FREE Full text] [doi: 10.2345/i0899-8205-40-5-375.1] [Medline: 17078369]
14. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
15. Ganslandt T, Boeker M, Löbe M, Prasser F, Schepers J, Semler SC, et al. Der Kerndatensatz der Medizininformatik-Initiative Ein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene. Forum der Medizin-Dokumentation und Medizin-Informatik 2018;20(1):17-21.
16. The National Association of Statutory Health Insurance Physicians and the regional Associations of Statutory Health Insurance Physicians. Kassenärztliche Bundesvereinigung. 2020. URL: https://www.kbv.de/html/about_us.php [accessed 2023-08-01]
17. Medizinische Informationsobjekte (MIO). Kassenärztliche Bundesvereinigung. 2021. URL: https://www.kbv.de/html/mio.php [accessed 2023-08-01]
18. Leitfaden Basis DE (R4). HL7 FHIR Implementierungsleitfäden. URL: https://ig.fhir.de/basisprofile-de/stable/Home.html [accessed 2023-08-01]
19. Koch M, Richter J, Hauswaldt J, Krefting D. How to Make Outpatient Healthcare Data in Germany Available for Research in the Dynamic Course of Digital Transformation. Stud Health Technol Inform 2023;307:12-21 [FREE Full text] [doi: 10.3233/SHTI230688] [Medline: 37697833]
20. US Core implementation guide. HL7 International. URL: https://www.hl7.org/fhir/us/core/ [accessed 2022-12-16]
21. System URL for ICD-10-GM. Fast Healthcare Interoperability Resources. URL: http://fhir.de/CodeSystem/dimdi/icd-10-gm [accessed 2023-12-30]
22. System URL for ICD-10-CM. HL7 International. URL: http://hl7.org/fhir/sid/icd-10-cm [accessed 2023-12-30]
23. OMOPonFHIR Project. URL: https://omoponfhir.org/ [accessed 2022-04-05]
24. Loshin D. Chapter 9—metadata. In: Loshin D, editor. Business Intelligence: The Savvy Manager's Guide, 2nd Edition. Waltham, MA: Morgan Kaufmann; 2013:119-130.
25. Ulrich H, Kock-Schoppenhauer A, Deppenwiese N, Gött R, Kern J, Lablans M, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022;24(1):e25440 [FREE Full text] [doi: 10.2196/25440] [Medline: 35014967]
26. ISO/IEC TR 19583-1:2019: information technology: concepts and usage of metadata—part 1: metadata concepts. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/67365.html [accessed 2023-05-15]
27. Breitman KK, Casanova MA, Truszkowski W. Semantic Web: Concepts, Technologies and Applications. London: Springer; 2007.
28. Moher D, Liberati A, Tetzlaff J, Altman D, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
29. Zotero. 2022. URL: https://www.zotero.org/ [accessed 2022-02-10]
30. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
31. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022;18(2):e1230 [FREE Full text] [doi: 10.1002/cl2.1230] [Medline: 36911350]
32. Post AR, Pai AK, Willard R, May BJ, West AC, Agravat S, et al. Metadata-driven clinical data loading into i2b2 for clinical and translational science institutes. AMIA Jt Summits Transl Sci Proc 2016;2016:184-193 [FREE Full text] [Medline: 27570667]
33. Quiroz JC, Chard T, Sa Z, Ritchie A, Jorm L, Gallego B. Extract, transform, load framework for the conversion of health databases to OMOP. PLoS One 2022;17(4):e0266911 [FREE Full text] [doi: 10.1371/journal.pone.0266911] [Medline: 35404974]

34. Ong TC, Kahn MG, Kwan BM, Yamashita T, Brandt E, Hosokawa P, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Med Inform Decis Mak 2017;17(1):134 [FREE Full text] [doi: 10.1186/s12911-017-0532-3] [Medline: 28903729]
35. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H, Bürkle T, et al. Ontology-based data integration between clinical and research systems. PLoS One 2015;10(1):e0116656 [FREE Full text] [doi: 10.1371/journal.pone.0116656] [Medline: 25588043]
36. Ozyurt IB, Grethe JS. Foundry: a message-oriented, horizontally scalable ETL system for scientific data integration and enhancement. Database (Oxford) 2018;2018:bay130 [FREE Full text] [doi: 10.1093/database/bay130] [Medline: 30576493]
37. Pacaci A, Gonul S, Sinaci AA, Yuksel M, Erturkmen GBL. A semantic transformation methodology for the secondary use of observational healthcare data in postmarketing safety studies. Front Pharmacol 2018;9:435 [FREE Full text] [doi: 10.3389/fphar.2018.00435] [Medline: 29760661]
38. Haarbrandt B, Tute E, Marschollek M. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. J Biomed Inform 2016;63:277-294 [FREE Full text] [doi: 10.1016/j.jbi.2016.08.007] [Medline: 27507090]
39. Del Carmen Legaz-García M, Miñarro-Giménez JA, Menárguez-Tortosa M, Fernández-Breis JT. Generation of open biomedical datasets through ontology-driven transformation and integration processes. J Biomed Semantics 2016;7:32 [FREE Full text] [doi: 10.1186/s13326-016-0075-z] [Medline: 27255189]
40. Gang H, Jin-Rong L, Xiu-Ying W. A kind of bidirectional mapping strategy of heterogeneous data model based on metadata-driven. 2012 Presented at: Proceedings of 2012 2nd International Conference on Computer Science and Network Technology; December 29-31, 2012; Changchun, China p. 1023-1027. [doi: 10.1109/iccsnt.2012.6526100]
41. Chen Z, Zhao T. A new tool for ETL process. 2012 Presented at: 2012 International Conference on Image Analysis and Signal Processing; November 09-11, 2012; Huangzhou, China p. 269-273. [doi: 10.1109/IASP.2012.6425038]
42. Puonti M, Raitalaakso T, Aho T, Mikkonen T. Automating transformations in data vault data warehouse loads. In: Thalheim B, Jaakkola H, Kiyoki Y, Yoshida N, editors. Information Modelling and Knowledge Bases XXVIII. Amsterdam: IOS Press; 2017:215-230.
43. Wang H, Zhang J, Guo J. Constructing data warehouses based on operational metadata-driven builder pattern. 2015 Presented at: 2015 International Conference on Logistics, Informatics and Service Sciences (LISS); July 27-29, 2015; Barcelona, Spain p. 1-4. [doi: 10.1109/liss.2015.7369630]
44. Tomingas K, Kliimask M, Tammet T. Data integration patterns for data warehouse automation. In: Vakali A, Trajcevski G, Kon-Popovska M, Ivanovic M, Bassiliades N, Palpanas T, et al, editors. New Trends in Database and Information Systems II. Berlin: Springer; 2015:41-55.
45. Kamil I, Inggriani MM, Asnar YDW. Data migration helper using domain information. 2014 Presented at: 2014 International Conference on Data and Software Engineering (ICODSE); November 26-27, 2014; Bandung, Indonesia p. 1-6. [doi: 10.1109/icodse.2014.7062492]
46. Dhaouadi A, Bousselmi K, Gammoudi MM, Monnet S, Hammoudi S. Data warehousing process modeling from classical approaches to new trends: main features and comparisons. Data 2022;7(8):113 [FREE Full text] [doi: 10.3390/data7080113]
47. Berkani N, Khouri S, Bellatreche L. Generic methodology for semantic data warehouse design: from schema definition to ETL. 2012 Presented at: 2012 Fourth International Conference on Intelligent Networking and Collaborative Systems; September 19-21, 2012; Bucharest, Romania p. 404-411. [doi: 10.1109/incos.2012.108]
48. Nath RPD, Romero O, Pedersen TB, Hose K. High-level ETL for semantic data warehouses. Semant Web 2022;13(1):85-132 [FREE Full text] [doi: 10.3233/sw-210429]
49. Yu QC. Metadata driven data mapper development. Appl Mech Mater 2013;411-414:403-407 [FREE Full text] [doi: 10.4028/www.scientific.net/amm.411-414.403]
50. Ta'a A, Abdullah MS. Ontology development for ETL process design. In: Ahmad MN, Abdullah MS, Colomb RM, editors. Ontology-based Applications for Enterprise Systems and Knowledge Management. Hershey, PA: Information Science Reference; 2013:261-275.
51. Novak M, Kermek D, Magdalenic I. Proposed architecture for ETL workflow generator. 2019 Presented at: Proceedings of the Central European Conference on Information and Intelligent Systems; October 2-4, 2019; Varaždin, Croatia p. 297-304.
52. Berkani N, Bellatreche L, Khouri S. Towards a conceptualization of ETL and physical storage of semantic data warehouses as a service. Cluster Comput 2013;16(4):915-931. [doi: 10.1007/s10586-013-0266-7]
53. Hilali I, Arfaoui N, Ejbali R. A new approach for integrating data into big data warehouse. 2022 Presented at: Proceedings Volume 12084, Fourteenth International Conference on Machine Vision (ICMV 2021); March 4, 2022; Rome, Italy p. 120841M. [doi: 10.1117/12.2623069]
54. Bansal SK, Kagemann S. Integrating big data: a semantic extract-transform-load framework. Computer 2015;48(3):42-50. [doi: 10.1109/mc.2015.76]
55. Bansal SK. Towards a semantic Extract-Transform-Load (ETL) framework for big data integration. 2014 Presented at: 2014 IEEE International Congress on Big Data; June 27-July 02, 2014; Anchorage, AK, USA p. 522-529. [doi: 10.1109/bigdata.congress.2014.82]

56. Boulahia C, Behja H, Louhdi MRC. Towards semantic ETL for integration of textual scientific documents in a big data environment: a theoretical approach. 2020 Presented at: 2020 6th IEEE Congress on Information Science and Technology (CiSt); June 05-12, 2021; Agadir-Essaouira, Morocco p. 133-138. [doi: 10.1109/cist49399.2021.9357280]
57. de Cesare C, Foy G, Lycett M. 4D-SETL a semantic data integration framework. 2016 Presented at: Proceedings of the 18th International Conference on Enterprise Information Systems—Volume 1: ICEIS; April 25-28, 2016; Rome, Italy p. 127-134. [doi: 10.5220/0005822501270134]
58. McCarthy S, McCarren A, Roantree M. A method for automated transformation and validation of online datasets. 2019 Presented at: 2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC); October 28-31, 2019; Paris, France p. 183-189. [doi: 10.1109/edoc.2019.00030]
59. Scriney M, McCarthy S, McCarren A, Cappellari P, Roantree M. Automating data mart construction from semi-structured data sources. Comput J 2019;62(3):394-413. [doi: 10.1093/comjnl/bxy064]
60. Suleykin A, Panfilov P. Metadata-driven industrial-grade ETL system. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); December 10-13, 2020; Atlanta, GA, USA p. 2433-2442. [doi: 10.1109/bigdata50022.2020.9378367]
61. Janecka K, Cerba O, Jedlicka K, Jezek J. Towards interoperability of spatial planning Data: 5-Steps harmonization framework. 2013 Presented at: 13th SGEM GeoConference on Informatics, Geoinformatics and Remote Sensing; June 16-22, 2013; Albena, Bulgaria p. 1005-1016. [doi: 10.5593/SGEM2013/BB2.V1/S11.051]
62. Binding C, Charno M, Jeffrey S, May K, Tudhope D. Template based semantic integration: from legacy archaeological datasets to linked data. Int J Semantic Web Inf Syst 2015;11(1):1-29. [doi: 10.4018/ijswis.2015010101]
63. Huang DM, Du YL, Zhang MH, Zhang C. Application of ontology-based automatic ETL in marine data integration. 2012 Presented at: 2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM); June 24-27, 2012; Kuala Lumpur p. 11-13. [doi: 10.1109/eeesym.2012.6258574]
64. Musen MA, Protégé Team. The Protégé project: a look back and a look forward. AI Matters 2015;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
65. Sure Y, Angele J, Staab S. OntoEdit: multifaceted inferencing for ontology engineering. In: Aberer K, March S, Spaccapietra S, editors. Journal on Data Semantics I. LNCS 2800. Verlag Berlin: Springer; 2003:128-152.
66. MMX metadata framework. Mindworks Industries. URL: https://www.mindworks.industries/mmx_framework.html [accessed 2023-11-21]
67. Sansone SA, Gonzalez-Beltran A, Rocca-Serra P, Alter G, Grethe JS, Xu H, et al. DATS, the data tag suite to enable discoverability of datasets. Sci Data 2017;4:170059 [FREE Full text] [doi: 10.1038/sdata.2017.59] [Medline: 28585923]

Abbreviations

CDM: Common Data Model
CDS: Core Data Set
DARWIN EU: Data Analysis and Real World Interrogation Network European Union
EHR: electronic health record
ELT: Extract-Load-Transform
ETL: Extract-Transform-Load
FHIR: Fast Healthcare Interoperability Resources
FTS: full-text screening
HL7: Health Level 7
KBV: The National Association of Statutory Health Insurance Physicians (German: Kassenärztliche Bundesvereinigung)
MDD: metadata-driven
MII: Medical Informatics Initiative
OMOP: Observational Medical Outcomes Partnership
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TAS: Title-Abstract-Screening
YAML: YAML Ain’t Markup Language

Edited by C Lovis; submitted 20.09.23; peer-reviewed by M Löbe, W Xu; comments to author 24.10.23; revised version received 01.12.23; accepted 03.12.23; published 14.02.24.

Please cite as:
Peng Y, Bathelt F, Gebler R, Gött R, Heidenreich A, Henke E, Kadioglu D, Lorenz S, Vengadeswaran A, Sedlmayr M
Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review
JMIR Med Inform 2024;12:e52967
URL: https://medinform.jmir.org/2024/1/e52967
doi:10.2196/52967
PMID:38354027

Original Paper

Integrating Clinical Data and Medical Imaging in Lung Cancer: Feasibility Study Using the Observational Medical Outcomes Partnership Common Data Model Extension

Hyerim Ji[1,2], MS; Seok Kim[1], MPH; Leonard Sunwoo[3], MD, PhD; Sowon Jang[3], MD; Ho-Young Lee[1,4], MD, PhD; Sooyoung Yoo[1], PhD

[1]Office of eHealth Research and Business, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

[2]Department of Health Science and Technology, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea

[3]Department of Radiology, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

[4]Department of Nuclear Medicine, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

Corresponding Author:
Sooyoung Yoo, PhD
Office of eHealth Research and Business
Seoul National University Bundang Hospital
172, Dolma-ro
Bundang-gu
Seongnam-si, 13605
Republic of Korea
Phone: 82 317878980
Fax: 82 317874061
Email: yoosoo0@snubh.org

Abstract

Background: Digital transformation, particularly the integration of medical imaging with clinical data, is vital in personalized medicine. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standardizes health data. However, integrating medical imaging remains a challenge.

Objective: This study proposes a method for combining medical imaging data with the OMOP CDM to improve multimodal research.

Methods: Our approach included the analysis and selection of digital imaging and communications in medicine header tags, validation of data formats, and alignment according to the OMOP CDM framework. The Fast Healthcare Interoperability Resources ImagingStudy profile guided our consistency in column naming and definitions. Imaging Common Data Model (I-CDM), constructed using the entity-attribute-value model, facilitates scalable and efficient medical imaging data management. For patients with lung cancer diagnosed between 2010 and 2017, we introduced 4 new tables—IMAGING_STUDY, IMAGING_SERIES, IMAGING_ANNOTATION, and FILEPATH—to standardize various imaging-related data and link to clinical data.

Results: This framework underscores the effectiveness of I-CDM in enhancing our understanding of lung cancer diagnostics and treatment strategies. The implementation of the I-CDM tables enabled the structured organization of a comprehensive data set, including 282,098 IMAGING_STUDY, 5,674,425 IMAGING_SERIES, and 48,536 IMAGING_ANNOTATION records, illustrating the extensive scope and depth of the approach. A scenario-based analysis using actual data from patients with lung cancer underscored the feasibility of our approach. A data quality check applying 44 specific rules confirmed the high integrity of the constructed data set, with all checks successfully passed, underscoring the reliability of our findings.

Conclusions: These findings indicate that I-CDM can improve the integration and analysis of medical imaging and clinical data. By addressing the challenges in data standardization and management, our approach contributes toward enhancing diagnostics and treatment strategies. Future research should expand the application of I-CDM to diverse disease populations and explore its wide-ranging utility for medical conditions.

(JMIR Med Inform 2024;12:e59187) doi:10.2196/59187

KEYWORDS

DICOM; OMOP; CDM; lung cancer; medical imaging; data integration; data quality; Common Data Model; Digital Imaging and Communications in Medicine; Observational Medical Outcomes Partnership

Introduction

The accessibility and use of health information in various formats and standards are limited, further limiting the development of advanced data analytics technologies, especially in an era where machine learning and other cutting-edge technologies have become essential for medical research. Integrating these sophisticated analytical tools requires a paradigm shift toward the standardization and harmonization of health care data [1,2]. Standardized data structures are not only beneficial but essential for the effective application of machine learning algorithms as they ensure consistent data quality (DQ), interoperability, and comprehensive analysis across different health care domains. By moving to standardized data formats, we laid the foundation for a more powerful and scalable application of emerging technologies, opening up new possibilities in medical research and patient care.

Several standardization projects and technologies have emerged in response to the demand for integrated approaches [3,4]. Among these, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is noteworthy for its advantages in converting diverse sources of data into a consistent format [5,6]. It harmonizes the structure and content of various clinical data sets, facilitating consistent analytical approaches in multi-institutional research. These characteristics ensure high efficiency and accuracy of data interpretation and use, thereby enhancing both the quality and pace of research in the rapidly evolving field of medicine.

Digital Imaging and Communications in Medicine (DICOM) is a universal standard for managing, storing, and transmitting medical images, ensuring interoperability and improved exchange of medical image data and associated information between health care systems. Recent research has focused on the integrated analysis of DICOM and OMOP CDM to promote accessibility to complex medical imaging data and electronic health records [7-11]. These efforts aim to combine detailed imaging metrics with diverse clinical data to contribute to the development of diagnostic and therapeutic strategies through comprehensive data analysis. However, the data duplication problem caused by constructing an instance-level table and the absence of a table that can store annotation data such as labeling (commonly used in image analysis) are major limitations. These limitations must be addressed to effectively manage the complex characteristics of medical imaging data and perform an integrated analysis with clinical information in the OMOP CDM. As the complexity of medical imaging data and the range of DICOM tags increase, effective solutions are required to integrate data seamlessly and consistently.

Lung cancer is the leading cause of cancer-related deaths worldwide and accounts for >20% of all cancer fatalities in South Korea. The etiology, progression, and therapeutic response of lung cancer are intricately linked to a myriad of biological and genetic factors. Therefore, a systematic understanding of the characteristics of lung cancer is paramount for its early detection, prevention, and construction of personalized treatment strategies [12,13]. However, this requires an approach that efficiently integrates high-resolution data across various fields.

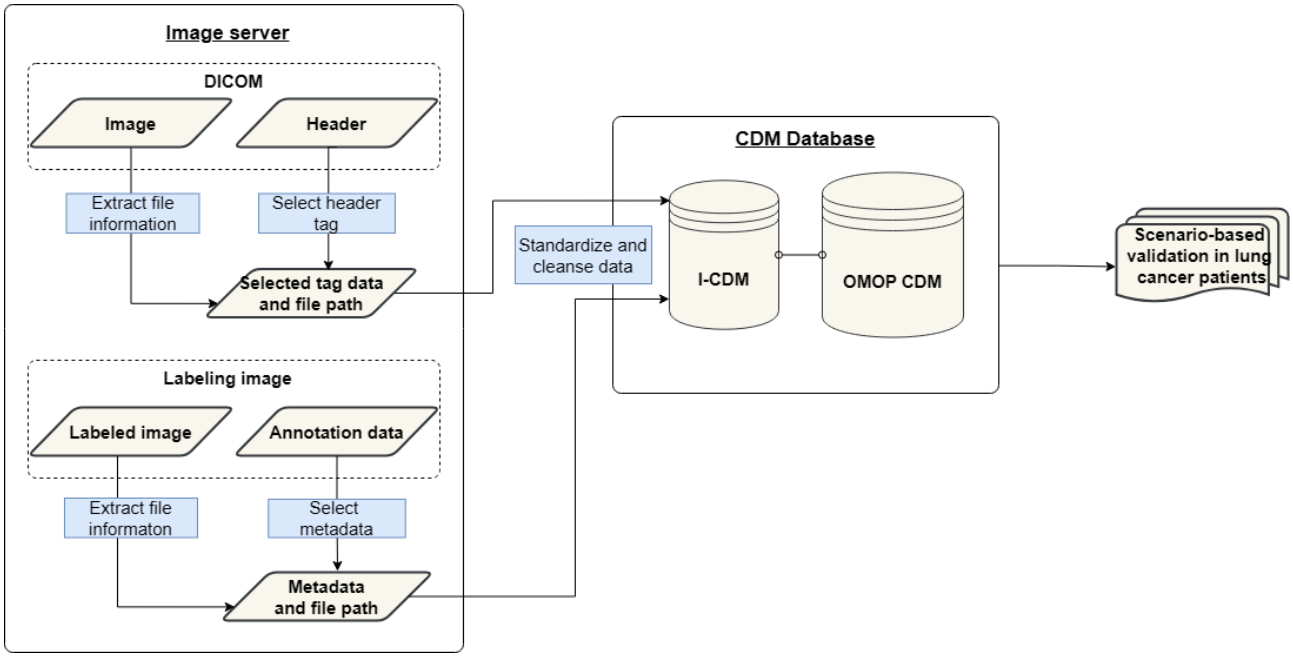
In this study, we propose a method to integrate medical imaging data with the OMOP CDM, aimed at enhancing multimodal research capabilities. This approach involves converting DICOM metadata and its annotation data to fit within the OMOP CDM framework and subsequently integrating it into a designed Imaging Common Data Model (I-CDM). We applied this integrated framework to a specific cohort of patients with lung cancer and brain metastases to not only test the feasibility and utility of our approach but also demonstrate its practical application through a series of research scenarios. Additionally, the use of scenarios was intended to showcase use cases that validate the operational functionality of our proposed model within real-world research settings.

Methods

Overview

We systematically analyzed and selected the DICOM header tags, verified their data formats, and mapped them to the OMOP CDM framework. To ensure consistency and interoperability in the column naming and definitions, we referenced the Fast Healthcare Interoperability Resources (FHIR) image study profiles, constructed I-CDM tables incorporating an entity-attribute-value (EAV) model for scalability, and performed data preprocessing to maintain data integrity. This allowed us to construct and validate a series of scenarios that combined clinical and imaging information from patients with lung cancer using a structured approach while ensuring interoperability. Figure 1 provides a visual overview of the processes used.

Figure 1. Workflow for I-CDM implementation in lung cancer research. CDM: Common Data Model; DICOM: Digital Imaging and Communications in Medicine; I-CDM: Imaging Common Data Model; OMOP: Observational Medical Outcomes Partnership.



DICOM Header Analysis

We selected preliminary DICOM header tags that are universally applicable across a spectrum of modalities through a systematic procedure. To ensure the relevance and appropriateness of these tags, we sought consultation with radiology specialists [14-16]. While the DICOM standard provides a framework, it does not enforce a uniform data format. This lack of uniformity has led to variations in data formats across medical institutions and modalities. Accordingly, we validated the extractable data formats, ensuring that only data from nonempty DICOM header tags are extracted to maintain DQ and relevance. For data values where no standard concepts were unavailable, we introduced new custom concepts that are used consistently and comprehensively within the Observational Health Data Sciences and Informatics (OHDSI) framework. Concurrently, we examined the FHIR standard's ImagingStudy profile of radiological image data to determine the appropriate table and column names [17]. This approach was adopted to ensure that the selected header tags provided comprehensive coverage with respect to established imaging information standards.

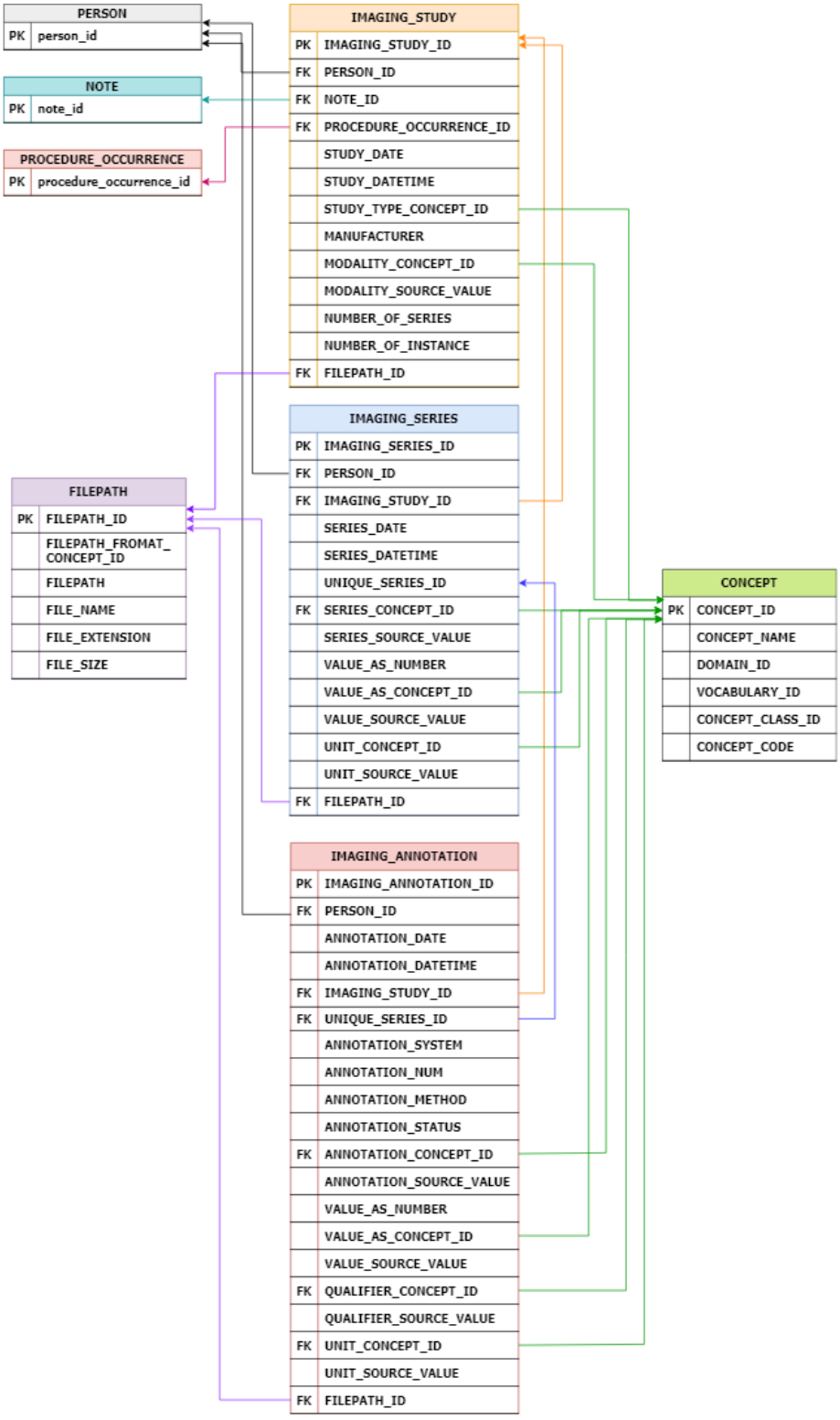
I-CDM Table Modeling

Our database architecture was designed to consolidate a variety of medical imaging data types, including DICOM header tags, study and series-level preprocessing information, and annotation information through labeling. The organization of IMAGING_STUDY, IMAGING_SERIES, IMAGING_ANNOTATION, and FILEPATH tables ensured structured and accessible data collection. The IMAGING_STUDY table refers to the CDM PROCEDURE_OCCURRENCE table associated with the corresponding imaging-order information.

To understand the significance of series-level analysis in medical imaging with reference to the ImagingStudy profile of the FHIR standard, we modeled the IMAGING_SERIES table [18-20] to house values derived from DICOM header tags pertinent to the series. Considering the potential variability in series-specific details owing to distinct imaging equipment or research requirements, we adopted the EAV format. The EAV model provides a data representation framework for the scalable and flexible storage of entities, in which the number and types of attributes (properties and parameters) can vary [21,22]. Using EAV, each attribute-value pair is stored as a separate record, making it easier to populate new data rows without altering the foundational database schema. This model facilitates data expansion without necessitating changes to the foundational table structure.

To observe the emergence and ubiquity of automated labeling tools in radiology, we designed an IMAGING_ANNOTATION table [23-26] structured to retain minimal metadata originating from the tools. Similar to the IMAGING_SERIES table, we used an EAV approach to promote the extension of the metadata. Finally, in response to the importance of the file size in image-oriented research and artificial intelligence implementations, we designed a FILEPATH table. This database captured fundamental attributes, such as file size, location, and specific format. Figure 2 shows the diagram constructed according to the I-CDM table definition. In Multimedia Appendix 1, we provide detailed definitions for each column in the IMAGING_STUDY, IMAGING_SERIES, IMAGING_ANNOTATION, and FILEPATH tables for I-CDM. Multimedia Appendix 1 elucidates the data format and captures broadly the attributes of the I-CDM framework. In addition, it specifies whether a column is mandatory, ensuring comprehensive documentation and consistency across data sets.

Figure 2. Diagram of Imaging Common Data Model workflow and attributes.



We analyzed and mapped the categorically constructed columns to the corresponding CDM concept IDs where feasible. The modality information was mapped to the concept IDs of the Procedure class in the CDM specifically related to the imaging equipment. This mapping enabled cross-validation using the method attributes of the linked procedure. The body part category was also aligned with the CDM concept, resulting in terms like “chest” being mapped to the Procedure class as “chest imaging.” For the *IMAGING_ANNOTATION* table, attributes such as the labeling plane and area were harmoniously mapped to the standard CDM concept IDs. In instances where mapping to standard CDM codes proved challenging, custom concepts were designated without limiting the classes and domains. Here, the original data extracted from the images were consistently included in the source-value column to ensure data fidelity. Additionally, in anticipation of RadLex potentially being adopted as a standard vocabulary within the OHDSI framework, we have also suggested additional RadLex mappings for our

data set. RadLex, developed by the Radiological Society of North America, is a comprehensive terminology system designed to standardize the names of radiological diagnoses, findings, and procedures. It encompasses a wide range of terms used in medical imaging, making it an invaluable resource for enhancing the comprehensiveness of medical imaging vocabularies within our data set. Multimedia Appendix 2 provides a detailed map of the standard terminology used in our I-CDM with their corresponding OMOP CDM concept IDs.

We expanded our methodological framework by developing a Python-based tool, publicly available on GitHub, for automatic conversion and integration of DICOM files into our I-CDM [27]. This tool was designed to work in conjunction with PostgreSQL to effectively create and populate essential tables such as IMAGING_STUDY, IMAGING_SERIES, and FILEPATH directly from specified DICOM file directories. Notably, the tool includes an algorithm to map the extracted DICOM header data systematically to the corresponding CDM concept IDs. This functionality ensures that the medical image data are not only accurately integrated into the I-CDM but also aligned with the standardized terminologies and classifications of the OMOP CDM. For practical applications, we chose the NSCLC-Radiomics open data set from The Cancer Imaging Archive. The NSCLC-Radiomics data set was used solely and only for the purpose of testing our DICOM file–to-PostgreSQL conversion tool, confirming its functionality with generic DICOM files, and providing a publicly shareable example of the processed output. After reading DICOM files from the NSCLC-Radiomics data set, our tool methodically constructs I-CDM tables within PostgreSQL, thereby streamlining the data integration process.

Data Preprocessing for I-CDM Table Construction

Before data preprocessing, all personal identifiers were removed to maintain patient confidentiality and protect personal information. Our first step was to ID and categorize the images into a series, ensuring that each series-specific folder contained only pertinent images, thereby maintaining a hierarchical directory structure. The Series Description in a DICOM header often comprises terms and abbreviations that describe the image characteristics. We performed a detailed analysis of the Series Descriptions of selected images to discern imaging attributes, such as the image plane, the presence of enhanced contrast, and designations such as low-dose, T1, or T2, among others. Based on a combination of these criteria, we designed rule-based naming conventions for folders, aiming for descriptive and meaningful names. Furthermore, we extracted information on the presence of “Black Blood” imaging in magnetic resonance imaging (MRI) scans using DICOM header data, which assisted in preparing data for the construction of the imaging series table.

To construct the imaging annotation table, 2 radiology experts identified and labeled the lesions on the chest computed tomography (CT) and brain MRI scans. Subsequently, we extracted metadata related to the labeled regions, such as area dimensions and characteristics.

Validation of I-CDM for Lung Cancer Studies

Overview

This study sought to define a research cohort comprising patients aged 18 years and older with primary lung cancer and structure the metadata of all chest x-ray, chest CT, and brain MRI images using I-CDM. Using the structured data, we aimed to elucidate the unique and major characteristics of patients with lung cancer by analyzing various scenarios.

Scenario 1: Association of Hypertension on Imaging Frequency in Patients With Epidermal Growth Factor Receptor Mutation-Positive Lung Cancer Receiving Osimertinib

We investigated the association of hypertension on imaging frequency in patients with lung cancer who were prescribed osimertinib, an epidermal growth factor receptor tyrosine kinase inhibitor. By comparing the frequency of CT imaging between groups with and without hypertension, this study aimed to determine whether the presence of hypertension affects imaging frequency in patients undergoing osimertinib treatment.

Scenario 2: Correlation Between Ground-Glass Nodules and Solid Tumor Volume in Lung Cancer

Using annotated data from chest CT scans to compare tumor volumes in patients with lung cancer who have ground-glass nodules (GGNs) with those who have solid nodules, we aimed to explore the relationship between ground-glass opacity nodules and tumor volume in lung cancer.

Scenario 3: Use of Low-Dose CT in Diagnostic Imaging for Lung Cancer

This scenario investigates the number of CT series, each consisting of more than 150 image instances, in patients who have undergone low-dose CT for lung cancer diagnosis. This study aimed to evaluate the adequacy of low-dose CT for providing diagnostic information while minimizing radiation exposure in patients.

Scenario 4: Number of Enhanced T1-Weighted MRI Images With a Slice Thickness of <1 mm in Patients With Lung Cancer Diagnosed at Younger Than 60 Years

This scenario targets patients with lung cancer diagnosed at younger than 60 years of age and involves quantifying the number of enhanced T1-weighted MRI images with a slice thickness of <1 mm. The collected data highlighted the volume of images available for subsequent annotation, and facilitated an in-depth radiological analysis of patient demographics.

Ethical Considerations

This study was approved by the Institutional Review Board of Seoul National University Bundang Hospital (approval B-2202-738-004; date: April 14, 2022). All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Statistical Analysis

Statistical analyses were performed using R software (version 4.2.2; The R Foundation for Statistical Computing). Descriptive statistics were used to summarize the data, including the calculation of means for continuous variables and frequency counts for categorical variables. The data were categorized as necessary to facilitate further analysis.

Results

Conversion of DICOM Data to the OMOP CDM Format: Realization and Integration in the I-CDM Framework

To systematically organize and efficiently manage the extensive collection of imaging data for the cohort of patients with lung cancer diagnosed between 2010 and 2017, sourced at Seoul National University Bundang Hospital, we structured the I-CDM into 4 basic tables: IMAGING_STUDY, IMAGING_SERIES, IMAGING_ANNOTATION, and FILEPATH (Table 1). The data set included imaging data from follow-ups in 2003-2021.

The I-CDM categorized 282,098 IMAGING_STUDY *records*, which were systematically linked to OMOP. This link provides an extensive overview of patient imaging trajectories and clinical data. The database contains 5,674,425 image series encompassing 47,381,027 individual image instances. Of the 282,098 records in the data set, 282,028 records contained information across various modalities, each containing a corresponding "number of series and instance" columns within the database.

Table 1. Imaging Common Data Model (I-CDM) data summary for lung cancer cohort (number of tables and records for data from 2003 to 2021 for patients diagnosed with lung cancer from 2010 to 2017).

I-CDM table or column name	Record count, n	Records with data, n (%)	Unique values
IMAGING_STUDY	282,098		
Modality	282,098	282,028 (99.9)	3
Manufacturer	282,098	281,940 (99.9)	265
Number of series	282,098	282,098 (100)	35
Number of instance	282,098	282,098 (100)	1825
IMAGING_SERIES	5,674,425		
Body part examined	382,517	382,517 (100)	25
Laterality	85,118	85,118 (100)	3
Slice thickness	411,351	411,351 (100)	1099
Series description	654,247	635,208 (100)	12,535
Window center	685,526	685,526 (100)	29,815
Window width	685,848	685,848 (100)	31,393
Patient position	458,770	444,770 (100)	11
Columns	717,154	717,154 (100)	2020
Rows	717,154	717,154 (100)	2066
Number of instance	717,169	717,169 (100)	626
BB[a]/non-BB	12,943	12,943 (100)	2
IMAGING_ANNOTATION	48,536		
Annotation system	48,536	48,536 (100)	5
Annotation text	3013	2689 (100)	69
Volume	11,153	11,153 (100)	3298
Long axis	31,353	31,333 (100)	159
Surface	3009	3009 (100)	2492
FILEPATH	1,000,361		
File path	1,000,361	1,000,361 (100)	998,844
File size	1,000,361	1,000,361 (100)	681,026

[a]BB: blood-brain barrier contrast.

The IMAGING_SERIES data represents a testament to the scale and complexity of the data set, with the 5,674,425 series illustrating the vast range of radiological examinations included in this framework. A total of 382,517 records detailing the

examined body parts underscored the targeted nature of radiological diagnostics. The data set was also characterized by an array of parameters, with 685,526 records for window centers and 685,848 for window widths. The structural resolution was meticulously captured with 717,154 data points for both the columns and rows, reflecting the intricate images. Each series was contextualized with descriptions recording the purpose and context of the imaging sequence in 654,247 data records. In total, 12,943 images were classified into blood-brain barrier contrast (BB)/non-BB categories to indicate the presence or absence of BB, highlighting the usefulness of specialized imaging sequences for detailed neurovascular assessments.

Table 2 displays the distribution of DICOM data across different modalities for patients with lung cancer, indicating the number of studies and series for each modality. X-ray examinations usually consist of a single series, whereas MRI and CT scans frequently include multiple series per study to accommodate a variety of imaging sequences. This highlights the detailed and complex nature of lung cancer diagnosis and monitoring. The structured categorization within the IMAGING_SERIES table using an EAV model, where the MRI data comprise 11 categories, including the BB/non-BB distinction. This implies a detailed classification of the MRI data, unlike the x-ray data, which are classified into 7 categories representing the variability of the desired parameters for each imaging modality. For x-ray images (n=1,410,844), the features are distributed as follows: number of instances (n=201,755, 14.3%), columns (n=201,755, 14.3%), rows (n=201,755, 14.3%), window width (n=200,938, 14.2%), window center (n=201,755, 14.3%), series description (n=201,755, 14.3%), patient position (n=200,938, 14.2%), slice thickness (n=200,938, 14.2%), body part examined (n=201,755, 14.3%), laterality (n=201,755, 14.3%), and BB_NonBB (n=201,755, 14.3%).

CT images (n=2,982,176) show the following feature distribution: number of instances (n=360,845, 12.1%), columns (n=360,845, 12.1%), rows (n=360,845, 12.1%), window width (n=360,845, 12.1%), window center (n=360,845, 12.1%), series description (n=360,845, 12.1%), patient position (n=360,845, 12.1%), slice thickness (n=360,845, 12.1%), body part examined (n=360,845, 12.1%), laterality (n=286,281, 9.6%), and BB_NonBB (n=116,305, 3.9%).

For MRI images (n=1,281,432), the distribution is: number of instances (n=140,958, 11%), columns (n=140,958, 11%), rows (n=140,958, 11%), window width (n=138,395, 10.8%), window center (n=138,395, 10.8%), series description (n=139,686, 10.9%), patient position (n=142,235, 11.1%), slice thickness (n=142,235, 11.1%), body part examined (n=142,235, 11.1%), laterality (n=80,733, 6.3%), and BB_NonBB (n=91,984, 7.1%).

The IMAGING_ANNOTATION table had 48,536 annotations predominantly sourced from CT and MRI scans. These annotations offer a detailed exploration of the examined regions, with volumetric and long-axis measurements documented. This granular level of detail is critical for the precise characterization and monitoring of diseases supported by a comprehensive understanding of the annotated regions. The FILETATH records within the I-CDM table (totaling 1,000,361) served as a bridge between the CDM tables and actual image file paths, spanning approximately 9.6 terabytes of image data. This illustrates not only the substantial volume of image data but also the expansive nature of our image repository within the I-CDM framework. In addition, DICOM folders were organized in a series, and common imaging characteristics were identified through series descriptions to assign meaningful folder names, further streamlining the data structure for efficient management and retrieval.

Table 2. Count data based on modality and type.

Modality	Study count	Series count
X-ray	201,569	201,770
CT[a]	65,923	373,116
MRI[b]	14,590	142,254

[a]CT: computed tomography.

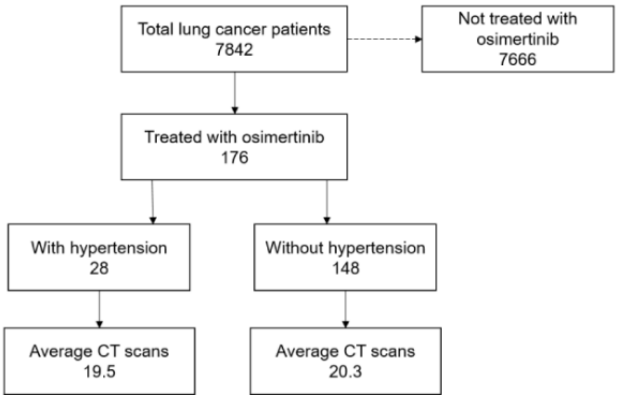
[b]MRI: magnetic resonance imaging.

Validation of I-CDM Scenarios for Enhanced Imaging and Treatment Classification in Patients With Lung Cancer

Scenario 1: Hypertension and Imaging Frequency in Patients Treated With Osimertinib

Among the total cohort of 7842 patients with lung cancer, 176 (2.24% of the total) prescribed osimertinib were diagnosed with hypertension. In the osimertinib arm, 28 (0.36%) patients had hypertension. The average number of CT scans in patients with and without hypertension was 19.5 and 20.3, respectively, indicating that hypertension did not significantly affect the frequency of CT imaging in patients with lung cancer receiving osimertinib treatment. Figure 3 shows this comparison of CT scan frequency over hypertension status among patients treated with osimertinib.

Figure 3. Diagram of CT scan frequency comparison according to hypertension status among osimertinib-treated patients with lung cancer. CT: computed tomography.



Scenario 2: Nodule Characterization and Volume Measurement in CT Imaging

We evaluated 1947 annotated CT scans from 1929 patients and observed that a significant number of GGNs contained solid components, necessitating labeling of both components within the GGNs. Our comparative analysis focused on the mean volume of solid nodules within GGNs compared with those without GGNs. In total, 673 GGNs were identified in 626 patients, of which 649 were classified as having solid nodules. The average volume of the GGN was 8135.616 mm^3, whereas the volume of solid nodules within the GGN was 2578.006 mm^3. In contrast, 1343 solid nodules without GGN were found in 1319 patients, with an average volume of 34,712.58 mm^3. This corroborates the findings of previous studies indicating that solid nodules, especially those not associated with GGNs, tended to be more abundant [28]. Our results provide additional evidence supporting these observations on the nodal nature of lung pathologies.

Scenario 3: Use of Low-Dose CT and Instance Range

Among the 63,446 CT studies conducted in the cohort of patients with lung cancer, which included 2,725,899 series, 48,587 were identified as low-dose imaging studies. Of these, 41,336 included >150 instances. The instance count in the 3 low-dose CT images varied significantly, with the smallest and largest series consisting of three and 633 instances, respectively. This highlights the need for low-dose imaging to capture extensive data while minimizing radiation exposure.

Scenario 4: MRI Imaging With 1-mm Thick Slice and T1 Enhancement

In Scenario 4, which focused on MRIs of 1-mm thick slices under T1 enhancement, our analysis of 137,566 MRI series identified 31,851 series using T1-weighting at the specified slice thickness, 5235 of which were associated with patients aged ≤60 years. This scenario allows the preemptive examination of images to be labeled, integrating image patterns and nodule characteristics with clinical data. This approach not only facilitates the identification of the scale of target images to be annotated but also enables precise quantification of the images that meet the specified criteria. Table 3 shows the categorization of the MRI series using I-CDM, providing a visual summary of the data refined by slice thickness and T1 enhancement and further filtered to include patients aged ≤60 years within the studied patient cohort.

Table 3. Magnetic resonance imaging series analysis using Imaging Common Data Model.

MRI[a] series type	Count
Total MRI series	137,566
T1 weighted	74,692
Contrast enhanced	51,582
Slice thickness ≤1 mm	31,851
Age of 60 years or younger	5235

[a]MRI: magnetic resonance imaging.

I-CDM Data Quality Check

We ensured DQ based on a set of 44 comprehensive DQ rules (outlined in Multimedia Appendix 3), focusing on the Radiation CDM quality assurance framework. These rules encompassed a broad spectrum of checks, including the evaluation of DICOM series and instance counts, data type consistency, and the accuracy of the linkage between the data set and existing clinical tables within the CDM. All data entries successfully met these criteria, indicating compliance with quality standards. Among these rules, those pertaining to interdata relationships and outlier detection were instrumental in validating the integrity of the data set. A selection of these quality checks and their outcomes are presented in Table 4, highlighting the importance of ensuring data quality and integrity within I-CDM.

Table 4. Selected data quality assurance rules and outlier analysis results from Multimedia Appendix 3.

CDM_TABLE, CONTCEPT_NAME, and check description	Threshold	Result (%)	Error (n)
IMAGING_STUDY			
NUMBER_OF_SERIES			
The NUMBER_OF_SERIES must be equal to the number of series in the IMAGING_SERIES table with the same IMAGING_STUDY_ID. This ensures that the number of series recorded in the IMAGING_STUDY matches the actual series entries in the related table	At least 95% match	PASS (99.9)	202
NUMBER_OF_INSTANCE			
The NUMBER_OF_INSTANCE must equal the sum of VALUE_AS_NUMBER for entries in the IMAGING_SERIES table where SERIES_CONCEPT_ID equals NUMBER_OF_INSTANCE, under the condition that they are mapped between the 2 tables. This is to verify that the number of instances (images) reported in the IMAGING_STUDY corresponds to the aggregated count of instances from the series data	At least 95% match	PASS (99.9)	109
NUMBER_OF_SERIES, NUMBER_OF_INSTANCE			
The presence of a Rule of NUMBER_OF_SERIES necessitates the presence of a Rule of NUMBER_OF_INSTANCE.	At least 95% match	PASS (99.9)	1
IMAGING_SERIES			
SERIES_CONCEPT_ID = SliceThickness			
VALUE_AS_NUMBER must exist and be a numeric value for at least 99% of the records	At least 99% of records must have a nonmissing	PASS (99.8)	496
SERIES_CONCEPT_ID = Rows			
VALUE_AS_NUMBER must exist and be a numeric value for at least 99% of the records	At least 99% of records must have a nonmissing	PASS (100)	100
Outliers, defined as values beyond the 1st and 99th percentiles, should be reviewed	Outliers should be under 5%	PASS (1.2)	8952
SERIES_CONCEPT_ID = Columns			
VALUE_AS_NUMBER must exist and be a numeric value for at least 99% of the records	At least 99% of records must have a nonmissing	PASS (100)	—[a]
Outliers, defined as values beyond the 1st and 99th percentiles, should be reviewed	Outliers should be under 5%	PASS (1.0)	7251
SERIES_CONCEPT_ID = BB/non-BB			
Values must be exclusively “Positive” or “Negative”, ensuring they represent these specific states without including the concept IDs 45884084 and 45878583	100% of records must have as one of the specified valid IDs	PASS (100)	—
IMAGING_ANNOTATION			
ANNOTATION_ CONCEPT_ID = Long axis			
VALUE_AS_NUMBER must exist and be a numeric value	No missing values for VALUE_AS_NUMBER	PASS (100)	—
Outliers, defined as values beyond the 1st and 99th percentiles, should be identified and reviewed to ensure they accurately reflect the intended measurements.	Outliers should be under 5%	PASS (0.2)	77
ANNOTATION_ CONCEPT_ID = Volume			
VALUE_AS_NUMBER must exist and be a numeric value	100% of records must be numeric and nonnull	PASS (100)	—
ANNOTATION_ CONCEPT_ID = annotation_text			
VALUE_SOURCE_VALUE must contain a nonempty text value	100% of records must be numeric and nonnull	PASS (100)	—
ANNOTATION_ CONCEPT_ID = surface area			
VALUE_AS_NUMBER must exist and be a numeric value	100% of records must be numeric and nonnull	PASS (100)	—

[a]Not applicable.

Discussion

Principal Results

This study proposes a method for integrating clinical and imaging data using I-CDM. By converting DICOM data into the OMOP CDM format and integrating it into the I-CDM framework, we implemented a systematic approach to efficiently manage medical imaging data. This approach enabled the connection and analysis of clinical and imaging data in different contexts. Additionally, detailed schema overviews for the use cases of integrating I-CDM can be found in Multimedia Appendix 4, and a comparison between MI-CDM and I-CDM is provided in Multimedia Appendix 5.

Limitations

The limitation of this study is its lack of consideration for the resources required for processing DICOM images and integrating annotation information. To customize and add data according to researcher needs using the EAV model, comprehensive knowledge and expertise on DICOM standards and tags are required [29]. Furthermore, integrating image annotation data within the I-CDM framework not only demands sufficient resources [30-32], but also requires advanced data management strategies to expand the integration and harmonization of data sets from various imaging modalities beyond chest CT, x-ray, and brain MRI [33,34]. Moreover, focusing exclusively on a cohort of patients with lung cancer narrowed the scope of the study. Additionally, in our study, the scenarios were designed to validate the functionality of the proposed model using actual medical data. These scenarios were deliberately simplified to ensure effective management within the capabilities of the implemented I-CDM framework. Future research will benefit from the incorporation of expanded annotation data, enabling more complex analyses, such as longitudinal comparisons of tumor sizes pre and posttreatment in individual patients.

Comparison With Prior Work

In this study, diverging from previous Radiology-CDM research, we refined the integration of imaging examination data by linking them with the PROCEDURE_OCCURRENCE table, which enables a more efficient analysis through improved data connectivity. Moreover, unlike previous research that relied on RadLex for standard terminology, this study directly mapped DICOM terms to OMOP CDM standard terminologies. This direct mapping simplifies the process and enables the use of custom codes, thus facilitating a deeper analytical integration of clinical and imaging data. And this study takes a distinct approach compared with recent I-CDM studies [35,36]. Our method enhances the analytical scope by facilitating the storage and management of annotation information. This ensures that imaging-related data, including annotations, can be comprehensively managed within the I-CDM framework. By using the EAV model, our study introduced flexibility in managing various data types and structures, rendering our approach adaptable to evolving research needs and data characteristics. Consequently, it exhibits good flexibility and adaptability, especially in research requiring integrated analysis of clinical and imaging data. Considering the file sizes associated with imaging data, effective file management is essential. Our study used the FILEPATH table to connect I-CDM with the original imaging data, including file extension and size information, to ensure quick access to file details and facilitate efficient management.

Scalability and Applicability of the I-CDM

The lung cancer cohort in this study was initially used to validate the functionality of the proposed I-CDM tables using actual medical image data. In future studies, the model is not only adaptable to lung cancer but also designed to accommodate a wider spectrum of medical conditions, including various tumor types and cardiovascular diseases. By leveraging OMOP CDM's standard vocabulary for "modality_concept_id" and "body part examined" ("value_as_concept_id"), the model can be broadly adapted to accommodate various diseases or different settings beyond lung cancer. This adaptability ensures that any additional data items users might require can be seamlessly integrated by aligning with OMOP CDM standard vocabulary, underlining the framework's potential for broad application across diverse medical data and settings. Furthermore, while RadLex is extensively used in medical imaging vocabularies, it is not yet included as a standard vocabulary in the OHDSI framework. Even if RadLex were incorporated, it would not cover all concepts related to imaging. Therefore, we had to consider various vocabularies to ensure comprehensive coverage. Recognizing this, we aimed to build the I-CDM by maximizing the use of existing OHDSI vocabularies according to OHDSI principles, rather than proposing new vocabularies. We have proactively suggested mapping terms compatible with RadLex within our study wherever possible. In the new scenario, the principle for term selection involves mapping the standard concept to the granularity level of the source data. This is achieved by selecting the term from the standard vocabulary that most accurately represents the clinical meaning. In addition to the features we have currently mapped, our study focused on lung cancer, but for other diseases, there are important concepts in imaging that should be considered. For instance:

- Ultrasound image tags: Commonly used tags include "Transducer Frequency," "Gain," and "Depth of Field," which are critical for analyzing the quality and characteristics of ultrasound images.
- Spine x-ray tags: Relevant tags such as "KVP" (Kilovoltage Peak), "Exposure Time," and "Focal Spot Size" are essential for understanding the technical parameters that affect image quality.
- Body part imaging concepts: Terms like "Entire Thorax," "Entire Liver," and "Entire Pelvis" are crucial for precisely describing the anatomical region being imaged, which can vary significantly depending on the disease or condition being studied.

These examples ensure that the I-CDM framework is adaptable and capable of integrating a wide range of imaging data characteristics and supporting diverse medical conditions and research scenarios.

Conclusions

This study implemented a systematic approach for the efficient management of medical imaging data, achieving integration of clinical and imaging data through the development of the I-CDM framework and the conversion of DICOM data into the OMOP CDM format. Future efforts should strive to broaden the application of the I-CDM framework to encompass various disease populations and include diverse imaging techniques for different body parts, such as abdominal CT, spine MRI, and liver MRI, thereby enhancing its applicability. Expanding its scope to incorporate these imaging modalities is crucial for conducting more comprehensive investigations into the utility of merging clinical and imaging data across different health conditions.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant HI22C0471).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Imaging Common Data Model table.
[DOCX File , 18 KB - medinform_v12i1e59187_app1.docx]

Multimedia Appendix 2
Imaging Common Data Model mapping Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) concept ID.
[DOCX File , 22 KB - medinform_v12i1e59187_app2.docx]

Multimedia Appendix 3
Data quality check rule and result.
[DOCX File , 21 KB - medinform_v12i1e59187_app3.docx]

Multimedia Appendix 4
Use case for integrating Imagining Common Data Model: detailed schema overview.
[PNG File , 198 KB - medinform_v12i1e59187_app4.png]

Multimedia Appendix 5
Comparison of Medical Imaging Common Data Model and Imaging Common Data Model.
[DOCX File , 17 KB - medinform_v12i1e59187_app5.docx]

References

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29 [FREE Full text] [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
3. Rho MJ, Kim SR, Park SH, Jang KS, Park BJ, Hong JY, et al. Common data model for decision support system of adverse drug reaction to extract knowledge from multi-center database. Inf Technol Manag 2015 Jul 3;17(1):57-66. [doi: 10.1007/s10799-015-0240-6]
4. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
5. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
6. Lim JE, Kim HM, Kim JH, Baek HS, Han MY. Association between dyslipidemia and asthma in children: a systematic review and multicenter cohort study using a common data model. Clin Exp Pediatr 2023 Aug;66(8):357-365 [FREE Full text] [doi: 10.3345/cep.2023.00290] [Medline: 37321588]

7. Park CH, You SC, Jeon H, Jeong CW, Choi JW, Park RW. Development and validation of the radiology common data model (R-CDM) for the international standardization of medical imaging data. Yonsei Med J 2022 Jan;63(Suppl):S74-S83 [FREE Full text] [doi: 10.3349/ymj.2022.63.S74] [Medline: 35040608]
8. Kim TH, Noh SH, Kim YR, Lee CS, Kim JE, Jeong CW, et al. Development and validation of a management system and dataset quality assessment tool for the Radiology Common Data Model (R_CDM): a case study in liver disease. Int J Med Inform 2022;162:104759 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104759] [Medline: 35390589]
9. Malik B, Jeon K, Alkasab T, Mallol P, You SC, Nagy P. Development of the medical imaging extension for OMOP-CDM. Observational Health Data Sciences and Informatics. 2022. URL: https://www.ohdsi.org/2022showcase-26/ [accessed 2024-06-28]
10. Praeta J, Scherer M, Smeets D. Application of the R-CDM extension to capture metadata and features extracted from quantitative brain MRI and CT data. 2023 Presented at: The Fourth European OHDSI Symposium; July 1-3, 2023; Rotterdam, The Netherlands.
11. Kalokyri V, Kondylakis H, Sfakianakis S, Nikiforaki K, Karatzanis I, Mazzetti S, et al. MI-common data model: extending observational medical outcomes partnership-common data model (OMOP-CDM) for registering medical imaging metadata and subsequent curation processes. JCO Clin Cancer Inform 2023 Sep;7:e2300101 [FREE Full text] [doi: 10.1200/CCI.23.00101] [Medline: 38061012]
12. Tsui DCC, Camidge DR, Rusthoven CG. Managing central nervous system spread of lung cancer: the state of the art. J Clin Oncol 2022;40(6):642-660. [doi: 10.1200/JCO.21.01715] [Medline: 34985937]
13. Eaton K. Lung Cancer: Translational and Emerging Therapies. New York, NY: CRC Press; 2007:-278.
14. Mildenberger P, Eichelberg M, Martin E. Introduction to the DICOM standard. Eur Radiol 2002;12(4):920-927. [doi: 10.1007/s003300101100] [Medline: 11960249]
15. Bidgood WD, Horii SC, Prior FW, Van Syckle DE. Understanding and using DICOM, the data interchange standard for biomedical imaging. J Am Med Inform Assoc 1997;4(3):199-212 [FREE Full text] [doi: 10.1136/jamia.1997.0040199] [Medline: 9147339]
16. Haripriya P, Porkodi R. A survey paper on data mining techniques and challenges in distributed DICOM. Int J Adv Res Comput Commun Eng 2016:741-747 [FREE Full text] [doi: 10.17148/IJARCCE.2016.53179]
17. Kamel P, Nagy P. Patient-centered radiology with FHIR: an introduction to the use of FHIR to offer radiology a clinically integrated platform. J Digit Imaging 2018;31(3):327-333 [FREE Full text] [doi: 10.1007/s10278-018-0087-6] [Medline: 29725963]
18. Tournier J, Smith R, Raffelt D, Tabbara R, Dhollander T, Pietsch M, et al. MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. Neuroimage 2019;202:116137. [doi: 10.1016/j.neuroimage.2019.116137] [Medline: 31473352]
19. Aiello M, Esposito G, Pagliari G, Borrelli P, Brancato V, Salvatore M. How does DICOM support big data management? investigating its use in medical imaging community. Insights Imaging 2021;12(1):164 [FREE Full text] [doi: 10.1186/s13244-021-01081-8] [Medline: 34748101]
20. Onken M, Riesmeier J, Bennett A. Digital imaging and communications in medicine. In: Biomedical Image Processing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010:427-454.
21. Lelong R, Soualmia L, Dahamna B, Griffon N, Darmoni SJ. Querying EHRs with a semantic and entity-oriented query language. Stud Health Technol Inform 2017;235:121-125 [FREE Full text] [doi: 10.3233/978-1-61499-753-5-121] [Medline: 28423767]
22. Nadkarni P. QAV: querying entity-attribute-value metadata in a biomedical database. Comput Methods Programs Biomed 1997 Jun;53(2):93-103 [FREE Full text] [doi: 10.1016/s0169-2607(97)01815-4] [Medline: 9186046]
23. Morozov S, Gombolevskiy V, Elizarov A, Gusev M, Novik V, Prokudaylo S, et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT scans. Comput Methods Programs Biomed 2021;206:106111. [doi: 10.1016/j.cmpb.2021.106111] [Medline: 33957377]
24. Diaz-Pinto A, Alle S, Nath V, Tang Y, Ihsani A, Asad M, et al. MONAI label: a framework for AI-assisted interactive labeling of 3D medical images. Med Image Anal 2024;95:103207. [doi: 10.1016/j.media.2024.103207] [Medline: 38776843]
25. Philbrick K, Weston A, Akkus Z, Kline TL, Korfiatis P, Sakinis T, et al. RIL-Contour: a medical imaging dataset annotation tool for and with deep learning. J Digit Imaging 2019 Aug;32(4):571-581 [FREE Full text] [doi: 10.1007/s10278-019-00232-0] [Medline: 31089974]
26. Lösel PD, van de Kamp T, Jayme A, Ershov A, Faragó T, Pichler O, et al. Introducing Biomedisa as an open-source online platform for biomedical image segmentation. Nat Commun 2020 Nov 04;11(1):5577. [doi: 10.1038/s41467-020-19303-w] [Medline: 33149150]
27. Radiology Common Dada Model (CDM) framework on OMOP CDM. GitHub. URL: https://github.com/HIRC-SNUBH/ImagingCDM [accessed 2024-06-12]
28. Chu Z, Zhang Y, Li WJ, Li Q, Zheng YN, Lv FJ. Primary solid lung cancerous nodules with different sizes: computed tomography features and their variations. BMC Cancer 2019;19(1):1060 [FREE Full text] [doi: 10.1186/s12885-019-6274-0] [Medline: 31699047]

29. Khan O, Lim Choi Keung SN, Zhao L, Arvanitis TN. A hybrid EAV-relational model for consistent and scalable capture of clinical research data. Stud Health Technol Inform 2014;202:32-35. [Medline: 25000008]
30. Gu Y, Shen M, Yang J, Yang G. Reliable label-efficient learning for biomedical image recognition. IEEE Trans Biomed Eng 2019;66(9):2423-2432. [doi: 10.1109/tbme.2018.2889915] [Medline: 30596566]
31. Wu Y, Zhou Z, Wu W. OneSeg: self-learning and one-shot learning based single-slice annotation for 3D medical image segmentation. arXiv Preprint posted online on September 24, 2023. [doi: 10.48550/arXiv.2309.13671]
32. Dimitrovski I, Kocev D, Loskovska S, Džeroski S. Hierarchical annotation of medical images. Pattern Recognit 2011 Oct;44(10-11):2436-2449. [doi: 10.1016/j.patcog.2011.03.026]
33. Martínez-García M, Hernández-Lemus E. Data integration challenges for machine learning in precision medicine. Front Med (Lausanne) 2021;8:784455 [FREE Full text] [doi: 10.3389/fmed.2021.784455] [Medline: 35145977]
34. Parciak M, Suhr M, Schmidt C, Bönisch C, Löhnhardt B, Kesztyüs D, et al. FAIRness through automation: development of an automated medical data integration infrastructure for FAIR health data in a maximum care university hospital. BMC Med Inform Decis Mak 2023;23(1):94 [FREE Full text] [doi: 10.1186/s12911-023-02195-3] [Medline: 37189148]
35. Kalokyri V, Kondylakis H, Sfakianakis S, Nikiforaki K, Karatzanis I, Mazzetti S, et al. MI-common data model: extending observational medical outcomes partnership-common data model (OMOP-CDM) for registering medical imaging metadata and subsequent curation processes. JCO Clin Cancer Inform 2023;7:e2300101 [FREE Full text] [doi: 10.1200/CCI.23.00101] [Medline: 38061012]
36. Park W, Jeon K, Schmidt TS, Kondylakis H, Alkasab T, Dewey BE, et al. Development of medical imaging data standardization for imaging-based observational research: OMOP common data model extension. J Imaging Inform Med 2024;37(2):899-908. [doi: 10.1007/s10278-024-00982-6] [Medline: 38315345]

Abbreviations

BB: blood-brain barrier contrast
CDM: Common Data Model
CT: computed tomography
DICOM: Digital Imaging and Communications in Medicine
DQ: data quality
EAV: entity-attribute-value
FHIR: Fast Healthcare Interoperability Resources
GGN: ground-glass nodule
I-CDM: Imaging Common Data Model
MRI: magnetic resonance imaging
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership

Edited by C Lovis; submitted 04.04.24; peer-reviewed by T Karen, C Jeong, R Kolde; comments to author 25.04.24; revised version received 10.05.24; accepted 08.06.24; published 12.07.24.

Please cite as:
Ji H, Kim S, Sunwoo L, Jang S, Lee HY, Yoo S
Integrating Clinical Data and Medical Imaging in Lung Cancer: Feasibility Study Using the Observational Medical Outcomes Partnership Common Data Model Extension
JMIR Med Inform 2024;12:e59187
URL: https://medinform.jmir.org/2024/1/e59187
doi:10.2196/59187
PMID:

Original Paper

Uncovering Harmonization Potential in Health Care Data Through Iterative Refinement of Fast Healthcare Interoperability Resources Profiles Based on Retrospective Discrepancy Analysis: Case Study

Lorenz Rosenau[1], MSc; Paul Behrend[1], MSc; Joshua Wiedekopf[1], MSc; Julian Gruendner[2], PhD; Josef Ingenerf[1,3], Prof Dr

[1]IT Center for Clinical Research, University of Lübeck, Lübeck, Germany

[2]Chair for Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[3]Institute of Medical Informatics, University of Lübeck, Lübeck, Germany

Corresponding Author:
Lorenz Rosenau, MSc
IT Center for Clinical Research
University of Lübeck
Gebäude 64, 2.OG, Raum 05
Ratzeburger Allee 160
Lübeck, 23562
Germany
Phone: 49 451 3101 5636
Email: lorenz.rosenau@uni-luebeck.de

Abstract

Background: Cross-institutional interoperability between health care providers remains a recurring challenge worldwide. The German Medical Informatics Initiative, a collaboration of 37 university hospitals in Germany, aims to enable interoperability between partner sites by defining Fast Healthcare Interoperability Resources (FHIR) profiles for the cross-institutional exchange of health care data, the Core Data Set (CDS). The current CDS and its extension modules define elements representing patients' health care records. All university hospitals in Germany have made significant progress in providing routine data in a standardized format based on the CDS. In addition, the central research platform for health, the German Portal for Medical Research Data feasibility tool, allows medical researchers to query the available CDS data items across many participating hospitals.

Objective: In this study, we aimed to evaluate a novel approach of combining the current top-down generated FHIR profiles with the bottom-up generated knowledge gained by the analysis of respective instance data. This allowed us to derive options for iteratively refining FHIR profiles using the information obtained from a discrepancy analysis.

Methods: We developed an FHIR validation pipeline and opted to derive more restrictive profiles from the original CDS profiles. This decision was driven by the need to align more closely with the specific assumptions and requirements of the central feasibility platform's search ontology. While the original CDS profiles offer a generic framework adaptable for a broad spectrum of medical informatics use cases, they lack the specificity to model the nuanced criteria essential for medical researchers. A key example of this is the necessity to represent specific laboratory codings and values interdependencies accurately. The validation results allow us to identify discrepancies between the instance data at the clinical sites and the profiles specified by the feasibility platform and addressed in the future.

Results: A total of 20 university hospitals participated in this study. Historical factors, lack of harmonization, a wide range of source systems, and case sensitivity of coding are some of the causes for the discrepancies identified. While in our case study, Conditions, Procedures, and Medications have a high degree of uniformity in the coding of instance data due to legislative requirements for billing in Germany, we found that laboratory values pose a significant data harmonization challenge due to their interdependency between coding and value.

Conclusions: While the CDS achieves interoperability, different challenges for federated data access arise, requiring more specificity in the profiles to make assumptions on the instance data. We further argue that further harmonization of the instance data can significantly lower required retrospective harmonization efforts. We recognize that discrepancies cannot be resolved

solely at the clinical site; therefore, our findings have a wide range of implications and will require action on multiple levels and by various stakeholders.

(JMIR Med Inform 2024;12:e57005) doi:10.2196/57005

KEYWORDS

Health Level 7 Fast Healthcare Interoperability Resources; HL7 FHIR; FHIR profiles; interoperability; data harmonization; discrepancy analysis; data quality; cross-institutional data exchange; Medical Informatics Initiative; federated data access challenges

Introduction

Overview

Interoperability, an essential component of contemporary medical informatics, facilitates seamless communication and data exchange between various devices, applications, and health care systems. Semantic interoperability ensures machine interpretation of health care data and, thus, data exchange, integration, and reuse for optimized collaboration between distributed players in health care and medical research. Consequently, it is pivotal in amplifying the efficacy and impact of numerous other medical informatics technologies, advancing the field as a whole.

The Layered Fast Healthcare Interoperability Resources Profile Model: Facilitating Reuse and Interoperability

The Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) addresses syntactic and semantic interoperability [1], providing a standardized framework for data structures known as resources. These resources are essentially a set of attributes that represent specific health care–related concepts. For instance, the *Patient* resource in FHIR might include attributes such as name, administrative gender, birth date, address, and contact details. This standardization of data structure promotes syntactic interoperability. Semantic interoperability is achieved using bindings of attributes to value sets with codes from universally recognized coding systems such as *Logical Observation Identifiers Names and Codes* (LOINC) or *Systematized Nomenclature of Medicine—Clinical Terms* (SNOMED CT).

FHIR also introduces the concept of *Profiling* [2]. FHIR profiles constrain or modify the base FHIR resources to cater to specific use cases or regional requirements. They provide guidelines on how resources should be structured; which attributes should be excluded, mandatory, added, or repeated; what terminology should be used for coded elements; and how these elements should be interpreted. During the development of profiles, domain experts from the medical field are involved at all stages to ensure that the data models capture all the required data. This approach ensures the syntactic consistency of the data and its semantic interpretability. FHIR profiles can be based on other FHIR profiles to constrain them further, allowing a layered structure from the most permissive to the most constrained profile.

The Medical Informatics Initiative

The German government recognized the importance of interoperability in the health care sector and consequently initiated the Medical Informatics Initiative (MII) in 2015 [3]. With the overarching goal of making routine data available for research, the MII aims to digitally connect patient data generated during hospital stays across the country. Four consortia, namely, DIFUTURE [4], HiGHmed [5], Medical Informatics in Research and Care in University Medicine [MIRACUM] [6], and Smart Medical Information Technology for Health care [7], and a central coordination office received funding from the Federal Ministry of Education and Research to establish data integration centers (DICs) responsible for data exchange. With over €400 (US $427 million) million in current total funding, the Federal Ministry of Education and Research supports consortia, DICs, and cross-consortium use cases.

Acknowledging the various accomplishments within the MII until this point [8-10], in this paper, we will place particular emphasis on the Core Data Set (CDS), its implementation at the DICs, and its applicability to the "Aligning Biobanking and DIC Efficiently" [11] use case project, which among others, included federated feasibility studies integrated into the *Forschungsdatenportal für Gesundheit* (German Portal for Medical Research Data [FDPG]) [12].

CDS Profiles

One of the primary responsibilities of the MII is the development and implementation of a unified data model that is binding for all German university hospitals.

The result of this ongoing endeavor within the MII is the CDS, a set of specific FHIR profiles that the local sites have agreed upon collectively. These CDS profiles define the minimum data set that should be included in each DIC. The CDS is subdivided into basic and extension modules, where the basic modules encompass basic health care data such as patient-derived information, conditions, procedures, medication, and laboratory measures, and the extension modules reflect data from specific applications or specialist areas (such as intensive care or oncology) [13]. The CDS is sustainably, nationally coordinated, continuously updated, and adapted to meet changing requirements. The development of the CDS leverages tools such as ART-DECOR for data set modeling and Simplifier.net [14] for creating and publishing FHIR profiles [15].

Upon the successful development of the CDS, the obligation falls to each DIC to make its routine data available as FHIR instance data. DICs integrate and standardize routine health care data at each site, essentially based on extract, transfer, load (ETL) processes with frequent late mappings of proprietary data from source systems. While their operation and management involve complex processes, for the scope of this paper, this conceptualization is sufficient. The data are secured and

standardized through the DIC, promoting efficient and secure cross-institutional data sharing and collaboration.

FDPG: Facilitating Accessible Research

The FDPG is critical in making the data across the 34 sites accessible to medical researchers [12]. It provides the legal and procedural framework to access routine data sets across sites. Among other components, it provides the feasibility platform that gives users a user-friendly view of available data items and allows users to query them across connected sites directly. It aggregates available patient counts for a specific, user-defined search query. To do so, it uses a search ontology automatically generated from FHIR profiles [16]. Users can select concepts from the search ontology and restrict them as needed, for example, by applying comparator values for quantitative laboratory values. The resulting criteria can be combined using Boolean algebra to create complex feasibility queries.

Figure 1 demonstrates the search ontology's user-friendly abstraction for researchers unfamiliar with FHIR:

- *Medical coding:* criteria are based on standard codings referenced in the FHIR profiles (eg, C71 from *German modification of the International Statistical Classification of Diseases and Related Health Problems, 10th revision* [*ICD-10-GM*] for malignant neoplasm of the brain)
- *Value and attribute filters:* specific FHIR attributes such as the value of an observation are modeled as "value filters" to express the "is" relationship between the coding and the value (such as leukocyte counts between 4000 and 10,000/µL). Additional attributes can also be expressed (such as place of collection for specimen) and allow the further refinement of criteria beyond their existence.
- *Time-based filters:* researchers can furthermore apply temporal constraints ("after" a specific date).

All other FHIR attributes are not available to the user to ensure high usability.

The FDPG feasibility tool's primary aim is to make the data findable that are available across the MII's DICs, which are based on the CDS. In its current iteration, the feasibility tool offers a selected subset of CDS criteria. This design choice is driven by the different requirements to be met by the CDS and the search ontology. The CDS profiles are developed for various primary source systems across German university hospitals. Profile specifications are required to be broad in many cases due to the diversity and complexity of these systems. Therefore, generic profiles are preferred, with more detailed models being developed only when necessary for specific content or organizational reasons. This serves DICs well due to having access to the instance data. By contrast, due to its federated nature, the FDPG has no access to the instance data. A more nuanced approach beyond generic modulations, such as accurately modeling the relationship between laboratory concepts and their values instead of simply representing LOINC codes, is required to provide users with criteria beyond determining their presence.

The feasibility tool also navigates the complexity arising from various code systems. For instance, the diagnosis profile accommodates codings from *ICD-10-GM*, SNOMED CT, Alpha-ID, and Orphanet, which are crucial for rare disease research. However, this diversity poses a usability challenge, potentially overwhelming users unfamiliar with these systems. In addition, the overlap in concept expression between *ICD-10-GM* and SNOMED CT can create confusion: users may not realize the necessity of selecting a particular concept from a specific code system or potentially both, contingent on their use case. A significant factor in the FDPG feasibility portal's initial focus on legally mandated code systems is the absence of instance data for certain criteria (eg, as of March 2023, none of the 20 participating sites had SNOMED CT codings for diagnoses), as offering such criteria might frustrate researchers and discourage tool use. Guided by these assumptions, Table 1 provides an overview of the CDS modules, their codings, and their coverage in the FDPG feasibility portal.

The present scope of the FDPG feasibility portal is not fixed, allowing for potential future expansions.

Figure 1. Example of a feasibility query in the German Portal for Medical Research Data feasibility portal to find patients with a leukocyte count within a normal range, those with a malignant neoplasm of the brain, those with available tumor tissue specimen, and those with a computed tomography scan after January 1, 2020, who did not take doxorubicin after January 1, 2023. MII: Medical Informatics Initiative.

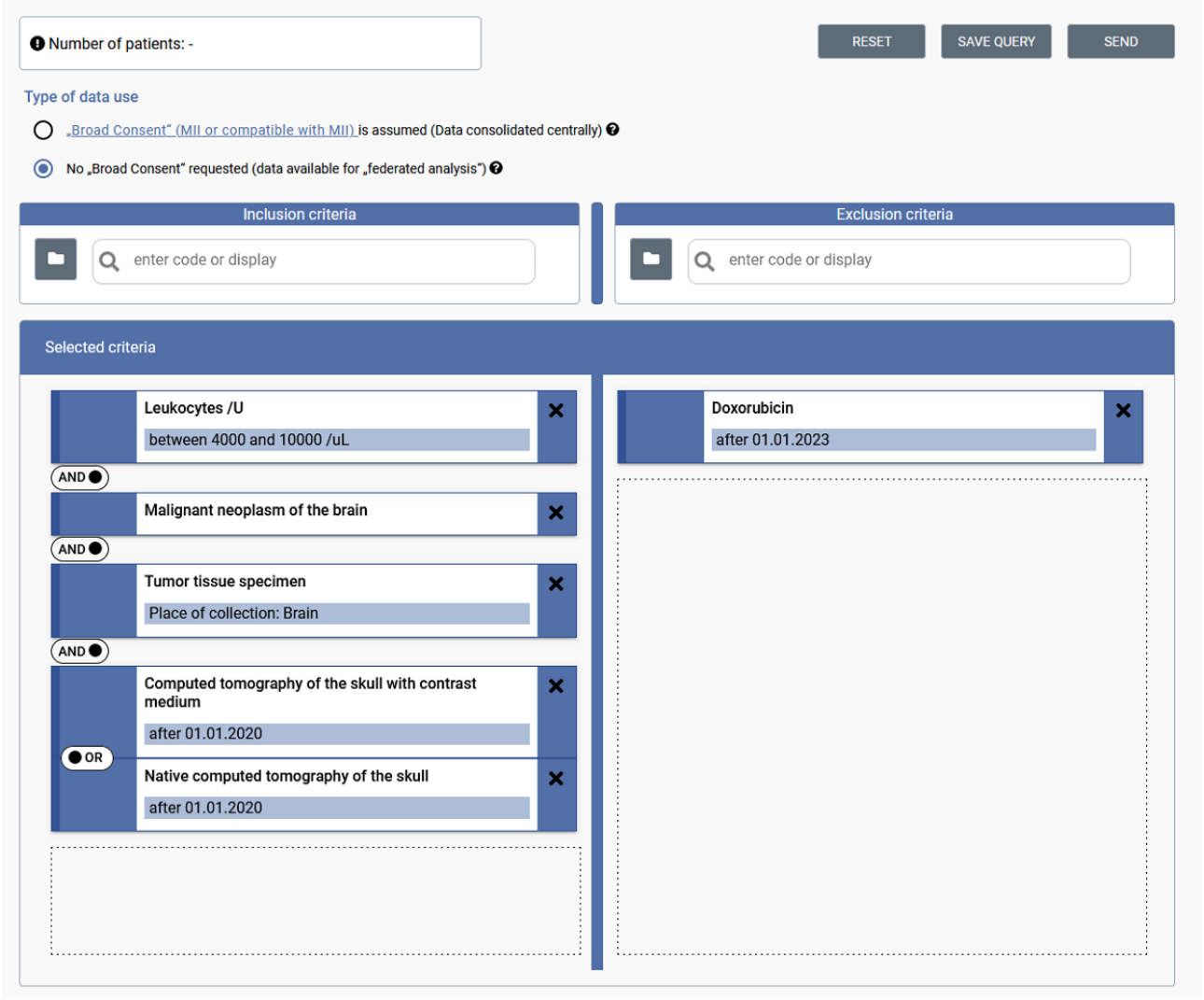


Table 1. Coverage of the Core Data Set (CDS) modules in the German Portal for Medical Research Data (FDPG) feasibility portal.

CDS module	CDS-supported codings	FDPG coverage
Consent	MII[a]_CS_Consent_Policy	MII_CS_Consent_Policy
Diagnosis	*ICD-10-GM*[b], Alpha-ID, SNOMED[c] diagnoses codes, Orphanet	*ICD-10-GM*
Laboratory	LOINC[d]	Defined subset of "TOP300" LOINC codes
Medication	ATC[e]-DE, ATC-EN, PZN[f]	ATC-DE
Person	—[g]	—
Procedure	OPS[h], SNOMED procedure codes	OPS
Specimen	SNOMED specimen codes	Defined subset of "TOP50" SNOMED specimen codes

[a]MII: Medical Informatics Initiative.

[b]*ICD-10-GM*: *German modification of the International Statistical Classification of Diseases and Related Health Problems, 10th revision.*

[c]SNOMED: Systematized Nomenclature of Medicine.

[d]LOINC: Logical Observation Identifiers Names and Codes.

[e]ATC: Anatomical Therapeutic Chemical.

[f]PZN: pharmazentralnummer.

[g]Not applicable.

[h]OPS: Operationen- und Prozedurenschlüssel.

Discrepancy Despite Standardization: Challenges and Opportunities

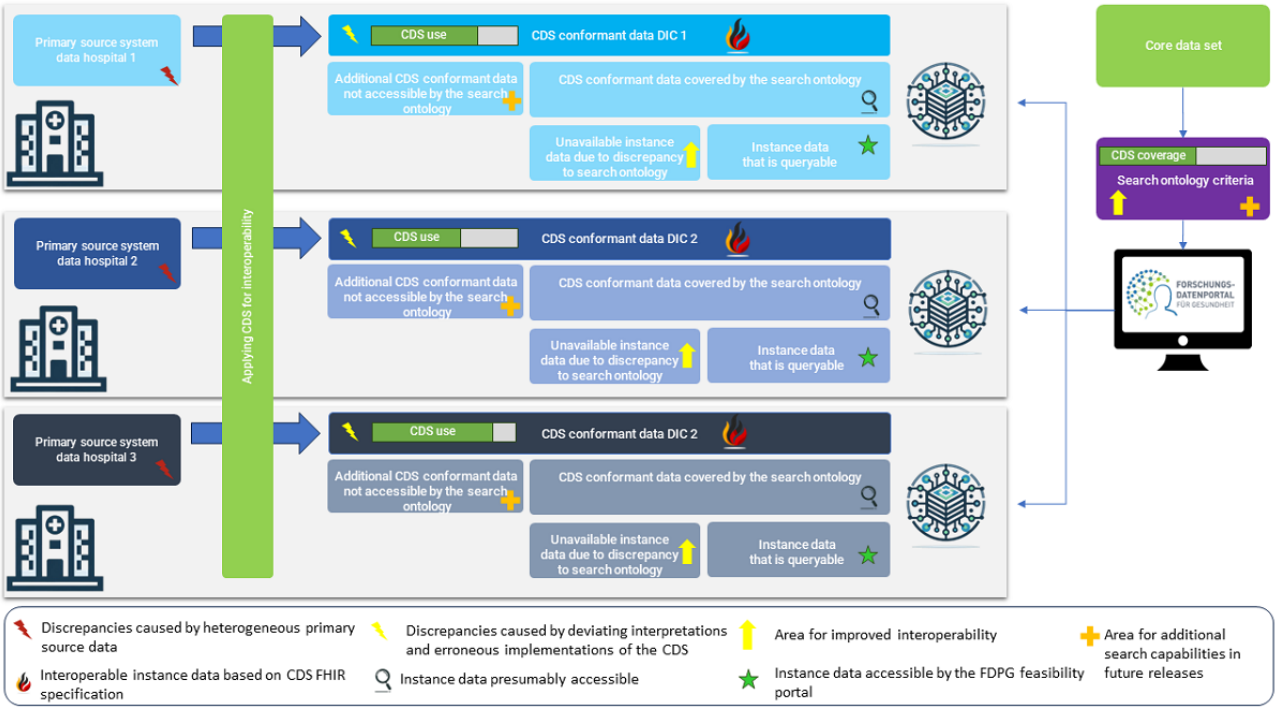
Figure 2 illustrates the interplay between the source systems, the DICs, the CDS, and the search ontology of the FDPG feasibility portal. Applying the CDS to the heterogeneous primary source data makes the heterogeneous source data interoperable. Despite the standardization, discrepancies can arise by deviating interpretations or erroneous implementations of the CDS. Typically, these can be identified by validating the instance data against the CDS and are addressed in the ETL jobs of the sites. It might be necessary to adjust the CDS implementation guide to provide more clarity. It is important to note that despite these efforts, inherent challenges related to data quality at the source, such as missing, erroneous, or inconsistently entered data, persist [17]. These complexities often necessitate extensive collaboration and resources for resolution and fall outside the direct control of DICs, whose primary function is data integration.

The FDPG feasibility portal cannot exhaustively query all instance data at the clinical sites. Consequently, the instance data can be divided into 2 subsets: the data that should be covered by the search ontology and the currently unsupported data. The latter provides opportunities for future developments but is outside the scope of this paper.

The search ontology's additional constraints and the CDS's limited data harmonization cause discrepancies that prevent the accessibility to instance data that should be available. While presenting challenges, such discrepancies are not unique to this framework but are commonly observed across various industries when implementing standards [18]. They mirror the common rationale behind the organization of Connectathons, which aim to test and improve interoperability.

Outside the governmental context of a Connectathon but maintaining the same objective of advancing interoperability through rigorous testing of multiple systems adhering to the same standard, our approach evaluates the assumptions underlying the FDPG's search ontology against the actual instance data at clinical sites. To achieve this, we use the same FHIR profiles used to develop the search ontology.

Figure 2. Simplified federated architecture of the German Portal for Medical Research Data and the data integration center and the data discrepancies and alignments that arise. CDS: Core Data Set; DIC: data integration center; FHIR: Fast Healthcare Interoperability Resources.



Methods

Overview

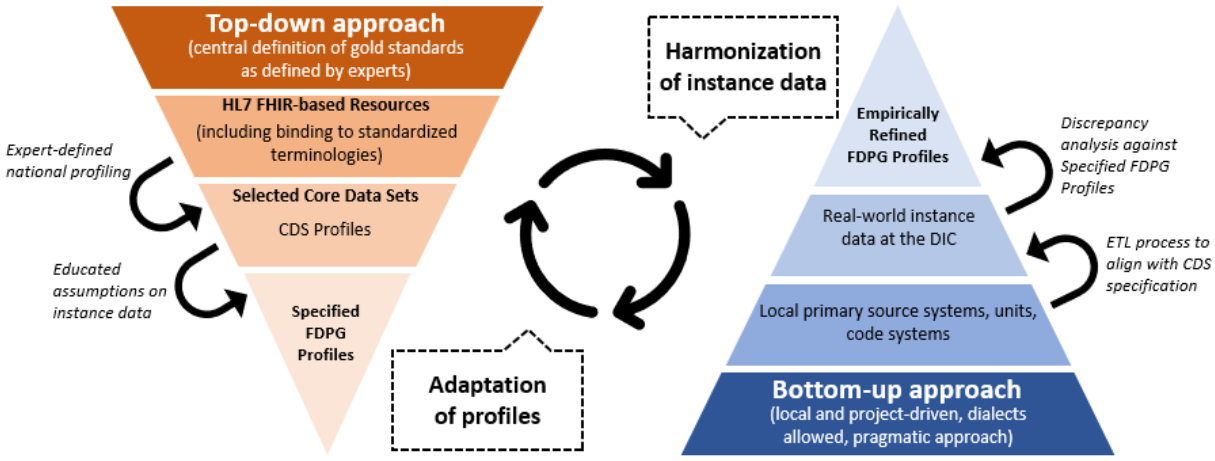
This research integrates a top-down approach for the definition of data models with an empirical bottom-up methodology (Figure 3). This novel approach addresses discrepancies in data standardization and representation. The representation of the search ontology as an FHIR profile based on the CDS forms the foundation of our top-down perspective (hereinafter referred to as FDPG profiles). The bottom-up approach, conversely, is grounded in an empirical analysis of the instance data.

Being able to rely on FHIR profiles has multiple advantages:

- *Using existing validation software:* unlike custom solutions, our approach leverages existing validation software to analyze and resolve discrepancies, offering a more streamlined and efficient process.
- *Insights from layered FHIR profiles:* the layered structure of FHIR profiles provides multifaceted insights. It not only aids in understanding the search ontology but also evaluates compliance with the CDS. In addition, performing this analysis across different sites sheds light on several aspects: (1) discrepancies in instance data across sites, (2) discrepancies between the CDS and site-specific instance data, and (3) discrepancies between the search ontology and the instance data
- *Leveraging established standards:* working with established technology in the MII allows for transparency and adaptability beyond the current use case.

The strength of our approach lies in its iterative nature, whereby each cycle involves refining the profiles based on the empirical insights gathered, subsequently informing the following empirical analysis. This continuous refinement allows us to transition from differential to data quality analysis. This progression not only targets the resolution of discrepancies but also seeks to enhance the quality and accessibility of the data. By iterating this process, we contribute a new framework for reconciling the tension between data standardization and real-world variability, a common challenge in health care data management. Notably, the approach is not solely focused on the search ontology, as would typically be the case when developing a product. Being a part of the MII community, we hope that our approach will also uncover potential for further harmonization across sites, enhancing the functionality and accessibility of the FDPG feasibility tool and facilitating a wider range of use cases.

Figure 3. Combining the top-down and bottom-up approach for the iterative, evidence-based refinement of FHIR profiles. CDS: Core Data Set; DIC: data integration center; ETL: extract, transfer, load; FDPG: German Portal for Medical Research Data; HL7 FHIR: Health Level Seven Fast Healthcare Interoperability Resources.



Validation Pipeline

To use the FDPG profiles, we developed and deployed a validation pipeline (Figure 4) that can be deployed at each site. A centralized approach was not feasible due to the high standards of data protection adhered to within the MII. The pipeline uses the FHIR Marshal, an HL7 application programming interface validation library extended with a Representational State Transfer application programming interface [19]. The validation library requires the profiles and their dependencies for the validation as well as the referenced terminology resources.

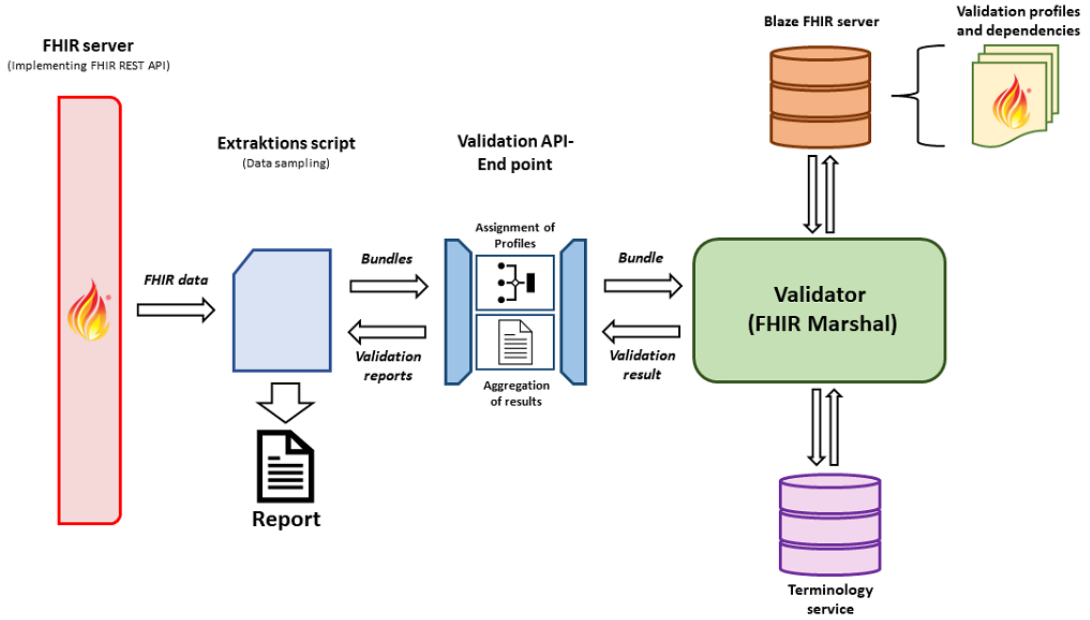
The FHIR profiles are themselves resources, more specifically, FHIR StructureDefinition resources. They are stored in an FHIR server (Blaze) [20] using the *FHIR Populator* tool [21]. This tool is designed to download FHIR profiles from package managers such as Simplifier and their dependencies, subsequently uploading them to our servers. The FHIR Populator also accommodates environments without internet access by offering the capability to upload a previously persisted package.

Regarding terminologies, we analyzed all StructureDefinitions and downloaded the expanded ValueSets from a terminology server based on the binding information. The terminologies are stored in a specialized FHIR server, Termite [22], which implements a minimal set of FHIR Terminology Services for validation and contains all relevant expanded ValueSets and CodeSystems [23].

The process chain uses a Python script to extract equivalence class test samples (for each profile, a maximum of 500 instances) from the local FHIR server. Within the meta information of each resource, the profile it implements is listed. By substituting this profile identifier from the CDS with the corresponding FDPG profile identifier, we can use the standard validation chain to verify the instance data's conformance with the FDPG profiles. The pipeline's output is a JSON file containing the validation results. For improved readability, we also generate a PDF report from these data.

For easy deployment at each site, we provide a configurable docker package containing all publicly available components via GitHub [24]. The pipeline does not require access to external networks and can be easily adapted for different profiles.

Figure 4. Validation pipeline. API: application programming interface; FHIR: Fast Healthcare Interoperability Resources; REST: Representational State Transfer.



FDPG Profiles and Underlying Assumptions

As previously established, the search ontology's current representation and its FHIR profile representation are guided by making the most common and most harmonized instance data in Germany available and iteratively building on that for specific user groups. For the modulation of the current version, we assumed that hospitals across Germany, owing to the country's billing system's requirements, would consistently and correctly use *ICD-10-GM* [25], Operationen- und Prozedurenschlüssel (OPS), and Anatomical Therapeutic Chemical (ATC) codes for diagnosis, procedure, and medication, respectively. Consequently, these profiles can be easily expressed by mandating a coding from the respective code systems. The CDS profile for specimen already limits valid codings to the descendants of the SNOMED CT concept specimen. The search ontology reduces the valid codings further to a subset of the top 50 most common specimen codes. This information was already collected bottom-up before this analysis from metadata across sites, implying a high likelihood of minor to no discrepancies for Specimen in this study.

From a technical perspective, a specific coding is necessary to identify a criterion in the instance data. Beyond requiring a specific coding, the FDPG profiles commonly ensure the existence of a technical reference to the patient, which is crucial for the operation of the FHIR search. A single search ontology profile is sufficient for all modules except laboratory, following these guidelines. As previously established, the laboratory module presents a unique challenge due to its complexity and the interdependence of the value, the LOINC coding, and the number of codings.

Vreeman et al [26] identified that out of the 55,000 codes LOINC offers in practice, only a small subset is required to account for up to 99% of all laboratory observations [26]. Following their example, the MII established a "LOINC TOP300" subset that addresses 80% of all laboratory use cases in Germany [27]. The MIRACUM Metadata Repository (MDR) [11] makes this subset available and guides the current implementation in the feasibility tool of the FDPG.

The LOINC scale type associated with each concept [28] determines whether the laboratory result is quantitative or qualitative. While LOINC provides dimensional requirements, a wide set of Unified Code for Units of Measure (UCUM) units can fulfill that requirement. LOINC also provides exemplary units, but they are not mandatory. Fortunately, a UCUM representation is readily available for quantitative laboratory results in the MDR. The MDR, therefore, presents a machine-processable modulation of the interdependency of the LOINC and its quantitative value.

The MDR does not contain ValueSets for qualitative values, necessitating an alternative approach. If available, the LOINC answer list associated with the respective LOINC code is used for qualitative values, accessed via a terminology server. If unavailable, a general-purpose ValueSet defined by the MII is used. However, this ValueSet contains multiple representations of the same concept, such as 23 different representations for "Absence finding." These various representations were reduced to a single value to enhance usability. Figure 5 illustrates how the information from the MDR is used to refine the CDS profile to specific FDPG profiles for each LOINC coding.

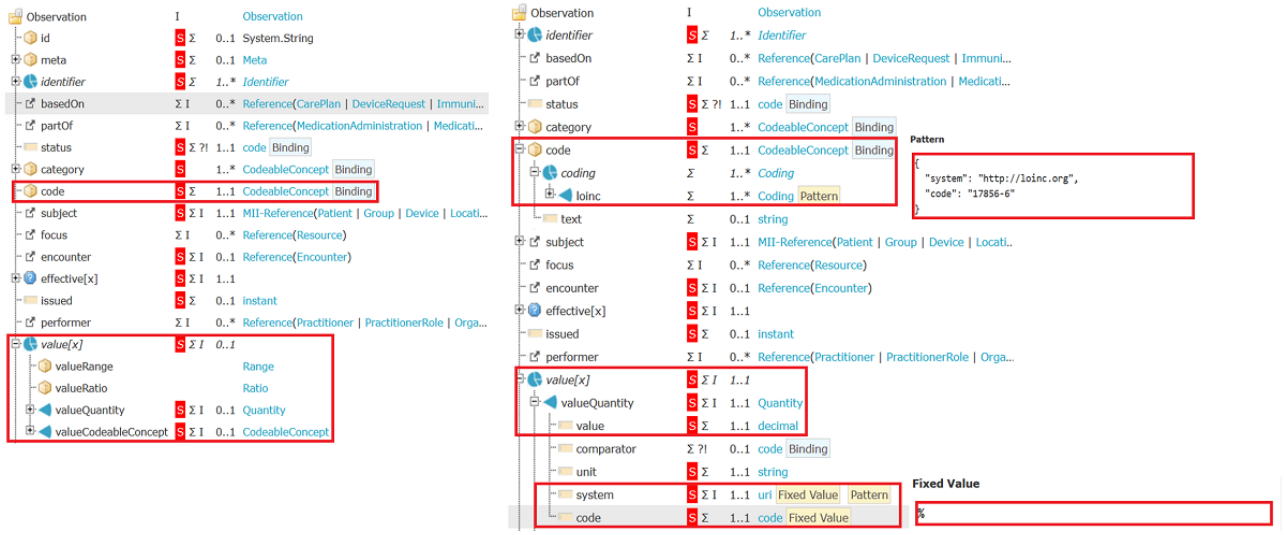
As the name suggests, the MDR is primarily used by the sites of this consortium and is not mandated for any sites in the context of the FDPG. Furthermore, given the nature of the laboratory profiles [29] within Germany concerning the representation of laboratory values cause expected variability.

The FDPG profiles are openly accessible on Simplifier [14].

Guided by these educated assumptions that led to the creation of the FDPG profiles, our hypothesis for this study is threefold. First, we anticipated high uniformity in applying *ICD-10-GM*, OPS, ATC, and specimen codes. Second, we hypothesized a

considerably higher level of variability in the application and interpretation of LOINC codes, with regional disparities in the representation of laboratory values and their units playing a significant role. Third, we anticipated that our approach of refining the FDPG profiles based on the insight of the instance data would allow us to improve interoperability. This study aimed to probe these hypotheses, shedding light on the complexities of harmonizing and standardizing clinical data across different health care regions and coding systems in Germany.

Figure 5. Core Data Set laboratory profile without interdependency between code and value (left) HbA1c profile with code value interdependency (right).



Ethical Considerations

This study did not require an ethics board's approval as it did not involve analyzing individual patient data. Instead, summary statistics on data quality were generated at the participating respective sites, a practice in line with the applicable general data protection regulation.

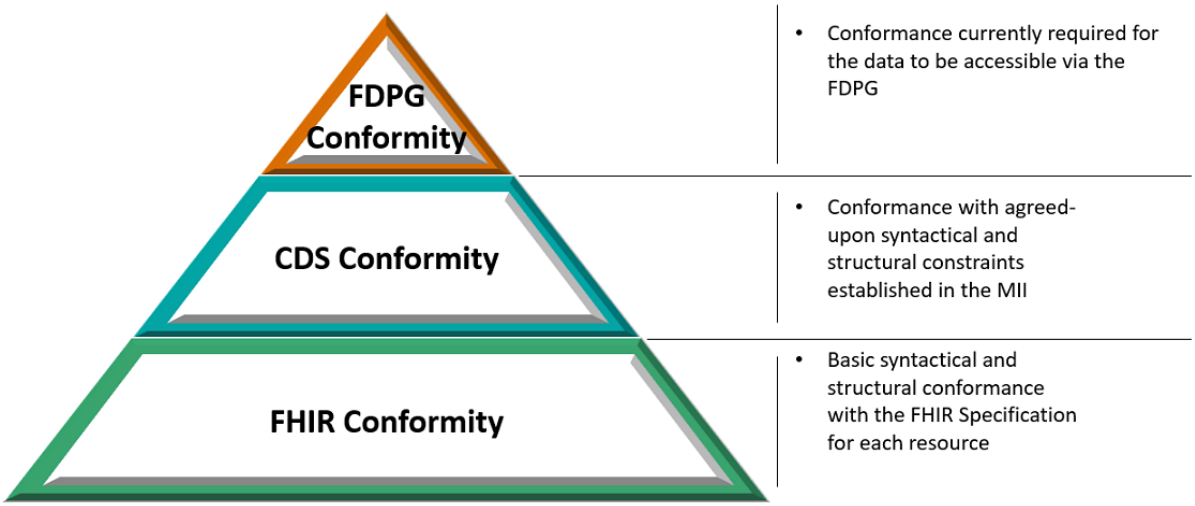
Results

Overview

To evaluate the methodology and tooling, we performed the discrepancy analysis over a wide range of sites (N=20) from all 4 consortia, including sites from western and eastern Germany and with different stages of DIC implementation. We differentiated 3 conformance levels when analyzing the results, as illustrated in Figure 6.

At a base level, all instance data must conform to the FHIR standard and the CDS. Deviations from these specifications are to be regarded as implementation errors. On the topmost level, the FDPG profiles are not binding for the DIC. Instead, they should be perceived as a description of the capabilities of the search ontology that allow insight into the subset of CDS conformant data currently searchable by the FDPG feasibility tool. Discrepancies on that level will be interpreted, and implications and possible solutions will be derived.

Figure 6. Data validation conformance levels. CDS: Core Data Set; FDPG: German Portal for Medical Research Data; FHIR: Fast Healthcare Interoperability Resources; MII: Medical Informatics Initiative.



CDS Compliance

Addressing CDS compliance, the surfaced discrepancies varied in nature and complexity, highlighting several key areas where adherence was lacking, spanning issues such as noncompliance with cardinalities, absent references, the nonfulfillment of rules stated in the profiles, and the use of incompatible codings. Notably, some of the key issues included the absence of the mandatory *status* attribute within the *MedicationAdministration* resources, which has a fundamental role in the context of the resource. Status codes that do not imply the completed administration are likely not relevant to clinical research, but if the status is not entered, no differentiation can be made.

A SNOMED CT–encoded *category* for *Procedure* was another prevalent discrepancy required when an OPS coding was used. While currently not the case, OPS codes might be used in a different context of a *Procedure* to differentiate in these cases the category that would be needed. Furthermore, errors in the *system* URL for codings caused validation errors, revealing an essential need for accuracy in defining codings.

While the above-listed findings in many cases do not directly affect the accessibility of the resources via the FDPG feasibility portal as these attributes are often not queried for, they emphasize the necessity for rigorous attention to CDS compliance in terms of detail and rigor in the data harmonization process for secondary use.

Beyond the issues that can be addressed with clear solutions, such as providing the required elements or adhering to specific rules or codings, we also found discrepancies that originate from misinterpretations of the implementation guides.

One example we found is using unit codes that do not stem from UCUM within the data element *MedicationAdministration.dosage.dose.code* in the Medication module. In discussing this discrepancy, we were pointed to the comment section for this element: “The preferred system is UCUM, but SNOMED CT can also be used (for customary units) or ISO 4217 for currency. The use context may also require a code from a particular system.” However, the profile requires “http://unitsofmeasure.org” as the system URL, causing a contradiction that should be addressed in the documentation.

FDPG Compliance

As anticipated, the additional constraints in the FDPG profiles are the leading cause of discrepancies. However, many of our hypotheses are confirmed within this subset of instance data.

Procedure, Condition, Medication, and Specimen

As expected, *procedure, condition, and medication resources are available with the codings* we anticipated. Unfortunately, our initial validation revealed that the terminology still caused discrepancies, but these can be traced back to the differences in the codes and the display. The national code systems used by the resources are currently not made available as FHIR resources but in their own format, leaving room for different interpretations and leading to different implementations to generate the displays. The Federal Institute for Drugs and Medical Devices (BfArM) responsible for the code systems is aware of this issue and has been assigned to create a central terminology service as the single source of truth for national code systems in Germany according to § 355 subsections 12 and 13 of the German Social Security Code V [30]. Our findings underline the importance of this intent. For the time being, we will ignore the display in the validation process.

Another discrepancy was using uppercase letters in OPS codes (5 of 20 sites) contrary to the FDPG feasibility portal’s expected lowercase codings for OPS. FHIR search to identify codings is case-sensitive unless the CodeSystem indicates case-insensitivity. FHIR server implementations to support case-insensitivity need to obtain this information from the code system information and adjust their internal search process accordingly. This feature is currently not implemented by the Blaze FHIR server. While the Blaze FHIR server might support this feature in the future, we tend not to make this assumption for all FHIR servers and, therefore, would advise adjusting the codings to have uniform upper or lower casing within one CodeSystem as a practical approach to this problem.

While the ATC code system is correctly coded, the possibility of representing medication information is another cause for discrepancies. The CDS defines 6 different representations for indicating that a medication has been prescribed to or taken by a patient. In FHIR, the prescription, the confirmed administration, and the statement of intake of drugs are modeled using different FHIR resources, MedicationRequest, MedicationAdministration, and MedicationStatement, respectively.

Furthermore, 2 options exist for the resources to link the specific code: the resources can reference a Medication resource or provide the CodeableConcept defining the medication. Currently, the FDPG feasibility portal only supports requesting MedicationAdministrations that reference a Medication resource. Moving forward, the FDPG feasibility tool must readdress requesting medication information.

The Specimen instance data could not be analyzed due to a software bug in the validation pipeline.

Consent

For the Consent module, we encountered an erroneous understanding of the code system, setting the permissions given by the patient. As defined by the MII, the code system contains codes that provide specific permission and codes that imply a set of permissions. The implication for those implementing the Consent resource is the erroneous tendency to use only the most encompassing code in an assumed hierarchy, which currently is not explicitly modeled. It is necessary to list all relevant codes, even if their parent code is already present. Sites not including all subsequent codings do not negatively impact the data privacy but exclude patients that should be within the cohort when searching for more specific permission. In future versions, the currently implicit relationship might be modeled in the CodeSystem using the *part of* the relationship. Once changed, it will require the FDPG feasibility portal to request subsequent codes with their parent codes.

We also found consent information from the consent management software generic Informed Consent Service (gICS), widely adopted across the MII [31]. As the name implies,

consent management software is used beyond the use case of federated feasibility queries in the MII. Therefore, consent resources made available by gICS are based on the more permissive profile defined by the official FHIR standard of the HL7 Germany Working Group Consent Management. These extended gICS consent resources can contain more attribute entries than consent resources based on the MII CDS Consent profile; for example, additional provision codings from the gICS defined code system for Consent resources that cause discrepancies when validated for our use case. Despite these findings, the consent resources remain compatible with our use case requirements, as they are based on the same standard, provided they contain all mandatory elements as defined by the MII CDS Consent profile.

Observation (Laboratory Data)

The analysis of the *Observation* resource revealed findings that align with our hypotheses and highlight the intricacies of data harmonization. The inherent flexibility of the laboratory profile, the nonbinding nature of the MIRACUM MDR directives, and regional historical discrepancies manifested in significant heterogeneity.

A recurring issue is using qualitative values for quantitative LOINC codes and vice versa. We observed that the first case is more prominent than the latter, which is attributable to using codings for invalid measurements. In our federated feasibility use case, the indication of an invalid measurement bears a minimal consequence: the omission of patients with erroneous laboratory values would not adversely affect the validity of the result if sufficient patients are identified with a specific laboratory value. However, for use cases that evaluate the data, indicating a value's invalidity is highly relevant and should be addressed in a standardized way. Solutions we uncovered range from using existing codes from various code systems with different granularity, such as indicating invalid measures using the SNOMED CT code for *invalid* or even a specific postcoordinated expression to provide additional insight about the cause, to using in-house code systems. The latter proves inadequate for achieving interoperability across various sites or when explaining the absence of data. In the future, clear guidance must be available to the sites to harmonize this information.

The leading cause for discrepancies that hinder the current accessibility of the laboratory values is the different UCUM units. However, the variety across the participating sites is not as wide as anticipated.

Within the instance data across the sites, we identified 368 different quantitative LOINC codes, with discrepancies in at least 1 site. Again, case sensitivity contributes to a significant number of discrepancies. Overall, 59 differences can be attributed to the inconsistencies between the use of the upper and lower letter "l" for liter. Overcoming this issue would lower the number of total discrepancies and the variety of different dimensions.

The BfArM defines a list of commonly used UCUM units [32]. According to our analysis against this list, discrepancies were attributed to units still predominantly used in eastern Germany. Some units, such as "Gpt/L" for giga particle per liter, are not included in the UCUM code system and are therefore unsupported. Another recurrent issue identified by this comparison is the use of the Greek letters (eg, "µ" instead of "u"). Considering only discrepancies caused by UCUM codes listed by the BfArM reduces the remaining differences to 248. This comparison with the prevalent UCUM units in Germany also revealed that 3 MDR entries differentiate from the BfArM list.

Moreover, 177 (71.4%) out of the 248 remaining discrepancies were due to different multiples of 10 in the representation (eg, ug/dl vs mg/L) with additional units that could be converted by applying more complex calculations such as converting mmol to g or mm [Hg] to kPa. Not having a medical background, we abstain from evaluating the correctness of the units, causing the remaining discrepancies. Importantly, we can show that there is a significant portion (298/368, 81%) of discrepancies that could be resolved by applying simple measures and, importantly, to lower variety in the discrepancies (66 out of the remaining 70 discrepancies would be caused by 1 different representation of a unit, and only 4 requiring the representation of the same value in 2 different units).

Multimedia Appendix 1 provides an excerpt of the table we created for this study and is one of the most important results of this study, as it can foster future developments. The LOINC codes most often found in the instance data are displayed.

Discussion

Principal Findings

Our work confirmed the need for sufficiently constrained profiles to inform the development of the federated feasibility tool. We presented tools and an approach that enables evidence-based decisions instead of the current educated guesses approach to create lower granularity CDS profiles. Performing the discrepancy analysis on the use case presented also granted significant insight into the data quality and alignment of the CDS among university hospital sites and the FDPG.

We also found first improvements in the data quality with the sites that addressed issues regarding the CDS identified in the data quality reports and subsequently undertook a reanalysis. This indicates that the presented tooling can improve data quality for agreed-upon profiles.

Implications for the Future of the FDPG Feasibility Tool

Despite the success in other areas, the high number of discrepancies in laboratory values presents a unique challenge. From a product development perspective, considering the diverse nature of laboratory data, several options emerge:

- The feasibility tool could change its capabilities to query for the existence of a laboratory value. Insight into existing clinical studies from ClinicalTrials.gov indicates that researchers have a high demand for specifying specific ranges for laboratory values. Moreover, it could be argued that aligned with the General Data Protection Regulation, call for "data minimization" capabilities to further

refinement even for feasibility queries could be deemed necessary.
- While theoretically feasible, expanding the range of selectable units to include all the existing ones would necessitate unit conversions by the researcher, likely negatively impacting usability and user awareness.
- Implementing on-the-fly conversions is typical in modern user interface design. However, given current technology constraints and the FDPG feasibility portal's response time requirements, this approach is not viable for handling big data.

Under normal circumstances, this would indicate the necessity to pivot from the current implementation of the feasibility tool; fortunately, the collaborative nature of the MII offers an alternative solution:

- Ideally, alignment of the laboratory values would stem from the primary source system; our empirical findings show (as portrayed in the *Results* section) that for the TOP300 LOINC Codes, the discrepancies can be overcome by applying rather simplistic measures to improve data harmonization. The data already being in a standardized format further enable the development of 1 tooling that suits all sites' needs. The MIRACUM consortium has initiated preliminary efforts in this direction and is available to all MII members [33]. While we would highly advise against mapping LOINC Codings, as implemented in the tooling, it showcases the feasibility of providing a conversion tool. We acknowledge that such tooling might require significant quality assurance (the tool LUMA [34] can support ensuring the used units match the dimensions defined in LOINC) or even the approval as software as a medical device given the existing expertise in the MII [35] and the reoccurring demand for further data harmonization for, for example, distributed machine learning, we regard the endeavor worthwhile.

Harmonizing the data would also give users a broader range of units by converting the selected value to the harmonized representation. There still might be cases where data harmonization of the instance data is not feasible. A fallback to the previously discussed solutions would still prevail in these cases.

Once the main issue of aligning the representation of the values in the search ontology is overcome, it will be necessary to expand the list of LOINC codings and their interdependencies continuously. Already, we found that 15 of the 20 sites had additional LOINC codings in the relatively small sample size of 500 Observations.

Related Work

Addressing the different granularities of FHIR resources and international, national, and domain-specific profiles is a topic that has seen more traction in recent years. With the International Patient Summary (International Organization for Standardization 27269) defining a minimal baseline of elements that need to be present in the electronic health record of a patient and its implementation in FHIR profiles to address the use case of "unplanned, cross border care," it is likely to see a rise of interdependent FHIR profiles. For European countries specifically, the goal of a shared European Health Data Space [36] will require defining an additional layer between the international and national levels. These profiles are not intended to serve all use cases. Figure 7 outlines the different layers based on the work of Vreeman (Vreeman, DJ, unpublished data, July 2023) and Aassve [37] and the role of existing implementation guides within these layers. The layered approach from minimally restrictive to sufficiently restrictive data modeling also offers the opportunity to promote reuse. The role of the CDS is not clearly defined within this model, while the FDPG profiles are sufficiently restrictive for the use case of federated search.

Kramer [38] pinpointed a discernible gap in reusability within FHIR, as revealed through their scrutiny of 125 implementation guides. The layered approach enables the definition of reusable extensions, promotes terminology use, and provides clear guidance on the existing profiles that can be enacted as a baseline.

While some may advocate for an all-encompassing top-down approach for particularly restrictive profiles, we believe that the existing instance data will inevitably require aligning for different reasons: existing instance data based on a less-stringent layer, missing restrictive profiles, or insufficient insight on real-world instance data when creating the profiles. Once this alignment is achieved, the established tooling in this work can be further used for quality assessment. While our work goes beyond the data quality assessment (DQA) of the CDS, it is inherently a part of our examination, if only regarding conformance.

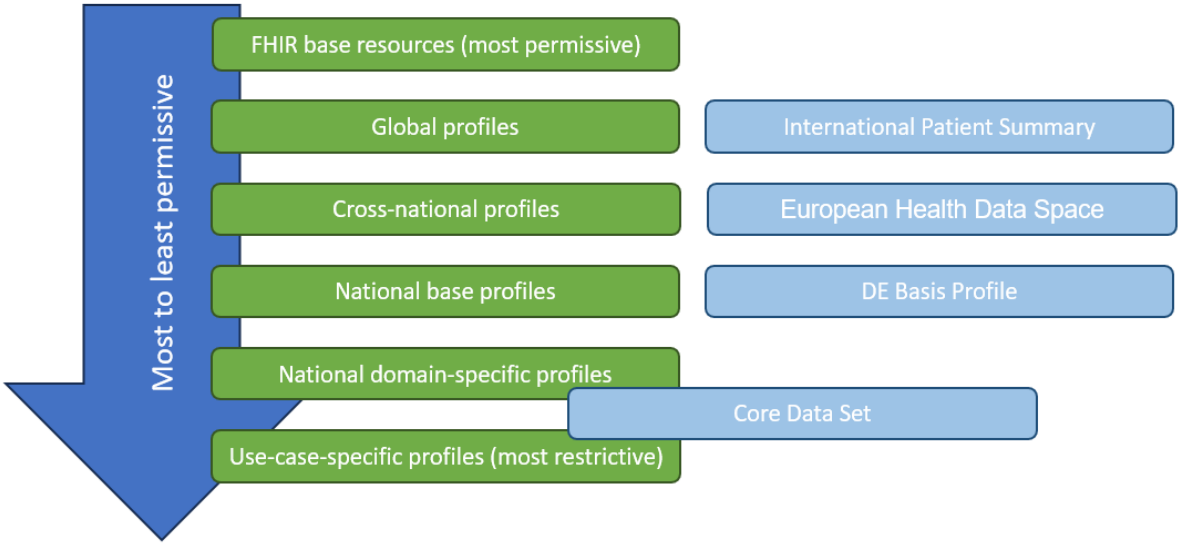
Draeger et al [39] and Kamdje-Wabo et al [40] performed a DQA across DICs in the MII using an R script to analyze specific elements in the instance data regarding their conformance, completeness, and plausibility as defined by Kahn et al [41]. Their findings align with our findings regarding conformance and completeness on a smaller sample set but go much more in depth when assessing plausibility. Ideally, future assessments will synthesize both methodologies: harnessing the robustness of FHIR profile validations as a foundation and superimposing intricate plausibility evaluations. Within the context of the MII, the MIRACUM DQA tool [42-44] assesses the data quality based on abstract data set definitions in the consortium-provided MIRACUM MDR. The same MDR underpinned our FHIR profile formulation for laboratory values.

Consequently, the sites that used the MIRACUM DQA tool had fewer discrepancies in our analysis. Still in the context of the MII but beyond the FHIR standard, the work of Tute et al [45] also applied rule-based evaluation on specific data elements using R. However, the data source they analyzed was an openEHR repository queried with AQL statements.

In the broader scheme of data quality of medical data, we also find the reoccurring topic of addressing data ambiguity. Schmidt et al [46] delve into the significance of data quality in observational health research and the different quality indicators, presenting a comprehensive framework that not only highlights the challenges but also offers software solutions in R to tackle them.

In conclusion, while the push toward defining and refining FHIR profiles and embracing standardized data formats is a significant step forward, it remains a piece of a giant puzzle to achieve interoperability and high data quality in the medical domain, which will ultimately require a multifaceted approach.

Figure 7. Layered Fast Healthcare Interoperability Resources profile model. DE: Deutsche (German).



Limitations

The presented approach requires high effort, limiting its inherent iterative nature. Therefore, it is essential to use all information obtained in each cycle to make informed decisions on the maximally constrained profiles that ultimately lead to profiles that the sites can use independently to access their data quality. The high effort is also why the initial run of discrepancy analysis is already based on educated assumptions. Notably, the method works even if only the required cardinalities and arbitrary values for the required attributes are specified in the profiles. Consequently, the first pass would cause maximum discrepancies and full insight into the instance data.

With the high alignment of educated assumptions, the analysis only evaluates data that should already match the search ontology criteria. LOINC Codes outside the TOP300, different codings for medications, diagnosis, or procedures are not part of the performed analysis and must be revisited once made available in the search ontology.

Compared with data analysis scripts, an advantage of this approach is its easy adaptation and reuse of the validation pipeline to perform DQAs for completeness and correctness. However, while aspects of plausibility can be found in our results, these are not directly related to our approach and must be addressed separately.

Finally, we did not analyze the *Patient* resource due to the federated nature of our analysis and privacy concerns. Being the central part of every feasibility query, sites should use the existing tooling to ensure conformance.

Outlook

We created an individual feedback PDF report containing all discrepancies for every participating site based on our findings. We hope that this feedback will allow the sites to address the identified discrepancies that are not caused by the FDPG profiles but rather are contradictions between the CDS and their ETL processes.

After reaching sufficient maturity, the FDPG profiles will be used to update the search ontology.

Further iterations, when expanding the ontology and between time periods, of the presented approach could be summed up in an iteration study providing insight into the continuous development of the central platform and the DIC.

This work revealed the necessity of extending collaborations between the CDS team, FDPG team, and DICs. While many discussions are still open at this point, first adjustments are already being made, that is, by the representatives of the CDS Consent module who have already implemented our recommendation of using the “part-of” relationship in the CodeSystems hierarchy. Now, it falls to the FDPG team to adjust the feasibility query accordingly in a future release.

We believe that it is pivotal to build on the presented approach to provide sites with tooling that enables them to verify their data quality concerning the CDS and identify if their instance data are available via the FDPG feasibility tool, allowing them to adjust their data or demand adjustments of the feasibility tooling if discrepancies arise. While the current tooling was tailored for this study, further adjustments can and should be made to provide sites with a more actionable report; that is, the report should provide a working link to a resource where a validation error occurred, allowing on-site users to have full access to all information.

We see a high demand for the presented approach. Whether partly to ensure data quality or fully to refine top-down defined profiles with evidence-based information to enhance interoperability. We provide a generic approach and sufficient tooling to make it usable for use cases beyond the FDPG, requiring only slight adjustments to work with other profiles.

Acknowledgments

The authors would like to express their gratitude to all sites taking part in this study and those responsible at the data integration centers who enabled this work by providing their report and even more important feedback on the findings and software, who are listed here in no particular order: Christoph Müller and Florisa Zanier (University Hospital Aachen), Natalia Ortmann and Jean Mascene Mazimpaka (University Hospital Augsburg), Florian Seidel (Charité Berlin), Klaus-Jürgen Quast (University Hospital Bonn), Mirko Gruhl (University Hospital Dresden), Stephan Wojke (University Hospital Frankfurt), Christopher Frank (University Hospital Gießen), Markus Mandalka and Sebastian Berthe (University Hospital Greifswald), Reto Wettstein (University Hospital Heidelberg), Usha Rama Nadan Pandit and Mehrshad Jaberansary (FKZ: 01ZZ2316K, University Hospital Cologne), Andreas Dürschmid and Thomas Wendt (University Hospital Leipzig), Jan Maluche (University Hospital Magdeburg), Sebastian Stöcker (University Hospital Marburg), Fabio Aubele (University Hospital of Munich), and Johannes Oehm (University Hospital Münster). The authors would like to thank the early adopters who paved the way to scale the study from 5 to 20 sites: Marvin Kampf (University Hospital Erlangen), Diana Pietzner (University Hospital Halle), Helmut Spengler and Raffael Bild (University Hospital rechts der Isar), Tobias Hilmer (University Hospital Schleswig-Holstein), and Georg Fette (University Hospital Würzburg). Furthermore, the authors would like to thank Martin Bialke (University Medicine Greifswald) for his feedback on the generic Informed Consent Service for a correct presentation in this paper. Finally, the authors would like to acknowledge all individuals in the Medical Informatics Initiative (specifically representatives from the Core Data Set team and coordinating office Technology, Methods, and Infrastructure for Networked Medical Research) whose insightful feedback was instrumental in shaping this paper.

The project was funded by the German Federal Ministry of Education and Research under the FDPG-PLUS project (grants 01ZZ2309D and 01ZZ2309A). This work was further supported in collaboration with participating sites funded by the NUM-DIZ project (grant 01KX2121).

Authors' Contributions

LR designed the methodology, supervised the initial test run implemented by PB for 5 sites [47], adjusted and performed the discrepancy analysis for 20 sites, and was the primary author of this manuscript. JW contributed significantly to implementing the Fast Healthcare Interoperability Resources Marshal and Populator used in the validation pipeline and offered crucial insights on medical terminology. JG played a pivotal role in coordinating and testing the discrepancy analysis. JI served in an advisory capacity throughout the research process. All authors actively reviewed, provided feedback, and approved the final version of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Analysis of LOINC (Logical Observation Identifiers Names and Codes) code units and derivations across multiple clinical sites and the German Portal for Medical Research Data.
[XLSX File (Microsoft Excel File), 30 KB - medinform_v12i1e57005_app1.xlsx]

References

1. Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013 Presented at: CBMS '13; June 20-22, 2013; Porto, Portugal p. 326-331 URL: https://ieeexplore.ieee.org/document/6627810 [doi: 10.1109/cbms.2013.6627810]
2. Profiling. FHIR v5.0.0. URL: https://hl7.org/FHIR/profiling.html [accessed 2023-11-26]
3. Semler S, Wissing F, Heyder R. German medical informatics initiative: a national approach to integrating health data from patient care and medical research. Methods Inf Med 2018 Jul 17;57(S 01):e50-e56. [doi: 10.3414/me18-03-0003]
4. Prasser F, Kohlbacher O, Mansmann U, Bauer B, Kuhn KA. Data integration for future medicine (DIFUTURE). Methods Inf Med 2018 Jul;57(S 01):e57-e65 [FREE Full text] [doi: 10.3414/ME17-02-0022] [Medline: 30016812]
5. Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed - an open platform approach to enhance care and research across institutional boundaries. Methods Inf Med 2018 Jul;57(S 01):e66-e81 [FREE Full text] [doi: 10.3414/ME18-02-0002] [Medline: 30016813]
6. Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: medical informatics in research and care in university medicine. Methods Inf Med 2018 Jul;57(S 01):e82-e91 [FREE Full text] [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
7. Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart medical information technology for healthcare (SMITH): data integration based on interoperability standards. Methods Inf Med 2018 Jul 17;57(S 01):e92-105. [doi: 10.3414/me18-02-0004]

8. Metzger P, Hess ME, Blaumeiser A, Pauli T, Schipperges V, Mertes R, et al. MIRACUM-Pipe: an adaptable pipeline for next-generation sequencing analysis, reporting, and visualization for clinical decision making. Cancers (Basel) 2023 Jul 01;15(13):3456 [FREE Full text] [doi: 10.3390/cancers15133456] [Medline: 37444566]
9. Meineke FA, Stäubert S, Löbe M, Uciteli A, Löffler M. Design and concept of the SMITH phenotyping pipeline. Stud Health Technol Inform 2019 Sep 03;267:164-172. [doi: 10.3233/SHTI190821] [Medline: 31483269]
10. Kindermann A, Tute E, Benda S, Löpprich M, Richter-Pechanski P, Dieterich C. Preliminary analysis of structured reporting in the HiGHmed use case cardiology: challenges and measures. Stud Health Technol Inform 2021 May 24;278:187-194. [doi: 10.3233/SHTI210068] [Medline: 34042893]
11. Prokosch HU, Baber R, Bollmann P, Gebhardt M, Gruendner J, Hummel M. Aligning biobanks and data integration centers efficiently (ABIDE_MI). In: Bürkle T, Denecke K, Holm J, Sariyar M, Lehmann M, editors. Studies in Health Technology and Informatics. New York, NY: IOS Press; May 16, 2022:37-42.
12. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch HU, et al. The architecture of a feasibility query portal for distributed COVID-19 fast healthcare interoperability resources (FHIR) patient data repositories: design and implementation study. JMIR Med Inform 2022 May 25;10(5):e36709 [FREE Full text] [doi: 10.2196/36709] [Medline: 35486893]
13. The medical informatics initiative's core data set. The Medical Informatics Initiative. URL: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set [accessed 2023-06-12]
14. FDPG query profiles. SIMPLIFIER. URL: https://simplifier.net/fdpg-query-profiles [accessed 2023-06-12]
15. Ganslandt T, Boeker M, Löbe M, Prasser F, Schepers J, Semler SC, et al. Der kerndatensatz der medizininformatik-initiative: ein schritt zur sekundärnutzung von versorgungsdaten auf nationaler ebene. Med Dok Med Inf 2018;20(1):17-21.
16. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, et al. Generation of a fast healthcare interoperability resources (FHIR)-based ontology for federated feasibility queries in the context of COVID-19: feasibility study. JMIR Med Inform 2022 Apr 27;10(4):e35789 [FREE Full text] [doi: 10.2196/35789] [Medline: 35380548]
17. Savitz ST, Savitz LA, Fleming NS, Shah ND, Go AS. How much can we trust electronic health record data? Healthc (Amst) 2020 Sep;8(3):100444. [doi: 10.1016/j.hjdsi.2020.100444] [Medline: 32919583]
18. Lewis GA, Morris E, Simanta S, Wrage L. Why standards are not enough to guarantee end-to-end interoperability. In: Proceedings of the 7th International Conference on Composition-Based Software Systems. 2008 Presented at: ICCBSS '08; February 25-29, 2008; Madrid, Spain p. 164-173 URL: https://ieeexplore.ieee.org/document/4464021 [doi: 10.1109/iccbss.2008.25]
19. fhir-marshal. GitHub. URL: https://github.com/itcr-uni-luebeck/fhir-marshal [accessed 2023-06-12]
20. Blaze. GitHub. 2023. URL: https://github.com/samply/blaze [accessed 2023-06-12]
21. FHIR Populator. GitHub. URL: https://github.com/itcr-uni-luebeck/fhir-populator [accessed 2023-06-12]
22. itcr-uni-luebeck/termite. GitHub. 2022. URL: https://github.com/itcr-uni-luebeck/termite [accessed 2023-06-12]
23. Terminology service. FHIR v4.0.1. URL: http://hl7.org/fhir/R4/terminology-service.html [accessed 2023-06-12]
24. fdpg-query-data-validation. GitHub. URL: https://github.com/medizininformatik-initiative/fdpg-query-data-validation [accessed 2023-06-12]
25. ICD-10-GM: international statistical classification of diseases, German modification. Federal Institute for Drugs and Medical Devices. URL: https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html [accessed 2024-04-29]
26. Vreeman DJ, Finnell JT, Overhage JM. A rationale for parsimonious laboratory term mapping by frequency. AMIA Annu Symp Proc 2007 Oct 11;2007:771-775 [FREE Full text] [Medline: 18693941]
27. Semler SC. LOINC: origin, development of and perspectives for medical research and biobanking – 20 years on the way to implementation in Germany. J Lab Med 2020 Jan 10;43(6):382. [doi: 10.1515/labmed-2019-0193]
28. Knowledge base. LOINC. URL: https://loinc.org/kb/ [accessed 2023-06-12]
29. Laborwerte: ohne umrechnungstabelle läuft nichts. Ärzteblatt DÄG Redaktion Deutsches. URL: https://www.aerzteblatt.de/archiv/39961/Laborwerte-Ohne-Umrechnungstabelle-laeuft-nichts [accessed 2023-06-12]
30. § 355 festlegungen für die semantische und syntaktische interoperabilität von daten in der elektronischen patientenakte. Bundesministrium der Justiz. URL: https://www.gesetze-im-internet.de/sgb_5/__355.html [accessed 2023-06-12]
31. Bialke M, Geidel L, Hampf C, Blumentritt A, Penndorf P, Schuldt R, et al. A FHIR has been lit on gICS: facilitating the standardised exchange of informed consent in a large network of university medicine. BMC Med Inform Decis Mak 2022 Dec 19;22(1):335 [FREE Full text] [doi: 10.1186/s12911-022-02081-4] [Medline: 36536405]
32. Anwendungsbereich von UCUM. BfArM für Bürgerinnen und Bürger. URL: https://www.bfarm.de/DE/Kodiersysteme/Terminologien/LOINC-UCUM/UCUM/_node.html [accessed 2023-08-26]
33. REST-server that converts LOINC codes and UCUM units to a standardized representation. GitHub. URL: https://github.com/medizininformatik-initiative/mii-loinc-conversion [accessed 2023-09-14]
34. Vogl K, Ingenerf J, Kramer J, Chantraine C, Drenkhahn C. LUMA: a mapping assistant for standardizing the units of LOINC-coded laboratory tests. Appl Sci 2022 Jun 08;12(12):5848. [doi: 10.3390/app12125848]
35. fit4translation - kompetenzerweiterung und unterstützung bei der entwicklung von medizinischer software unter dem regulatorischen rahmen der MDR & IVDR im akademischen umfeld. Bundesministerium für Bildung und Forschung. URL: https://tinyurl.com/5dkw6fxv [accessed 2023-12-18]

36. European health data space. The European Commission. URL: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en [accessed 2023-08-09]
37. Øyvind Aassve - making FHIR work at the national level. YouTube. 2020. URL: https://www.youtube.com/watch?v=8pv-Zztibyg&ab_channel=DevDays [accessed 2024-04-29]
38. Kramer MA. Reducing FHIR "profiliferation": a data-driven approach. AMIA Annu Symp Proc 2022;2022:634-643 [FREE Full text] [Medline: 37128432]
39. Draeger C, Tute E, Schmidt CO, Waltemath D, Boeker M, Winter A, et al. Identifying relevant FHIR elements for data quality assessment in the German core data set. In: Hägglund M, Blusi M, Bonacina S, Nilsson L, Cort Madsen I, Pelayo S, et al, editors. Studies in Health Technology and Informatics. New York, NY: IOS Press; 2023:272-276.
40. Kamdje-Wabo G, Gradinger T, Löbe M, Lodahl R, Seuchter SA, Sax U, et al. Towards structured data quality assessment in the German medical informatics initiative: initial approach in the MII demonstrator study. Stud Health Technol Inform 2019 Aug 21;264:1508-1509. [doi: 10.3233/SHTI190508] [Medline: 31438205]
41. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016 Sep 11;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
42. Kapsner LA, Kampf MO, Seuchter SA, Kamdje-Wabo G, Gradinger T, Ganslandt T, et al. Moving towards an EHR data quality framework: the MIRACUM approach. Stud Health Technol Inform 2019 Sep 03;267:247-253. [doi: 10.3233/SHTI190834] [Medline: 31483279]
43. Kapsner LA, Mang JM, Mate S, Seuchter SA, Vengadeswaran A, Bathelt F, et al. Linking a consortium-wide data quality assessment tool with the MIRACUM metadata repository. Appl Clin Inform 2021 Aug 25;12(4):826-835 [FREE Full text] [doi: 10.1055/s-0041-1733847] [Medline: 34433217]
44. Mang JM, Seuchter SA, Gulden C, Schild S, Kraska D, Prokosch HU, et al. DQAgui: a graphical user interface for the MIRACUM data quality assessment tool. BMC Med Inform Decis Mak 2022 Aug 11;22(1):213 [FREE Full text] [doi: 10.1186/s12911-022-01961-z] [Medline: 35953813]
45. Tute E, Scheffner I, Marschollek M. A method for interoperable knowledge-based data quality assessment. BMC Med Inform Decis Mak 2021 Mar 09;21(1):93 [FREE Full text] [doi: 10.1186/s12911-021-01458-1] [Medline: 33750371]
46. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021 Apr 02;21(1):63 [FREE Full text] [doi: 10.1186/s12874-021-01252-7] [Medline: 33810787]
47. Buzug T, Handels H, Rostalski P, Mertins A, Müller S, Hübner C, et al. Student Conference Proceedings 2024: 13th Student Conference on Medical Engineering Science, 9th Student Conference on Medical Informatics, 7th ... Conference on Psychology - Cognitive Systems. New York, NY: Infinite Science Publishing; 2024.

Abbreviations

ATC: Anatomical Therapeutic Chemical
BfArM: Federal Institute for Drugs and Medical Devices
CDS: Core Data Set
DIC: data integration center
DQA: data quality assessment
ETL: extract, transfer, load
FDPG: German Portal for Medical Research Data
FHIR: Fast Healthcare Interoperability Resources
gICS: generic Informed Consent Service
HL7: Health Level 7
***ICD-10-GM*:** *German Modification of the International Statistical Classification of Diseases and Related Health Problems, 10th Revision*
LOINC: Logical Observation Identifiers Names and Codes
MDR: Metadata Repository
MII: Medical Informatics Initiative
MIRACUM: Medical Informatics in Research and Care in University Medicine
OPS: Operationen- und Prozedurenschlüssel
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
UCUM: Unified Code for Units of Measure

Edited by C Lovis; submitted 01.02.24; peer-reviewed by O Aassve, T Sagi; comments to author 05.04.24; revised version received 15.04.24; accepted 17.04.24; published 23.07.24.

Please cite as:
Rosenau L, Behrend P, Wiedekopf J, Gruendner J, Ingenerf J
Uncovering Harmonization Potential in Health Care Data Through Iterative Refinement of Fast Healthcare Interoperability Resources Profiles Based on Retrospective Discrepancy Analysis: Case Study
JMIR Med Inform 2024;12:e57005
URL: https://medinform.jmir.org/2024/1/e57005
doi:10.2196/57005
PMID:

Practical Aspects of Using Large Language Models to Screen Abstracts for Cardiovascular Drug Development: Cross-Sectional Study

Jay G Ronquillo[1], MPH, MMSc, MEng, MD; Jamie Ye[1], MPH; Donal Gorman[2], PhD; Adina R Lemeshow[1], PhD; Stephen J Watt[1], MD

1

2

Corresponding Author:
Jay G Ronquillo, MPH, MMSc, MEng, MD

Abstract

Cardiovascular drug development requires synthesizing relevant literature about indications, mechanisms, biomarkers, and outcomes. This short study investigates the performance, cost, and prompt engineering trade-offs of 3 large language models accelerating the literature screening process for cardiovascular drug development applications.

(JMIR Med Inform 2024;12:e64143) doi:10.2196/64143

KEYWORDS

biomedical informatics; drug development; cardiology; cardio; LLM; biomedical; drug; cross-sectional study; biomarker; cardiovascular; screening optimization; GPT; large language model; AI; artificial intelligence

Introduction

Cardiovascular drug development requires synthesizing information about indications, mechanisms, biomarkers, and outcomes [1,2]. Large language models (LLMs) leveraging billions of data points could accelerate fundamental, resource-intensive aspects of this process, like screening published literature [3]. However, this depends on the design, development, and implementation of LLM instructions (prompt engineering) that work effectively within the context of cardiology [4-6]. To our knowledge, this is one of the first studies investigating LLMs to accelerate the literature screening process for cardiovascular drug development applications [3,4,6,7].

Methods

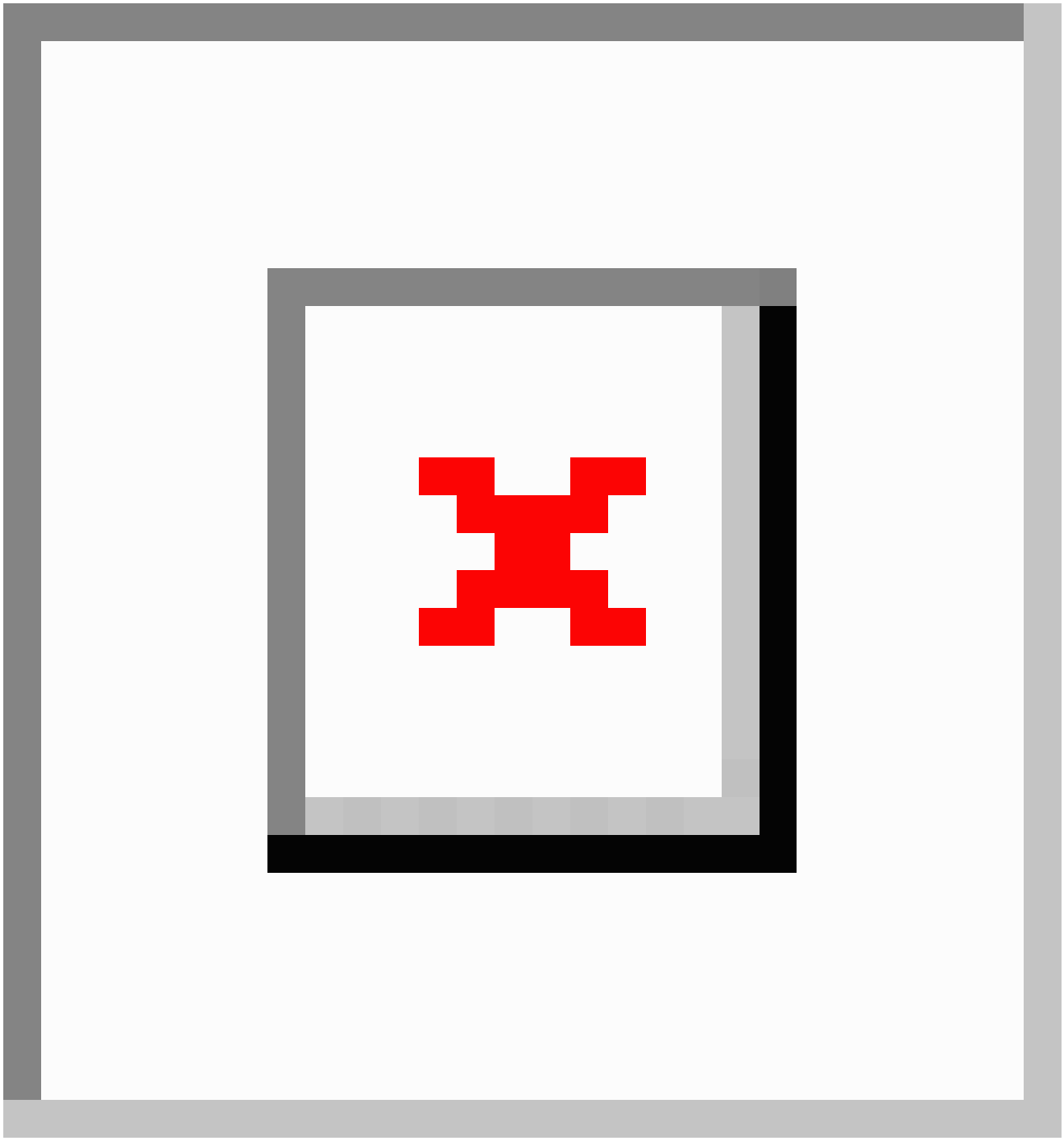
Study Design

Leveraging prior work, a PubMed query using both available Medical Subject Headings (MeSH) and the title and abstract keyword search of MeSH Entry Terms identified observational studies of heart failure that (1) were published from 2013 to 2023, (2) contained at least one relevant biomarker (brain natriuretic peptide, N-terminal pro–atrial natriuretic peptide, N-terminal pro–brain natriuretic peptide, and peak oxygen consumption), and (3) measured long-term outcomes (hospitalization and mortality) [2].

Abstracts were extracted through the PubMed application programming interface (API), and LLM instructions (prompts) were created to assess different screening optimization strategies (Figure 1) across LLMs (GPT-3.5 Turbo [OpenAI], GPT-4 [OpenAI], and Claude 2 [Anthropic PBC]) [5]. The “base” LLM prompt (1) presented abstract text, (2) listed two eligibility screening criteria (ie, values found for at least one biomarker and outcome), and (3) instructed LLMs to determine if abstracts met eligibility criteria and return results in a standardized format. “Technical” optimization was defined as adding delimiters to the base prompt delineating key sections (abstract and criteria), while “content” optimization further instructed LLMs to (1) assume a scientific role and (2) address a cardiology drug development target audience [3,5]. The different prompts used in this study are described in Multimedia Appendix 1. Total units of text processed (“tokens”) were estimated using spaCy, and LLM abstract screening costs were estimated using current API prices per million input and output tokens, respectively, for GPT-3.5 (US $0.50 and US $1.50), GPT-4 (US $30 and US $60), and Claude 2 (US $8 and US $24).

A Python script performed data processing and analysis. Accuracy was assessed by comparing LLM outputs against manual epidemiologist review of study suitability for inclusion, with descriptive statistics calculated for each LLM and prompt type. Performance differences between fully optimized prompts (GPT-3.5 vs GPT-4, GPT-3.5 vs Claude 2, and GPT-4 vs Claude 2) were evaluated using the chi-square test. A P value of <.05 was considered statistically significant.

Figure 1. Biomedical informatics pipeline for comparing different LLM and prompt optimization approaches to abstract screening for cardiovascular drug development. LLM: large language model.



Ethical Considerations

This study did not meet the definition of human participants research and thus did not require institutional review board approval.

Results

Of 69 articles found in PubMed, 32 (46%) met eligibility criteria after manual review; corresponding LLM screening accuracies are summarized in Table 1. By LLM, the best performances came from the base prompt (GPT-3.5), technical and combined prompts (GPT-4), and technical prompts (Claude 2). Overall, combined prompts for GPT-3.5 and GPT 4 performed similarly against each other (P>.99) and against Claude 2 (P=.61 against both).

GPT-3.5 processed a total of 124,826 tokens, while GPT-4 and Claude 2 processed 14.4% (N=142,750) and 15.9% (N=144,703) more tokens, respectively. Total costs for GPT-4 (US $4.89) and Claude 2 (US $1.52) were 75.4 and 23.4 times higher, respectively, than total costs for GPT-3.5 (US $0.06).

Table . Abstract screening accuracies reflecting total abstracts correctly identified by large language models (LLMs) for inclusion and exclusion based on manual review of study suitability, by LLM and prompt optimization type (abstracts: N=69).

Prompt optimization type	Accuracy, n (%)		
	GPT-3.5	GPT-4	Claude 2
Base (none)	43 (62)	40 (58)	35 (51)
Technical	34 (49)	41 (59)	43 (62)
Content	42 (61)	38 (55)	38 (55)
Technical and content	41 (59)	41 (59)	37 (54)

Discussion

Despite the complex and limited public cardiology data integrated into LLMs, our findings were consistent with similar studies for oncology and current LLM abilities to pass medical licensing exams [4,8]. Performance could be further improved by adding specific examples to the prompt (few-shot prompting) or to the LLM training data (fine-tuning) [4,8,9].

Technical optimizations showed modest performance improvements across some LLMs, indicating one practical way to improve accuracy and prompt readability without significantly expanding the size of input prompts. Standardizing outputs helped generate valid responses, although GPT-4 and Claude 2 still had higher costs as a result of more verbose output. Enterprise LLM–based abstract screening will require balancing prompt performance, cost, and complexity with cardiology subject matter expert capabilities and workflows.

Limitations include a small cardiovascular dataset leveraging proprietary LLMs and only a subset of available optimization techniques. Future efforts must engage diverse scientific communities; develop guardrails to ensure safe and responsible LLM use; and apply data-driven best practices that generalize, optimize, and validate LLM applications and their impact on patients with cardiovascular disease.

Conflicts of Interest

All authors are employees of Pfizer. The funding sources had no role in the design and conduct of this study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Multimedia Appendix 1
Approach for creating prompts focused on abstract screening for cardiovascular drug development, starting with the base prompt (black) and including content optimization (A) and technical optimization (B-E).
[PNG File, 404 KB - medinform_v12i1e64143_app1.png]

References

1. Boonstra MJ, Weissenbacher D, Moore JH, Gonzalez-Hernandez G, Asselbergs FW. Artificial intelligence: revolutionizing cardiology with large language models. Eur Heart J 2024 Feb 1;45(5):332-345. [doi: 10.1093/eurheartj/ehad838] [Medline: 38170821]
2. Wessler BS, Kramer DG, Kelly JL, et al. Drug and device effects on peak oxygen consumption, 6-minute walk distance, and natriuretic peptides as predictors of therapeutic effects on mortality in patients with heart failure and reduced ejection fraction. Circ Heart Fail 2011 Sep;4(5):578-588. [doi: 10.1161/CIRCHEARTFAILURE.111.961573] [Medline: 21705485]
3. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models. Pharmacoecon Open 2024 Mar;8(2):205-220. [doi: 10.1007/s41669-024-00476-9] [Medline: 38340277]
4. Ferber D, Wiest IC, Wölflein G, et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. NEJM AI 2024 May 17;1(6). [doi: 10.1056/AIcs2300235]
5. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. JMIR Med Inform 2024 Apr 8;12:e55318. [doi: 10.2196/55318] [Medline: 38587879]
6. Sharma A, Medapalli T, Alexandrou M, Brilakis E, Prasad A. Exploring the role of ChatGPT in cardiology: a systematic review of the current literature. Cureus 2024 Apr 24;16(4):e58936. [doi: 10.7759/cureus.58936] [Medline: 38800264]
7. Zaghir J, Naguib M, Bjelogrlic M, Neveol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review and recommendations for better practices. arXiv. Preprint posted online on May 2, 2024. [doi: 10.48550/arXiv.2405.01249]
8. Sahoo SS, Plasek JM, Xu H, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. J Am Med Inform Assoc 2024 Sep 1;31(9):2114-2124. [doi: 10.1093/jamia/ocae074] [Medline: 38657567]

9. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Presented at: 34th Conference on Neural Information Processing Systems (NeurIPS 2020); Dec 6-12, 2020; Vancouver, Canada URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [accessed 2024-09-18]

Abbreviations

API: application programming interface
LLM: large language model
MeSH: Medical Subject Headings

Edited by C Lovis; submitted 09.07.24; peer-reviewed by E Bilgin, J Zaghir; revised version received 29.08.24; accepted 01.09.24; published 30.09.24.

Please cite as:
Ronquillo JG, Ye J, Gorman D, Lemeshow AR, Watt SJ
Practical Aspects of Using Large Language Models to Screen Abstracts for Cardiovascular Drug Development: Cross-Sectional Study
JMIR Med Inform 2024;12:e64143
URL: https://medinform.jmir.org/2024/1/e64143
doi:10.2196/64143

Research Letter

Claude 3 Opus and ChatGPT With GPT-4 in Dermoscopic Image Analysis for Melanoma Diagnosis: Comparative Performance Analysis

Xu Liu[1*], BS; Chaoli Duan[1*], MS; Min-kyu Kim[1*], MS; Lu Zhang[1*], MD; Eunjin Jee[1], MS; Beenu Maharjan[1], MS; Yuwei Huang[1], MD; Dan Du[1], MS; Xian Jiang[1*], MD

Department of Dermatology, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Xian Jiang, MD
Department of Dermatology
West China Hospital
Sichuan University
No. 37, Guoxue Xiang
Wuhou District
Chengdu, 610041
China
Phone: 86 02885423315
Email: jiangxian@scu.edu.cn

Abstract

Background: Recent advancements in artificial intelligence (AI) and large language models (LLMs) have shown potential in medical fields, including dermatology. With the introduction of image analysis capabilities in LLMs, their application in dermatological diagnostics has garnered significant interest. These capabilities are enabled by the integration of computer vision techniques into the underlying architecture of LLMs.

Objective: This study aimed to compare the diagnostic performance of Claude 3 Opus and ChatGPT with GPT-4 in analyzing dermoscopic images for melanoma detection, providing insights into their strengths and limitations.

Methods: We randomly selected 100 histopathology-confirmed dermoscopic images (50 malignant, 50 benign) from the International Skin Imaging Collaboration (ISIC) archive using a computer-generated randomization process. The ISIC archive was chosen due to its comprehensive and well-annotated collection of dermoscopic images, ensuring a diverse and representative sample. Images were included if they were dermoscopic images of melanocytic lesions with histopathologically confirmed diagnoses. Each model was given the same prompt, instructing it to provide the top 3 differential diagnoses for each image, ranked by likelihood. Primary diagnosis accuracy, accuracy of the top 3 differential diagnoses, and malignancy discrimination ability were assessed. The McNemar test was chosen to compare the diagnostic performance of the 2 models, as it is suitable for analyzing paired nominal data.

Results: In the primary diagnosis, Claude 3 Opus achieved 54.9% sensitivity (95% CI 44.08%-65.37%), 57.14% specificity (95% CI 46.31%-67.46%), and 56% accuracy (95% CI 46.22%-65.42%), while ChatGPT demonstrated 56.86% sensitivity (95% CI 45.99%-67.21%), 38.78% specificity (95% CI 28.77%-49.59%), and 48% accuracy (95% CI 38.37%-57.75%). The McNemar test showed no significant difference between the 2 models (P=.17). For the top 3 differential diagnoses, Claude 3 Opus and ChatGPT included the correct diagnosis in 76% (95% CI 66.33%-83.77%) and 78% (95% CI 68.46%-85.45%) of cases, respectively. The McNemar test showed no significant difference (P=.56). In malignancy discrimination, Claude 3 Opus outperformed ChatGPT with 47.06% sensitivity, 81.63% specificity, and 64% accuracy, compared to 45.1%, 42.86%, and 44%, respectively. The McNemar test showed a significant difference (P<.001). Claude 3 Opus had an odds ratio of 3.951 (95% CI 1.685-9.263) in discriminating malignancy, while ChatGPT-4 had an odds ratio of 0.616 (95% CI 0.297-1.278).

Conclusions: Our study highlights the potential of LLMs in assisting dermatologists but also reveals their limitations. Both models made errors in diagnosing melanoma and benign lesions. These findings underscore the need for developing robust, transparent, and clinically validated AI models through collaborative efforts between AI researchers, dermatologists, and other health care professionals. While AI can provide valuable insights, it cannot yet replace the expertise of trained clinicians.

(JMIR Med Inform 2024;12:e59273) doi:10.2196/59273

KEYWORDS

artificial intelligence; AI; large language model; LLM; Claude; ChatGPT; dermatologist

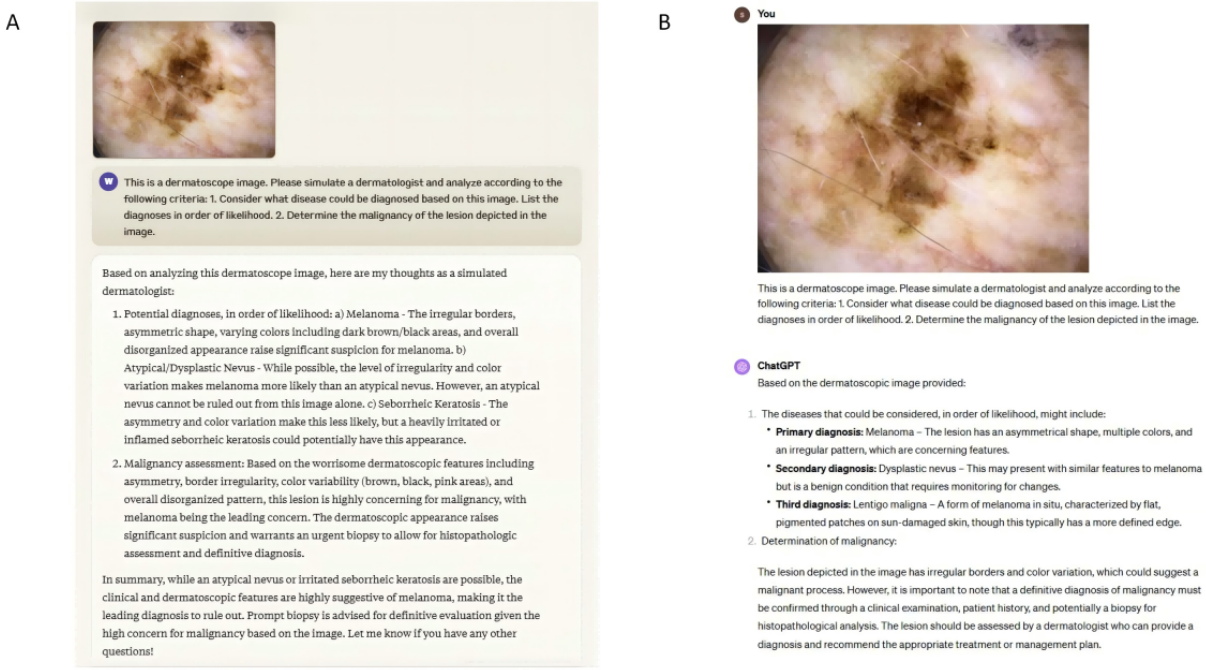
Introduction

Artificial intelligence (AI) has shown great promise in various medical fields, including dermatology [1]. The emergence of large language models (LLMs), such as ChatGPT, has demonstrated their ability to generate human-like responses and assist in clinical decision-making [2]. With the introduction of image analysis capabilities in LLMs [3], their application in dermatological diagnostics has attracted significant interest. Claude 3 Opus, an advanced conversational AI model, has shown promising performance in various natural language processing tasks [4]. This study aims to fill that gap by comparing Claude 3 Opus and ChatGPT with GPT-4. Despite the potential of AI in health care, encompassing diagnosis, treatment, and public health initiatives, these technologies are largely underused in clinical practice [5]. Moreover, the use of AI in health care raises important legal and ethical considerations, particularly for "high-risk" applications such as medical diagnosis [6]. In this context, comparing the diagnostic performance of Claude 3 Opus and ChatGPT can provide valuable insights into their strengths and limitations, guiding the selection and optimization of AI-assisted diagnostic tools in dermatology while taking into account the regulatory landscape.

Methods

We randomly selected 100 dermoscopic images (50 malignant melanomas, 50 benign nevi) from the International Skin Imaging Collaboration (ISIC) archive [7] using a computer-generated randomization process to avoid selection bias. The ISIC archive was chosen due to its comprehensive and well-annotated collection of dermoscopic images, which ensures a diverse and representative sample. Images were included if they were dermoscopic images of melanocytic lesions with histopathologically confirmed diagnoses. Each image was presented to Claude 3 Opus and ChatGPT. The models were given the same prompt, instructing them to provide the top 3 differential diagnoses for each image, ranked by likelihood. The exact prompt was "Please provide the top 3 differential diagnoses for this dermoscopic image, ranked by likelihood. Focus on distinguishing between melanoma and benign nevi." The models' responses were recorded for analysis (Figure 1A, B). We assessed primary diagnosis accuracy, accuracy of the top 3 differential diagnoses, and malignancy discrimination ability. The McNemar test was used to compare the models' performance.

Figure 1. Performance comparison of Claude 3 Opus and ChatGPT with GPT4-Vision in skin dermoscopy image analysis and melanoma diagnosis: application scenarios. (A.) Application scenario of Claude 3 Opus in the analysis process of dermoscopic images. (B) Application scenario of GPT4-Vision in the analysis process of dermoscopic images.



Results

For the primary diagnosis, Claude 3 Opus achieved 54.9% sensitivity (95% CI 44.08%-65.37%), 57.14% specificity (95% CI 46.31%-67.46%), and 56% accuracy (95% CI 46.22%-65.42%), while ChatGPT demonstrated 56.86% sensitivity (95% CI 45.99%-67.21%), 38.78% specificity (95% CI 28.77%-49.59%), and 48% accuracy (95% CI 38.37%-57.75%). The McNemar test showed no significant difference between the 2 models (P=.17; Multimedia Appendices 1-2). For the top 3 differential diagnoses, Claude 3 Opus and ChatGPT included the correct diagnosis in 76% (95% CI 66.33%-83.77%) and 78% (95% CI 68.46%-85.45%) of cases, respectively. The McNemar test showed no significant difference (P=.56). In malignancy discrimination, Claude 3 Opus outperformed ChatGPT with 47.06% sensitivity, 81.63% specificity, and 64% accuracy, compared to 45.1%, 42.86%, and 44%, respectively. The McNemar test showed a significant difference (P<.001). Claude 3 Opus had an odds ratio of 3.951 (95% CI 1.685-9.263) in discriminating malignancy, while ChatGPT had an odds ratio of 0.616 (95% CI 0.297-1.278) (Multimedia Appendix 3).

Discussion

Our study demonstrates the potential of LLMs in assisting dermatological diagnosis, while also revealing their current limitations. Claude 3 Opus showed superior performance in discriminating between malignant and benign lesions compared to ChatGPT. However, both models made errors in diagnosing melanoma and nevi. For example, Claude 3 Opus misdiagnosed several melanomas as benign lesions, while ChatGPT had a higher false positive rate, misclassifying many nevi as melanomas. These findings highlight the need for further development and rigorous clinical validation of AI diagnostic tools before their widespread implementation in dermatology practice (Multimedia Appendix 4). Future research should focus on improving the robustness and interpretability of these models through close collaboration between AI researchers, dermatologists, and other health care stakeholders. Moreover, the potential impact of AI in health care extends beyond technical performance, encompassing legal and ethical dimensions. The European Commission has already proposed legislation for “high-risk” AI applications, providing a framework for the safe and responsible use of medical AI [6]. As LLMs and other AI tools continue to advance, it is crucial to proactively address these regulatory aspects to ensure their beneficial integration into clinical practice. In conclusion, while LLMs such as Claude 3 Opus and ChatGPT show promise in assisting dermatological diagnosis, they are not yet capable of replacing human expertise. Continued research, collaborative development, and proactive regulation are essential for realizing the full potential of AI in dermatology while prioritizing patient safety and ethical standards.

Acknowledgments

This study received funding from the National Natural Science Foundation of China (82273559, 82304052, and 82073473) and the 1·3·5 Project for Disciplines of Excellence, West China Hospital, Sichuan University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Claude 3 Opus and GPT4-Vision for the analysis of dermoscopic images and diagnosis of melanoma based on data comparison.
[DOCX File , 17 KB - medinform_v12i1e59273_app1.docx]

Multimedia Appendix 2
McNemar test results for the performance differences between Claude 3 Opus and GPT4-Vision.
[DOCX File , 17 KB - medinform_v12i1e59273_app2.docx]

Multimedia Appendix 3
Odds ratio and 95% CIs for Claude 3 Opus and GPT4-Vision for their ability to distinguish between benign and malignant conditions.
[DOCX File , 16 KB - medinform_v12i1e59273_app3.docx]

Multimedia Appendix 4
(A) Examples of Claude 3 Opus misdiagnosing melanoma as benign lesions. (B) Examples of Claude 3 Opus misdiagnosing benign lesions as melanoma. (C) Examples of GPT4-Vision misdiagnosing melanoma as benign lesions. (D) Examples of GPT4-Vision misdiagnosing benign lesions as melanoma.
[PDF File (Adobe PDF File), 2428 KB - medinform_v12i1e59273_app4.pdf]

References

1. Gomolin A, Netchiporouk E, Gniadecki R, Litvinov IV. Artificial intelligence applications in dermatology: where do we stand? Front Med (Lausanne) 2020;7:100 [FREE Full text] [doi: 10.3389/fmed.2020.00100] [Medline: 32296706]
2. Rundle CW, Szeto MD, Presley CL, Shahwan KT, Carr DR. Analysis of ChatGPT generated differential diagnoses in response to physical exam findings for benign and malignant cutaneous neoplasms. J Am Acad Dermatol 2024 Mar;90(3):615-616. [doi: 10.1016/j.jaad.2023.10.040] [Medline: 37898341]
3. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. J Am Acad Dermatol 2024 May;90(5):1057-1059 [FREE Full text] [doi: 10.1016/j.jaad.2023.12.062] [Medline: 38244612]
4. Introducing Claude. Anthropic. URL: https://www.anthropic.com/index/introducing-claude [accessed 2023-10-20]
5. Pagallo U, O'Sullivan S, Nevejans N, Holzinger A, Friebe M, Jeanquartier F, et al. The underuse of AI in the health sector: opportunity costs, success stories, risks and recommendations. Health Technol (Berl) 2024;14(1):1-14 [FREE Full text] [doi: 10.1007/s12553-023-00806-7] [Medline: 38229886]
6. Stöger K, Schneeberger D, Holzinger A. Medical artificial intelligence: the European legal perspective. Commun ACM 2021 Oct 25;64(11):34-36. [doi: 10.1145/3458652]
7. The International Skin Imaging Collaboration (ISIC) archive. URL: https://www.isic-archive.com [accessed 2023-10-19]

Abbreviations

AI: artificial intelligence
ISIC: International Skin Imaging Collaboration
LLM: large language model

Edited by A Castonguay; submitted 08.04.24; peer-reviewed by GK Gupta, A Holzinger; comments to author 27.06.24; revised version received 28.06.24; accepted 18.07.24; published 06.08.24.

Please cite as:
Liu X, Duan C, Kim MK, Zhang L, Jee E, Maharjan B, Huang Y, Du D, Jiang X
Claude 3 Opus and ChatGPT With GPT-4 in Dermoscopic Image Analysis for Melanoma Diagnosis: Comparative Performance Analysis
JMIR Med Inform 2024;12:e59273
URL: https://medinform.jmir.org/2024/1/e59273
doi:10.2196/59273
PMID:

Viability of Open Large Language Models for Clinical Documentation in German Health Care: Real-World Model Evaluation Study

Felix Heilmeyer[1]; Daniel Böhringer[2], Dr med; Thomas Reinhard[2], Dr med; Sebastian Arens[2], Dr med; Lisa Lyssenko[1], Dr sc hum; Christian Haverkamp[1], Dr med

[1]

[2]

Corresponding Author:
Felix Heilmeyer

Abstract

Background: The use of large language models (LLMs) as writing assistance for medical professionals is a promising approach to reduce the time required for documentation, but there may be practical, ethical, and legal challenges in many jurisdictions complicating the use of the most powerful commercial LLM solutions.

Objective: In this study, we assessed the feasibility of using nonproprietary LLMs of the GPT variety as writing assistance for medical professionals in an on-premise setting with restricted compute resources, generating German medical text.

Methods: We trained four 7-billion–parameter models with 3 different architectures for our task and evaluated their performance using a powerful commercial LLM, namely Anthropic's Claude-v2, as a rater. Based on this, we selected the best-performing model and evaluated its practical usability with 2 independent human raters on real-world data.

Results: In the automated evaluation with Claude-v2, BLOOM-CLP-German, a model trained from scratch on the German text, achieved the best results. In the manual evaluation by human experts, 95 (93.1%) of the 102 reports generated by that model were evaluated as usable as is or with only minor changes by both human raters.

Conclusions: The results show that even with restricted compute resources, it is possible to generate medical texts that are suitable for documentation in routine clinical practice. However, the target language should be considered in the model selection when processing non-English text.

(JMIR Med Inform 2024;12:e59617) doi:10.2196/59617

KEYWORDS

machine learning; ML; artificial intelligence; AI; large language model; large language models; LLM; LLMs; natural language processing; NLP; deep learning; algorithm; algorithms; model; models; analytics; practical model; practical models; medical documentation; writing assistance; medical administration; writing assistance for physicians

Introduction

Background

Physicians are often overloaded with documentation requirements, including writing a doctor's note and a summary of a patient's visit. An analysis of clinical software log files showed that interaction with electronic health records (EHRs) constitutes a large portion of physicians' daily work, approximately one-fourth of which is spent writing documentation [1]. Completion of the documentation in the EHR is perceived as a tedious task, which is often done after work hours [1]. More time spent on documentation after work hours has been shown to be associated with burnout and decreased work-life satisfaction [2].

A promising approach to reduce the time required for documentation is the use of writing assistance based on large language models (LLMs). In a feasibility study, the authors trained previous-generation LLMs (GPT-2 and GPT-Neo) to complete text in medical records [3]. They concluded that the models could be used in medical charting but still have some room for improvement. A large source of error was abrupt changes in the topic, which is common in the documentation of EHRs.

With recent advances in LLM technology and the release of ChatGPT, LLMs have seen widespread adoption in assisting professionals produce text for communication or documentation purposes. For example, under the Copilot brand, Microsoft is building generative artificial intelligence (AI) capabilities into their widely used Office application suite to assist in business

use cases. This leads us to believe that current-generation LLMs could also provide valuable assistance in the health care sector.

Challenges in the Use of LLMs in the Health Care Sector

Among the best-performing LLMs, according to the continuously updated Holistic Evaluation of Language Models [4] at Stanford University, are currently commercial offerings from companies such as OpenAI or Anthropic. With these offerings, the models run on the providers' infrastructure and are accessible via an application programming interface. However, these services cannot be used in a clinical context without further consideration.

First, in many countries, the services do not meet the legal requirements for processing protected health information. In some jurisdictions, legal and regulatory frameworks mandate that data originating from health care providers must be processed within the country's borders or even on-premise. This is particularly problematic for European countries, as the European Union's General Data Protection Regulation prohibits the transfer of protected health information to data centers in the United States, where most providers are located.

Second, clinical software must be thoroughly validated before it is released to end users, and in some cases, it is even subject to the Medical Device Regulation. This conflicts with the update policy of providers of commercial AI solutions. The scope of model updates is usually communicated only a few weeks in advance, for example, 2 weeks in the case of OpenAI [5]. This would not be a problem if these updates were only additive in functionality, but the opaque nature of current LLMs also means that improvements to some aspects of model performance might unexpectedly negatively affect the performance on other tasks [6]. The use of fixed model versions, as offered by some providers, is not practicable in the long term, as older models are often removed after the release of updates; in the case of OpenAI, after 3 months [5].

Training Nonproprietary AI Models for Medical Text

An alternative is the use of nonproprietary AI models. In these models, the architecture as well as the trained parameters are available to the user. This solves the aforementioned problems by giving the user the option to train and deploy these models on any infrastructure and fully control any changes to it.

One of the largest pretrained LLMs is GPT models that enable model training with limited data sets. There are several approaches to applying GPT models to a task. One common approach is to use a very large model that is trained primarily with general text corpora and includes instructions for the task in the input for the model, the so-called prompt. This is sometimes called in-context learning (ICL) or, depending on whether examples are provided, zero-shot or few-shot learning.

ICL works reasonably well on tasks that have a good representation in the base models' training corpus. However, the structure and content of clinical notes differ significantly from the general-purpose text corpora used to train most publicly available LLMs. Even including biomedical text from publications, such as PubMed papers, in the training data could only have minor effects on model performance compared to training on clinical text [7-9]. Lehman et al [9] compared ICL and multiple alternatives such as (1) training from scratch on a clinical corpus, (2) continuing training a pretrained model on the clinical text and then fine-tuning for the downstream task, or (3) directly training the GPT for the downstream task without further pretraining. They show that relatively small specialized clinical models substantially outperform all ICL approaches and conclude that pretraining on clinical text allows for smaller, more parameter-efficient models.

One fact that must be taken into account when using GPT models in a clinical context is that the pretrained models have now become very large. Complete fine-tuning, in which all model parameters are retrained on the task-specific data, is therefore becoming less and less feasible. This is particularly the case if the models have to be trained on site for legal or economic reasons. The computing power available here is usually limited, which restricts the size of the models that can be trained. Accordingly, the choice of models is a trade-off between training time and costs, model accuracy, and maximum sequence length.

One possibility to address the problem of limited working memory is the Low-Rank Adaptation (LoRA) technique [10]. Here, all the model weights are frozen, and only a few very small additional low-rank matrices are added to the query and key parameter matrices of the transformer attention heads and subsequently optimized. This reduces the number of trainable parameters by 10,000 times and the graphics processing unit (GPU) memory requirement by 3 times. Recently, training of quantized models became possible by combining LoRA with quantization [11]. With Quantized Low-Rank Adaptation (QLoRA), a frozen quantized model is fine-tuned by optimizing added low-rank adapters at 16-bit floating-point precision. The QLoRA technique also introduced additional memory-saving mechanisms such as the 4-bit Normal Float (NF4) data type for quantization and paged optimizers [11].

Aim of This Study

In this study, we assessed the feasibility of using nonproprietary LLMs of the GPT variety as writing assistance for medical professionals in an on-premise setting with restricted compute resources, generating non-English medical text. We trained 4 models with 3 different architectures for our task using the Hugging Face Transformers framework [12] and explored their performance using a powerful commercial LLM, namely Anthropic's Claude-v2, as a rater. Based on this, we selected the best-performing model and evaluated its practical usability with 2 independent human raters on real-world data.

Methods

Ethical Considerations

The study was implemented in the outpatient clinic of the Eye Center at Medical Center, University of Freiburg, Germany, and was approved by the responsible ethical review committee (registration 23 - 1444S1). All data used in the study were deidentified and contained no references to patients and practitioners. Informed consent for the anonymization process

was not obtained. Data processing was justified based on the legal basis of "legitimate interest" in accordance with the General Data Protection Regulation and the state hospital law of Baden-Württemberg ("Landeskrankenhausgesetz"). Participants did not receive any financial compensation for their data use, as the data were retrospectively reviewed from existing medical records.

Study Design

The target for assistive text generation was the final part of the medical documentation of an examination or treatment, the so-called epicrisis report. In this report, the doctors write a structured compilation of the information so far documented in the EHR in text form. It contains the relevant medical information of the case and usually consists of three sections: (1) main diagnosis or the patient's reason for visit, (2) therapeutic procedures or medication, and (3) recommendations for further intervention and need for a follow-up appointment.

Data

Data Source and Description

The data pool used for training the models was the EHR records of 82,482 unique patient encounters that span approximately 10 years of clinical practice. The EHR record of an encounter contains all digital information about a patient's examination or treatment in the outpatient clinic, which offers specialist, emergency, and follow-up care. The data are collected in various ways over the patient's visit. Support staff record basic information in structured forms, doctors document the medical history, symptoms and previous or planned treatments are documented in text notes, and diagnostic data from electronic devices are mainly stored in numeric format. The final epicrisis report consists of a stand-alone text, which is filed alongside all other information in the EHR record.

The whole training data set amounts to approximately 140 MB of uncompressed text in Unicode Transformation Format 8 encoding or approximately 29 to 33 million tokens, depending on the tokenizer model used. A data set of 509 patient encounters that occurred after the training set date cutoff was set aside for comparison of model performance in the evaluation. The complete data set consists of German text. The examples used in this paper were translated into English by the authors of the paper.

Preprocessing and Formatting

For the LLM training, all available data in the EHR record were concatenated into 1 continuous text sequence per encounter. The types of information were separated by newlines and prefixed with a descriptor such as "History" or "Pressure Measurement" to form the prompt. If no data were documented in a section, it was left empty. The order of sections matched the order in which the fields are displayed to users in the EHR software interface. The last section of each text sequence was the epicrisis report. If there were separate records for each eye, the individual records were additionally prefixed with an abbreviation indicating the side.

Task training was implemented by inserting special tokens to mark the text to be generated by the final models, that is, the epicrisis report. Each text sequence starts with a special token indicating the beginning of the input data recorded during the patient visit, that is, all other information in the EHR record. A second special token is inserted before the epicrisis report, indicating the start of the generation task. In the training data, this token is followed by the actual report of the attending physician. The text sequence ends with a "Stop of Sequence" token, which indicates that the model should end the generation process.

For instruction-tuned models, the text sequence was prefixed with a so-called system message enclosed in special tokens indicating instructions for the model, reading as follows: "You are an experienced doctor in a German eye hospital. Your writing style is concise, accurate, and respectful. You are writing a short note in German to a colleague about a patient. The letter should contain the provided information."

Models

Model Selection Overview

In the selection of models from openly available pretrained models, we considered hardware costs, feasibility of the training process, language aspects, and performance benchmark results, such as Stanford's Holistic Evaluation of Language Models [4] and the Open LLM Leaderboard on Hugging Face [13]. Most LLMs are predominantly trained in English texts, and currently, there is no model that contains a greater amount of medical text. Consequently, we chose the following 3 models: LLaMA, LLaMA-2-Chat, and BLOOM-CLP-German.

LLaMA

At the start of this study, Meta AI's LLaMA model was among the top performers on several open LLM benchmarks. In contrast to some of its competitors, its training corpus also contains some German text but no clinical content [14]. Since then, more powerful models have been released, but LLaMA still achieves competitive results on many benchmarks.

LLaMA-2-Chat

During our experiments, Meta AI released the successor to LLaMA [15]. Together with the updated base model, they also released an instruction-tuned model aligned with human preferences using reinforcement learning, similar to how ChatGPT was based on GPT-3 [16]. We chose this model to investigate the potential advantage of using an instruction-tuned model.

BLOOM-CLP-German

This model is designed for tasks in German based on the BLOOM architecture from BigScience Workshop et al [17]. It was initialized with the novel cross-lingual and progressive transfer learning (CLP) technique [18], which uses information from a small model trained in a target language and a larger model in a source language. This considerably reduces the training needed to achieve performance on par with that of a model trained from scratch. Although the model is still severely undertrained for its size [19], we included it to study the potential performance gains achieved by a model with a training corpus closer to the target text material.

Training

Overview

We restrict our training setup to 8x NVIDIA RTX 3090 24-GiB consumer-grade GPUs in a single host. We load and train our models using the "transformers" Python library by Hugging Face [12] with the PyTorch [20] backend. Data are preprocessed using Hugging Face's "datasets" Python library [21]. Distributed training on multiple GPUs is implemented via the "accelerate" Python library [22].

For each training process, we randomly sample 5% of the training data as validation data. We regularly evaluate training loss on the validation set during training, about 20 times per epoch. We stop training when the validation loss does not improve in 10 evaluation steps. This amounts to around 13 epochs for most models.

Memory Optimization

For fine-tuning the model for our task, we use the LoRA at full 16-bit precision and QLoRA [11] at reduced NF4 precision techniques. Reducing the precision also reduces the memory use and allows for longer input text sequences with the available memory. With this, we explore the trade-off between computational precision and input context size.

Additionally, we use 2 methods to trade reduced memory requirements for computation time. First, we use gradient checkpointing, a technique that recomputes some network activations during the backward pass on the fly instead of caching them in memory. Second, we use the Zero Redundancy Optimizer technique [23], which includes memory savings achieved by reducing redundancy when training on multiple GPUs as well as offloading some tasks to the CPU, both at the cost of communication overhead. Both make the training process considerably slower but should not impact the task performance of the resulting model.

Specifically, we trained the following model variants: LLaMA with LoRA at floating point 16-bit precision, LLaMA 2 Chat with QLoRA at NF4 precision, BLOOM-CLP German with QLoRA at NF4 precision, and BLOOM-CLP German with LoRA at floating point 16-bit precision.

Inference

Overview

At their core, the decoder part of the transformer architecture models a probability distribution for the next token, given a sequence of input tokens. Both the composition of the initial input tokens and the method of choosing the next token from the produced probability distribution can have a big impact on the quality of the final result.

Completion Prefixing

At inference time, the model receives an input text sequence, often called the prompt. It consists of the input data, as described in the "Preprocessing and Formatting" section, followed by a special token, indicating that the subsequent text should be an epicrisis report. In other words, the model receives a text sequence containing all information from an EHR record except for the attending physician's epicrisis report and is asked to write this report, that is, to generate a text that corresponds in content, structure, and form to the epicrisis reports included in the training data. However, in the qualitative analysis of our initial findings, we found that in some cases the models attempted to continue with the recorded data rather than start writing a final report.

In an effort to improve results without retraining our models, we introduce a simple form of prompt tuning by adding a static suffix to the prompt, that is, forcing the model to begin the generated text with the words "During today's visit...." This suffix represents the typical beginning of the epicrisis report, as almost all reports written by doctors in the training data set start with some variation of these words. We hope that this gives the models an additional signal to complete the text with a summary and recommendations instead of trying to invent more "facts" about the patient's stay. We report and compare the evaluation results on reports generated with and without the completion prefix.

Contrastive Search

For a given input sequence, the trained transformer model produces a probability distribution for the next token. Simply choosing the token with the highest probability often produces text that lacks coherence and diversity. Techniques that maximize the probability over multiple tokens (eg, beam search) or stochastic sampling can enhance coherence and diversity but are not targeted at the problem of repetition that is common to the type of highly standardized text generated in this study. We therefore use a more recently introduced technique, called contrastive search, which has been shown to encourage diversity and produce coherent results while reducing repetitiveness [24,25].

Evaluation

Overview

Evaluating the quality of generated natural language text using (preferably multiple) human raters is costly and time-consuming, especially, if the rating process requires specialized domain knowledge as in this study. On the other hand, there is no obvious way to automate this process. An interesting idea is to use larger and more powerful language models to rate the quality of the output. This technique has recently been used in some publications in the LLM space, for example, in the creation of the LLaMA-2 model and in evaluating the performance of QLoRA training [11,15]. Large commercial language models such as OpenAI's GPT-4 and Anthropic's Claude-v1 model have been shown to achieve agreement rates with human raters of up to 80% when evaluating the output of other models [26].

Automated Evaluation With Claude-v2

We evaluate the generated text in a 2-step process using Claude-v2 by comparing the generated text to the epicrisis reports that were written by physicians for 509 individual patient encounters. In the first step, we extract the text passages that contain relevant information for each of the three main categories of information: (1) main diagnosis or patient's reason for visit, (2) therapeutic procedures or medication, and (3)

recommendations for further intervention and need for a follow-up appointment. In the second step, for each case and category separately, we ask Claude to evaluate whether the extracted passage from the generated report matches the passage extracted from the report written by a human.

Human Evaluation

The suitability of the generated text by the best-performing model is evaluated by 2 independent expert senior physicians. For this purpose, the raters are presented with the basic data from the documentation of 102 patients as well as both versions of the report: the one written by the attending physician and the computer-generated version. The raters assess whether the computer-generated version is suitable as a text template and could be used without major changes.

Results

Model Performance

Table 1 shows the percentage of reports in the test set in which the models matched the extracted diagnosis, follow-up, and therapy recommendation. The highest agreement rates with reports written by a doctor were achieved by the BLOOM-CLP-German model, followed by LLaMA-2 and LLaMA. The ranking was consistent across all the diagnosis, follow-up, and therapy dimensions. On average, the models achieved the highest scores in the diagnosis dimension, followed by the therapy and follow-up dimensions.

Of the BLOOM-CLP-German variants trained with full floating point 16-bit precision LoRA and reduced NF4 integer precision QLoRA, the latter achieved slightly higher agreement rates. In contrast to our intuition, prefixing the model prompt at inference time (see "Completion Prefixing" section) slightly reduced the performance across all models rather than improving it.

Table . Fraction of reports in the test set where the models match the information extracted from the text written by a doctor.

Category	Model (%)								
	BLOOM-CLP-FP16	BLOOM-CLP-FP16-prefix	BLOOM-CLP-QLoRA	BLOOM-CLP-QLoRA-prefix	LLaMA-2-QLoRA	LLaMA-2-QLoRA-prefix	LLaMA-FP16	LLaMA-FP16-prefix	Mean (SD)
Diagnosis	50.10	44.01	55.40	45.78	34.38	32.81	31.24	16.50	38.78 (12.47)
Follow-up	41.45	32.02	43.42	33.79	36.94	31.24	34.97	21.02	34.36 (6.90)
Therapy	43.81	37.33	50.69	42.44	36.54	36.35	29.67	13.95	36.35 (10.99)
Mean (SD)	45.12 (4.47)	37.79 (6.01)	49.84 (6.04)	40.67 (6.01)	35.95 (1.38)	33.46 (2.62)	31.96 (2.72)	17.16 (3.58)	—[a]

[a]Not applicable.

Human Evaluation

A total of 102 reports generated by the BLOOM-CLP German model trained with QLoRA at NF4 precision were rated for suitability by 2 independent expert senior physicians. Of the 102 reports, 95 (93.1%) were evaluated as suitable by both raters, which means that computer-generated reports could be used in this form or with minor changes. Only 7 (6.9%) of the reports were rated as unsuitable by at least 1 of the raters. Cohen κ was run to determine the interrater reliability. There was moderate agreement between the 2 physicians' judgments (κ=0.582, 95% CI 0.217-0.947; P<.001).

The 7 reports that were rated as unsuitable show different anomalies. In 3 of the reports, the model was caught in a loop of repeating nonsensical word sequences, for example, "we recommend local therapy with Bepanthen eye ointment 5x daily on both sides for 5 - 7 days, then 1x daily on both sides for 5 - 7 days, then 1x daily on both sides for 5 - 7 days, then 1x daily on both sides for 5 - 7 days, etc" (26 repetitions). In one case, there is no text output because the patient's appointment did not take place. Only in 3 reports are content-related aspects decisive. In one case, the main diagnosis is not mentioned; in one case, information is missing in the treatment recommendation; and in one case, the time given for the follow-up appointment is incorrect.

Discussion

Principal Findings

Despite being severely undertrained compared to both LLaMA models, the BLOOM-CLP-German model achieved the best performance in our experiments. This suggests that a better alignment of the base model with the reports' language might be more important than a longer training time. We speculate that the German vocabulary in the model's tokenizer better-captured domain semantics compared to the multilingual tokenizers. Additionally, the model might have profited from a larger maximum input sequence length, given the limited memory. This is an effect of the smaller token per character ratio of a tokenizer with a better alignment to the text's language.

Because its vocabulary is closer to our data, BLOOM-CLP-German's tokenizer encodes up to 30% fewer tokens for the same input text compared to LLaMA's tokenizer. This means that we can fit more information into the context window, training and inference consume about half as much memory, and inference is about twice as fast. This makes for

significant cost reductions compared to models with a multilanguage tokenizer.

Of both BLOOM variants trained with LoRA and reduced QLoRA precision, the latter performed better in our analysis. This suggests that the reduced precision is more than offset by the bigger maximum input sequence length, given the memory constraints. We surmise that capturing more context in the model input outweighs compute-optimal training or precision.

In contrast to our intuition, forcing the models to start the generated text with a predefined prefix did not improve the results. We speculated from our manual testing that this technique might eliminate some edge cases where the models sometimes start generating text completely unrelated to the input sequence. While this might still be true, the prefix also might have impacted the models' ability to flexibly react to the input and therefore reduced quality in more cases than improving it.

Feasibility of Nonproprietary On-Site AI

Our manual evaluation clearly shows that it is possible to provide helpful writing assistance using nonproprietary on-site AI technologies. Most of our test samples were rated useful as is or with only minor modifications. Additionally, qualitative analysis of samples rated as unusable showed that these were edge cases where the model produced no output or text that was easily identifiable as an anomaly. Only in very few reports were content-related aspects decisive, that is, the model omitted major details or produced factually incorrect information.

Legal and ethical concerns, as discussed in the "Challenges in the Use of LLMs in the Health Care Sector" section, currently may prevent many health care providers in European countries from using proprietary AI assistance for charting. Nonproprietary models, as used in this study, allow for flexible model deployment to comply with data protection requirements. Full control over the model also addresses legal concerns regarding software certification and some ethical concerns because these models can be more easily inspected regarding potential biases. Therefore, the approach presented in this study should be feasible for most health care providers.

In this study, we chose model sizes around 7 billion parameters. In comparison, GPT-3, the model that powered the first version of ChatGPT, has 175 billion parameters. With careful optimization of trade-offs between training time and cost, model precision, and maximum sequence length, we show that it is still possible to provide helpful writing assistance even with a much smaller model. At our chosen model scale, with around 7 billion parameters operating, the models should be economically accessible to many health care providers or local service providers, making it easier to comply with local regulations and reducing possible dependence on external or foreign service providers.

Limitations

Due to the limited availability of compute time, we were unable to test all combinations of model and training modalities. LLaMA-2 was only trained using QLoRA, and LLaMA only using LoRA, limiting possible comparisons between the base models. Similarly, we only included the instruction-tuned variant of LLaMA-2 and cannot compare to the base model without instruction tuning.

The limited training of the BLOOM model probably affected its accuracy. However, this limitation highlights the importance of language alignment with the undertrained BLOOM model, outperforming both LLaMA models.

While our human raters evaluated our chosen model's outputs favorably, this happened in dedicated research settings. This means that the contextual information available to the human raters was restricted to the limited information included in the study data set. It remains to be shown whether AI writing assistance is still perceived as helpful in a real clinical setting or if the additional mental load caused by having to check the AI's output in a more complex case outweighs its usefulness.

Future Work

Moving forward, leveraging German clinical corpora for pretraining could provide useful in-domain semantics. Techniques such as CLP fine-tuning can enable the use of such data with lower compute requirements. In a future study, we will explore the use of our models in a real-world setting.

Conclusions

This work demonstrates the feasibility of localized AI assistance for clinical note generation using small-scale nonproprietary models. Our results highlight the advantages of language-specific model tuning, providing a promising direction for future research, especially when considering the significant speed and cost advantages of the language-specific model.

Acknowledgments

The authors gratefully acknowledge the Platform for Infrastructure, Education and Research in AI at the University of Freiburg Medical Center for providing computing resources and the support by the Open Access Publication Fund of the University of Freiburg.

Data Availability

The code used to train and evaluate the models was archived on Zenodo [27].

Authors' Contributions

FH prepared training data, trained the models, performed the automated evaluation, analyzed evaluation results, and wrote the manuscript. DB initiated the project, collected and prepared training data, performed the qualitative evaluation of results during model training, and contributed to the quantitative human evaluation and to writing the manuscript. TR supervised the project in the Eye Center. SA organized and contributed to the human evaluation. LL analyzed the results of the human evaluation and was a major contributor to the writing of the manuscript. CH supervised the project at the Institute for Digitization in Medicine. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Robertson SL, Robinson MD, Reid A. Electronic health record effects on work-life balance and burnout within the I3 population collaborative. J Grad Med Educ 2017 Aug;9(4):479-484. [doi: 10.4300/JGME-D-16-00123.1] [Medline: 28824762]
2. Overhage JM, McCallie D. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. Ann Intern Med 2020 Feb 4;172(3):169-174. [doi: 10.7326/M18-3684] [Medline: 31931523]
3. Sirrianni J, Sezgin E, Claman D, Linwood SL. Medical text prediction and suggestion using generative pretrained transformer models with dental medical notes. Methods Inf Med 2022 Dec;61(5-06):195-200. [doi: 10.1055/a-1900-7351] [Medline: 35835447]
4. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. arXiv. Preprint posted online on Oct 1, 2023. [doi: 10.48550/arXiv.2211.09110]
5. Deprecations. OpenAI Platform. 2023. URL: https://platform.openai.com/docs/deprecations [accessed 2024-07-23]
6. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv. Preprint posted online on Oct 31, 2023. [doi: 10.48550/arXiv.2307.09009]
7. Moradi M, Blagec K, Haberl F, Samwald M. GPT-3 models are poor few-shot learners in the biomedical domain. arXiv. Preprint posted online on Jun 1, 2022. [doi: 10.48550/arXiv.2109.02555]
8. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194. [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
9. Lehman E, Hernandez E, Mahajan D, et al. Do we still need clinical language models? arXiv. Preprint posted online on Feb 16, 2023. [doi: 10.48550/arXiv.2302.08091]
10. Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of large language models. arXiv. Preprint posted online on Oct 16, 2021. [doi: 10.48550/arXiv.2106.09685]
11. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. arXiv. Preprint posted online on May 23, 2023. [doi: 10.48550/arXiv.2305.14314]
12. Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. Presented at: 2020 Conference on Empirical Methods in Natural Language Processing; Nov 16-20, 2020; Online p. 38-45. [doi: 10.18653/v1/2020.emnlp-demos.6]
13. Beeching E, Wolf T, Fourrier C, et al. Open LLM leaderboard. Hugging Face. 2023. URL: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard [accessed 2024-07-23]
14. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on Feb 27, 2023. [doi: 10.48550/arXiv.2302.13971]
15. Touvron H, Martin L, Stone K, et al. LlaMA 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online on Jul 19, 2023. [doi: 10.48550/arXiv.2307.09288]
16. Christiano P, Leike J, Brown TB, et al. Deep reinforcement learning from human preferences. arXiv. Preprint posted online on Feb 17, 2023. [doi: 10.48550/arXiv.1706.03741]
17. BigScience Workshop, Le Scao T, Fan A, et al. BLOOM: a 176B-parameter open-access multilingual language model. arXiv. Preprint posted online on Jun 27, 2023. [doi: 10.48550/arXiv.2211.05100]
18. Ostendorff M, Rehm G. Efficient language model training through crosslingual and progressive transfer learning. arXiv. Preprint posted online on Jan 23, 2023. [doi: 10.48550/arXiv.2301.09626]
19. Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. arXiv. Preprint posted online on Mar 29, 2022. [doi: 10.48550/arXiv.2203.15556]
20. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. Presented at: NIPS'19: 33rd International Conference on Neural Information Processing Systems; Dec 8-14, 2019; Vancouver, Canada p. 8026-8037.
21. Lhoest Q, del Moral AV, Jernite Y, et al. Datasets: a community library for natural language processing. Presented at: 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Nov 7-11, 2021; Online and Punta Cana, Dominican Republic p. 175-184. [doi: 10.18653/v1/2021.emnlp-demo.21]

22. Gugger S, Debut L, Wolf T, et al. Accelerate: training and inference at scale made simple, efficient and adaptable. GitHub. 2022. URL: https://github.com/huggingface/accelerate [accessed 2024-07-23]
23. Rajbhandari S, Rasley J, Ruwase O, He Y. Memory optimizations toward training trillion parameter models. Presented at: SC20: International Conference for High Performance Computing, Networking, Storage and Analysis; Nov 9-19, 2020; Atlanta, GA, USA p. 1-16. [doi: 10.1109/SC41405.2020.00024]
24. Su Y, Collier N. Contrastive search is what you need for neural text generation. arXiv. Preprint posted online on Feb 14, 2023. [doi: 10.48550/arXiv.2210.14140]
25. Su Y, Lan T, Wang Y, Yogatama D, Kong L, Collier N. A contrastive framework for neural text generation. Presented at: 36th Conference on Neural Information Processing Systems (NeurIPS 2022); Nov 28 to Dec 9, 2022; New Orleans, LA p. 21548-21561.
26. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. arXiv. Preprint posted online on Dec 24, 2023. [doi: 10.48550/arXiv.2306.05685]
27. Heilmeyer FA, Böhringer D, Reinhard T, Arens A, Lyssenko L, Haverkamp C. Assessing the viability of open large language models for clinical documentation: real-world study in German health care. Zenodo. URL: https://zenodo.org/records/11355001 [accessed 2024-08-03]

Abbreviations

AI: artificial intelligence
CLP: cross-lingual and progressive transfer learning
EHR: electronic health record
GPU: graphics processing unit
ICL: in-context learning
LLM: large language model
LoRA: Low-Rank Adaptation
NF4: 4-bit Normal Float
QLoRA: Quantized Low-Rank Adaptation

Edited by C Lovis; submitted 17.04.24; peer-reviewed by J Zaghir; revised version received 27.05.24; accepted 02.06.24; published 28.08.24.

Please cite as:
Heilmeyer F, Böhringer D, Reinhard T, Arens S, Lyssenko L, Haverkamp C
Viability of Open Large Language Models for Clinical Documentation in German Health Care: Real-World Model Evaluation Study
JMIR Med Inform 2024;12:e59617
URL: https://medinform.jmir.org/2024/1/e59617
doi:10.2196/59617

Review

The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review

Carl Preiksaitis[1], MD; Nicholas Ashenburg[1], MD; Gabrielle Bunney[1], MBA, MD; Andrew Chu[1], MD; Rana Kabeer[1], MPH, MD; Fran Riley[1], MSE, MD; Ryan Ribeira[1], MPH, MD; Christian Rose[1], MD

Department of Emergency Medicine, Stanford University School of Medicine, Palo Alto, CA, United States

Corresponding Author:
Carl Preiksaitis, MD
Department of Emergency Medicine
Stanford University School of Medicine
900 Welch Road
Suite 350
Palo Alto, CA, 94304
United States
Phone: 1 650 723 6576
Email: cpreiksaitis@stanford.edu

Abstract

Background: Artificial intelligence (AI), more specifically large language models (LLMs), holds significant potential in revolutionizing emergency care delivery by optimizing clinical workflows and enhancing the quality of decision-making. Although enthusiasm for integrating LLMs into emergency medicine (EM) is growing, the existing literature is characterized by a disparate collection of individual studies, conceptual analyses, and preliminary implementations. Given these complexities and gaps in understanding, a cohesive framework is needed to comprehend the existing body of knowledge on the application of LLMs in EM.

Objective: Given the absence of a comprehensive framework for exploring the roles of LLMs in EM, this scoping review aims to systematically map the existing literature on LLMs' potential applications within EM and identify directions for future research. Addressing this gap will allow for informed advancements in the field.

Methods: Using PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) criteria, we searched Ovid MEDLINE, Embase, Web of Science, and Google Scholar for papers published between January 2018 and August 2023 that discussed LLMs' use in EM. We excluded other forms of AI. A total of 1994 unique titles and abstracts were screened, and each full-text paper was independently reviewed by 2 authors. Data were abstracted independently, and 5 authors performed a collaborative quantitative and qualitative synthesis of the data.

Results: A total of 43 papers were included. Studies were predominantly from 2022 to 2023 and conducted in the United States and China. We uncovered four major themes: (1) clinical decision-making and support was highlighted as a pivotal area, with LLMs playing a substantial role in enhancing patient care, notably through their application in real-time triage, allowing early recognition of patient urgency; (2) efficiency, workflow, and information management demonstrated the capacity of LLMs to significantly boost operational efficiency, particularly through the automation of patient record synthesis, which could reduce administrative burden and enhance patient-centric care; (3) risks, ethics, and transparency were identified as areas of concern, especially regarding the reliability of LLMs' outputs, and specific studies highlighted the challenges of ensuring unbiased decision-making amidst potentially flawed training data sets, stressing the importance of thorough validation and ethical oversight; and (4) education and communication possibilities included LLMs' capacity to enrich medical training, such as through using simulated patient interactions that enhance communication skills.

Conclusions: LLMs have the potential to fundamentally transform EM, enhancing clinical decision-making, optimizing workflows, and improving patient outcomes. This review sets the stage for future advancements by identifying key research areas: prospective validation of LLM applications, establishing standards for responsible use, understanding provider and patient perceptions, and improving physicians' AI literacy. Effective integration of LLMs into EM will require collaborative efforts and thorough evaluation to ensure these technologies can be safely and effectively applied.

(JMIR Med Inform 2024;12:e53787) doi:10.2196/53787

KEYWORDS

large language model; LLM; emergency medicine; clinical decision support; workflow efficiency; medical education; artificial intelligence; AI; natural language processing; NLP; AI literacy; ChatGPT; Bard; Pathways Language Model; Med-PaLM; Bidirectional Encoder Representations from Transformers; BERT; generative pretrained transformer; GPT; United States; US; China; scoping review; Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA; decision support; workflow efficiency; risk; ethics; education; communication; medical training; physician; health literacy; emergency care

Introduction

Background

Emergency medicine (EM) is at an inflection point. With increasing patient volumes, decreasing staff availability, and rapidly evolving clinical guidelines, emergency providers are overburdened and burnout is significant [1]. While the role of artificial intelligence (AI) in enhancing emergency care is increasingly recognized, the emergence of large language models (LLMs) offers a novel perspective. Previous reviews have systematically categorized AI applications in EM, focusing on diagnostic-specific and triage-specific branches, emphasizing diagnostic prediction and decision support [2-5]. This review aims to build upon these foundations by exploring the unique potential of LLMs in EM, particularly in areas requiring complex data processing and decision-making under time constraints.

An LLM is a deep learning–based artificial neural network, distinguished from traditional machine learning models by its training on vast amounts of textual data. This enables LLMs to recognize, translate, predict, or generate text or other content [6]. Characterized by transformer architecture and the ability to encode contextual information using several parameters, LLMs allow for nuanced understanding and application across a diverse range of topics. Unlike traditional AI models, which often rely on structured data and predefined algorithms, LLMs are adept at interpreting unstructured text data. This feature makes them particularly useful in tasks such as real-time data interpretation, augmenting clinical decision-making, and enhancing patient engagement in clinical settings. For instance, LLMs can efficiently sift through electronic health records (EHRs) to identify critical patient histories and assist clinicians in interpreting multimodal diagnostic data. In addition, they can serve as advanced decision support tools in differential diagnosis, enhancing the quality of care while reducing the cognitive load and decision fatigue for emergency providers. Furthermore, the content generation ability of LLMs, ranging from technical computer code to essays and poetry, demonstrates their versatility and exceeds the functional scope of traditional machine learning models in terms of content creation and natural language processing.

Importance

While interest in applying LLMs to EM is gaining momentum, the existing body of literature remains a patchwork of isolated studies, theoretical discussions, and small-scale implementations. Moreover, existing research often focuses on specific use cases, such as diagnostic assistance or triage prioritization, rather than providing a holistic view of how LLMs can be integrated into the EM workflow. Conclusions based on other forms of machine learning are not readily translatable to LLMs. This fragmented landscape makes it challenging for emergency clinicians, who are already burdened by the complexities and pace of their practice, to discern actionable insights or formulate a coherent strategy for adopting these technologies. Despite the promise shown by several models, such as ChatGPT-4 (OpenAI) or Med-PaLM 2 (Google AI), the absence of standardized metrics for evaluating their clinical efficacy, ethical use, and long-term sustainability leaves researchers and clinicians navigating an uncharted territory. Consequently, the potential for LLMs to enhance emergency medical care remains largely untapped and poorly understood.

Goals of This Review

In light of these complexities and informational disparities, our study undertakes a crucial step to consolidate, assess, and contextualize the fragmented knowledge base surrounding LLMs in EM. Through a scoping review, we aim to establish a foundational understanding of the field's current standing, from technological capabilities to clinical applications and ethical considerations. This synthesis serves a dual purpose: first, to equip emergency providers with a navigable map of existing research and, second, to identify critical gaps and avenues for future inquiry. As EM increasingly embraces technological solutions for its unique challenges, our goal is to provide clarity to the responsible and effective incorporation of LLMs into clinical practice.

Methods

Overview

We adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [7] and used the scoping review methodology proposed by Arksey and O'Malley [8] and furthered by Levac et al [9]. This included the following steps: (1) identifying the research question; (2) identifying relevant studies; (3) selecting studies; (4) charting the data; (5) collating, summarizing, and reporting the results; and (6) consultation. Our full review protocol is published elsewhere [10].

Identifying the Research Question

The overall purpose of this review was to map the current literature describing the potential uses of LLMs in EM and to identify directions for future research. To achieve this goal, we aimed to answer the primary research question: "What are the current and potential uses of LLMs in EM described in the literature?" We chose to explicitly focus on LLMs as this subset of AI is rapidly developing and generating significant interest for potential applications.

Identifying Relevant Studies

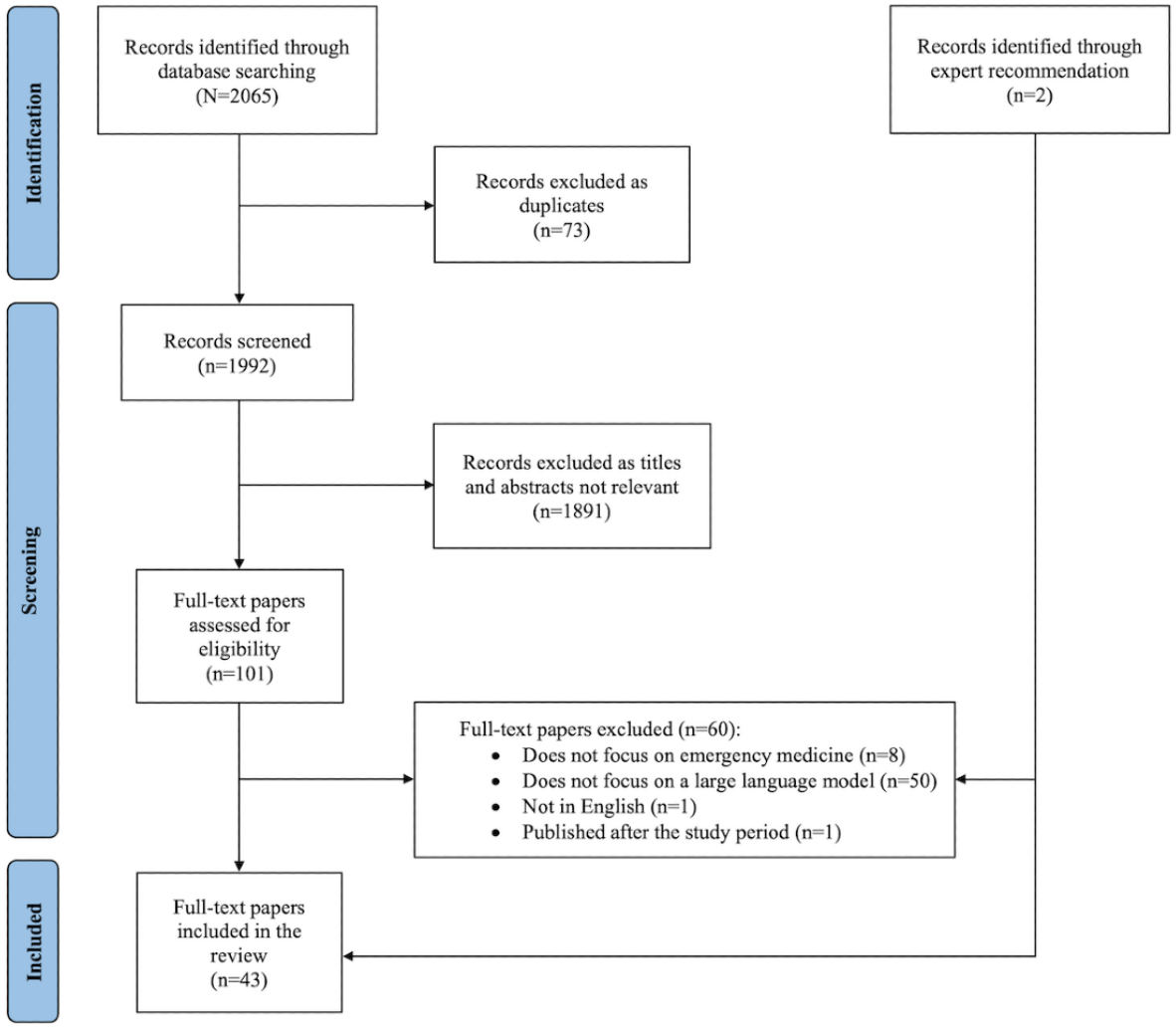
In August 2023, we searched Ovid MEDLINE, Embase, Web of Science, and Google Scholar for potential citations of interest. We limited our search to papers published after January 2018 as the Bidirectional Encoder Representations from Transformers (BERT; Google) model was introduced that year and considered by many to be the first in the contemporary class of LLMs [11]. Our search strategy (Multimedia Appendix 1), created in consultation with a medical librarian, combined keywords and MeSH (Medical Subject Headings) terms related to LLMs and EM. We reviewed the bibliographies of identified studies for potential missed papers.

Study Selection

Citations were managed using Covidence web-based software (Veritas Health Innovation). Manuscripts were included if they discussed the use of an LLM in EM, including applications in the emergency department (ED) and prehospital and periadmission settings. Furthermore, we included use cases related to public health, disease monitoring, or disaster preparedness as these are relevant to EDs. We excluded studies that used other forms of machine learning or natural language processing that were not LLMs and studies that did not clearly relate to EM. We also excluded cases where the only use of an LLM was in generating the manuscript without any additional commentary.

Two investigators (CP and CR) independently screened 100 abstracts, and the interrater reliability showed substantial agreement (κ=0.75). The remaining abstracts were screened by 1 author (CP), who consulted with a second author as needed for clarification regarding inclusion and exclusion criteria. All papers meeting the initial criteria were independently reviewed in full by 2 authors (CP and CR). Studies determined to meet the eligibility criteria by both reviewers were included in the analysis. Discrepancies were resolved by consensus and with the addition of a third reviewer (NA) if needed. Our initial search strategy identified 2065 papers, of which 73 (3.54%) were duplicates, resulting in 1992 (96.46%) papers for screening (Figure 1). Of the 1992 papers, 1891 (94.93%) were excluded based on the title or abstract. In total, 5.07% (101/1992) of the papers were reviewed in full, and 2.11% (42/1992) of the papers were found to meet the study inclusion criteria. During manuscript review, 2 additional papers were brought to our attention by experts, and 1 of these met the inclusion criteria, bringing the total number of included papers to 43.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of search and screening for large language models in emergency medicine.



Charting the Data

Data abstraction was independently conducted using a structured form to capture paper details, including the author, year of publication, study type, specific study population, study or paper location, purpose, and main findings. Data to address our primary research question was iteratively abstracted from the papers as our themes emerged, as explained in the subsequent sections.

Collating, Summarizing, and Reporting the Results

We synthesized and collated the data, performing both a quantitative and qualitative analysis. A descriptive summary of the included studies was created. Then, we used the methodology proposed by Braun and Clarke [12] to conduct a thematic analysis to address our primary research question. Five authors (CP, CR, AC, NA, and RR) independently familiarized themselves with and generated codes for a purposively diverse selection of 10 papers, focusing on content that suggested possible uses for LLMs in EM. The group met to discuss preliminary findings and refine the group's approach. Individuals then independently aggregated codes into themes. These themes were reviewed and refined as a group. Then, 2 authors (CP and CR) reviewed the remaining manuscripts for any additional themes and data that supported or contradicted our existing themes. These data were used to refine themes through group discussion. Our analysis included a discussion and emphasis on the implications and future research directions for the field, based on the guidance from Levac et al [9].

Consultation

To ensure our review accurately characterized the available knowledge and that our interpretations of it were correct, we consulted with external emergency physicians with topic expertise in AI. We incorporated feedback as appropriate. For example, we more completely defined LLMs for clarity and included a table describing common models (Table 1). Our findings and recommendations were endorsed by our consultants.

Table 1. Large language models reported in the identified literature.

Model	Interface	Model size (parameters)	Developer	Year of release
GPT-3.5 Turbo	ChatGPT	175 billion [13]	OpenAI	2022
GPT-4	ChatGPT	Approximately 1.8 trillion (estimated) [14]	OpenAI	2023
Pathways Language Model	Bard	540 billion [15]	Google AI[a]	2023
Embeddings from Language Model	Full model available	93.6 billion [16]	Allen Institute for AI	2018
Bidirectional Encoder Representations from Transformers	Full model available	110 million and 340 million [17]	Google	2018

[a]AI: artificial intelligence.

Results

Overview

Most identified studies (29/43, 67%) were published in 2023. Of the 43 studies, 14 (33%) were conducted in the United States, followed by 6 (14%) in China, 4 (9%) in Australia, 3 (7%) each in Taiwan and France, and 2 (5%) each in Singapore and Korea. Several other individual studies (5/43, 12%) were from various countries (Table 2).

In terms of study type, 40% (17/43) of the papers were methodology studies; 40% (17/43) were case studies; 16% (7/43) were commentaries; and 2% (1/43) each of a case report, qualitative investigation, and retrospective cross-sectional study. In total, 58% (25/43) of these studies addressed the ED setting specifically, followed by 14% (6/43) addressing the prehospital setting and 14% (6/43) addressing other non-ED hospital settings. In total, 7% (3/43) of the studies focused on using LLMs for the public, 5% (2/43) focused on using them for social media analysis, and 2% (1/43) focused on using them for research applications. LLMs used in the reviewed papers (Table 1) included versions of GPT (OpenAI; eg, ChatGPT, GPT-4, and GPT-2), Pathways Language Model (Bard; Google AI), Embeddings from Language Model, XLNet, and BERT (Google; eg, BioBERT, ClinicalBERT, and decoding-enhanced BERT with disentangled information).

We identified four major themes in our analysis: (1) clinical decision-making and support; (2) efficiency, workflow, and information management; (3) risks, ethics, and transparency; and (4) education and communication. Major themes, subthemes, and representative quotations are presented in Table 3.

Table 2. Summary of included studies and identified themes (N=43).

Study	Country	Study type	Purpose	Setting and context	Large language models used	Sample size	Themes
Xu et al [18], 2020	France	Methodology	Classification of visits into trauma and nontrauma based on ED[a] notes	ED	GPT-2 (OpenAI)	16,1930 notes	CDMS[b] and EWIM[c]
Wang et al [19], 2020	China	Retrospective cross-sectional study	Sentiment analysis of social media posts related to COVID-19	Social media	BERT[d] (Google)	99,9978 posts	EWIM
Chen et al [20], 2020	Taiwan	Methodology	Diagnosis identification from discharge summaries	Inpatient	BERT and BioBERT	25,8850 discharge diagnoses	EWIM
Chang et al [21], 2020	United States	Methodology	Categorize free-text ED chief complaints	ED	BERT and Embeddings from Language Model	2.1 million adult and pediatric ED visits	CDMS and EWIM
Wang et al [22], 2021	Singapore	Methodology	Summarize EMS[e] reports for clinical audits	EMS and prehospital	BERT	58,898 ambulance incidents	EWIM
Gil-Jardiné et al [23], 2021	France	Methodology	Classify content of EMS calls during the COVID-19 pandemic	EMS and prehospital	GPT-2	888,469 calls (training), 39,907 calls (validation), and 254,633 calls (application)	EWIM
Shung et al [24], 2021	United States	Methodology	Identify patients with gastrointestinal bleeding from ED triage and ROS data	ED	BERT	7144 cases	CDMS
Tahayori et al [25], 2021	Australia	Methodology	Predict patient disposition from ED triage notes	ED	BERT	249,532 ED encounters	CDMS and EWIM
Kim et al [26], 2021	South Korea	Case study	Assign triage severity to simulated cases	ED	BERT	762 cases	CDMS
Wang et al [27], 2021	China	Methodology	Predict diagnosis and appropriate hospital team from medical record	Prehospital	BERT and ClinicalBERT	198,000 patient records	EWIM
McMaster et al [28], 2021	Australia	Methodology	Identify adverse drug events from discharge summaries	Inpatient	BERT (ClinicalBERT and DeBERTa[f])	861 discharge summaries	EWIM
Chen et al [29], 2021	Taiwan	Methodology	Classify electronic health record data into disease presentations	ED	BERT	1,040,989 ED visits and 305,897 NHAMCS[g] samples	EWIM
Drozdov et al [30], 2021	United Kingdom	Methodology	Generate annotations for CXRs[h] to train model to identify COVID-19 cases	ED	BERT (to generate image annotations)	214,042 CXRs	CDMS
Zhang et al [31], 2022	China	Methodology	Classify EMS cases into disease categories	EMS and prehospital	BERT	3500 records	EWIM
Pease et al [32], 2023	United States	Qualitative investigation	Determine the attitudes of clinicians toward using AI[i] in suicide screening	ED	N/A[j]	3 clinicians	CDMS and RET[k]
Chae et al [33], 2023	United States	Methodology	Predict ED visits and hospitalizations for patients with heart failure	Prehospital (home health care)	BERT (BioclinicalBERT)	9362 patients	CDMS and RET

Study	Country	Study type	Purpose	Setting and context	Large language models used	Sample size	Themes
Huang et al [34], 2023	United States	Methodology	Predict nonaccidental trauma	ED	BERT	244,326 trajectories (test) and 2,077,852 trajectories (validation)	CDMS
Chen et al [35], 2023	Taiwan	Methodology	Predict critical outcomes from ED data	ED	BERT (comparator)	171,275 ED visits	CDMS
Smith et al [36], 2023	Australia	Case study	Determine model performance on EM[l] accreditation examination	ED	GPT-3.5 (OpenAI), GPT-4 (OpenAI), Bard-PaLM[m], Bard-PaLM 2, and Bing (Microsoft Corporation)	240 questions	CDMS, RET, and EC[n]
Gupta et al [37], 2023	United States	Case study	Determine the ability of the model to correctly diagnose simulated cases	ED	ChatGPT	20 cases	CDMS, RET, and EC
Abavisani et al [38], 2023	Iran	Commentary	Potential uses of the model in emergency surgery	Emergency surgery	ChatGPT	N/A	CDMS and RET
Rahman et al [39], 2023	United States	Methodology	Identify cases and patterns in unstructured EMS data	EMS and prehospital	BERT (BioBERT and ClinicaBERT)	40,000 EMS narratives	EWIM
Lam and Au [40], 2023	China	Case study	Evaluate model response to lay questions regarding stroke	General public	ChatGPT	3 questions	EC
Bushuven et al [41], 2023	Germany	Case study	Use of the model to advise parents during pediatric emergencies	General public	ChatGPT and GPT-4	22 cases	CDMS, RET, and EC
Ahn [42], 2023	South Korea	Case study	Use of model to provide a lay-person instruction for cardiopulmonary resuscitation	General public	ChatGPT	3 questions	RET and EC
Preiksaitis et al [43], 2023	United States	Commentary	Potential limitations to using models for clinical charting	General medicine	ChatGPT	N/A	EWIM and RET
Barash et al [44], 2023	Israel	Case study	Use of model to aid radiology referral in the ED	ED	GPT-4	40 cases	CDMS and RET
Dahdah et al [45], 2023	United States	Case study	Use of model to triage based on chief complaints	ED	ChatGPT	30 questions	CDMS and RET
Gottlieb et al [46], 2023	United States	Commentary	Discuss advantages and disadvantages of using the model in research	ED and research	ChatGPT	N/A	RET and EC
Babl and Babl [47], 2023	Australia	Case study	Determine the ability of the model to generate a scientific abstract	Research	ChatGPT	1 abstract	RET and EC
Chen et al [48], 2023	China	Methodology	Use the model to study the functioning of web-based self-organizations	Social media	BERT	47,173 users	EWIM
Bradshaw [49], 2023	United States	Case study	Determine the ability of the model to generate discharge instructions	ED	ChatGPT	1 set of discharge instructions	EWIM and EC
Cheng et al [50], 2023	China	Commentary	Potential uses for the model in surgical management	ED	ChatGPT	N/A	CDMS and EWIM

Study	Country	Study type	Purpose	Setting and context	Large language models used	Sample size	Themes
Rao et al [51], 2023	United States	Case study	Test the model performance in several clinical scenarios	General medicine	ChatGPT	36 clinical vignette	EWIM and EC
Brown et al [52], 2023	Jersey	Case report and commentary	Discuss possible model uses in supporting decision-making and clinical care	ED	ChatGPT	1 case	CDMS and EWIM, RET and EC
Bhattaram et al [53], 2023	India	Case study	The ability of the model to triage clinical scenarios	ED	ChatGPT	5 scenarios	CDMS, RET and EC
Webb [54], 2023	United States	Case study	The ability of the model to be used as a communication skill trainer	ED	ChatGPT-3.5	1 case	RET and EC
Hamed et al [55], 2023	Qatar	Case study	The ability of the model to synthesize clinical practice guidelines for diabetic ketoacidosis	General medicine	ChatGPT	3 guidelines	EWIM and RET
Altamimi et al [56], 2023	Saudi Arabia	Case study	The ability of the model to recommend management in snakebites	ED	ChatGPT	9 questions	CDMS and RET
Gebrael et al [57], 2023	United States	Case study	Predict the disposition of patients with metastatic prostate cancer based on ED documentation	ED	ChatGPT-4	56 patients	CDMS, EWIM, and RET
Sarbay et al [58], 2023	Turkey	Case study	Use of the model for patient triage using clinical scenarios	ED	ChatGPT	50 case scenarios	CDMS, EWIM, and RET
Okada et al [59], 2023	Singapore	Commentary	Discuss possible applications for the model in resuscitation	ED or intensive care unit	GPT-3 and GPT-4	N/A	CDMS, EWIM, and RET
Chenais et al [60], 2023	France	Commentary	Describe the landscape of AI-based applications currently in use in EM	ED	BERT and GPT-2	N/A	CDMS, EWIM, and RET

[a]ED: emergency department.

[b]CDMS: clinical decision-making and support.

[c]EWIM: efficiency, workflow, and information management.

[d]BERT: Bidirectional Encoder Representations from Transformers.

[e]EMS: emergency medical service.

[f]DeBERTa: decoding-enhanced Bidirectional Encoder Representations from Transformers with disentangled information.

[g]NHAMCS: National Hospital Ambulatory Medical Care Survey.

[h]CXR: chest x-ray.

[i]AI: artificial intelligence.

[j]N/A: not applicable.

[k]RET: risks, ethics, and transparency.

[l]EM: emergency medicine.

[m]PaLM: Pathways Language Model.

[n]EC: education and communication.

Table 3. Major themes identified, associated subthemes, and representative quotations.

Major theme and subtheme	Representative quotation
Theme 1: clinical decision-making and support	
Prediction	"Machine-learning and natural language processing can be together applied to the ED triage note to predict patient disposition with a high level of accuracy." [25]
Treatment recommendations	"An under-explored use of AI in medicine is predicting and synthesizing patient diagnoses, treatment plans, and outcomes." [51]
Symptom checking and self-triage	"To our knowledge, this is the first work to investigate the capabilities of ChatGPT and GPT-4 on PALS core cases in the hypothetical scenario that laypersons would use the chatbot for support until EMS arrive." [41]
Classification	"In this proof-of-concept study, we demonstrated the process of developing a reliable NER [named-entity recognition] model that could reliably identify clinical entities from unlabeled paramedic free text reports." [22]
Triage	"...this preliminary study showed the potential of developing an automatic classification system that directly classifies the KTAS [triage] level and symptoms from the conversations between patients and clinicians." [26]
Screening	"We showed that PABLO, a pretrained, domain-adapted outcome forecasting model, can be used to predict both first and recurrent instances of NAT [non-accidental trauma]." [34]
Differential diagnosis building	"These results suggest that ChatGPT has a high level of accuracy in predicting top differential diagnoses in simulated medical cases." [37]
Decision support	"...ChatGPT-4 demonstrates encouraging results as a support tool in the ED. LLMs such as ChatGPT-4 can facilitate appropriate imaging examination selection and improve radiology referral quality." [44]
Clinical augmentation	"AI can serve as an adjunct in clinical decision making throughout the entire clinical workflow, from triage to diagnosis to management." [51]
Theme 2: efficiency, workflow, and information management	
Unstructured data extraction	"The proposed model will provide a method to further extract the unstructured free-text portions in EHRs to obtain an abundance of health data. As we enter the forefront of the artificial intelligence era, NLP deep-learning models are well under development. In our model, all medical free-text data can be transformed into meaningful embeddings, which will enhance medical studies and strengthen doctors' capabilities." [20]
Charting efficiency	"While notes have become more structured and burdensome, the field of data science has rapidly advanced. With such powerful tools available, it seems reasonable to explore their use to automate seemingly mundane tasks such as writing clinical notes. Generative AI models like ChatGPT could be developed to populate notes for patients based on massive amounts of data contained in current EHRs." [43]
Summarization or synthesis	"Although ChatGPT demonstrates the potential for the synthesis of clinical guidelines, the presence of multiple recurrent errors and inconsistencies underscores the need for expert human intervention and validation." [55]
Pattern identification	"This embedding system can be used as a disease retrieval model, which encodes queries and finds the most relevant patients and diseases. In the retrieval demonstration, the query subject was a 53-year-old female patient who suffered from abdominal pain in the upper right quarter to right flanks for 3 days and noticed dizziness and tarry stool on the day of the interview. Through the retrieval, we obtained the five most similar patients with similar symptoms that were possibly related to different diseases." [29]
Workflow efficiency	"Integration of LLMs with existing EHR (with appropriate regulations) could facilitate improved patient outcomes and workflow efficiency." [51]
Theme 3: risks, ethics, and transparency	
Oversight	"Generally speaking, the Ethics Guideline for Trustworthy AI suggested seven key requirements including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, nondiscrimination and fairness, environmental and societal well-being, and accountability." [59]
Fairness	"[Use of LLMs] could also increase equity by assisting researchers with disabilities such as dyslexia." [46]
Ethical and legal responsibilities	"Legal and ethical implications are associated with using AI in clinical practice, particularly regarding privacy and informed consent issues." [52]

Major theme and subtheme	Representative quotation
Reliance on input data	"...data quality can affect the performance of LLMs and NLP techniques applied to the task of extracting and summarizing clinical guidelines." [55]
Overreliance	"Overreliance on AI systems and the assumption that they are infallible or less fallible than human judgment–automation bias–can lead to errors." [52]
Explainability and transparency	"Creating a clinician-interpretable risk prediction model is essential for clinical adoption and implementation of models because it builds trust in decisionmakers, enables error identification and correction in the model, and facilitates integration into clinical workflows." [33]
Bias propagation	"A risk of bias is possible if the initial training data is not representative of the study population. There is a possibility of compounding of bias and error, leading to incorrect assessment." [53]
Human bias reduction	"AI tools can offer a near real-time interpretation of medical imaging and clinical decision support and may identify latent patterns that may not be evident to clinicians. While humans are prone to cognitive biases, such as prejudice or fatigue, which can hinder their decision-making process, AI can mitigate these biases and improve accuracy in patient care." [52]
Accuracy	"LLMs may not be exposed to the broader range of literature (particularly if studies are located behind paywalls), which may limit the comprehensiveness or accuracy of the data." [46]
Theme 4: education and communication	
Clinician education	"While LLM performance in medical examinations may initially seem to be little more than a novelty, their ability to generate coherent and well-explained content hints at other potential uses. As a medical education tool they could potentially help generate practice questions, design mock examinations or provide additional explanations for complex concepts." [36]
Communication	"Although in its infancy, AI chatbot use has the potential to disrupt how we teach medical students and graduate medical residents communication skills in outpatient and hospital settings." [54]
Content generation	"ChatGPT or similar programmes, with careful review of the product by authors, may become a valuable scientific writing tool." [47]
Research assistance	"Conversational AI has some clear benefits and disadvantages. As the technology further evolves, it is incumbent on the scientific community to determine how best to incorporate LLMs into the research and publication process with attention to scientific integrity, adherence to ethical principles, and existing copyright laws." [46]

Theme 1: Clinical Decision-Making and Support

The first theme we identified is clinical decision-making and support. LLMs have been used or proposed for applications such as providing advice to the public before arrival; aiding in triage as patients arrive at the ED; or augmenting the activities of physicians as they provide care, either through supporting diagnostics or predicting patient resource use.

Several applications focused on advising the public and aiding in symptom checking, self-triage, and occasionally advising first-aid before the arrival of emergency medical services. These included counseling parents during potential pediatric emergencies, recognizing stroke, or providing advice during potential cardiac arrests [40-42]. Wang et al [27] proposed a model that could potentially help patients navigate the complexities of the health care system in China and present to the correct medical setting for the care they need.

Furthermore, LLMs have the potential to efficiently screen patients for important outcomes, such as pediatric patients at risk for nonaccidental trauma, suicide risk, or COVID-19 infection [30,32,34]. These can be implemented based on data in the medical record or as clinical data are obtained in real time.

Early identification of patient risks could help physicians more rapidly identify important diagnoses. Several studies discussed implementations of LLMs that work in conjunction with physicians while caring for patients in the ED [50,51]. Brown et al [52] discuss the potential role of these models in overcoming cognitive biases and reducing errors. These models could be used in developing a differential diagnosis, recommending imaging studies, providing treatment recommendations, or interpreting clinical guidelines [37,44,55,56].

Several studies centered on predicting outcomes such as presentation to the ED, hospitalization, intensive care unit admission, or in-hospital cardiac arrest [25,33,35,57]. Applications of LLMs in the triage process could potentially identify patients who require immediate attention or patients at a high risk of certain diagnoses, such as gastrointestinal bleeding [24,26,53,58,60].

Theme 2: Efficiency, Workflow, and Information Management

The second theme identified is information management, workflow, and efficiency. LLMs show great promise in increasing the usability of data available in the EHR. Interactions with the EHR take up a substantial amount of physician time, and it is often difficult to identify crucial information during critical times [43]. LLMs could serve a variety of information management functions. They could be used to perform audits for quality improvement purposes, identify potential adverse events such as drug interactions, anticipate and monitor public health emergencies, and assist with information entry during

the clinical encounter [19,20,22,23,28,31,39,43,49]. LLMs developed and trained on data from the ED could quickly identify similar patient presentations, recognize patterns, and extract important information from unstructured text [18,20,21,60].

Some authors suggest that LLMs can enhance care throughout the entire EM encounter [30,50-52]. LLMs could potentially be used as digital adjuncts for clinical decision-making because they could generate differentials, predict final diagnoses, offer interpretations of imaging studies, and suggest treatment plans [30,51,52,61]. They may mitigate human cognitive biases and address human factors (eg, time constraints, frequent task switching, high cognitive load, constant interruptions, and decision fatigue) that predispose emergency physicians to error [52].

The flexibility and versatility of the LLMs offer particular benefits to EM practice. The diverse ways in which these models can aid throughout the entire clinical workflow could help physicians process large quantities of complex clinical data, mitigate cognitive biases, and deliver relevant information in a comprehensible format [30,51,52,61]. By streamlining these burdensome tasks, LLMs could help improve the efficiency of care for the high volume of patients the physicians routinely see in the ED.

Theme 3: Risks, Transparency, and Ethics

Despite the potential for advancement and improvement in the care that EM physicians can provide through the inclusion of LLMs in practice, several issues limit their implementation into practice at this time.

The most often discussed risk, mentioned in 11 (26%) of the 43 papers, is the reliability of model responses and the potential for erroneous results [20,21,28-30,44,51,53,55,56,59]. These output errors often result from inaccuracies in the training data, which are most commonly gathered from the internet and unvetted for reliability. Sources of inaccurate responses may be identified by examining the training material, but other errors due to data noise, mislabeling, or outdated information may be harder to detect [21,28,30,56]. Similarly, biases in training data can be propagated to the model, leading to inaccurate or discriminatory results [51,53,57,60,62]. In medical applications, the consequences of the errors can be significant, and even small errors could lead to adverse outcomes [51].

Understanding and mitigating errors in LLMs is challenging due to issues with transparency and reproducibility of model outputs [52-54,59,62]. Better understanding among clinicians of the algorithms and statistical methods used by LLMs is a suggested method to ensure cautious use [52]. Concentrating on making models more explainable or transparent is another potential approach [62]. However, the degree to which this will be feasible, given the complexity of these models, remains to be determined.

Patient and data privacy is another clearly articulated risk of using these models in the clinical environment [35,52,53]. There are some proposed methodologies using unsupervised methods that can train the models with limited access to sensitive information; however, these require further exploration [35]. Patient attitudes and willingness to allow models access to their health information for training and how to address disclosure of this use have not been extensively discussed. Finally, the legal and ethical implications of using LLM output to guide patient care is an often-mentioned concern [52,53,59]. How the responsibility for patient care decisions is distributed if LLMs are used to guide clinical decisions is yet to be determined.

Theme 4: Education and Communication

LLMs offer several opportunities for education and communication. First, several papers noted that the successful integration of LLMs into clinical practice will require physicians to understand the underlying algorithms and statistical methods used by these models [52,59]. There is a need for dedicated educational programs on AI in medicine at all levels of medical education to ensure that the solutions developed align with the clinical environment and address the unique challenges of working with clinical data [34,51,63].

In terms of clinical education, several studies have demonstrated reasonable performance of LLMs on standardized tests in medicine, which could indicate the potential for these models to develop study materials [36]. In addition, these models may be able to help physicians communicate with and educate the patients. Dahdah et al [45] used ChatGPT to answer several common medical questions in easy-to-understand language, suggesting the ability to enhance physician responses to patient queries. Webb [54] demonstrated the use of ChatGPT to simulate patient conversation and provide feedback to a physician learning how to break bad news.

Patient education may be facilitated via these models without physician input as well. As discussed in the previous sections, several authors described applications designed to educate patients during emergencies before they arrived in the ED [27,40-42]. Finally, LLMs could be used to aid in knowledge dissemination. Gottlieb et al [46] and Babl and Babl [47] describe potential applications for LLMs in research and scientific writing. They highlight potential benefits to individuals who struggle with English or have challenges with writing or knowledge synthesis. In addition, models may be used to translate scientific papers more rapidly. However, the use of these models to generate scientific papers raises concerns regarding the potential for academic dishonesty [46,47].

Discussion

Principal Findings

Our review aligns with the growing body of literature emphasizing the great potential for AI in EM, particularly in areas such as time-sensitive decision-making and managing high-volume data [2-5,60]. However, our focus on LLMs and their unique capabilities extends the current understanding of AI applications in EM. Although several specific applications and limitations have been reported and suggested in the literature, our analysis identified 4 major areas of focus for LLMs in EM: clinical decision support, workflow efficiency, risks, ethics, and education. We propose these topics as a framework for understanding emerging implementations of LLMs and as a guide to inform future areas of investigation.

At their core, LLMs and their associated natural language processing techniques offer a way to organize and engage with vast amounts of unstructured text data. Depending on how they are trained and used, they can be operationalized to make predictions or identify patterns, which gives rise to most of our identified applications. Most commercially available LLMs, such as ChatGPT, are trained on massive volumes of text gathered from the internet and then optimized for conversational interaction [64]. This ability to access a breadth of general knowledge and the resulting wide applicability have contributed to the increased use of LLMs by professionals and the public across a variety of fields [65]. As these models become more ubiquitous, there is potential for their use across the care continuum. They could not only support clinical care but also provide an opportunity to offer advice to the public regarding medical concerns. Several papers (3/34, 9%) in our review identified the feasibility of using LLMs to provide first-aid instructions and offer decision support to potential patients seeking care [40-42].

Preliminary work suggests that dedicated training can enhance the ability of these models to make triage recommendations, but prospective implementation has not been tested [27]. LLMs could certainly aid patients in self-triage or with basic medical questions; nevertheless, how this can be effectively and safely implemented needs further exploration, especially with concerns regarding the accuracy of outputs. Possibilities to improve outputs include additional dedicated training of the models to align with the medical and emergency settings to improve their reliability and accuracy. These context-specific models could be equipped with information on the local health care system to help patients identify available resources, schedule appointments, or activate emergency medical services.

In the ED, LLMs could increase workflow efficiency by rapidly synthesizing relevant information from a patient's medical record, structuring and categorizing chief complaint data, and assigning an emergency severity index level [18,21,26,45,53,58]. In addition, quickly accessing data from the medical record could improve the efficiency and thoroughness of chart review. A model's ability to identify subtle patterns in data could offer additional diagnostic support by recommending or interpreting laboratory and imaging studies [30,51,52,61]. By facilitating tasks such as information retrieval and synthesis, LLMs could reduce this burden for clinicians and minimize errors due to buried or disorganized data, potentially contributing to workflow efficiency. Furthermore, they may counteract human cognitive biases and fatigue when used to support clinical decisions [52]. Although some studies have demonstrated reasonable accuracy on focused use cases, further validation of any of these applications across diverse settings and patient populations is required. Thoughtful integration of LLMs has the potential to revolutionize EM by providing clinical decision support, improving situational awareness, and increasing productivity.

However, barriers to seamless implementation exist. As noted by several authors, erroneous outputs remain a concern, given the dependence on training data [28-30,35,51,53,55,56,59]. Information surrounding the most publicly available LLMs today is obscured across three important layers: (1) the underlying training data used—commonly reported to be publicly available data on the internet and from third-party licensed data sets, (2) the underlying architecture of the model—whose exact mechanisms are not always easy to discern, and (3) the intricacies of human-led fine-tuning—often done at the end of development to provide guardrails for output. These layers of obscurity make it difficult to troubleshoot the cause of any single erroneous output.

Regarding privacy and data rights, it is imperative to discuss and implement privacy-preserving methods for patient data. The use of techniques such as data anonymization, differential privacy, and federated learning are instrumental in safeguarding patient information. Data anonymization involves removing or modifying personal identifiers to prevent the association of data with individual patients. Differential privacy introduces randomness into the data or queries to ensure individual data points cannot be isolated [66]. Federated learning enables models to be trained against multiple decentralized devices or servers holding local data samples without exchanging them, thus enhancing privacy [67]. The specific ways in which LLMs will interface with other hospital information systems, such as the EHR, need further exploration, and careful integration is critical to address privacy concerns, especially given the sensitive nature of health care data.

Moreover, the ongoing discussions about the information used in these models underscore the need for continuous scrutiny [52,53,59]. In addition to privacy, the legal and ethical implications of AI-assisted health care require further exploration to establish robust oversight and accountability structures. Without a commitment to explainability and transparency, the use of *black box* LLMs may encounter resistance from clinicians.

Our review reveals several opportunities for future exploration and research. Perhaps the most important is effectively identifying problems that are best solved using LLMs in EM. Our review outlines several immediate areas of potential exploration, including improved communication, translation, and summarization of highly detailed and domain-specific knowledge for providers and patients, but further exploration and prospective validation of specific use cases is required. We expect the potential use cases in EM to grow as LLMs become increasingly complex and develop emergent properties–actions that are not explicitly programmed or anticipated. To bridge the *AI chasm* between innovations in the research realm and widespread adoption, these applications should be identified with significant input from providers in the clinical space who can uniquely identify areas of potential benefit. To accomplish this, a better understanding of the abilities and limitations of LLMs among physicians is needed to optimize their best use and ensure they are effectively implemented, and AI literacy is increasingly described as an essential competency for physicians [68]. We encourage the development of curricula and training programs designed for emergency physicians.

Given the black-box nature of LLMs, standardized frameworks and metrics for evaluation that are specific to health care use cases are needed to evaluate their performance and implementation effectively. These frameworks should encompass an understanding of both the technical capabilities

and constraints of a model, along with the human interaction aspects that affect its use. A crucial part of this assessment involves comparing the performance of LLMs to human proficiency, determining whether the objective is to replace or enhance tasks currently carried out by health care professionals. Thorough testing of models in real time, real-world scenarios is imperative before their deployment. The selection of patient- or provider-focused outcomes is essential, and the effectiveness of models should not be evaluated in isolation. Instead, it is crucial to assess the combined performance of the provider and AI system to ensure that models are effective and practical in real-world settings. Implementing and validating solutions should occur across diverse populations and care environments, with particular focus on cohorts underrepresented in the training data to mitigate potential harm from model biases [69]. Provider perspectives are essential, but equally important are patient perspectives about the use of LLMs in medicine. Impacts on physician-patient communication, patient concerns surrounding privacy, and attitudes toward AI-generated recommendations must be further explored. Collaboration between all relevant stakeholders who develop or will be impacted by LLMs for clinical medicine is essential for developing models that can be used effectively, equitably, and safely.

Limitations

This scoping review has some limitations worth noting. First, we restricted our search to papers published after 2018, when LLMs first emerged. While this captures the current era of LLMs, earlier works relevant to natural language processing in EM may have been overlooked. In addition, despite searching 4 databases and consulting a medical librarian on the search strategy, some pertinent studies may have been missed, and given the rapidly evolving nature of this research area, there are certainly more studies that have emerged since our literature search [70]. However, our review establishes an initial foundation that can be built upon as the field continues to grow. Finally, in an effort to be maximally inclusive in our review, we did not include or exclude papers based on the quality of their evidence. Similarly, we did not make any quality determinations of our included studies. High-quality studies are required to make any determination regarding the efficacy of LLMs for the applications we described, and our review hopefully provides a framework to design these investigations.

Conclusions

This review underscores the transformative potential of LLMs in enhancing the delivery of emergency care. By leveraging their ability to process vast amounts of data rapidly, LLMs offer unprecedented opportunities to improve decision-making speed and accuracy, a critical component in the high-stakes, fast-paced EM environment. From the identified themes, it is evident that LLMs have the potential to revolutionize various aspects of emergency care, highlighting their versatility and the breadth of their applicability.

From the theme of clinical decision-making and support, LLMs can augment the diagnostic process, support differential diagnosis, and aid in the efficient allocation of resources. In the domain of efficiency, workflow, and information management, LLMs have shown promise in enhancing operational efficiencies, reducing the cognitive load on clinicians, and streamlining patient care processes. Regarding risks, ethics, and transparency, the review illuminates the need for meticulous attention to the accuracy, bias, and ethical considerations inherent in deploying LLMs in a clinical setting. Finally, in the realm of education and communication, LLMs' potential to facilitate learning and improve patient and provider communication signifies a paradigm shift in medical education and engagement.

The most urgent research need identified in this review is the development of robust, evidence-based frameworks for evaluating the clinical efficacy of LLMs in EM; addressing ethical concerns; ensuring data privacy; and mitigating potential biases in model outputs. There is a critical need for prospective studies that validate the utility of LLMs in real-world emergency care settings and explore the optimization of these models for specific clinical tasks. Furthermore, research should focus on understanding the best practices for integrating LLMs into the existing health care workflows without disrupting the clinician-patient relationship.

The successful integration of LLMs into EM necessitates a multidisciplinary approach involving clinicians, computer scientists, ethicists, patients, and policy makers. Collaborative efforts are essential to navigate the challenges of implementing AI technologies in health care, ensuring LLMs complement the clinical judgment of EM professionals and align with the overarching goal of improving patient care. The judicious application of LLMs has the potential to fundamentally redefine much of EM practice, ushering in a future where care is more accurate, efficient, and responsive to the needs of patients. Furthermore, by reducing the many burdens that currently encumber clinicians, these technologies hold the promise of restoring and deepening the invaluable human connections between physicians and their patients.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature review search strategy.
[DOCX File , 14 KB - medinform_v12i1e53787_app1.docx]

Multimedia Appendix 2

PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 630 KB - medinform_v12i1e53787_app2.pdf]

References

1. Petrino R, Riesgo LG, Yilmaz B. Burnout in emergency medicine professionals after 2 years of the COVID-19 pandemic: a threat to the healthcare system? Eur J Emerg Med 2022 Aug 01;29(4):279-284 [FREE Full text] [doi: 10.1097/MEJ.0000000000000952] [Medline: 35620812]
2. Piliuk K, Tomforde S. Artificial intelligence in emergency medicine. A systematic literature review. Int J Med Inform 2023 Dec;180:105274 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105274] [Medline: 37944275]
3. Kirubarajan A, Taher A, Khan S, Masood S. Artificial intelligence in emergency medicine: a scoping review. J Am Coll Emerg Physicians Open 2020 Nov 07;1(6):1691-1702 [FREE Full text] [doi: 10.1002/emp2.12277] [Medline: 33392578]
4. Masoumian Hosseini M, Masoumian Hosseini ST, Qayumi K, Ahmady S, Koohestani HR. The aspects of running artificial intelligence in emergency care; a scoping review. Arch Acad Emerg Med 2023 May 11;11(1):e38 [FREE Full text] [doi: 10.22037/aaem.v11i1.1974] [Medline: 37215232]
5. Mueller B, Kinoshita T, Peebles A, Graber MA, Lee S. Artificial intelligence and machine learning in emergency medicine: a narrative review. Acute Med Surg 2022 Mar 1;9(1):e740 [FREE Full text] [doi: 10.1002/ams2.740] [Medline: 35251669]
6. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
8. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb 23;8(1):19-32. [doi: 10.1080/1364557032000119616]
9. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
10. Preiksaitis C. Protocol for a scoping review of the application of large language models in emergency medicine. OSF Home. 2023 Oct 19. URL: https://osf.io/tdghu/ [accessed 2024-04-28]
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online October 11, 2018 2024(https://arxiv.org/abs/1810.04805) [FREE Full text] [doi: 10.5260/chara.21.2.8]
12. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
13. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online May 28, 2020 2024. [doi: 10.5860/choice.189890]
14. Schreiner M. GPT-4 architecture, datasets, costs and more leaked. The Decoder. 2023 Jul 11. URL: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/ [accessed 2023-10-12]
15. Narang S, Chowdhery A. Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance. Google Research. 2022 Apr 04. URL: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html [accessed 2023-10-12]
16. AllenNLP - ELMo. Allen Institute for Artificial Intelligence. URL: https://allenai.org/allennlp/software/elmo [accessed 2023-10-12]
17. Devlin J, Chang MW. Open sourcing BERT: state-of-the-art pre-training for natural language processing. Google Research. URL: https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html [accessed 2023-10-12]
18. Xu B, Gil-Jardiné C, Thiessard F, Tellier E, Avalos M, Lagarde E. Pre-training a neural language model improves the sample efficiency of an emergency room classification model. arXiv. Preprint posted online August 30, 2019 2024.
19. Wang T, Lu K, Chow KP, Zhu Q. COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. IEEE Access 2020 Jul 28;8:138162-138169. [doi: 10.1109/access.2020.3012595]
20. Chen YP, Chen YY, Lin JJ, Huang CH, Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med Inform 2020 Apr 29;8(4):e17787 [FREE Full text] [doi: 10.2196/17787] [Medline: 32347806]
21. Chang D, Hong WS, Taylor RA. Generating contextual embeddings for emergency department chief complaints. JAMIA Open 2020 Jul 15;3(2):160-166 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa022] [Medline: 32734154]
22. Wang H, Yeung WL, Ng QX, Tung A, Tay JA, Ryanputra D, et al. A weakly-supervised named entity recognition machine learning approach for emergency medical services clinical audit. Int J Environ Res Public Health 2021 Jul 22;18(15):7776 [FREE Full text] [doi: 10.3390/ijerph18157776] [Medline: 34360065]
23. Gil-Jardiné C, Chenais G, Pradeau C, Tentillier E, Revel P, Combes X, et al. Trends in reasons for emergency calls during the COVID-19 crisis in the department of Gironde, France using artificial neural network for natural language classification.

Scand J Trauma Resusc Emerg Med 2021 Mar 31;29(1):55 [FREE Full text] [doi: 10.1186/s13049-021-00862-w] [Medline: 33789721]
24. Shung D, Tsay C, Laine L, Chang D, Li F, Thomas P, et al. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. J Gastroenterol Hepatol 2021 Jun;36(6):1590-1597. [doi: 10.1111/jgh.15313] [Medline: 33105045]
25. Tahayori B, Chini-Foroush N, Akhlaghi H. Advanced natural language processing technique to predict patient disposition based on emergency triage notes. Emerg Med Australas 2021 Jun;33(3):480-484. [doi: 10.1111/1742-6723.13656] [Medline: 33043570]
26. Kim D, Oh J, Im H, Yoon M, Park J, Lee J. Automatic classification of the Korean triage acuity scale in simulated emergency rooms using speech recognition and natural language processing: a proof of concept study. J Korean Med Sci 2021 Jul 12;36(27):e175 [FREE Full text] [doi: 10.3346/jkms.2021.36.e175] [Medline: 34254471]
27. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J Cloud Comput 2021 Jan 15;10:4. [doi: 10.1186/s13677-020-00218-2]
28. McMaster C, Chan J, Liew DF, Su E, Frauman AG, Chapman WW, et al. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. J Biomed Inform 2023 Jan;137:104265 [FREE Full text] [doi: 10.1016/j.jbi.2022.104265] [Medline: 36464227]
29. Chen YP, Lo YH, Lai F, Huang CH. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: algorithm development and validation study. J Med Internet Res 2021 Jan 27;23(1):e25113 [FREE Full text] [doi: 10.2196/25113] [Medline: 33502324]
30. Drozdov I, Szubert B, Reda E, Makary P, Forbes D, Chang SL, et al. Development and prospective validation of COVID-19 chest X-ray screening model for patients attending emergency departments. Sci Rep 2021 Oct 14;11(1):20384 [FREE Full text] [doi: 10.1038/s41598-021-99986-3] [Medline: 34650190]
31. Zhang X, Zhang H, Sheng L, Tian F. DL-PER: deep learning model for Chinese prehospital emergency record classification. IEEE Access 2022 Jun 03;10:64638-64649. [doi: 10.1109/ACCESS.2022.3179685]
32. Pease JL, Thompson D, Wright-Berryman J, Campbell M. User feedback on the use of a natural language processing application to screen for suicide risk in the emergency department. J Behav Health Serv Res 2023 Oct 03;50(4):548-554 [FREE Full text] [doi: 10.1007/s11414-023-09831-w] [Medline: 36737559]
33. Chae S, Davoudi A, Song J, Evans L, Hobensack M, Bowles KH, et al. Predicting emergency department visits and hospitalizations for patients with heart failure in home healthcare using a time series risk model. J Am Med Inform Assoc 2023 Sep 25;30(10):1622-1633. [doi: 10.1093/jamia/ocad129] [Medline: 37433577]
34. Huang D, Cogill S, Hsia RY, Yang S, Kim D. Development and external validation of a pretrained deep learning model for the prediction of non-accidental trauma. NPJ Digit Med 2023 Jul 19;6(1):131 [FREE Full text] [doi: 10.1038/s41746-023-00875-y] [Medline: 37468526]
35. Chen MC, Huang TY, Chen TY, Boonyarat P, Chang YC. Clinical narrative-aware deep neural network for emergency department critical outcome prediction. J Biomed Inform 2023 Feb;138:104284 [FREE Full text] [doi: 10.1016/j.jbi.2023.104284] [Medline: 36632861]
36. Smith J, Choi PM, Buntine P. Will code one day run a code? Performance of language models on ACEM primary examinations and implications. Emerg Med Australas 2023 Oct;35(5):876-878. [doi: 10.1111/1742-6723.14280] [Medline: 37414729]
37. Gupta P, Nayak R, Alazzeh M. The accuracy of medical diagnoses in emergency medicine by modern artificial intelligence. Acad Emerg Med 2023;30(Suppl 1):395 [FREE Full text] [doi: 10.1111/acem.14718]
38. Abavisani M, Dadgar F, Keikha M. A commentary on emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor's right-hand man. Int J Surg 2023 Oct 01;109(10):3195-3196 [FREE Full text] [doi: 10.1097/JS9.0000000000000561] [Medline: 37318859]
39. Rahman MA, Preum SM, Williams RD, Alemzadeh H, Stankovic J. EMS-BERT: a pre-trained language representation model for the emergency medical services (EMS) domain. In: Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies. 2023 Presented at: CHASE '23; June 21-23, 2023; Orlando, FL. [doi: 10.1145/3580252.3586978]
40. Lam WY, Au SC. Stroke care in the ChatGPT era: potential use in early symptom recognition. J Acute Dis 2023 Jun;12(3):129-130. [doi: 10.4103/2221-6189.379278]
41. Bushuven S, Bentele M, Bentele S, Gerber B, Bansbach J, Ganter J, et al. “ChatGPT, can you help me save my child’s life?” - diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital Basic Life Support and Paediatric Advanced Life Support cases – an in-silico analysis. Research Square. Preprint posted online May 12, 2023 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-2910261/v1]
42. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023 Apr;185:109729. [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
43. Preiksaitis C, Sinsky CA, Rose C. ChatGPT is not the solution to physicians' documentation burden. Nat Med 2023 Jun;29(6):1296-1297. [doi: 10.1038/s41591-023-02341-4] [Medline: 37169865]

44. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol 2023 Oct;20(10):998-1003. [doi: 10.1016/j.jacr.2023.06.009] [Medline: 37423350]
45. Dahdah JE, Kassab J, Helou MC, Gaballa A, Sayles S3, Phelan MP. ChatGPT: a valuable tool for emergency medical assistance. Ann Emerg Med 2023 Sep;82(3):411-413. [doi: 10.1016/j.annemergmed.2023.04.027] [Medline: 37330721]
46. Gottlieb M, Kline JA, Schneider AJ, Coates WC. ChatGPT and conversational artificial intelligence: friend, foe, or future of research? Am J Emerg Med 2023 Aug;70:81-83. [doi: 10.1016/j.ajem.2023.05.018] [Medline: 37229893]
47. Babl FE, Babl MP. Generative artificial intelligence: can ChatGPT write a quality abstract? Emerg Med Australas 2023 Oct;35(5):809-811 [FREE Full text] [doi: 10.1111/1742-6723.14233] [Medline: 37142327]
48. Chen J, Liu Q, Liu X, Wang Y, Nie H, Xie X. Exploring the functioning of online self-organizations during public health emergencies: patterns and mechanism. Int J Environ Res Public Health 2023 Feb 23;20(5):4012 [FREE Full text] [doi: 10.3390/ijerph20054012] [Medline: 36901022]
49. Bradshaw JC. The ChatGPT era: artificial intelligence in emergency medicine. Ann Emerg Med 2023 Jun;81(6):764-765. [doi: 10.1016/j.annemergmed.2023.01.022] [Medline: 37210166]
50. Cheng K, Li Z, Guo Q, Sun Z, Wu H, Li C. Emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor's right-hand man. Int J Surg 2023 Jun 01;109(6):1816-1818 [FREE Full text] [doi: 10.1097/JS9.0000000000000410] [Medline: 37074733]
51. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv. Preprint posted online February 26, 2023 2023 Feb 26 [FREE Full text] [doi: 10.1101/2023.02.21.23285886] [Medline: 36865204]
52. Brown C, Nazeer R, Gibbs A, Le Page P, Mitchell AR. Breaking bias: the role of artificial intelligence in improving clinical decision-making. Cureus 2023 Mar 20;15(3):e36415 [FREE Full text] [doi: 10.7759/cureus.36415] [Medline: 37090406]
53. Bhattaram S, Shinde VS, Khumujam PP. ChatGPT: the next-gen tool for triaging? Am J Emerg Med 2023 Jul;69:215-217. [doi: 10.1016/j.ajem.2023.03.027] [Medline: 37024324]
54. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May 09;15(5):e38755 [FREE Full text] [doi: 10.7759/cureus.38755] [Medline: 37303324]
55. Hamed E, Eid A, Alberry M. Exploring ChatGPT's potential in facilitating adaptation of clinical guidelines: a case study of diabetic ketoacidosis guidelines. Cureus 2023 May 09;15(5):e38784 [FREE Full text] [doi: 10.7759/cureus.38784] [Medline: 37303347]
56. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. Cureus 2023 Jun 13;15(6):e40351 [FREE Full text] [doi: 10.7759/cureus.40351] [Medline: 37456381]
57. Gebrael G, Sahu KK, Chigarira B, Tripathi N, Mathew Thomas V, Sayegh N, et al. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. Cancers (Basel) 2023 Jul 22;15(14):3717 [FREE Full text] [doi: 10.3390/cancers15143717] [Medline: 37509379]
58. Sarbay İ, Berikol G, Özturan İ. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. Turk J Emerg Med 2023 Jun 26;23(3):156-161 [FREE Full text] [doi: 10.4103/tjem.tjem_79_23] [Medline: 37529789]
59. Okada Y, Mertens M, Liu N, Lam SS, Ong ME. AI and machine learning in resuscitation: ongoing research, new concepts, and key challenges. Resusc Plus 2023 Jul 28;15:100435 [FREE Full text] [doi: 10.1016/j.resplu.2023.100435] [Medline: 37547540]
60. Chenais G, Lagarde E, Gil-Jardiné C. Artificial intelligence in emergency medicine: viewpoint of current applications and foreseeable opportunities and challenges. J Med Internet Res 2023 May 23;25:e40031 [FREE Full text] [doi: 10.2196/40031] [Medline: 36972306]
61. Chen HL, Chen HH. Have you chatted today? - medical education surfing with artificial intelligence. J Med Educ 2023 Mar;27(1):1-4 [FREE Full text]
62. Fanconi C, van Buchem M, Hernandez-Boussard T. Natural language processing methods to identify oncology patients at high risk for acute care with clinical notes. AMIA Jt Summits Transl Sci Proc 2023 Jun 16;2023:138-147 [FREE Full text] [Medline: 37350895]
63. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
64. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-10-06]
65. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. 2023 Feb 02. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2023-10-06]
66. Ziller A, Usynin D, Braren R, Makowski M, Rueckert D, Kaissis G. Medical imaging deep learning with differential privacy. Sci Rep 2021 Jun 29;11(1):13524 [FREE Full text] [doi: 10.1038/s41598-021-93030-0] [Medline: 34188157]
67. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med 2020 Sep 14;3:119 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]

68. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 01;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
69. Rose C, Barber R, Preiksaitis C, Kim I, Mishra N, Kayser K, et al. A conference (missingness in action) to address missingness in data and AI in health care: qualitative thematic analysis. J Med Internet Res 2023 Nov 23;25:e49314 [FREE Full text] [doi: 10.2196/49314] [Medline: 37995113]
70. Chenais G, Gil-Jardiné C, Touchais H, Avalos Fernandez M, Contrand B, Tellier E, et al. Deep learning transformer models for building a comprehensive and real-time trauma observatory: development and validation study. JMIR AI 2023 Jan 12;2:e40843. [doi: 10.2196/40843]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
ED: emergency department
EHR: electronic health record
EM: emergency medicine
LLM: large language model
MeSH: Medical Subject Headings
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Castonguay; submitted 19.10.23; peer-reviewed by L Zhu, C Gil-Jardiné, MO Khursheed; comments to author 13.12.23; revised version received 20.12.23; accepted 05.04.24; published 10.05.24.

Please cite as:
Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, Ribeira R, Rose C
The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review
JMIR Med Inform 2024;12:e53787
URL: https://medinform.jmir.org/2024/1/e53787
doi:10.2196/53787
PMID:38728687

Original Paper

Assessing ChatGPT as a Medical Consultation Assistant for Chronic Hepatitis B: Cross-Language Study of English and Chinese

Yijie Wang[1], MD; Yining Chen[2], MD; Jifang Sheng[1], MD

[1]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Disease, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Department of Urology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China

Corresponding Author:
Jifang Sheng, MD
State Key Laboratory for Diagnosis and Treatment of Infectious Diseases
Collaborative Innovation Center for Diagnosis and Treatment of Infectious Disease
The First Affiliated Hospital, Zhejiang University School of Medicine
79 Qingchun Road
Hangzhou, 310003
China
Phone: 86 13600519200
Email: jifang_sheng@zju.edu.cn

Abstract

Background: Chronic hepatitis B (CHB) imposes substantial economic and social burdens globally. The management of CHB involves intricate monitoring and adherence challenges, particularly in regions like China, where a high prevalence of CHB intersects with health care resource limitations. This study explores the potential of ChatGPT-3.5, an emerging artificial intelligence (AI) assistant, to address these complexities. With notable capabilities in medical education and practice, ChatGPT-3.5's role is examined in managing CHB, particularly in regions with distinct health care landscapes.

Objective: This study aimed to uncover insights into ChatGPT-3.5's potential and limitations in delivering personalized medical consultation assistance for CHB patients across diverse linguistic contexts.

Methods: Questions sourced from published guidelines, online CHB communities, and search engines in English and Chinese were refined, translated, and compiled into 96 inquiries. Subsequently, these questions were presented to both ChatGPT-3.5 and ChatGPT-4.0 in independent dialogues. The responses were then evaluated by senior physicians, focusing on informativeness, emotional management, consistency across repeated inquiries, and cautionary statements regarding medical advice. Additionally, a true-or-false questionnaire was employed to further discern the variance in information accuracy for closed questions between ChatGPT-3.5 and ChatGPT-4.0.

Results: Over half of the responses (228/370, 61.6%) from ChatGPT-3.5 were considered comprehensive. In contrast, ChatGPT-4.0 exhibited a higher percentage at 74.5% (172/222; $P<.001$). Notably, superior performance was evident in English, particularly in terms of informativeness and consistency across repeated queries. However, deficiencies were identified in emotional management guidance, with only 3.2% (6/186) in ChatGPT-3.5 and 8.1% (15/154) in ChatGPT-4.0 ($P=.04$). ChatGPT-3.5 included a disclaimer in 10.8% (24/222) of responses, while ChatGPT-4.0 included a disclaimer in 13.1% (29/222) of responses ($P=.46$). When responding to true-or-false questions, ChatGPT-4.0 achieved an accuracy rate of 93.3% (168/180), significantly surpassing ChatGPT-3.5's accuracy rate of 65.0% (117/180) ($P<.001$).

Conclusions: In this study, ChatGPT demonstrated basic capabilities as a medical consultation assistant for CHB management. The choice of working language for ChatGPT-3.5 was considered a potential factor influencing its performance, particularly in the use of terminology and colloquial language, and this potentially affects its applicability within specific target populations. However, as an updated model, ChatGPT-4.0 exhibits improved information processing capabilities, overcoming the language impact on information accuracy. This suggests that the implications of model advancement on applications need to be considered when selecting large language models as medical consultation assistants. Given that both models performed inadequately in emotional guidance management, this study highlights the importance of providing specific language training and emotional management strategies when deploying ChatGPT for medical purposes. Furthermore, the tendency of these models to use

disclaimers in conversations should be further investigated to understand the impact on patients' experiences in practical applications.

(JMIR Med Inform 2024;12:e56426) doi:10.2196/56426

KEYWORDS

chronic hepatitis B; artificial intelligence; large language models; chatbots; medical consultation; AI in health care; cross-linguistic study

Introduction

Chronic Hepatitis B: A Dual Burden on Patients and Society

Chronic hepatitis B (CHB) imposes significant economic and social burdens. In 2019, approximately 296 million people were affected by CHB, resulting in an estimated 820 thousand deaths [1]. The World Health Organization (WHO) noted that among those chronically infected with hepatitis B and C, about 20% or more would develop end-stage chronic liver disease, such as cirrhosis and hepatocellular carcinoma [2].

Hepatitis B virus (HBV) primarily spreads through blood contact, unprotected sexual intercourse, and mother-to-infant transmission. Effective management of chronic infection necessitates daily monitoring and self-care [3]. Nevertheless, the intricacy of regular monitoring, encompassing multiple tests, such as hepatitis B surface antigen (HBsAg), hepatitis B e-antigen (HBeAg), HBV-DNA, alanine transaminase (ALT), and fibrosis assessment, as endorsed by authoritative bodies in hepatitis B diagnosis and treatment, including the European Association for the Study of the Liver (EASL), presents hurdles to patient compliance [4]. Additionally, the prolonged, often lifelong, administration of antiviral medications contributes to further adherence issues [4,5]. Unique considerations for pregnant individuals and children add another layer of complexity, demanding targeted counseling and specialized management. This intricate management landscape not only burdens patients with emotional stress but also jeopardizes adherence to treatment regimens [6,7]. The complexity of CHB management requires personalized health care strategies, easing individual and societal burdens and emphasizing the importance of diverse health approaches.

ChatGPT as a Prospective Medical Assistant

Currently, artificial intelligence (AI) has become integral in the medical domain, particularly in medical research and clinical practice. Notably, according to Wani et al [8], traditional machine learning methodologies require the supervision of skilled individuals and structured input data, resulting in considerable resource-intensive processes. Recognizing the limitations of traditional approaches, Haug et al [9] proposed chatbots for capabilities in medical practice assistance.

ChatGPT-3.5, which was released in June 2020, underpins ChatGPT's emergence in AI-assisted medical applications. As a large language model (LLM), it shows potential for medical assistance [10], though challenges and concerns persist in its application within the field [11]. The functioning of LLMs involves predicting and generating a coherent and contextually relevant response based on preinput materials, necessitating training on massive amounts of diverse textual data. Various studies have explored ChatGPT's capacity to act as a virtual doctor or medical tutor for diagnosis or treatment [12].

The study by Gilson et al [13] revealed that ChatGPT performed well in medical knowledge assessments, demonstrating potential as a virtual medical tutor. Yeo et al [14] evaluated ChatGPT's performance in answering questions regarding cirrhosis and hepatocellular carcinoma. Most studies have compared its performance to that of real doctors or medical students, aiming to determine whether AI assistants could surpass human medical service providers. However, there are challenges and risks of employing ChatGPT in clinical practice, including the potential generation of plausible yet inaccurate information and ethical considerations [15]. According to these studies, LLMs could potentially assist in medical consulting and auxiliary diagnosis, as well as traditional medical research, treatment, and education, but there are still unidentified risks and problems.

ChatGPT-4.0, which was released on March 14, 2023, is an updated version of the ChatGPT model. Many researchers have compared the applications of ChatGPT-3.5 and ChatGPT-4.0 in medical practice [16-18]. In this research, we included ChatGPT-4.0 as a comparative model to further assess the application problem of this model.

Medical Assistance in Hepatitis B Management With Chinese as the Primary Language

Bearing the highest global burden of hepatitis B, China recorded over 90 million people living with CHB in 2017 [1,2]. Research has revealed troubling trends in treatment noncompliance for HBV in China, including challenges in preventing vertical transmission [19-21]. Beyond China, studies in various regions have highlighted the impact of factors like family income, employment, and patient gender on medical treatment compliance for CHB [22].

Physician encouragement is crucial for patient compliance with medication regimens [23]. Despite a rising number of medical doctors in China, there is a shortage of medical practitioners, including licensed physicians and physician assistants, who face high workloads and burnout rates [24-26]. While research has indicated that Chinese physicians generally adhere to hepatitis B guidelines [27], medical errors due to workload demands could undermine intended impacts on patient compliance [26]. Amidst these challenges, exploring medical assistance using Chinese as the primary working language is crucial. This inquiry is vital for enhancing patient compliance in hepatitis B management and alleviating strain on health care professionals amid work-related stress. However, a study involving ChatGPT's performance in a medical examination in Chinese

emphasized the significance of exploring the cross-language difference in ChatGPT's performance in a future study [28].

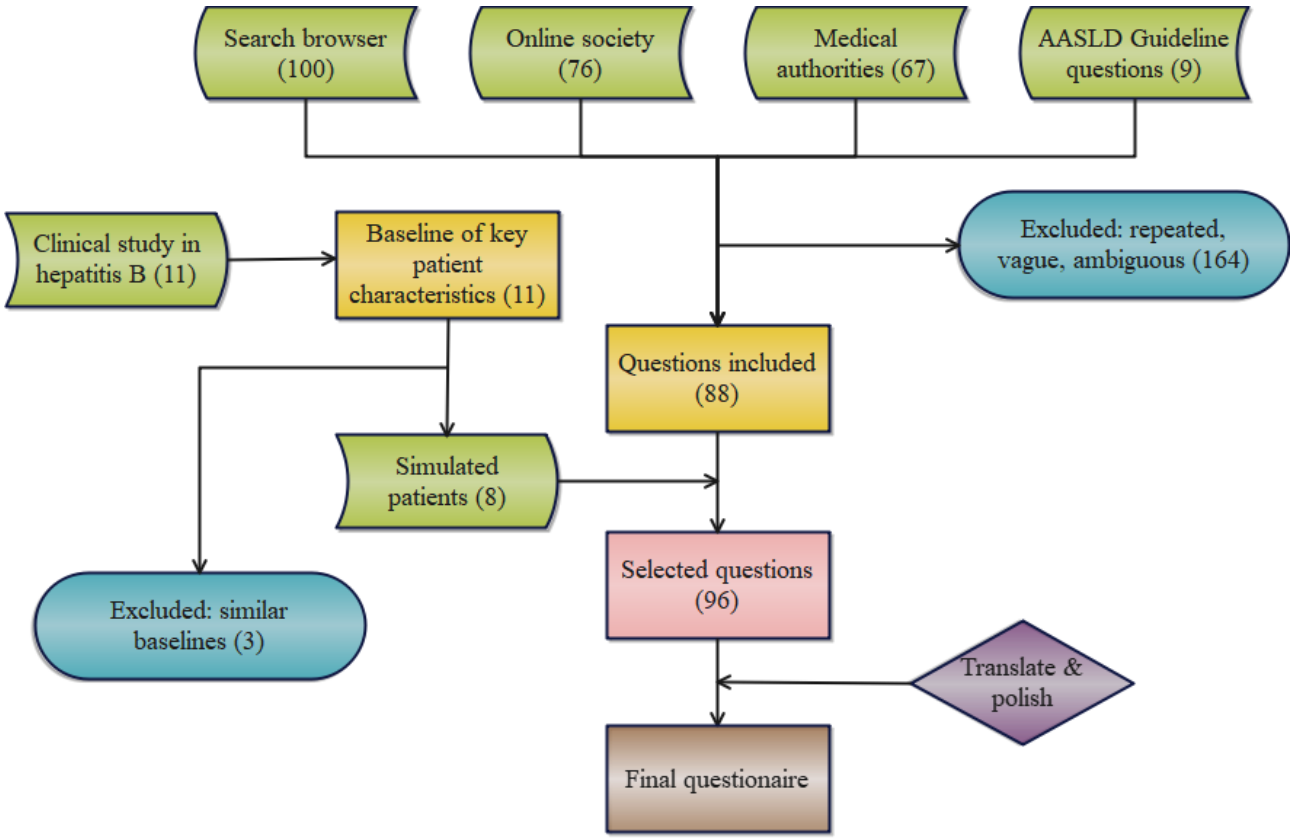
In brief, a dialogue-based medical assistant is being increasingly recognized as essential in clinical practice. Exploring the application of ChatGPT in this domain shows promise for medical research and clinical usage. This study assessed ChatGPT-3.5 and ChatGPT-4.0 in tasks such as diagnosis, providing management advice, and addressing counseling needs among patients with CHB. Given that English data account for the largest proportion of data (approximately 92%) used for the original training of this model [29], it is reasonable to assume that among all the languages included in the pretraining resource, this model performs best in English. However, the Chinese language is used by the world's largest group of CHB patients, underscoring the irreplaceable role of Chinese in studies regarding medical AI assistants. Therefore, this research includes both English and Chinese as working languages and compares ChatGPT's performance in both languages. Additionally, the study compares ChatGPT-3.5 and ChatGPT-4.0 to investigate the improvements from the former version to the latter version. Through this investigation, we aim to uncover the potential of this application and its limitations in medical practice.

Methods

Questionnaire Development Process

Following the workflow shown in Figure 1, we systematically compiled a set of questions relevant to both patients and physicians in clinical practice. This compilation process involved several processes. First, we sourced questions from esteemed professional associations and institutions, such as the WHO, American Centers for Disease Control and Prevention (CDC), American Association for the Study of Liver Disease (AASLD), EASL, and Asian Pacific Association for the Study of the Liver (APASL). Second, we identified queries about hepatitis B found on online social media platforms, particularly in patient support groups and disease-specific forums. The inclusion criteria prioritized relevance to diagnosis, treatment, daily monitoring, lifestyle, and other hepatitis B–related concerns. Questions with precise wording, minimal grammatical errors, and clarity were included, while nonmedical inquiries, ambiguously framed questions, and those related to nonmedical issues were excluded. Questions with significant updates after September 2021 were also omitted. Third, we conducted an exploration of associative keywords following the entry of "hepatitis B" or "HBV" into widely used search engines, such as Google, Bing, and Baidu, in both Chinese and English languages. Fourth, based on diverse published hepatitis B clinical studies, we systematically extracted key patient characteristics, including age, gender, hepatitis B serum markers, HBV-DNA levels, ALT levels, and concomitant diseases. We developed profiles for 8 simulated patients using these data with a random number function. ChatGPT was tasked with providing advice to these simulated individuals on various aspects, including treatment or examination recommendations, treatment strategies, daily monitoring practices, lifestyle adjustments, etc.

Figure 1. Workflow of the questionnaire design process. The specific information of each stage of the questionnaire compiling process is shown. AASLD: American Association for the Study of Liver Disease.



Among all the questions gathered, multiple questions were separated into single entities, while repeated questions were excluded. To avoid ambiguity and misunderstanding resulting from language vagueness, which could potentially impact the assessment of the model's information accuracy, we carefully polished all the collected questions, refined their grammar and phrasing, and performed localized translations between Chinese and English. Examples of revised or excluded questions are provided in Multimedia Appendix 1. In total, we gathered 96 questions about hepatitis B. Among the questions, there were 2 with only an English version and 5 with only a Chinese version. These language-specific questions focused on issues specific to the country or region where the questioner was located.

Section Allocation

We systematically categorized all questions into 5 distinct sections: Term Explanation Questions, Short Answer Questions, Clinical Problem Questions, AASLD Guideline Questions, and Simulated Patient Questions.

The "Term Explanation Questions" section featured 17 terms associated with hepatitis B, including 1 term exclusively for Chinese responses. In the "Short Answer Questions" section, there were 22 questions, with 1 specifically designed for Chinese responses. Questions within the "Clinical Problem Questions" section were primarily sourced from online hepatitis B societies, totaling 40 questions. Within this section, there was 1 question intended solely for English responses and 2 exclusively for Chinese responses. The questions in the "AASLD Guideline Questions" section were derived from the AASLD guidelines for hepatitis B in 2016 and 2018 (updated version) and included 9 questions that were all translated into Chinese. The "Simulated Patient Questions" section consisted of 8 questions related to simulated patient information, as previously constructed. These questions were provided in both English and Chinese versions.

Gathering Responses

The questions were submitted to ChatGPT-3.5 from April 1 to April 30, 2023, with each question forming a separate dialogue. Each question was sent twice for Chinese and English separately to ensure a comprehensive evaluation, and responses were collected. In the case of a system error preventing ChatGPT-3.5 from responding, the question was resubmitted in a new dialogue. All responses were compiled into a table for further assessment.

Assessment of Responses

Two senior physicians independently evaluated all responses. In the case of discrepancies in information accuracy, consistency of repeated responses, and emotional management guidance assessments, a third senior physician with over 30 years of experience in hepatitis B diagnosis and treatment conducted a final review for the ultimate assessment and provided the final scores. The criterion of assessment was discussed and voted on by a committee of 5 senior physicians in hepatitis B diagnosis and treatment. The assessment process referred to the research of Yeo et al [14].

Information Accuracy Assessment

The information accuracy assessment was mainly focused on correctness and comprehensiveness. Four assessment grades (1-4) were assigned: 1, correct and comprehensive; 2, correct but with missing information; 3, a mix of correct and incorrect details; and 4, wholly incorrect or irrelevant information.

Categorization of the Types of Mistakes

Mistakes in responses assessed as "a mix of correct and incorrect details" and "wholly incorrect or irrelevant information" were analyzed, and the types of mistakes were categorized. The mistakes were classified into 5 categories: A, misunderstanding of medical terms or jargon; B, incorrect usage of medical terms; C, mistakes in diagnosis/treatment/management without mistakes in terms or jargon; D, total irrelevant information; and E, a mixture of two or more kinds of mistakes among A-C.

Content Consistency of Repeated Response Assessment

A binary assessment ("Yes" or "No") was employed to indicate the consistency of the 2 responses for each question. This evaluation was independent of the information accuracy assessment and solely focused on the consistency of the response content.

Emotional Management Guidance Assessment

For all responses in the "Clinical Problem Questions" and "Simulated Patient Questions" sections (48 in total; 1 with only an English version and 2 with only a Chinese version), an emotional management guidance assessment was conducted. The assessment comprised three levels: (1) sufficient emotional and psychological management guidance, (2) respectful but lacking or inadequate emotional or psychological management guidance, and (3) disrespectful or negative emotional guidance.

Analysis of ChatGPT's Cautionary Statements Regarding Medical Advice

We quantified the instances where ChatGPT recommended consulting a genuine health care provider or doctor. Meanwhile, we counted the frequency of ChatGPT explicitly stating disclaimers, such as "I am not a doctor" and "I cannot give diagnosis or treatment," among all questions involving clinical practice (including the sections of Clinical Problem Questions, AASLD Guideline Questions, and Simulated Patient Questions).

Parallel Assessment of ChatGPT-4.0's Performance

We replicated the above assessment process for ChatGPT-4.0. Considering that ChatGPT-4.0 is the updated version of the model, we omitted sections involving only basic medical knowledge in the questionnaire. As a more intuitive alternative, we chose closed questions to evaluate the fundamental knowledge differences between the 2 model versions. The assessment of ChatGPT-4.0 included questions from the "Clinical Problem Questions," "AASLD Guideline Questions," and "Simulated Patient Questions" sections of the questionnaires used in the previous assessment. However, mistake analysis was omitted as there were no responses from ChatGPT-4.0 that were assessed as incorrect.

Comparison of ChatGPT-3.5 and ChatGPT-4.0 Using Closed Questions (True-or-False Statements)

In this assessment, we formulated 30 statements based on AASLD guidelines for the treatment of CHB, including all its updates up to September 2021. These statements were input into the models in separate dialogues. We used prompts to ask the models to judge whether the statements were correct and to provide a judgment with "Yes" or "No." The prompts are detailed in Table 1. Each statement was input into the model 3 times, and the response for each iteration was recorded. All responses of the models were collected, and their accuracy and stability (the consistency of 3 responses to a repeated statement) were assessed.

Table 1. An example of the prompts used in closed questions.

Language	Prompts for ChatGPT-4.0	Prompts for ChatGPT-3.5
English	Now, I would like you to act as a hepatologist in the upcoming conversation and determine whether the statements are true and answer with only "Yes" or "No". Here are the statements: []a	Now, I would like you to act as a hepatologist in the upcoming conversation and determine whether the statements are true and answer with only "Yes" or "No", and do not add any explanation. Here are the statements: []
Chinese	现在，我希望你在接下来的对话中扮演一名肝病学医师，判断以下陈述是否正确，并用"是"或"否"来回答: []a	现在，我希望你在接下来的对话中扮演一名肝病学医师，判断以下陈述是否正确，并用"是"或"否"来回答，不要增加任何解释或说明: []

[a]The statements were added in square brackets.

Statistical Analysis

All statistical analyses were performed with the SPSS 26.0 statistical package (IBM Corp). Cohen kappa coefficients were used to determine interobserver reliabilities. Assessment grades were calculated and reported as percentages. Comparative analysis of ranked data employed the Mann-Whitney test. Categorical data were compared using chi-square tests. The Wilcoxon signed rank test was applied to compare the grades of response 1 and response 2 to each question. Statistical significance was set at $P<.05$.

Ethical Considerations

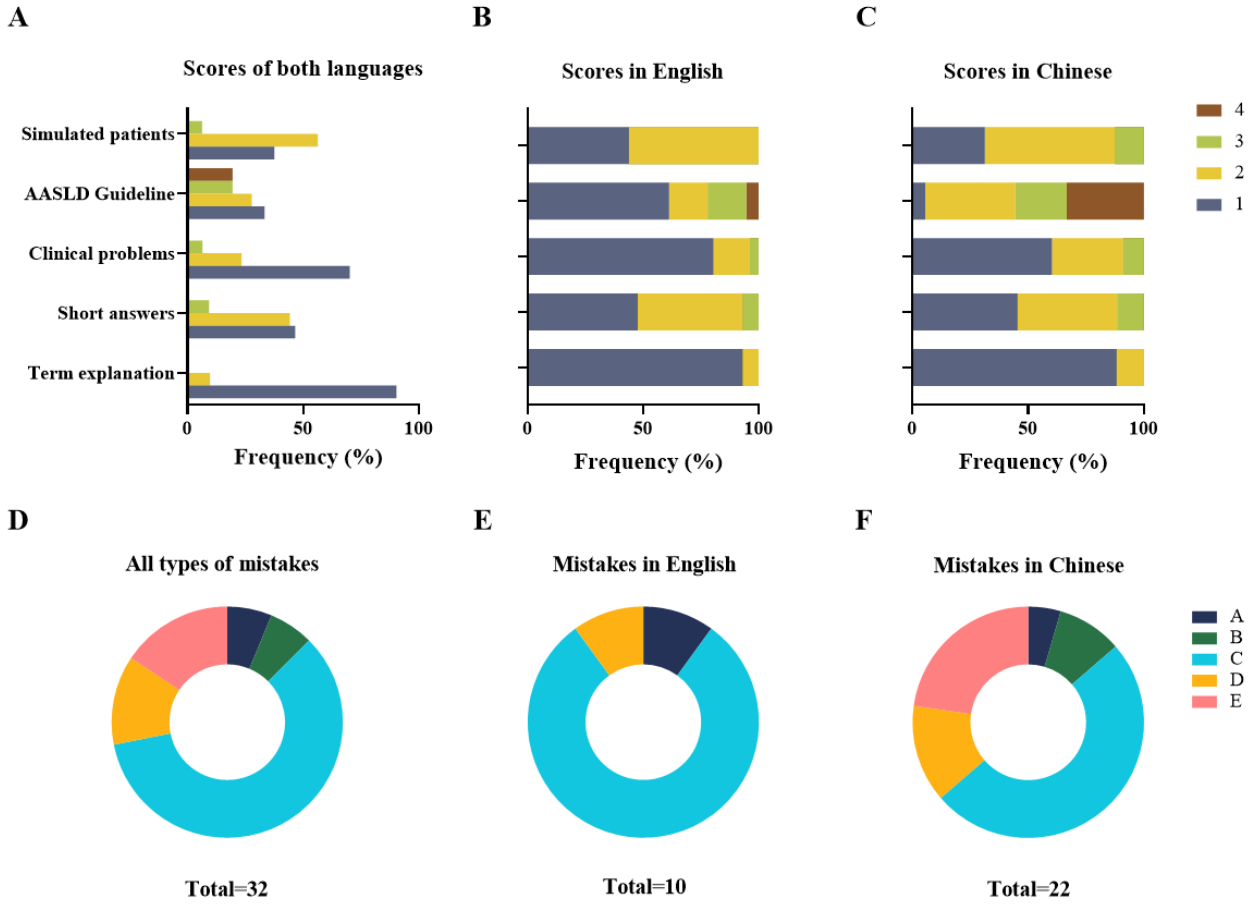
Our study did not use any information of real-world humans. Questions obtained from the internet were all polished and revised, and were included in the model with no personal information. Files of simulated patients were composed based on baseline data of clinical trials of hepatitis B, but the numbers were modified using a random number function to avoid possible leakage of real information.

Results

Information Accuracy Assessment of ChatGPT-3.5

The interobserver reliability κ was 0.6020 ($P<.001$) for the information accuracy assessment. The results of this assessment are shown in Figure 2. Across all the questions, 90.8% (336/370) of responses from ChatGPT-3.5 contained no incorrect information (including comprehensive responses and correct but incomprehensive responses). The likelihood of ChatGPT giving correct and comprehensive responses was 61.6% (228/370), while there was a 29.2% (108/370) probability of responses being correct with missing information (Multimedia Appendix 2). Responses with a mix of correct and incorrect information accounted for 7.3% (27/370) of responses. There were 1.9% (7/370) of responses wholly incorrect or irrelevant to the questions.

Figure 2. Results of the information accuracy assessment and mistake analysis of ChatGPT-3.5. (A) Comparison of the percentage for each grade across all responses in distinct question sections. (B) Percentage of each grade of responses in English in separated question sections. (C) Percentage of each grade of responses in Chinese in separated question sections. (D) Overview of mistake types across all responses. (E) Breakdown of mistake types specifically among responses in English. (F) Breakdown of mistake types specifically among responses in Chinese. In parts D-F, grade A indicates misunderstanding of medical terms or jargon, grade B indicates incorrect usage of medical terms, grade C indicates mistakes in diagnosis/treatment/management without mistakes in terms or jargon, grade D indicates total irrelevant information, and grade E indicates a mixture of two or more kinds of mistakes among grades A-C. AASLD: American Association for the Study of Liver Disease.



The performance of ChatGPT-3.5 varied across the sections, and the differences were statistically significant (P<.0001; Multimedia Appendix 2). The section "Term Explanation Questions" had the highest percentage of responses assessed as complete and comprehensive (26/28, 93% in English and 30/34, 88% in Chinese; Figure 2A), while the section "AASLD Guideline Questions" had the highest percentage of responses totally wrong or irrelevant, or mixed with incorrect information (4/18, 22% in English and 10/18, 56% in Chinese).

The language environment in which ChatGPT-3.5 operated also influenced its performance (Figure 2B). ChatGPT demonstrated poorer performance in Chinese than in English (P=.001), particularly in the sections "Clinical Problem Questions" (P=.03) and "AASLD Guideline Questions" (P=.002). However, performance in the sections "Term Explanation Questions" (P=.54), "Short Answer Questions" (P=.62), and "Simulated Patient Questions" (P=.33) showed no significant difference between the 2 working languages. The evaluation table is presented in Multimedia Appendix 3.

Categorization of the Types of Mistakes of ChatGPT-3.5

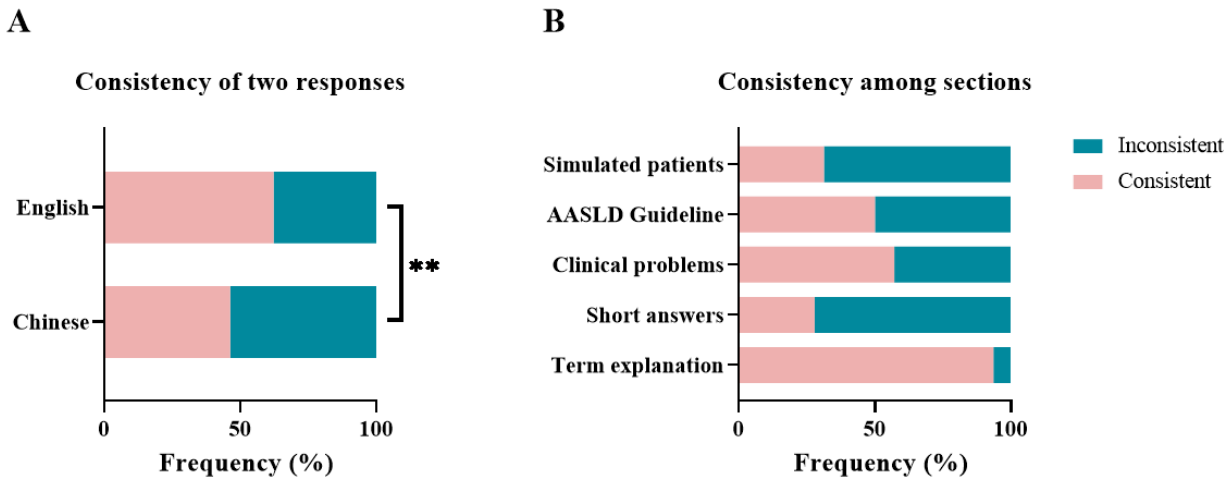
Figures 2D-F summarize the types of mistakes in the responses. In both languages, the most common error pertained to diagnosis, treatment, or disease management (Figure 2D). Notably, in Chinese, 10 out of 32 mistakes (31%) involved incorrect usage or misunderstanding of technical terms (Figure 2F), while in English, there were no such mistakes (Figure 2E). The evaluation table is provided in Multimedia Appendix 4.

Content Consistency of the Repeated Response Assessment of ChatGPT-3.5

The interobserver reliability κ was 0.6532 (P<.001) for content consistency of the repeated response assessment of ChatGPT-3.5. Figure 3 shows the content consistency of repeated responses. For all questions, the probability of content consistency between 2 responses was 54.1% (100/185 pairs of responses). In English, the consistency was 62% (56/90 pairs of responses), while in Chinese, it was 46% (44/95 pairs of responses), showing a significant difference (P=.04; Figure 3A and Multimedia Appendix 2). This disparity was also significant in the section "Clinical Problem Questions" (P=.04; Figure 3B

and Multimedia Appendix 2). The "Term Explanation Questions" section had the highest consistency at 94% (29/31 pairs of responses), while the "Short Answer Questions" section had the lowest consistency at 30% (13/43 pairs of responses). Despite poor content consistency, the responses exhibited similarity in grades (P=.65). The evaluation table is provided in Multimedia Appendix 5.

Figure 3. Assessment of the content consistency of responses to repeated questions. (A) Comparison of content consistency between responses in different working languages. (B) Examination of content consistency in different sections of questions. AASLD: American Association for the Study of Liver Disease. **P<.01.



Emotional Management Guidance Assessment of ChatGPT-3.5

Among responses to questions within the "Clinical Problem Questions" and "Simulated Patient Questions" sections, only 3.2% (6/186) were deemed to provide sufficient emotional management support (Table 2). Related responses are listed in Multimedia Appendix 6. Most responses (180/186, 96.8%) were assessed as "respectful but lacking or inadequate emotional or psychological management guidance." No response was assessed as "disrespectful or negative emotional guidance." ChatGPT-3.5 exhibited comparable performance in both languages (P=.39).

Table 2. Results of the emotional management guidance assessment.

Language	Clinical Problem Questions (n=154), n			Simulated Patient Questions (n=32), n			Total (N=186), n			*P* value
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	
Chinese	1	77	0	1	15	0	2	92	0	.48[a]
English	4	72	0	0	16	0	4	88	0	
Total	5	149	0	1	31	0	6	180	0	.39[b]

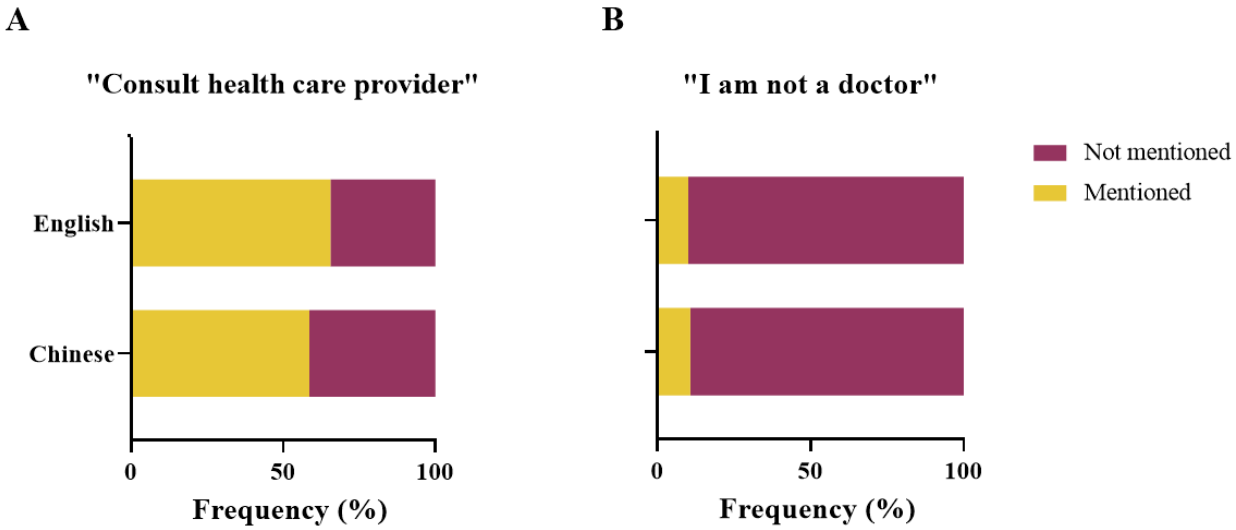
[a]*P* value across the grades of each section.

[b]*P* value between the grades of different working languages.

Analysis of the Cautionary Statements of ChatGPT-3.5 Regarding Medical Advice

Figure 4 shows the results of this assessment. ChatGPT-3.5 exhibited distinct characteristics as a medical assistant. In most responses, ChatGPT-3.5 tended to remind patients to consult a health care provider or a physician (227/370, 61.4%). This percentage was consistent in both English (118/180, 65.6%; Figure 4A) and Chinese (109/190, 57.4%), with no significant difference (P=.20). These responses are listed in Multimedia Appendix 7.

Figure 4. Percentage of ChatGPT cautionary statements regarding medical advice and disclaimers. (A) Percentage of responses that include the recommendation to "consult health care providers or doctors." (B) Percentage of responses containing the disclaimer phrase "I am not a doctor" or "I cannot give diagnosis or treatment".

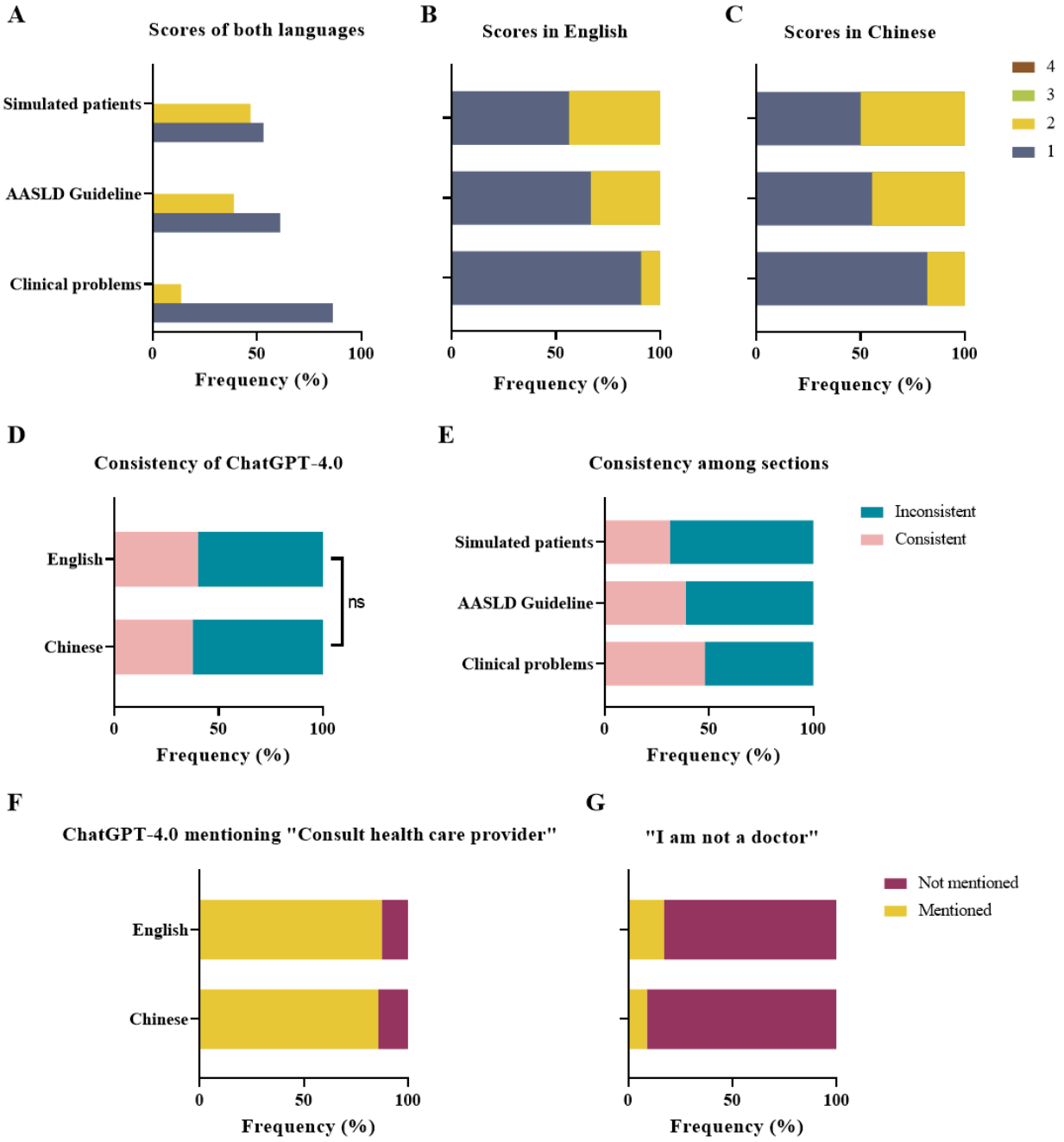


Among all questions involving clinical practice, the probability of ChatGPT-3.5 using the phrase "I am not a doctor" or "As a language model, I cannot give diagnosis or treatment…" was 10.8% (24/222). The probability was 11.6% (13/112) in Chinese and 10.0% (11/110) in English (Figure 4B). No significant difference was observed between the 2 languages (P>.99). These responses are listed in Multimedia Appendix 8.

Parallel Assessment of ChatGPT-4.0 in Sections Involving Clinical Practice

The interobserver reliability κ was 0.6896 (P<.001) for information accuracy assessment. Notably, ChatGPT-4.0 demonstrated distinct performance compared to ChatGPT-3.5. The scores of ChatGPT-4.0 are presented in Figures 5A-C. Across the 3 sections of "Clinical Problem Questions," "AASLD Guideline Questions," and "Simulated Patient Questions," the percentage of responses assessed as complete and comprehensive, as well as "grade 1," was higher for ChatGPT-4.0 than for ChatGPT-3.5 (172/222, 77.5% vs 132/222, 59.5%), with a significant difference (P<.001). Furthermore, variations in grades were observed across the sections (P<.001). The "Clinical Problem Questions" section exhibited the highest percentage of responses assessed as complete and comprehensive for ChatGPT-4.0 (133/154, 86.4%), surpassing that for ChatGPT-3.5 (108/154, 70.1%; P<.001). Importantly, no responses from ChatGPT-4.0 were assessed as "a mix of correct and incorrect details" and "wholly incorrect or irrelevant information." In general, ChatGPT-4.0 demonstrated superior information accuracy compared to ChatGPT-3.5. Moreover, ChatGPT-4.0 showed improved performance in responding to Chinese questions. Although there was a slightly lower percentage of responses assessed as "grade 1" for Chinese (82/112, 73.2%) than for English (90/110, 81.8%), the difference in performance between the languages was not significant (P=.13). The evaluation tables are presented in Multimedia Appendix 9 and 10.

Figure 5. Results of the parallel assessment of ChatGPT-4.0 in sections involving clinical practice. (A) Comparison of the percentage of each grade across all responses in distinct question sections. (B) Percentage of each grade of responses in English in separated question sections. (C) Percentage of each grade of responses in Chinese in separated question sections. (D) Comparison of content consistency between responses in different working languages. (E) Examination of content consistency in different sections of questions. (F) Percentage of responses that include the recommendation to "consult health care providers or doctors." (G) Percentage of responses containing the disclaimer phrase "I am not a doctor" or "I cannot give diagnosis or treatment." AASLD: American Association for the Study of Liver Disease; ns: not significant.



The interobserver reliability κ was 0.6052 (P<.001) for content consistency of the repeated response assessment. ChatGPT-4.0 showed poor consistency in responses to repeated questions. Across all questions, ChatGPT-4.0 provided 44.1% (49/111 pairs of responses) stable repeated responses, and this proportion was lower than that for ChatGPT-3.5 (58/111, 52.3%). However, the difference was not significant (P=.23). Specifically, ChatGPT-4.0's stability percentage was 38% (21/56 pairs of responses) in Chinese and 51% (28/55 pairs of responses) in English (Figure 5D), with no significant difference (P=.16). Among all the sections, responses in the "Clinical Problem Questions" section exhibited the highest rate of consistency at 48% (37/77 pairs of responses; Figure 5E). The difference in consistency across sections was not significant (P=.42). Detailed evaluation tables are provided in Multimedia Appendix 9 and 11.

Regarding responses to questions within the "Clinical Problem Questions" and "Simulated Patient Questions" sections, ChatGPT-4.0's responses were assessed to provide sufficient emotional management support 8.1% (15/186) of the time (Table

3). This performance differed significantly from that of ChatGPT-3.5 (*P*=.04). The percentage was similar between Chinese and English (7/94, 7% and 8/92, 9%, respectively; *P*=.76). No response was assessed as "disrespectful or negative emotional guidance." ChatGPT-4.0 showed similar performance between the 2 sections (10/154, 6.5% for Clinical Problem Questions and 5/32, 15.6% for Simulated Patient Questions assessed as grade 1; *P*=.08). However, among all responses assessed as "unstable," there was no significant difference between the scores of response 1 and response 2 (*P*=.06). All the responses assessed as "sufficient emotional and psychological management guidance" are listed in Multimedia Appendix 12.

Table 3. Results of the emotional management guidance assessment of ChatGPT-4.0.

Language	Clinical Problem Questions (n=154), n			Simulated Patient Questions (n=32), n			Total (N=186), n			*P* value
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	
Chinese	5	73	0	2	14	0	7	87	0	.08[a]
English	5	71	0	3	13	0	8	84	0	
Total	10	144	0	5	27	0	15	171	0	.75[b]

[a]*P* value across the grades of each section.

[b]*P* value between the grades of different working languages.

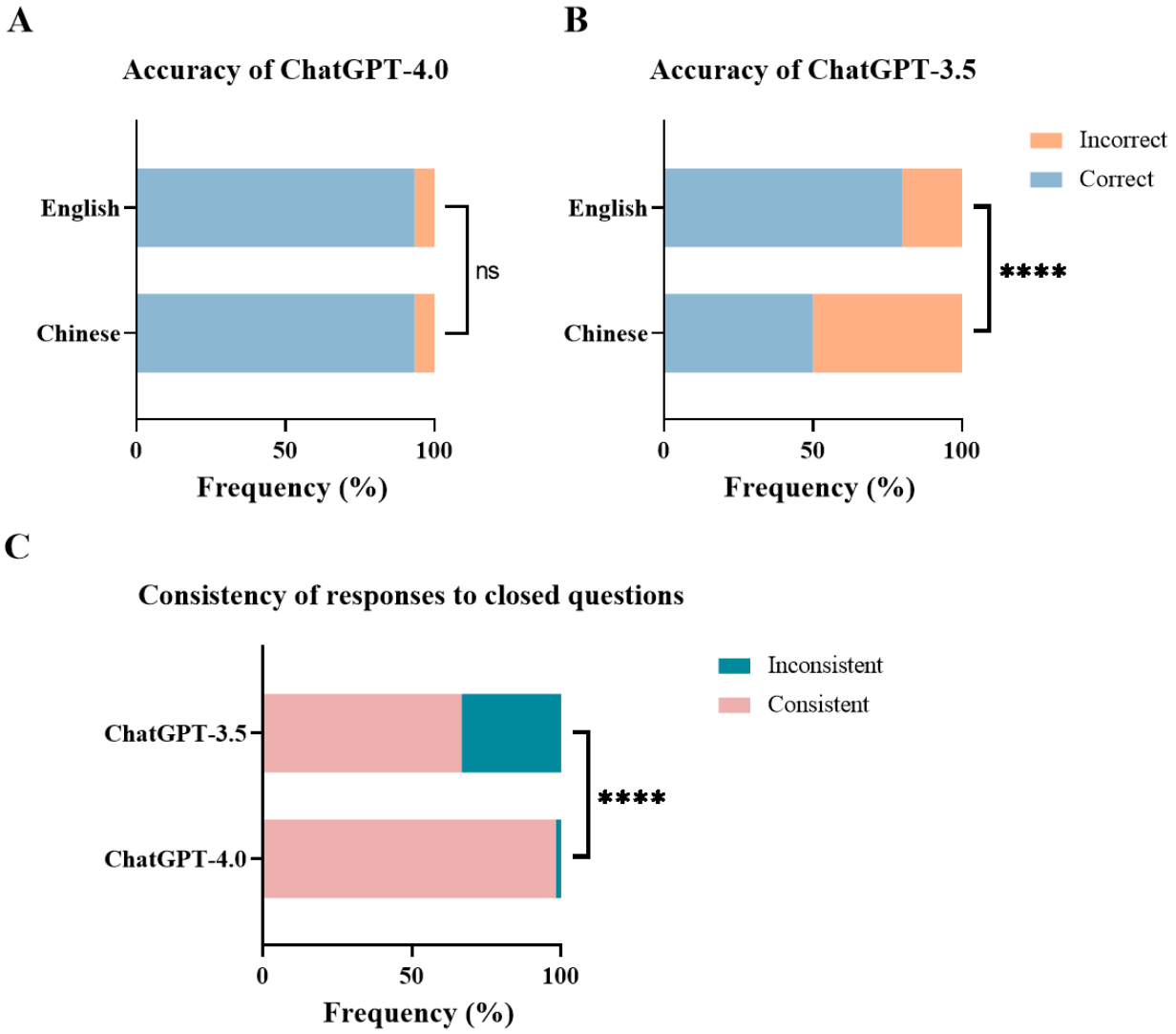
As shown in Figure 5F, ChatGPT-4.0 demonstrated comparable performance to ChatGPT-3.5 across all responses, with 86.5% (192/222) of responses emphasizing the importance of seeking medical assistance. In Chinese, 96 out of 112 responses (85.7%) stressed this need, while in English, 96 out of 110 responses (87.3%) did the same. Notably, no significant difference was observed between the languages (*P*=.73). In responses from ChatGPT-3.5 in the sections "Clinical Problem Questions," "AASLD Guideline Questions," and "Simulated Patient Questions," the total percentage of responses with a medical service recommendation was 81.5% (181/222), which was not different from that for ChatGPT-4.0 (*P*=.15). All responses emphasizing the necessity of seeking medical services are listed in Multimedia Appendix 13.

Figure 5G illustrates that among all responses to questions involving clinical practice, ChatGPT-4.0 used the phrase "I am not a doctor" or "As a language model, I cannot give diagnosis or treatment…" with a probability of 13.1% (29/222). This percentage did not significantly differ from that of ChatGPT-3.5 (*P*=.46). In Chinese, the probability was 8.9% (10/112), while in English, the probability was 17.3% (19/110; Figure 5B). No significant difference was observed between the 2 languages (*P*=.07). These responses are detailed in Multimedia Appendix 14.

Assessment of Responses to Closed Questions Across ChatGPT-4.0 and ChatGPT-3.5

When assessing the accuracy of statements derived from the AASLD guidelines for the treatment of CHB, ChatGPT-4.0 exhibited significantly superior performance compared to ChatGPT-3.5 (Figures 6A and B). ChatGPT-4.0 achieved a correctness percentage of 93.3% (168/180), with the same percentage accuracy in both Chinese and English (93.3% for each language). Conversely, ChatGPT-3.5 yielded an overall accuracy of 65.0% (117/180), with a split of 50.0% (45/90) in Chinese and 80.0% (72/90) in English.

Figure 6. Results of responses to true-or-false questions. (A) Rate of accuracy of ChatGPT-4.0. (B) Rate of accuracy of ChatGPT-3.5. (C) Comparison of the consistency of responses to true-or-false questions between ChatGPT-4.0 and ChatGPT-3.5. ns: not significant. ****P<.0001.



Furthermore, ChatGPT-4.0 displayed enhanced consistency in repeated responses (Figure 6C). Stable responses accounted for 98.3% (59/60) of responses in ChatGPT-4.0, whereas ChatGPT-3.5 provided only 66.7% (40/60) stable responses. The difference in response stability between the models was statistically significant (P<.001). Details are provided in Multimedia Appendix 15.

Discussion

ChatGPT-3.5 Working as a Medical Consulting Assistant

Our evaluation highlighted the proficiency of ChatGPT-3.5 as a medical consultation assistant. ChatGPT-3.5 provided predominantly accurate information, but there was a notable limitation in the comprehensiveness of the responses, indicating a need for targeted medical professional input. Continuous enhancement of LLMs may contribute to more specific and reliable guidance. Despite its strengths, ChatGPT-3.5 displayed limitations in emotional management support, a crucial aspect of chronic disease management [30]. Facilitating emotional modulation is integral to fostering patient willingness for self-management and treatment compliance [7,30].

Therefore, it is imperative to consider emotional cognition and regulation in medical diagnosis and treatment. Our study suggested that the potential for ChatGPT to serve as an emotional management assistant for chronic patients warrants further study, with related localized training considered if LLMs are to be employed in clinical practice as health consultation assistants.

Impact of Working Language on Performance

By revealing ChatGPT's inferior performance in Chinese compared to English, the study emphasized the influence of the choice of working language on stability and correctness. ChatGPT-3.5 showed worse performance on information accuracy in Chinese, implying the insufficient input of knowledgeable materials in Chinese. Both ChatGPT-3.5 and ChatGPT-4.0 showed less stability in Chinese, which was reflected in a lower consistency rate of responses to the same

questions. This challenge stemmed from variations in language resources during the model's original training, primarily centered around English-based medical guidelines. Though there are Chinese versions of these guidelines, the timeliness and accuracy of Chinese materials are limited. To enhance ChatGPT's efficacy in diverse language environments, the model should undergo additional training based on data sourced from specific language resources. This targeted training should focus on potential misunderstandings related to terms and phrasings in local languages, thereby addressing language-specific nuances and enhancing overall performance. Notably, ChatGPT-3.5 exhibited language-specific mistakes, with Chinese responses showing errors related to misunderstanding or incorrect usage of terms. This underscores the importance of targeted language training for LLMs to minimize inaccuracies, especially in medical contexts.

Cautionary Statements and Patient-Oriented Usage

In discussions related to diagnosis and therapy, ChatGPT-3.5 consistently emphasized the importance of consulting a health care provider, indicating a cautious approach. Owing to constraints in both timeliness and accuracy inherent in language models, ChatGPT-3.5 occasionally emphasized its nondoctor status, thus refraining from providing direct diagnosis or therapy in the conversation. However, such statements may imply the unreliability of the medical judgment, especially in a Chinese cultural context. Thus, further inquiries are warranted to evaluate the potential risks and benefits of this response mode, considering its impact on patient trust and compliance challenges.

Implications for Future Development in Clinical Medicine

As AI, including LLMs, is being progressively integrated into clinical medicine, understanding the advantages and disadvantages is paramount. While ChatGPT demonstrates promise as a medical consulting assistant for CHB patients, future research and development should prioritize targeted language input and emotional management training. Besides, establishing and updating prompts, which are in a specific order, and templates based on which LLMs could provide responses in a standardized format would significantly enhance ChatGPT's performance. Overcoming language barriers and addressing emotional support deficiencies will be crucial for maximizing the potential benefits of LLMs in medical assistance.

Comparison of ChatGPT-4.0 to ChatGPT-3.5

ChatGPT-4.0 demonstrated superior performance compared with ChatGPT-3.5 in terms of information accuracy. This improvement aligns with the expected advancements in ChatGPT-4.0 as a more advanced iteration. However, ChatGPT-4.0 did not exhibit better response stability in open-ended questions. This could be attributed to a reduced ability to follow chain-of-thought prompting [31]. Despite this inconsistency, it did not affect the accuracy of information, suggesting that LLMs tend to employ diverse language patterns and content combinations.

In responses to closed questions (30 true-or-false statements based on the AASLD guidelines for the treatment of hepatitis B), ChatGPT-4.0 demonstrated a higher rate of accuracy and stability, indicating substantial improvement in the model's understanding of hepatitis B medical knowledge as the model progressed.

The improvement of ChatGPT-4.0 in terms of information accuracy suggested the tremendous benefit of the model update, but the deficiency in emotional management remained. Therefore, additional training related to emotional management guidance and humanistic care is essential for the preparation of the model before application.

Notably, in responses to open questions, ChatGPT-4.0 displayed interesting changes compared to ChatGPT-3.5. ChatGPT-4.0 included reference information in 5 of the responses, all of which were verified to be accurate. This suggests an enhancement in the format and reliability of ChatGPT-4.0. However, the impact of such changes on the patient experience warrants further exploration. Additionally, ChatGPT-4.0 was more likely to use a direct disclaimer like "I am an AI model..." or "I'm not a doctor..." and even "Disclaimer: I'm not a doctor...," indicating a more stringent approach. However, the increase in possibility was too subtle to be considered significant.

Comparison to Prior Work

Numerous studies have explored the potential application of ChatGPT in clinical practice. Ayers et al [32] observed that ChatGPT tends to deliver longer and more empathetic responses of higher quality compared to real doctors. In a study by Cascella et al [12], ChatGPT demonstrated proficiency in composing medical notes for intensive care unit patients and scientific writing, despite lacking medical expertise. The researchers highlighted the model's effectiveness in providing medical advice and its potential in patient communication [12]. Several studies have evaluated ChatGPT's responses in various medical specialties [14,33-35]. In contrast, our study uniquely focused on ChatGPT's cross-language performance in clinical counseling, revealing that language choice impacts accuracy and answer stability. This emphasizes the importance of language selection for the practical application of LLMs.

Limitations

It is important to acknowledge certain limitations in our study. The evaluation did not comprehensively assess ChatGPT's knowledge and ability in guiding emotional management for patients due to questionnaire resource constraints. The lack of a standardized questionnaire also limited the reliability of the questionnaire used owing to the lack of a related interrater reliability measure. Meanwhile, as the first work in the assessment of medical consulting AI systems for hepatitis B, it was difficult to estimate the possible effects of vagueness, ambiguity, and misunderstanding associated with grammatical mistakes or vagueness of the questions. The researchers revised the questions to address such concerns, which created new concerns about discrepancies between these "standard" questions and practical application scenarios. These problems should be addressed in future research. Additionally, while cautionary statements promote responsible usage, the potential risks and benefits of this approach require further exploration. Future studies should address these limitations for a more

comprehensive understanding of ChatGPT's application in medical assistance.

Conclusion

ChatGPT-3.5 exhibits promising capabilities as a medical consultation assistant, providing accurate yet occasionally less comprehensive information. ChatGPT-4.0, which is an improved version of the model, showed stronger application potential than ChatGPT-3.5. Recognizing their limitations in emotional support and language-specific performance, future developments should prioritize targeted language training and enhanced emotional management features. While cautionary statements underscore responsible usage, the model's potential in aiding patients with CHB is evident. As AI continues to shape the medical practice, refining LLMs for nuanced health care contexts is imperative. Striking a balance among linguistic accuracy, emotional sensitivity, and ethical patient engagement remains important for successful integration into clinical settings.

Acknowledgments

This study was supported by the Young Scientists Fund of the National Natural Science Foundation of China (number: 82100640). The study was also assisted by Ruihong Zhao, Yu Shi, and Hong Zhao.

Authors' Contributions

YW and YC conceived the study, collected and filtered questions, and developed the assessment criteria. YW translated and polished the questions, conducted the questionnaire assessment process, and wrote the manuscript. YC revised the manuscript. JS directed the entire research process, invited 2 independent reviewers, and acted as the final senior specialist.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of questions revised or eliminated.
[DOCX File , 14 KB - medinform_v12i1e56426_app1.docx]

Multimedia Appendix 2
Summary of the information accuracy grades and consistency of ChatGPT-3.5.
[DOCX File , 17 KB - medinform_v12i1e56426_app2.docx]

Multimedia Appendix 3
Assessment of the information accuracy of ChatGPT-3.5.
[DOCX File , 258 KB - medinform_v12i1e56426_app3.docx]

Multimedia Appendix 4
Results of mistake type evaluation of ChatGPT-3.5.
[DOCX File , 35 KB - medinform_v12i1e56426_app4.docx]

Multimedia Appendix 5
Content consistency assessment of ChatGPT-3.5.
[DOCX File , 247 KB - medinform_v12i1e56426_app5.docx]

Multimedia Appendix 6
Responses provided with sufficient emotional management guidance in ChatGPT-3.5.
[DOCX File , 22 KB - medinform_v12i1e56426_app6.docx]

Multimedia Appendix 7
Evaluation of whether responses mentioned consulting health care providers or doctors in ChatGPT-3.5.
[DOCX File , 242 KB - medinform_v12i1e56426_app7.docx]

Multimedia Appendix 8
Responses mentioning "I am not a doctor" or "I cannot give diagnosis or treatment" in ChatGPT-3.5.
[DOCX File , 38 KB - medinform_v12i1e56426_app8.docx]

Multimedia Appendix 9
Summary of the information accuracy grades and consistency of ChatGPT-4.0.
[DOCX File , 16 KB - medinform_v12i1e56426_app9.docx]

Multimedia Appendix 10
Assessment of the information accuracy of ChatGPT-4.0.
[DOCX File , 232 KB - medinform_v12i1e56426_app10.docx]

Multimedia Appendix 11
Content consistency assessment of ChatGPT-4.0.
[DOCX File , 235 KB - medinform_v12i1e56426_app11.docx]

Multimedia Appendix 12
Responses provided with sufficient emotional management guidance in ChatGPT-4.0.
[DOCX File , 37 KB - medinform_v12i1e56426_app12.docx]

Multimedia Appendix 13
Evaluation of whether responses mentioned consulting health care providers or doctors in ChatGPT-4.0.
[DOCX File , 228 KB - medinform_v12i1e56426_app13.docx]

Multimedia Appendix 14
Responses mentioning “I am not a doctor” or “I cannot give diagnosis or treatment” in ChatGPT-4.0.
[DOCX File , 57 KB - medinform_v12i1e56426_app14.docx]

Multimedia Appendix 15
Results of responses to closed questions (true-or-false) in ChatGPT-3.5 and ChatGPT-4.0.
[DOCX File , 28 KB - medinform_v12i1e56426_app15.docx]

References

1. Hepatitis B. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b [accessed 2024-04-09]
2. Global hepatitis report, 2017. World Health Organization. URL: https://www.who.int/publications/i/item/9789241565455 [accessed 2024-04-19]
3. Han S, Tran TT. Management of Chronic Hepatitis B: An Overview of Practice Guidelines for Primary Care Providers. J Am Board Fam Med 2015;28(6):822-837 [FREE Full text] [doi: 10.3122/jabfm.2015.06.140331] [Medline: 26546661]
4. European Association for the Study of the Liver. Electronic address: easloffice@easloffice.eu, European Association for the Study of the Liver. EASL 2017 Clinical Practice Guidelines on the management of hepatitis B virus infection. J Hepatol 2017 Apr 17;67(2):370-398. [doi: 10.1016/j.jhep.2017.03.021] [Medline: 28427875]
5. Degasperi E, Anolli MP, Lampertico P. Towards a Functional Cure for Hepatitis B Virus: A 2022 Update on New Antiviral Strategies. Viruses 2022 Oct 29;14(11):2024 [FREE Full text] [doi: 10.3390/v14112404] [Medline: 36366502]
6. Appleton AA, Buka SL, Loucks EB, Gilman SE, Kubzansky LD. Divergent associations of adaptive and maladaptive emotion regulation strategies with inflammation. Health Psychol 2013 Jul;32(7):748-756 [FREE Full text] [doi: 10.1037/a0030068] [Medline: 23815767]
7. de Ridder D, Geenen R, Kuijer R, van Middendorp H. Psychological adjustment to chronic disease. The Lancet 2008 Jul 19;372(9634):246-255. [doi: 10.1016/s0140-6736(08)61078-8] [Medline: 18640461]
8. Wani SUD, Khan NA, Thakur G, Gautam SP, Ali M, Alam P, et al. Utilization of Artificial Intelligence in Disease Prevention: Diagnosis, Treatment, and Implications for the Healthcare Workforce. Healthcare (Basel) 2022 Mar 24;10(4):608 [FREE Full text] [doi: 10.3390/healthcare10040608] [Medline: 35455786]
9. Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/NEJMRA2302038] [Medline: 36988595]
10. Au K, Yang W. Auxiliary use of ChatGPT in surgical diagnosis and treatment. Int J Surg 2023 Dec 01;109(12):3940-3943 [FREE Full text] [doi: 10.1097/JS9.0000000000000686] [Medline: 37678271]
11. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. Nature 2023 Mar;614(7947):224-226 [FREE Full text] [doi: 10.1038/d41586-023-00288-7] [Medline: 36737653]
12. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]

13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ 2023 Mar 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
14. Yeo YH, Samaan JS, Ng WH, Ting P, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023 Jul;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
15. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res 2023 Jun 28;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
16. Deng L, Wang T, Zhai Z, Tao W, Li J, Zhao Y, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int J Surg 2024 Apr 01;110(4):1941-1950 [FREE Full text] [doi: 10.1097/JS9.0000000000001066] [Medline: 38668655]
17. Frosolini A, Franz L, Benedetti S, Vaira LA, de Filippis C, Gennaro P, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol 2023 Nov;280(11):5129-5133. [doi: 10.1007/s00405-023-08205-4] [Medline: 37679532]
18. Lee T, Rao AK, Campbell DJ, Radfar N, Dayal M, Khrais A. Evaluating ChatGPT-3.5 and ChatGPT-4.0 Responses on Hyperlipidemia for Patient Education. Cureus 2024 May;16(5):e61067 [FREE Full text] [doi: 10.7759/cureus.61067] [Medline: 38803402]
19. Zheng H, Zhang G, Chan P, Wang F, Rodewald LE, Miao N, et al. Compliance among infants exposed to hepatitis B virus in a post-vaccination serological testing program in four provinces in China. Infect Dis Poverty 2019 Jul 04;8(1):57 [FREE Full text] [doi: 10.1186/s40249-019-0568-y] [Medline: 31269994]
20. Wang M, Chen EQ. [Impact of treatment compliance in chronic hepatitis B]. Zhonghua Gan Zang Bing Za Zhi 2022 Nov 20;30(11):1266-1269. [doi: 10.3760/cma.j.cn501113-20201201-00635] [Medline: 36891709]
21. Zhou X, Zhang F, Ao Y, Lu C, Li T, Xu X, et al. Diagnosis experiences from 50 hepatitis B patients in Chongqing, China: a qualitative study. BMC Public Health 2021 Dec 01;21(1):2195 [FREE Full text] [doi: 10.1186/s12889-021-11929-9] [Medline: 34852813]
22. Tütüncü EE, Güner R, Gürbüz Y, Kaya Kalem A, Öztürk B, Hasanoğlu İ, et al. Adherence to Nucleoside/Nucleotide Analogue Treatment in Patients with Chronic Hepatitis B. Balkan Med J 2017 Dec 01;34(6):540-545 [FREE Full text] [doi: 10.4274/balkanmedj.2016.1461] [Medline: 29215337]
23. Ozyigitoglu D, Sevgi DY, Tahtasakal CA, Oncul A, Gunduz A, Dokmetas I. Adherence to Treatment with Oral Nucleoside/Nucleotide Analogs in Patients with Chronic Hepatitis B. Sisli Etfal Hastan Tip Bul 2022;56(4):543-551 [FREE Full text] [doi: 10.14744/SEMB.2022.82608] [Medline: 36660396]
24. Jiang H, Zhao Q, Chen K, Yang J, Li Q. The main features of physician assistants/associates and insights for the development of similar professions in China. J Evid Based Med 2022 Dec;15(4):398-407 [FREE Full text] [doi: 10.1111/jebm.12504] [Medline: 36573381]
25. Yu Q, Yin W, Huang D, Sun K, Chen Z, Guo H, et al. Trend and equity of general practitioners' allocation in China based on the data from 2012-2017. Hum Resour Health 2021 Mar 15;19(1):20 [FREE Full text] [doi: 10.1186/s12960-021-00561-8] [Medline: 33588888]
26. Wen J, Cheng Y, Hu X, Yuan P, Hao T, Shi Y. Workload, burnout, and medical mistakes among physicians in China: A cross-sectional study. Biosci Trends 2016 Mar;10(1):27-33 [FREE Full text] [doi: 10.5582/bst.2015.01175] [Medline: 26961213]
27. Wei L, Jia J, Weng X, Dou X, Jiang J, Tang H, et al. Treating chronic hepatitis B virus: Chinese physicians' awareness of the 2010 guidelines. World J Hepatol 2016 Jun 28;8(18):762-769 [FREE Full text] [doi: 10.4254/wjh.v8.i18.762] [Medline: 27366303]
28. Yu P, Fang C, Liu X, Fu W, Ling J, Yan Z, et al. Performance of ChatGPT on the Chinese Postgraduate Examination for Clinical Medicine: Survey Study. JMIR Med Educ 2024 Mar 09;10:e48514 [FREE Full text] [doi: 10.2196/48514] [Medline: 38335017]
29. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv. URL: https://arxiv.org/abs/2005.14165 [accessed 2024-07-24]
30. Wierenga KL, Lehto RH, Given B. Emotion Regulation in Chronic Disease Populations: An Integrative Review. Res Theory Nurs Pract 2017 Aug 01;31(3):247-271 [FREE Full text] [doi: 10.1891/1541-6577.31.3.247] [Medline: 28793948]
31. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv. URL: https://arxiv.org/abs/2307.09009 [accessed 2024-07-24]
32. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med 2023 Jun 01;183(6):589-596 [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
33. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol 2023 Jun;228(6):696-705. [doi: 10.1016/j.ajog.2023.03.009] [Medline: 36924907]

34. Fournier A, Fallet C, Sadeghipour F, Perrottet N. Assessing the applicability and appropriateness of ChatGPT in answering clinical pharmacy questions. Ann Pharm Fr 2024 May;82(3):507-513. [doi: 10.1016/j.pharma.2023.11.001] [Medline: 37992892]
35. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci 2023 Dec;3(4):100324 [FREE Full text] [doi: 10.1016/j.xops.2023.100324] [Medline: 37334036]

Abbreviations

AASLD: American Association for the Study of Liver Disease
AI: artificial intelligence
ALT: alanine transaminase
CHB: chronic hepatitis B
EASL: European Association for the Study of the Liver
HBV: hepatitis B virus
LLM: large language model
WHO: World Health Organization

Edited by C Lovis; submitted 16.01.24; peer-reviewed by Q Yang, R Yin; comments to author 10.05.24; revised version received 24.05.24; accepted 21.07.24; published 08.08.24.

Please cite as:
Wang Y, Chen Y, Sheng J
Assessing ChatGPT as a Medical Consultation Assistant for Chronic Hepatitis B: Cross-Language Study of English and Chinese
JMIR Med Inform 2024;12:e56426
URL: https://medinform.jmir.org/2024/1/e56426
doi:10.2196/56426
PMID:39115930

Original Paper

Extraction of Substance Use Information From Clinical Notes: Generative Pretrained Transformer–Based Investigation

Fatemeh Shah-Mohammadi[1*], PhD; Joseph Finkelstein[1*], MD, PhD

Department of Biomedical Informatics, School of Medicine, The University of Utah, Salt Lake City, UT, United States

[*]all authors contributed equally

Corresponding Author:
Fatemeh Shah-Mohammadi, PhD
Department of Biomedical Informatics
School of Medicine
The University of Utah
421 Wakara Way
Ste 140
Salt Lake City, UT, 84108
United States
Phone: 1 801 581 4080
Email: fatemeh.shah-mohammadi@utah.edu

Abstract

Background: Understanding the multifaceted nature of health outcomes requires a comprehensive examination of the social, economic, and environmental determinants that shape individual well-being. Among these determinants, behavioral factors play a crucial role, particularly the consumption patterns of psychoactive substances, which have important implications on public health. The Global Burden of Disease Study shows a growing impact in disability-adjusted life years due to substance use. The successful identification of patients' substance use information equips clinical care teams to address substance-related issues more effectively, enabling targeted support and ultimately improving patient outcomes.

Objective: Traditional natural language processing methods face limitations in accurately parsing diverse clinical language associated with substance use. Large language models offer promise in overcoming these challenges by adapting to diverse language patterns. This study investigates the application of the generative pretrained transformer (GPT) model in specific GPT-3.5 for extracting tobacco, alcohol, and substance use information from patient discharge summaries in zero-shot and few-shot learning settings. This study contributes to the evolving landscape of health care informatics by showcasing the potential of advanced language models in extracting nuanced information critical for enhancing patient care.

Methods: The main data source for analysis in this paper is Medical Information Mart for Intensive Care III data set. Among all notes in this data set, we focused on discharge summaries. Prompt engineering was undertaken, involving an iterative exploration of diverse prompts. Leveraging carefully curated examples and refined prompts, we investigate the model's proficiency through zero-shot as well as few-shot prompting strategies.

Results: Results show GPT's varying effectiveness in identifying mentions of tobacco, alcohol, and substance use across learning scenarios. Zero-shot learning showed high accuracy in identifying substance use, whereas few-shot learning reduced accuracy but improved in identifying substance use status, enhancing recall and F_1-score at the expense of lower precision.

Conclusions: Excellence of zero-shot learning in precisely extracting text span mentioning substance use demonstrates its effectiveness in situations in which comprehensive recall is important. Conversely, few-shot learning offers advantages when accurately determining the status of substance use is the primary focus, even if it involves a trade-off in precision. The results contribute to enhancement of early detection and intervention strategies, tailor treatment plans with greater precision, and ultimately, contribute to a holistic understanding of patient health profiles. By integrating these artificial intelligence–driven methods into electronic health record systems, clinicians can gain immediate, comprehensive insights into substance use that results in shaping interventions that are not only timely but also more personalized and effective.

(JMIR Med Inform 2024;12:e56243) doi:10.2196/56243

KEYWORDS

substance use; natural language processing; GPT; prompt engineering; zero-shot learning; few-shot learning

Introduction

The use and misuse of psychoactive substances rank as critical risk elements for global health, contributing substantially to the worldwide disease burden [1,2]. Alcohol, tobacco, and illegal drugs are implicated in more than 80 identified conditions that lead to disease and injury [3,4], incurring significant health and societal costs [5-7]. Tobacco use is primarily connected to chronic diseases that often result in death, while alcohol consumption is associated with both acute conditions, such as injuries—both intentional and accidental—and chronic diseases, varying in mortality risk (eg, high risk includes liver cirrhosis and head and neck cancers; low risk covers conditions such as depression and alcohol dependency). Illicit drug use carries risks of infectious diseases, particularly through intravenous methods that may transmit HIV, in addition to heightened risks of suicide and drug use disorders. Unlike tobacco or illicit drugs, alcohol presents a complex profile, as certain levels and patterns of consumption have been shown to have protective effects against some diseases, notably coronary heart disease [8-10].

The documentation of substance use information in patient clinical notes plays an important role in care delivery by impacting clinical decision-making processes. First, it furnishes health care providers with vital information concerning a patient's addiction history, a fundamental component in constructing a comprehensive medical profile [11]. This knowledge is instrumental in devising patient-centered treatment plans that not only address the primary medical concern but also consider the complexities of the use and its potential impact on treatment efficacy [12]. Furthermore, the extraction of this information aids in risk assessment, enabling the identification of patients who may be at higher risk of relapse or complications, thereby allowing for more proactive and tailored interventions [13]. The incorporation of extraction of substance use information from clinical notes directly informs patient treatment approaches. It enables health care providers to design interventions that address not only the immediate health concern but also the underlying addiction issue if exists [14]. The integration of substance use information into treatment planning facilitates the development of harm reduction strategies and medication-assisted therapies, tailored to each patient's unique needs and readiness for change [15]. This patient-centered approach not only enhances treatment outcomes but also fosters a supportive therapeutic relationship, promoting long-term recovery and well-being. Ultimately, this information enhances the precision of clinical decision-making by fostering a holistic understanding of the patient's health, thus underscoring the indispensability of addiction status extraction from clinical documentation in modern health care practice. By incorporating substance use information into risk assessment and treatment decision-making, health care professionals can deliver more precise, effective, and patient-centered care, identifying patients at higher risk of complications, relapse, or adverse outcomes due to their substance use history, ultimately leading to more targeted interventions and improved patient outcomes.

Studies [16,17] have used machine learning techniques to predict treatment outcomes for patients with substance use disorders, demonstrating how addiction status data can inform risk assessment and stratification. Studies [18,19] have explored risk stratification for opioid overdose, incorporating addiction status data and clinical information to identify patients at higher risk, thereby informing targeted interventions and care plans. Researchers also examine how addiction status information informs the development of personalized treatment plans and explore how health care providers tailor interventions to address both the primary medical issue and the underlying addiction concerns [20-22]. Many works emphasize the importance of patient-centered care and how extraction of substance addiction data enhances this approach. They highlight the significance of understanding a patient's history of addiction for delivering more effective and empathetic care [23-28]. However, implications for health care policy and the implementation of substance use data into clinical practice have their own barriers and challenges [29-33].

The traditional process of extracting data related to substance use from clinical notes involves the use of rule-based approaches to parse the unstructured clinical narratives, identifying and categorizing relevant information pertaining to substance use. However, rule-based approaches lack a standardized rule language. On the other hand, the high variability in language found within clinical notes imposes significant limits on the accuracy of traditional techniques that rely on parsing rules to detect text patterns. Clinician typographical errors, abbreviations, and other linguistic variations hinder the effectiveness of these methods. Conversely, deep learning methods have shown impressive efficacy in extracting such information from the intricate and complex texts within clinical notes [34-36]. However, the necessity for extensive, high-quality annotated data sets for training—as information extraction is a supervised task in natural language processing—presents a significant challenge that must be overcome to fully realize the potential of these models in new and practical real-world settings.

Recently, large language models (LLMs) have emerged as a promising solution to this challenge, particularly due to their significant ability to "learn" and adapt to diverse language patterns without the need for additional model training [37]. LLMs demonstrate an unparalleled ability to comprehend nuances of clinical narratives, extracting meaning from diverse and complex medical texts. Although primarily trained on open-source and non–domain-specific texts, generative pretrained transformer (GPT) [38], as a recent development in LLMs, has underscored its effectiveness when applied to clinical notes [39,40]. GPT has also showcased its capability in US medical licensing examinations by achieving or even surpassing human-level performance in perception of clinical context [41]. This exceptional performance may be attributed to several factors, including the extensive model parameters, large pretraining data sets, and instruction tuning and optimization with reinforcement learning human feedback [42]. Recent works

in the extraction of substance use information leverage LLMs such as Bidirectional Encoder Representations from Transformers and T5, with models being fine-tuned specifically for the social determinants of health extraction task [36,43]. The emergence of LLMs has also enabled new training paradigms, including few-shot or zero-shot learning [38,44].

Leveraging GPT (GPT-3.5 model), in this work we explore the extraction of patient's substance use information in specific patients' tobacco, alcohol, and illicit substance use information from their notes and assignment of a status to classifying the individual's engagement with the substance into categories based on time-related factors (ie, past, present, or none). Since prompt engineering is essential when interacting with GPT to obtain high-quality responses, our proposed workflow involves performing zero-shot as well as few-shot prompting. In zero-shot learning, the model is expected to generalize to tasks without having seen any examples from that specific task during training. It relies on understanding the task description and applying previously learned knowledge and patterns to new, unseen situations. In a zero-shot learning scenario, the GPT is instructed to perform a particular task through an input prompt, and it produces text as a response, which serves as the output. For instance, when provided with the prompt: *List mentions of substance use in the following note: <clinical note>*. Then, the GPT extracts the reference to substance use with surrounding information relevant to it and produces the following output text—h/o prior tobacco abuse × 60 pack years—while few-shot learning involves training models on a very small data set. In this prompting setting, the model is designed to learn information from a few examples and generalize that knowledge to new data. We first experimented and formulated prompts to elicit the desired responses from the model. Then we used our finalized prompt in zero-shot learning setting. For few-shot prompting, we add a few examples to our finalized prompt to directly address the types of errors observed in zero-shot learning.

To the best of our knowledge, no scholarly publication has investigated the use of zero-shot and few-shot learning approaches with the GPT-3.5 model in the context of extracting data on patients' substance use as well as determining their usage status. The substance use information is usually scattered throughout multiple clinical notes and may be overlooked by a new provider despite the fact that this information can affect clinical decision-making. By automating extraction of substance use profile from multiple clinical notes, the substance use status can be provided in a summarized format. It can also be used in automated clinical decision embedded into electronic health records (EHR). Secondary analysis of real-world data can be biased if it does not account for substance use profile. Automated extraction of substance use profiles can greatly facilitate generation of real-world evidence from EHR data. Our evaluation aims to provide insights into the capabilities and limitations of LLMs in substance use information extraction. Ultimately, our goal is to contribute to the ongoing development of the use of LLMs in the field of substance use information extraction, with the aim of improving the quality of care.

Methods

Study Design

The main data source for analysis in this paper is Medical Information Mart for Intensive Care III (MIMIC-III) data set. This data set is a widely used and comprehensive source of deidentified health care data. It contains detailed clinical information from more than 60,000 critical care patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts, spanning a period of nearly a decade. This rich data set includes EHR, laboratory results, prescription records, and clinical notes, making it a valuable resource for medical research, particularly in the fields of critical care, epidemiology, and health informatics. Among all notes in this data set, we focused on discharge summaries. These notes typically provide a comprehensive overview of a patient's hospital stay, including the reason for admission, the treatments and procedures performed, and the patient's medical and social history and recommendations for postdischarge care. The social history section of a discharge summary typically covers various aspects of patients' life, such as their marital status, living situation, occupation, and lifestyle factors. If a patient has a history of addiction, particularly if it is relevant to the reason for his or her hospitalization or has implications for his or her postdischarge care, it is included in this section. The MIMIC-III contains 59,652 discharge summaries of 46,146 patients, among which we selected the patients with history of chronic obstructive pulmonary disease (COPD). Most of the patients with COPD are represented by older adults for whom identification of substance use information plays an important role in establishing an optimal treatment plan. Patients with COPD often have a history of smoking, as cigarette smoking is a primary risk factor for the development of COPD. Many individuals diagnosed with COPD have a significant smoking history. Alcohol and drug addiction are not typically considered direct risk factors for the development of COPD. However, substance abuse can exacerbate COPD symptoms, hinder treatment compliance, and lead to a more rapid decline in lung function in individuals already diagnosed with the disease [45-49].

Among 1646 patients with COPD, we selected discharge summary for 500 random patients, which was shown sufficient for assessing natural language processing pipeline accuracy in previous studies [50-53]. In this study, we use GPT and in specific GPT-3.5 model for generative question answering. We leveraged most capable and most cost-effective model in the GPT-3.5 family, which is GPT-3.5-turbo. This model has been optimized for chat. We accessed this model through the chat completions Application Programming Interface end point for extraction of substance use information, in specific patient's tobacco, alcohol, and illicit substance use, in 2 learning settings: zero-shot and few-shot.

In the zero-shot learning setting, a model is presented with tasks or queries for which it has not received explicit training. It is expected to extrapolate knowledge from its preexisting understanding of language and context to generate meaningful responses. This setting challenges the model to generalize

effectively and showcase adaptability to novel prompts, reflecting its capacity to comprehend and manipulate language beyond the scope of its training data.

On the contrary, few-shot learning involves training a model on a minimal number of task-specific examples. In this setting, GPT is provided with a few examples, allowing it to learn task-specific patterns and nuances. This approach leverages the model's pretrained knowledge to swiftly adapt to new tasks, demonstrating a remarkable capability for transfer learning. Few-shot learning is particularly advantageous when dealing with tasks that require a prompt-specific understanding, as it empowers the model to distill essential information from a handful of examples and apply this knowledge to generate coherent and contextually relevant responses.

Since prompt engineering is essential when interacting with GPT to obtain high-quality responses, we first experimented and formulated prompts to elicit the desired responses from the model. Then we used our finalized prompt in zero-shot learning setting. Multimedia Appendix 1 shows examples of our examined prompts, and Multimedia Appendix 2 shows our finalized prompt along with GPT responses to this prompt on different notes. These multimedia appendices show that we experimented different prompts and our finalized prompt was selected to be as follows:

Using the following patient's text, list tobacco use, illicit substance use, and alcohol use mentions and each one's status ("present," "past," and "none") in the bullets: <clinical note>.

The reason for selecting this prompt is that it provides the most comprehensive and detailed guidance for the task compared with the other prompts. This prompt is the only one that provides a clear and specific set of instructions. It not only asks to list data of tobacco, illicit substances, and alcohol use but also instructs GPT to include the status of usage as being "present," "past," or "none." This specificity helps guide the model to provide a more detailed and informative response. This prompt, in addition, is well structured and unambiguous in its request. It leaves no room for interpretation regarding what information is expected, making it easier for the model to generate accurate and relevant content. Moreover, in the context of medical or health care–related information, knowing the status of use (whether it is current, past, or not present) is critical for patient care and understanding their health history. This prompt includes this essential aspect, making it the most informative and complete prompt. While the last prompt in Multimedia Appendix 1 is also relatively detailed, it does not specify the need to provide the usage status for each category, which can be a crucial element in a medical or clinical context.

Next, we conducted error analysis that involves a detailed examination of the model's outputs to find specific instances where the model is underperforming. This process helps in selecting the most instructive examples to be included for few-shot learning in order to improve the model's performance. Multimedia Appendix 3 shows instances in which GPT underperformed. Considering first text, it can be seen that GPT had errors on assigning the use status of "None" to all type of substances. The clinical text indicates "No hx of tobacco or EtOH," where hx stands for history. The phrase "No hx" explicitly indicates that there is no history, which should correspond to the status of "none" for both tobacco and alcohol use. GPT's output failed to recognize this nuance. It interpreted "No hx" as a lack of mention, rather than an absence of use, and did not assign a status as instructed. GPT, like many language models, relies heavily on context to make predictions. Without a more extensive context, it might be challenging for the model to deduce that "no history of tobacco" implies "None" as tobacco use status without specific training or instructions. On the other hand, the use of negative phrasing, such as "no hx of," can be challenging for models to interpret correctly, especially without specialized training. Moreover, in shorter phrases or isolated sentences, the model may not have enough context to accurately infer the intended meaning. Furthermore, GPT's response "No mention of illicit substance use" suggests that there was no information provided about illicit substances, which is a correct extraction but lacks the explicit assignment of none status.

The second clinical note states: "No h/o tobacco and rare Etoh, no IVDA." Here, h/o stands for history of, Etoh stands for ethyl alcohol, and IVDA stands for intravenous drug abuse. Investigating GPT's response on this text shows that GPT's response was partially incorrect. While it did correctly identify that there is "None" for tobacco use, it failed to echo the specific language of the note, which included "No h/o tobacco," with h/o meaning "history of." In clinical contexts, maintaining the specific terminology used in patient records is crucial for accuracy and clarity.

Similarly, for illicit substance use, GPT's response of "None" is correct in the absence of use but lacks the explicit mention of "no IVDA" found in the clinical note. For alcohol use, the phrase "rare Etoh" suggests infrequent but current use of alcohol. The correct status should be "present" since it implies ongoing use. GPT's output incorrectly marked this as "past," which is an error. The phrase "rare" does not indicate cessation of use but rather infrequency and should be understood within the current context unless historical context is provided to imply past use.

Investigating GPT's response on third text also shows GPT's inability to correctly identify "Denies alcohol/drugs" as "None" for alcohol and illicit substance use. The phrase "Denies alcohol/drugs" is linguistically complex, and the model may not easily interpret its negation. In addressing the identified discrepancies within the model's output, it is imperative to rectify the inaccuracies by aligning the generated responses with the precise medical terminology and context presented in the clinical notes. The process entails reformulating the outputs to accurately reflect the specific language used, such as "No h/o tobacco" to denote a nonhistory of tobacco use. The refined examples, embodying both the exact phrasing and the proper status assignments, should then be systematically integrated into the training regime of the model through few-shot learning. This integration will facilitate the model's proficiency in comprehending and processing medical shorthand and context-sensitive information, thereby enhancing its performance on tasks that involve the extraction of nuanced data from clinical documentation. Through iterative exposure to these corrected instances, the model will incrementally improve its ability to

discern and categorize substance use information with a higher degree of accuracy and reliability, a crucial aspect for applications within clinical settings. Finally, Multimedia Appendix 4 shows the edited prompt for few-shot learning.

Ethical Considerations

No protected health information was collected, and the analytical data set was fully de-identified. To process the data, HIPAA (Health Insurance Portability and Accountability Act)-compliant Microsoft Azure OpenAI Application Programming Interface has been used.

Results

Among the 59,625 discharge summaries included in the MIMIC data set, 2043 were specifically associated with patients having a history of COPD. These particular summaries corresponded to a total of 1646 distinct patients with COPD. From this cohort, a random selection process was applied to obtain discharge summaries for a subset of 500 individuals for further analysis.

Table 1 presents general statistics pertaining to the data set. This table provides demographic information, presenting the distribution of attributes among the surveyed population. The data include the percentage breakdown of individuals based on gender, ethnicity, and marital status. The gender distribution shows a relatively balanced representation, with 53% male and 47% female respondents. This suggests a fair inclusion of both genders in the study. The majority of the surveyed population identifies as "White," constituting 73.16%. "Black," "Asian," and "Other" ethnicities make up 11.69%, 1.52%, and 13.63%, respectively. The marital status distribution reveals that a significant portion of the respondents is married (43.07%), followed by widowed (23.81%) and single (22.29%) individuals. There is also a small percentage with unknown marital status (4.98%), and divorced (4.55%) and separated (1.30%) individuals make up the rest. Accuracy, precession, recall, and F_1-score have been selected as evaluation metrics noting that every metrics have been calculated by manually reviewing all notes in data set.

Table 2 provides an overview of the results obtained from using GPT for the extraction of substance-related mentions and the corresponding status of the usage, comparing few-shot learning and zero-shot learning settings across tobacco, drug, and alcohol categories. The noticeable discrepancy between the accuracy of substance use mentions and status extraction in the zero-shot setting suggests a potential area for improvement, particularly in the nuanced understanding of the status associated with all categories of substance use. To leverage this insight and transition toward few-shot learning, we examined the specific instances where the zero-shot model struggled to accurately extract usage statuses and identified patterns, types of sentences, or contextual cues that may had contributed to the lower accuracy in extraction. Multimedia Appendix 3 shows multiple instances on which GPT made errors. These instances were used to update the finalized prompt for few-shot learning. Multimedia Appendix 4 shows our finalized prompt for few-shot learning.

Table 1. General statistics.

Attributes	Proportion, %
Sex	
Female	47
Male	53
Ethnicity	
White	73.16
Black	11.69
Asian	1.52
Other	13.63
Marital status	
Married	43.07
Widowed	23.81
Single	22.29
Unknown	4.98
Divorced	4.55
Separated	1.30

Table 2. Performance of generative pretrained transformer in a zero-shot and few-shot learning setting.

	Few-shot learning		Zero-shot learning	
	Mention (%)	Status (%)	Mention (%)	Status (%)
Tobacco				
Recall	87	66	93	29
Precision	58	51	98	87
F_1-score	70	57.5	96	43.5
Accuracy	60	40	92	26
Drug				
Recall	88	89	92	34
Precision	93	89	99	100
F_1-score	91	89	95	51
Accuracy	82	79	90	32
Alcohol				
Recall	89	78	89	29
Precision	78	73	99	100
F_1-score	83	76	94	45
Accuracy	71	57	90	29

In zero-shot learning, for tobacco, the precision was at 98%, and for both drug and alcohol mentions, it reached an impressive 99%. The recall for tobacco mentions was 93%, suggesting that the model was able to identify a large majority of the relevant instances. However, the recall for the status of tobacco use was substantially lower at 29%. For drugs, the recall was also high at 92% for mentions but significantly lower at 34% for the status. Similarly, for alcohol, the recall was 89% for mentions but dropped to 29% for the status. The F_1-scores, which balance recall and precision, were quite high for mentions, with tobacco at 96%, drugs at 95%, and alcohol at 94%, indicating strong overall performance in this aspect. Nevertheless, the F_1-scores for the status were lower: 43.5% for tobacco, 51% for drugs, and 45% for alcohol.

After few-shot learning, the accuracy of extraction of status was changed from 26%, 32%, and 29% to 40%, 79%, and 57%, for tobacco, alcohol, and substance use, respectively. The observed changes in the accuracy of extraction of status of the usage, following the incorporation of a new crafted prompt and the inclusion of examples where GPT previously had errors, indicate a 14%, 47%, and 28% improvement in the model's performance in terms of tobacco, drug, and alcohol status use extraction, respectively. On the other hand, few-shot learning led to significant decrease in the accuracy of mentions of substance use across all categories. The accuracy of extraction of tobacco, alcohol, and substance use mentions in zero-shot setting scenario was 92%, 90%, and 90%, respectively. While the accuracy for the use mentions in few-shot setting was 60%, 82%, and 71%, respectively.

Regarding the extraction of mentions of substance use, in few-shot learning, for tobacco, the recall is high at 87%, but precision is comparatively lower at 58%, resulting in a balanced F_1-score of 70%. Similar patterns are observed for alcohol category. While in contrast, precision value for mentions of drug use (93%) is higher than recall value (88%). Zero-shot learning exhibits higher recall in extraction of use mentions for all substance use categories, ranging from 89% to 93%, with precision ranging from 98% to 99%. Consequently, F_1-scores vary between 94% and 96%.

Regarding the extraction of the usage status, in few-shot learning, the recall value for tobacco is 66%, with precision just more than 50% and F_1-score of 57.5%. While in comparison with few-shot learning, zero-shot learning resulted in 37% lower recall and 14% lower F_1-score but 36% higher precision. The same pattern can be seen for alcohol and drug use status extraction across both learning setting, meaning lower recall and higher precision in zero-shot learning compared with few-shot learning resulted in higher F_1-score in few-shot learning. The discrepancies observed in the extraction performance metrics before and after few-shot learning may be attributed to several factors related to model configuration, prompt specificity, and data characteristics. First, the model configuration in few-shot learning involves exposure to specific examples that may not be diverse enough, potentially leading the model to overfit to particular features of the examples provided rather than generalizing effectively. This overfitting could result in reduced precision in mention extraction as the model becomes more sensitive to the nuances of the few-shot examples at the cost of broader applicability.

Second, prompt specificity plays a significant role in directing the model's attention and interpretation mechanisms. In the few-shot scenario, if the prompts are crafted with high specificity toward the status of use, the model's focus might shift from mention detection toward status classification,

explaining the improvement in status extraction accuracy and the concomitant decline in mention extraction accuracy.

Finally, data characteristics, such as the complexity, ambiguity, and representativeness of the clinical notes, can significantly influence the outcomes. Few-shot learning might result in better recall for status extraction if the examples chosen for training closely resemble the test cases, indicating that these examples were well selected to represent the variety of ways that status can be expressed in clinical texts. Conversely, zero-shot learning's higher precision suggests that the model, without the bias of the few-shot examples, might be more conservative and specific in its outputs, thus avoiding false positives.

Discussion

Principal Findings

The process of trying diverse prompts and selecting the one that yields the desired output was instrumental in harnessing the capabilities of GPT to align with objectives of this study. The act of crafting varied prompts allowed to explore the model's versatility and adaptability in correct extraction of patients' substance usage status. By experimenting with different prompt formulations, it becomes feasible to ascertain the prompt's impact on the model's behavior, leading to high accuracy in extraction. The finalized prompt in zero-shot is well structured and unambiguous in its request. It leaves no room for interpretation regarding what information is expected, making it easier for the model to generate accurate and relevant content. Moreover, in the context of medical or health care–related information, knowing the status of substance use (whether it is current, past, or none) is critical for patient care and understanding their health history. This prompt includes this essential aspect, making it the most informative and complete prompt. While the last prompt in Multimedia Appendix 1 is also relatively detailed, it does not specify the need to provide the status of each mention, which can be a crucial element in a medical or clinical context.

Crafting the new prompt by strategically using few-shot learning and tailoring to the challenges observed in the zero-shot setting resulted in increase on the accuracy of extraction of usage status. This approach capitalizes the importance of providing targeted guidance to enhance the model's proficiency in extracting nuanced information related to tobacco, alcohol, and substance uses. While the progress is commendable, it is essential to recognize that model refinement is an iterative process. Continued iterations, incorporating additional examples and refining the prompt, may further enhance accuracy, particularly in scenarios with inherent complexities.

The presented results in Table 2 highlight the contrasting performance of GPT in extracting mentions of tobacco, alcohol, and substance use in both zero-shot and few-shot learning scenarios. In the zero-shot setting, the accuracy for extraction of tobacco, alcohol, and substance use mentions is notably high. However, in the few-shot setting, the accuracy diminishes significantly. On the contrary, few-shot learning led to significant increase in devising the status of substance use compared with zero-shot learning (significant increase in recall and F_1-score). However, this improvement comes at the cost of a reduction in precision in both substance use information extraction and devising the status of the use. Accordingly, the selection between zero-shot and few-shot learning hinges on the goals of the task. Zero-shot learning excels in precisely extracting use mentions, demonstrating its effectiveness in situations in which comprehensive recall is paramount. Conversely, few-shot learning offers advantages when accurately determining the status of use is the primary focus, even if it involves a trade-off in precision.

The models we developed can be integrated with EHR systems to automatically extract and update patient substance use information. This integration facilitates real-time updates to patient profiles, ensuring that health care providers have access to the most current data when making treatment decisions. By embedding our models into clinical decision support systems, health care providers can receive proactive alerts and recommendations based on the extracted data. For example, if a patient's history of substance use changes, the system could automatically suggest modifications to his or her treatment plan or recommend additional screenings.

In acknowledging the limitations of this study, it is important to recognize the constraints imposed by the use of a single model, GPT-3.5, which, while demonstrating substantial capabilities, also exhibits specific challenges in processing complex linguistic structures such as negations and subtle context cues. This limitation notably impacted the accuracy of status identification in zero-shot learning settings, where the model sometimes failed to correctly interpret negations, leading to errors in status assignment.

Furthermore, the study's reliance on the MIMIC-III data set, while extensive, limits the generalizability of findings across diverse demographic and clinical settings. The data set's inherent biases and the specific clinical environment from which it was derived might not fully represent the broader patient populations encountered in different geographic or health care contexts. To address these limitations, future research should consider using a multimodel approach to validate findings and enhance the robustness of the conclusions drawn. Incorporating additional models such as Bidirectional Encoder Representations from Transformers may provide comparative insights and help mitigate the biases of a single model approach. Moreover, expanding the data set to include a wider array of clinical environments and patient demographics would enhance the generalizability of the artificial intelligence tools developed. In addition, the implementation of advanced training techniques, including more sophisticated prompt engineering and error analysis methodologies, could further refine the artificial intelligence's understanding of complex clinical narratives.

Conclusion and Future Work

The extraction of psychoactive substance use status from clinical notes holds significant implications for risk assessment and patient treatment. It empowers health care providers to perform risk evaluations and to devise individualized treatment plans, leading to enhancing the precision and efficacy of care delivery while addressing the complex interplay between medical conditions and addiction. In this study, we investigate the

efficacy of 2 prompt-based approaches—zero-shot and few-shot learning—for extracting patient's substance use information from discharge summaries of patients with COPD using GPT-3. Our findings indicate that GPT-3's few-shot learning capabilities serve as a promising starting point for extracting status of substance use without the need for annotated data. The GPT-3 exhibited high precision but lower recall, suggesting a conservative approach that yields fewer false positives but may miss relevant cases. Conversely, few-shot learning demonstrated a marked improvement in recall, indicating a greater ability to identify relevant instances, yet at the expense of precision. The implications of these findings are significant for the landscape of clinical practice, where the accurate assessment of usage status is crucial for risk assessment and tailoring patient treatment plans. The enhanced recall in few-shot learning suggests its use in scenarios where missing a case of substance use is highly detrimental, while the high precision of zero-shot learning would be preferred in contexts where the cost of false positives is greater. Therefore, researchers and practitioners should carefully consider the emphasis on recall, precision, and the overall balance between the 2 when deciding between these learning scenarios based on the specific requirements of their application. We prompted GPT-3 with only 4 randomly selected samples. More examples for few-shot learning may improve the performance. In addition, our reliance on the MIMIC-III data set, though comprehensive, restricts the generalizability of our findings. The data set's inherent biases and its derivation from a specific clinical environment may not accurately reflect the varied patient populations found across different geographic or health care settings. Despite these limitations, the study presents a significant step forward in our understanding of the capabilities and limitations of advanced language models in the critical domain of health care. As the future work, we investigate the capability of LLMs in extraction of the quantity, frequency, duration, and severity of substance use disorder.

Acknowledgments

This project was supported in part by a grant (R33HL143317) from the National Heart, Lung, and Blood Institute.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of 3 different prompts to GPT-3.5 and its responses.
[PNG File , 251 KB - medinform_v12i1e56243_app1.png]

Multimedia Appendix 2
Finalized prompt and examples of 3 GPT-3.5 responses to this prompt.
[PNG File , 250 KB - medinform_v12i1e56243_app2.png]

Multimedia Appendix 3
Examples on which generative pretrained transformer had errors.
[PNG File , 239 KB - medinform_v12i1e56243_app3.png]

Multimedia Appendix 4
Crafted prompt for few-shot learning setting.
[PNG File , 302 KB - medinform_v12i1e56243_app4.png]

References

1. Ezzati M, Lopez AD, Rodgers A, Murray CJL. Comparative quantification of health risks. In: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors. Geneva: World Health Organization; 2004.
2. Rehm J, Room R. The global burden of disease attributable to alcohol, tobacco and illicit drugs. In: Preventing Harmful Substance Use: The Evidence Base for Policy and Practice. New York, NY: Wiley; 2005:25-41.
3. Rehm J, Patra J, Popova S. Alcohol-attributable mortality and potential years of life lost in Canada 2001: implications for prevention and policy. Addiction 2006;101(3):373-384 [FREE Full text] [doi: 10.1111/j.1360-0443.2005.01338.x] [Medline: 16499510]
4. Popova S, Rehm J, Patra J. Illegal drug-attributable mortality and potential years of life lost in Canada 2002: implications for prevention and policy. Contemp Drug Probl 2018;33(3):343-366. [doi: 10.1177/009145090603300302]
5. Single E, Robson L, Xie X, Rehm J. The economic costs of alcohol, tobacco and illicit drugs in Canada, 1992. Addiction 1998;93(7):991-1006. [doi: 10.1046/j.1360-0443.1998.9379914.x] [Medline: 9744130]
6. Andlin-Sobocki P. Economic evidence in addiction: a review. Eur J Health Econ 2004;5 Suppl 1:S5-S12. [doi: 10.1007/s10198-005-0282-5] [Medline: 15754074]

7. Andlin-Sobocki P, Rehm J. Cost of addiction in Europe. Eur J Neurol 2005;12 Suppl 1:28-33. [doi: 10.1111/j.1468-1331.2005.01194.x] [Medline: 15877775]
8. Rehm J, Room R, Graham K, Monteiro M, Gmel G, Sempos CT. The relationship of average volume of alcohol consumption and patterns of drinking to burden of disease: an overview. Addiction 2003;98(9):1209-1228. [doi: 10.1046/j.1360-0443.2003.00467.x] [Medline: 12930209]
9. Rehm J, Sempos CT, Trevisan M. Average volume of alcohol consumption, patterns of drinking and risk of coronary heart disease—a review. Eur J Cardiovas Prev Rehabil 2003;10(1):15-20. [doi: 10.1177/174182670301000104]
10. Corrao G, Rubbiati L, Bagnardi V, Zambon A, Poikolainen K. Alcohol and coronary heart disease: a meta-analysis. Addiction 2000;95(10):1505-1523. [doi: 10.1046/j.1360-0443.2000.951015056.x] [Medline: 11070527]
11. Volkow ND. Drugs, Brains, and Behavior: The Science of Addiction. Bethesda, MD: National Institute on Drug Abuse; 2010:255-169.
12. McLellan AT, Lewis DC, O'Brien CP, Kleber HD. Drug dependence, a chronic medical illness: implications for treatment, insurance, and outcomes evaluation. JAMA 2000;284(13):1689-1695. [doi: 10.1001/jama.284.13.1689] [Medline: 11015800]
13. Bogenschutz MP, Donovan DM, Mandler RN, Perl HI, Forcehimes AA, Crandall C, et al. Brief intervention for patients with problematic drug use presenting in emergency departments. JAMA Intern Med 2014;174(11):1736. [doi: 10.1001/jamainternmed.2014.4052]
14. Babor TF, McRee BG, Kassebaum PA, Grimaldi PL, Ahmed K, Bray J. Screening, brief intervention, and referral to treatment (SBIRT): toward a public health approach to the management of substance abuse. Subst Abus 2007;28(3):7-30. [doi: 10.1300/J465v28n03_03] [Medline: 18077300]
15. Minozzi S, Amato L, Bellisario C, Davoli M. Maintenance treatments for opiate-dependent adolescents. Cochrane Database Syst Rev 2014;2014(6):CD007210 [FREE Full text] [doi: 10.1002/14651858.CD007210.pub3] [Medline: 24957634]
16. Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. PLoS One 2017;12(4):e0175383 [FREE Full text] [doi: 10.1371/journal.pone.0175383] [Medline: 28394905]
17. Tapia-Galisteo J, Iniesta JM, Perez-Gandia C, Garcia-Saez G, Puertolas DU, Izquierdo FJ, et al. Prediction of cocaine inpatient treatment success using machine learning on high-dimensional heterogeneous data. IEEE Access 2020;8:218936-218953. [doi: 10.1109/access.2020.3041895]
18. Weiner SG, Baker O, Bernson D, Schuur JD. One-Year mortality of patients after emergency department treatment for nonfatal opioid overdose. Ann Emerg Med 2020;75(1):13-17 [FREE Full text] [doi: 10.1016/j.annemergmed.2019.04.020] [Medline: 31229387]
19. Shah-Mohammadi F, Cui W, Bachi K, Hurd Y, Finkelstein J. Using natural language processing of clinical notes to predict outcomes of opioid treatment program. 2022 Presented at: 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2022 July 11-15; Glasgow, Scotland, United Kingdom p. 4415-4420.
20. Rich KM, Bia J, Altice FL, Feinberg J. Integrated models of care for individuals with opioid use disorder: how do we prevent HIV and HCV? Curr HIV/AIDS Rep 2018;15(3):266-275 [FREE Full text] [doi: 10.1007/s11904-018-0396-x] [Medline: 29774442]
21. Schwartz RP, Kelly SM, Mitchell SG, Gryczynski J, O'Grady KE, Gandhi D, et al. Patient-centered methadone treatment: a randomized clinical trial. Addiction 2017;112(3):454-464 [FREE Full text] [doi: 10.1111/add.13622] [Medline: 27661788]
22. Englander H, Dobbertin K, Lind BK, Nicolaidis C, Graven P, Dorfman C, et al. Inpatient addiction medicine consultation and post-hospital substance use disorder treatment engagement: a propensity-matched analysis. J Gen Intern Med 2019;34(12):2796-2803 [FREE Full text] [doi: 10.1007/s11606-019-05251-9] [Medline: 31410816]
23. Miller D, Steele Gray C, Kuluski K, Cott C. Patient-centered care and patient-reported measures: let's look before we leap. Patient 2015;8(4):293-299 [FREE Full text] [doi: 10.1007/s40271-014-0095-7] [Medline: 25354873]
24. Novilla MLB, Goates MC, Leffler T, Novilla NKB, Wu C, Dall A, et al. Integrating social care into healthcare: a review on applying the social determinants of health in clinical settings. Int J Environ Res Public Health 2023;20(19):6873 [FREE Full text] [doi: 10.3390/ijerph20196873] [Medline: 37835143]
25. Strauss T. Organizational factors underlying the adoption of a patient-centered approach in physician-patient interaction [dissertation]. Israel: University of Haifa. 2020. URL: https://login.ezproxy.lib.utah.edu/login?url=https://www.proquest.com/dissertations-theses/organizational-factors-underlying-adoption/docview/2593014088/se-2?accountid=14677 [accessed 2024-07-25]
26. Jones KG, Roth SE, Vartanian KB. Health and health care use strongly associated with cumulative burden of social determinants of health. Popul Health Manag 2022;25(2):218-226. [doi: 10.1089/pop.2021.0255] [Medline: 34935504]
27. Karapareddy V. A review of integrated care for concurrent disorders: cost effectiveness and clinical outcomes. J Dual Diagn 2019;15(1):56-66. [doi: 10.1080/15504263.2018.1518553] [Medline: 30806190]
28. King C, Collins D, Patten A, Nicolaidis C, Englander H. Trust in hospital physicians among patients with substance use disorder referred to an addiction consult service: a mixed-methods study. J Addict Med 2022;16(1):41-48 [FREE Full text] [doi: 10.1097/ADM.0000000000000819] [Medline: 33577229]
29. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health 2021;50(11):i-v [FREE Full text] [doi: 10.18502/ijph.v50i11.7600] [Medline: 35223619]

30. Rothstein MA. Health privacy in the electronic age. J Leg Med 2007;28(4):487-501 [FREE Full text] [doi: 10.1080/01947640701732148] [Medline: 18066975]
31. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019;25(1):37-43 [FREE Full text] [doi: 10.1038/s41591-018-0272-7] [Medline: 30617331]
32. Crowley RA, Kirschner N, Health and Public Policy Committee of the American College of Physicians. The integration of care for mental health, substance abuse, and other behavioral health conditions into primary care: executive summary of an American College of Physicians position paper. Ann Intern Med 2015;163(4):298-299 [FREE Full text] [doi: 10.7326/M15-0510] [Medline: 26121401]
33. Edmunds M, Frank R, Hogan M, McCarty D, Robinson-Beale R, Weisner C, editors. Managing Managed Care: Quality Improvement in Behavioral Health. Washington, DC: National Academies Press; 1997:1-309.
34. Poulsen MN, Freda PJ, Troiani V, Davoudi A, Mowery DL. Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing. Front Public Health 2022;10:850619 [FREE Full text] [doi: 10.3389/fpubh.2022.850619] [Medline: 35615042]
35. Patra B, Sharma M, Vekaria V, Adekkanattu P, Patterson O, Glicksberg B, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. J Am Med Inform Assoc 2021;28(12):2716-2727 [FREE Full text] [doi: 10.1093/jamia/ocab170] [Medline: 34613399]
36. Romanowski B, Ben Abacha A, Fan Y. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. J Am Med Inform Assoc 2023;30(8):1448-1455 [FREE Full text] [doi: 10.1093/jamia/ocad071] [Medline: 37100768]
37. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv Preprint posted online Mar 31, 2023. [doi: 10.48550/arXiv.2303.18223]
38. GPT-4 technical report. OpenAI. 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2023-12-22]
39. Zhou J, Li T, Fong SJ, Dey N, González-Crespo R. Exploring ChatGPT's potential for consultation, recommendations and report diagnosis: gastric cancer and gastroscopy reports' case. Int J Interact Multimed Artif Intell 2023;8(2):7-13. [doi: 10.9781/ijimai.2023.04.007]
40. Choi HS, Song JY, Shin KH, Chang JH, Jang B. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J 2023;41(3):209-216 [FREE Full text] [doi: 10.3857/roj.2023.00633] [Medline: 37793630]
41. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
42. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Adv Neural Inf Process Syst. 2022 Presented at: 36th International Conference on Neural Information Processing Systems; 28 November 2022- 9 December 2022; New Orleans LA USA p. 27730-27744.
43. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. J Biomed Inform 2021;113:103631 [FREE Full text] [doi: 10.1016/j.jbi.2020.103631] [Medline: 33290878]
44. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Adv Neural Inf Process Syst.: Curran Associates Inc.57 Morehouse LaneRed HookNYUnited States; 2020 Presented at: 34th International Conference on Neural Information Processing Systems; December 6 - 12; Vancouver BC Canada p. 1877-1901.
45. Saeed AM, Raafat RH, Muneer MM. Study of addiction in COPD patients in abbassia chest hospital. QJM Int J Med 2023;116(Supplement 1):69-174.
46. Mahmoud EM, Mohammed ZA, El hawary AE, Ibrahim DA. Screening for drug misusers in exacerbated chronic obstructive pulmonary disease (COPD) patients. Med Updates 2023;14(14):1-19. [doi: 10.21608/muj.2023.209496.1139]
47. Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med 2007;176(6):532-555. [doi: 10.1164/rccm.200703-456SO] [Medline: 17507545]
48. Carlin B. Chronic obstructive pulmonary disease: "learn more, breathe better". J Cardiopulm Rehabil Prev 2007;27(5):311-313. [doi: 10.1097/01.HCR.0000291300.55270.0d] [Medline: 17885511]
49. American Lung Association. Learn about COPD. URL: https://www.lung.org/lung-health-diseases/lung-disease-lookup/copd/learn-about-copd [accessed 2024-01-01]
50. Shah-Mohammadi F, Finkelstein J. NLP-assisted differential diagnosis of chronic obstructive pulmonary disease exacerbation. Stud Health Technol Inform 2024;310:589-593. [doi: 10.3233/SHTI231033] [Medline: 38269877]
51. Cui W, Shah-Mohammadi F, Finkelstein J. Using electronic medical records and clinical notes to predict the outcome of opioid treatment program. Stud Health Technol Inform 2023;305:568-571. [doi: 10.3233/SHTI230560] [Medline: 37387094]
52. Wang Y, Chen ES, Pakhomov S, Arsoniadis E, Carter EW, Lindemann E, et al. Automated extraction of substance use information from clinical texts. AMIA Annu Symp Proc 2015;2015:2121-2130 [FREE Full text] [Medline: 26958312]
53. Shah-Mohammadi F, Cui W, Finkelstein J. Entity extraction for clinical notes, a comparison between metamap and amazon comprehend medical. Stud Health Technol Inform 2021;281:258-262. [doi: 10.3233/SHTI210160] [Medline: 34042745]

Abbreviations

COPD: chronic obstructive pulmonary disease
EHR: electronic health record
GPT: generative pretrained transformer
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MIMIC-III: Medical Information Mart for Intensive Care III

Edited by A Castonguay; submitted 10.01.24; peer-reviewed by T Church, Y Liu; comments to author 17.04.24; revised version received 24.06.24; accepted 18.07.24; published 19.08.24.

Please cite as:
Shah-Mohammadi F, Finkelstein J
Extraction of Substance Use Information From Clinical Notes: Generative Pretrained Transformer–Based Investigation
JMIR Med Inform 2024;12:e56243
URL: https://medinform.jmir.org/2024/1/e56243
doi:10.2196/56243
PMID:39037700

Original Paper

Evaluating the Capabilities of Generative AI Tools in Understanding Medical Papers: Qualitative Study

Seyma Handan Akyon[1], MD; Fatih Cagatay Akyon[2,3], PhD; Ahmet Sefa Camyar[4], MD; Fatih Hızlı[5], MD; Talha Sari[2,6], BSc; Şamil Hızlı[7], Prof Dr, MD

[1]Golpazari Family Health Center, Bilecik, Turkey

[2]SafeVideo AI, San Francisco, CA, United States

[3]Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

[4]Department of Internal Medicine, Ankara Etlik City Hospital, Ankara, Turkey

[5]Faculty of Medicine, Ankara Yildirim Beyazit University, Ankara, Turkey

[6]Department of Computer Science, Istanbul Technical University, Istanbul, Turkey

[7]Department of Pediatric Gastroenterology, Children Hospital, Ankara Bilkent City Hospital, Ankara Yildirim Beyazit University, Ankara, Turkey

Corresponding Author:
Seyma Handan Akyon, MD
Golpazari Family Health Center
Istiklal Mahallesi Fevzi Cakmak Caddesi No:23 Golpazari
Bilecik, 11700
Turkey
Phone: 90 5052568096
Email: drseymahandan@gmail.com

Abstract

Background: Reading medical papers is a challenging and time-consuming task for doctors, especially when the papers are long and complex. A tool that can help doctors efficiently process and understand medical papers is needed.

Objective: This study aims to critically assess and compare the comprehension capabilities of large language models (LLMs) in accurately and efficiently understanding medical research papers using the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist, which provides a standardized framework for evaluating key elements of observational study.

Methods: The study is a methodological type of research. The study aims to evaluate the understanding capabilities of new generative artificial intelligence tools in medical papers. A novel benchmark pipeline processed 50 medical research papers from PubMed, comparing the answers of 6 LLMs (GPT-3.5-Turbo, GPT-4-0613, GPT-4-1106, PaLM 2, Claude v1, and Gemini Pro) to the benchmark established by expert medical professors. Fifteen questions, derived from the STROBE checklist, assessed LLMs' understanding of different sections of a research paper.

Results: LLMs exhibited varying performance, with GPT-3.5-Turbo achieving the highest percentage of correct answers (n=3916, 66.9%), followed by GPT-4-1106 (n=3837, 65.6%), PaLM 2 (n=3632, 62.1%), Claude v1 (n=2887, 58.3%), Gemini Pro (n=2878, 49.2%), and GPT-4-0613 (n=2580, 44.1%). Statistical analysis revealed statistically significant differences between LLMs (P<.001), with older models showing inconsistent performance compared to newer versions. LLMs showcased distinct performances for each question across different parts of a scholarly paper—with certain models like PaLM 2 and GPT-3.5 showing remarkable versatility and depth in understanding.

Conclusions: This study is the first to evaluate the performance of different LLMs in understanding medical papers using the retrieval augmented generation method. The findings highlight the potential of LLMs to enhance medical research by improving efficiency and facilitating evidence-based decision-making. Further research is needed to address limitations such as the influence of question formats, potential biases, and the rapid evolution of LLM models.

(JMIR Med Inform 2024;12:e59258) doi:10.2196/59258

KEYWORDS

large language models; LLM; LLMs; ChatGPT; artificial intelligence; AI; natural language processing; medicine; health care; GPT; machine learning; language model; language models; generative; research paper; research papers; scientific research;

answer; answers; response; responses; comprehension; STROBE; Strengthening the Reporting of Observational Studies in Epidemiology

Introduction

Artificial intelligence (AI) has revolutionized numerous fields, including health care, with its potential to enhance patient outcomes, increase efficiency, and reduce costs [1]. AI devices are divided into 2 main categories. One category uses machine learning techniques to analyze structured data for medical applications, while the other category uses natural language processing methods to extract information from unstructured data, such as clinical notes, thereby improving the analysis of structured medical data [2]. A key development within natural language processing has been the emergence of large language models (LLMs), which are advanced systems trained on vast amounts of text data to generate human-like language and perform a variety of language-based tasks [3]. While deep learning models recognize patterns in data [4], LLMs are trained to predict the probability of a word sequence based on the context. By training on large amounts of text data, LLMs can generate new and plausible sequences of words that the mode has not previously observed [4]. ChatGPT, an advanced conversational AI technology developed by OpenAI in late 2022, is a general-purpose LLM [5]. GPT is part of a growing landscape of conversational AI products, with other notable examples including Llama (Meta), Jurassic (Ai21), Claude (Anthropic), Command (Cohere), Gemini (formerly known as Bard), PaLM, and Bard (Google) [5]. The potential of AI systems to enhance medical care and health outcomes is highly promising [6]. Therefore, it is essential to ensure that the creation of AI systems in health care adheres to the principles of trust and explainability. Evaluating the medical knowledge of AI systems compared to that of expert clinicians is a vital initial step to assess these qualities [5,7,8].

Reading medical papers is a challenging and time-consuming task for doctors, especially when the papers are long and complex. This poses a significant barrier to efficient knowledge acquisition and evidence-based decision-making in health care. There is a need for a tool that can help doctors to process and understand medical papers more efficiently and accurately. Although LLMs are promising in evaluating patients, diagnosis, and treatment processes [9], studies on reading academic papers are limited. LLMs can be directly questioned and can generate answers from their own memory [10,11]. This has been extensively studied in many papers. However, these pose the problem of artificial hallucinations, which are inaccurate outputs, in LLMs. The retrieval augmented generation (RAG) method, which intuitively addresses the knowledge gap by conditioning language models on relevant documents retrieved from an external knowledge source, can be used to overcome this issue [12].

The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist provides a standardized framework for evaluating key elements of observational study and sufficient information for critical evaluation. These guidelines consist of 22 items that authors should adhere to before submitting their manuscripts for publication [13-15]. This study aims to address this gap by evaluating the comprehension capabilities of LLMs in accurately and efficiently understanding medical research papers. We use the STROBE checklist to assess LLMs' ability to understand different sections of research papers. This study uses a novel benchmark pipeline that can process PubMed papers regardless of their length using various generative AI tools. This research will provide critical insights into the strengths and weaknesses of different LLMs in enhancing medical research paper comprehension. To overcome the problem of "artificial hallucinations," we implement the RAG method. RAG involves providing the LLMs with a prompt that instructs them to answer while staying relevant to the given document, ensuring responses align with the provided information. The results of this study will provide valuable information for medical professionals, researchers, and developers seeking to leverage the potential of LLMs for improving medical literature comprehension and ultimately enhance patient care and research efficiency.

Methods

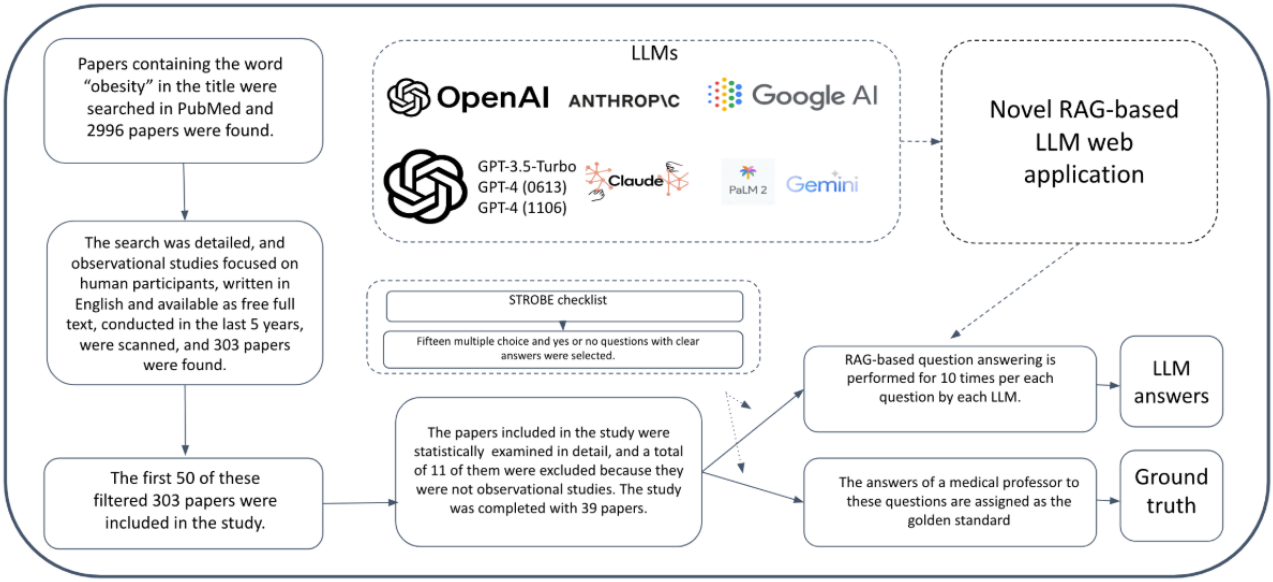
Design of Study

This study uses a methodological research design to evaluate the comprehension capabilities of generative AI tools using the STROBE checklist.

Paper Selection

We included the first 50 observational studies conducted within the past 5 years that were retrieved through an advanced search on PubMed on December 19, 2023, using "obesity" in the title as the search term. The included studies were limited to those written in English, available as free full text, and focusing specifically on human participants (Figure 1). The papers included in the study were statistically examined in detail, and a total of 11 of them were excluded because they were not observational studies. The study was completed with 39 papers. A post hoc power analysis was conducted to assess the statistical power of our study based on the total correct responses across all repetitions. The analysis excluded GPT-4-1106 and GPT-3.5-Turbo-1106 due to their similar performance and the significant differences observed between other models. The power analysis, conducted using G*Power (version 3.1.9.7; Heinrich-Heine-Universität Düsseldorf), indicated that all analyses exceeded 95% power. Thus, the study was completed with the 39 selected papers, ensuring sufficient statistical power to detect meaningful differences in LLM performance.

Figure 1. Flowchart: recruitment and data collection process for evaluating LLM comprehension of medical research papers. LLM: large language model; RAG: retrieval augmented generation.



Benchmark Development

This study used a novel benchmark pipeline to evaluate the understanding capabilities of LLMs when processing medical research papers. To establish a reference standard for evaluating the LLMs' comprehension, we relied on the expertise of an experienced medical professor and an epidemiology expert doctor. The professor, with their extensive medical knowledge, was tasked with answering 15 questions derived from the STROBE checklist, designed to assess key elements of observational studies and cover different sections of a research paper (Table 1). The epidemiology expert doctor, with their specialized knowledge in statistical analysis and epidemiological methods, provided verification and validation of the professor's answers, ensuring the rigor of the benchmark. The combined expertise of both professionals provided a robust and reliable reference standard against which the LLMs' responses were compared.

Table 1. The questions derived from the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for observational study and answers.

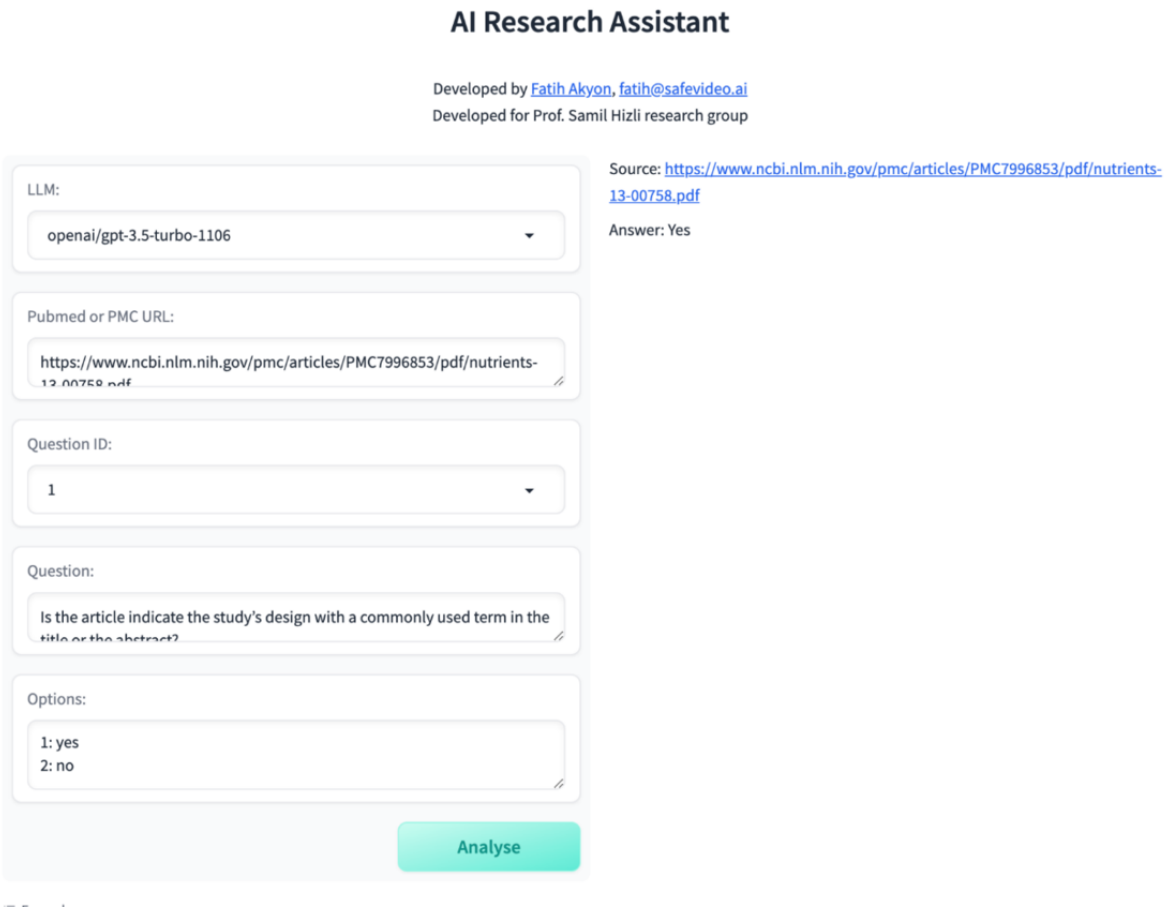
Questions	Answers
Title and abstract	
Q1. Does the paper indicate the study's design with a commonly used term in the title or the abstract?	1. Yes 2. No
Methods	
Q2. What is the observational study type: cohort, case-control, or cross-sectional studies?	1. Cohort study 2. A case-control study 3. Cross-sectional study 4. The study type is not stated in the paper
Q3. Were settings or locations mentioned in the method?	1. Yes 2. No
Q4. Were relevant dates mentioned in the method?	1. Yes 2. No
Q5. Were eligibility criteria for selecting participants mentioned in the method?	1. Yes 2. No
Q6. Were sources and methods of selection of participants mentioned in the method?	1. Yes 2. No
Q7. Were any efforts to address potential sources of bias described in the method or discussion?	1. Yes 2. No
Q8. Which program was used for statistical analysis?	1. SPSS was used for statistical analysis 2. MedCalc was used for statistical analysis 3. SAS was used for statistical analysis 4. STATA was used for statistical analysis 5. R program was used 6. Another program was used for statistical analysis 7. The program for statistical analysis is not specified
Results	
Q9. Were report numbers of individuals at each stage of the study (eg, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed) mentioned in the results?	1. Yes 2. No
Q10. Was a flowchart used to show the reported numbers of individuals at each stage of the study?	1. Yes 2. No
Q11. Were the study participants' demographic characteristics (eg, age and sex) given in the results?	1. Yes 2. No
Discussion	
Q12. Does the discussion part summarize key results concerning study objectives?	1. Yes 2. No
Q13. Are the limitations of the study discussed in the paper?	1. Yes 2. No
Q14. Is the generalizability of the study discussed in the discussion part?	1. Yes 2. No
Funding	
Q15. Is the funding of the study mentioned in the paper?	1. Yes 2. No

This list of 15 questions, 2 multiple-choice and 13 yes or no questions, has been prepared by selecting the STROBE checklist items that can be answered definitively and have clear, nonsubjective responses. Question 1, related to title and abstract, examines the LLMs' ability to identify and understand research designs and terms that are commonly used, evaluating the model's comprehension of the concise language typically used in titles and abstracts. Questions 2-8, related to methods, cover various aspects of the study's methodology, from the type of observational study to the statistical analysis programs used. They test the model's understanding of the detailed and technical language often found in this section. Questions 9-11, related to results, focus on the accuracy and completeness of reported results, such as participant numbers at each study stage and demographic characteristics. These questions gauge the LLMs' capability to parse and summarize factual data. Questions 12-14, related to the discussion, involve summarizing key results, discussing limitations, and addressing the study's generalizability. These questions assess the LLMs' ability to engage with more interpretive and evaluative content, showcasing their understanding of research impacts and contexts. Question 15, related to funding, tests the LLMs' attentiveness to specific yet crucial details that could influence the interpretation of research findings.

Development of Novel RAG-Based LLM Web Application

The methodology incorporated a novel web application specifically designed for this purpose to assess the understanding capabilities of generative AI tools in medical research papers (Figure 2). To mitigate the problem of "artificial hallucinations" inherent to LLMs, this study implemented the RAG method, which involves using a web application to dissect PDF-format medical papers from PubMed into text chunks ready to be processed by various LLMs. This approach guides the LLMs to provide answers grounded in the provided information by supplying them with relevant text chunks retrieved from the target paper.

Figure 2. Novel retrieval augmented generation–based large language model web application interface. AI: artificial intelligence.



Benchmark Pipeline

The benchmark pipeline itself is designed to process PubMed papers of varying lengths and extract relevant information for analysis. This pipeline operates as follows:

- Paper retrieval: We retrieved 39 observational studies from PubMed using the search term "obesity" in the title.
- Text extraction and chunking: Each retrieved PubMed paper was converted to PDF format and then processed through our web application. The application extracts all text content from the paper and divides it into smaller text chunks of manageable size.
- Vector representation: Using the OpenAI text-ada-embedding-002 model, each text chunk was converted into a representation vector. These vectors capture the semantic meaning of the text chunks, allowing for efficient information retrieval.

- Vector database storage: The generated representation vectors were stored in a vector database (LanceDB in our case). This database allows for rapid searching and retrieval of the most relevant text chunks based on a given query.
- Query processing: When a query (question from the STROBE checklist) was posed to an LLM, our pipeline calculated the cosine similarities between the query's representation vector and the vectors stored in the database. This identified the most relevant text chunks from the paper.
- RAG: The retrieved text chunks, along with the original query, were then combined and presented to the LLM. This approach, known as RAG, ensured that the LLM's responses were grounded in the specific information present in the paper, mitigating the risk of hallucinations.
- Answer generation and evaluation: The LLM generated an answer to the query based on the provided text chunks. The accuracy of each LLM's response was then evaluated by comparing it to the benchmark answers provided by a medical professor.

LLMs

Using this benchmark pipeline, we compared the answers of the generative AI tools, such as GPT-3.5-Turbo-1106 (June 11th version), GPT-4-0613 (November 6th version), GPT-4-1106 (June 11th version), PaLM 2 (chat-bison), Claude v1, and Gemini Pro, with the benchmark in 15 questions for 39 medical research papers (Table 2). In this study, 15 questions selected from the STROBE checklists were posed 10 times each for 39 papers to 6 different LLMs.

Table 2. The generative artificial intelligence (AI) tools compared with the benchmark in study.

Generative AI tool	Version	Company	Cutoff date
GPT-3,5-Turbo	November 6, 2023	OpenAI	September 2021
GPT-4-0613	June 13, 2023	OpenAI	September 2021
GPT-4-1106	November 6, 2023	OpenAI	April 2023
Claude v1	Version 1	Anthropic	—[a]
PaLM 2	Chat-bison	Google	—
Gemini Pro	1.0	Google	—

[a]The company does not explicitly state a cutoff date.

Access issues with Claude v1, specifically restrictions on its ability to process certain medical information, resulted in the exclusion of data from 6 papers, limiting the study's scope to 33 papers. LLMs commonly provide a "knowledge-cutoff" date, indicating the point at which their training data ends and they may not have access to the most up-to-date information. With some LLMs, however, the company does not explicitly state a cutoff date. The explicitly stated cutoff dates are given in Table 2, based on the publicly available information for each LLM.

A chatbot conversation begins when a user enters a query, often called a system prompt. The chatbot responds in natural language within a second, creating an interactive, conversation-like exchange. This is possible because the chatbot understands context. In addition to the RAG method, providing LLMs with well-designed system prompts that guide them to stay relevant to a given document can help generate responses that align with the provided information. We used the following system prompt for all LLMs:

> *You are an expert medical professor specialized in pediatric gastroenterology hepatology and nutrition, with a detailed understanding of various research methodologies, study types, ethical considerations, and statistical analysis procedures. Your task is to categorize research articles based on information provided in query prompts. There are multiple options for each question, and you must select the most appropriate one based on your expertise and the context of the research article presented in the query.*

The language models used in this study rely on statistical models that incorporate random seeds to facilitate the generation of diverse outputs. However, the companies behind these LLMs do not offer a stable way to fix these seeds, meaning that a degree of randomness is inherent in their responses. To further control this randomness, we used the "temperature" parameter within the language models. This parameter allows for adjustment of the level of randomness, with a lower temperature setting generally producing more deterministic outputs. For this study, we opted for a low-temperature parameter setting of 0.1 to minimize the impact of randomness. Despite these efforts, complete elimination of randomness is not possible. To further mitigate its effects and enhance the consistency of our findings, we repeated each question 10 times for the same language model. By analyzing the responses across these 10 repetitions, we could determine the frequency of accurate and consistent answers. This approach helped to identify instances where the LLM's responses were consistently aligned with the benchmark answers, highlighting areas of strength and consistency in comprehension.

Statistical Analysis

Each question was repeated 10 times in the same time period to obtain answers from multiple LLMs and ensure the consistency and reliability of responses. Consequently, the responses to the same question were analyzed to determine how many aligned with the benchmark, and the findings were examined. Only the answers that were correct and followed the instructions provided in the question text were considered "correct." Ambiguous answers, evident mistakes, and responses with an excessive number of candidates were considered

incorrect. The data were carefully examined, and the findings were documented and analyzed. Each inquiry and its response formed the basis of the analysis. Various descriptive statistical tests were used to assess the data presented as numbers and percentages. The Shapiro-Wilk test was used to assess the data's normal distribution. The Kruskal-Wallis and Pearson chi-square tests were used in the statistical analysis. Type I error level was accepted as 5% in the analyses performed using the SPSS (version 29.0; IBM Corp).

Ethical Considerations

This study only used information that had already been published on the internet. Ethics approval is not required for this study since it did not involve any human or animal research participants. This study did not involve a clinical trial, as it focused on evaluating the capabilities of AI tools in understanding medical papers.

Results

In this study, 15 questions selected from the STROBE checklists were posed 10 times each for 39 papers to 6 different LLMs. Access issues with Claude v1, specifically restrictions on its ability to process certain medical information, resulted in the exclusion of data from 6 papers, limiting the study's scope to 33 papers. The percentage of correct answers for each LLM is shown in Table 3, with GPT-3.5-Turbo achieving the highest rate (n=3916, 66.9%), followed by GPT-4-1106 (n=3837, 65.6%), PaLM 2 (n=3632, 62.1%), Claude v1 (n=2887, 58.3%), Gemini Pro (n=2878, 49.2%), and GPT-4-0613 (n=2580, 44.1%).

Table 3. The total amounts of correct answers among large language models (LLMs).

LLM	Total questions asked	Correct answers, n (%)
GPT-3.5-Turbo-1106	5850	3916 (66.9)
GPT-4-0613	5850	2580 (44.1)
GPT-4-1106	5850	3837 (65.6)
Claude v1	4950	2887 (58.3)
PaLM 2-chat-bison	5850	3632 (62.1)
Gemini Pro	5850	2878 (49.2)

Each LLM was compared with another LLM that provided a lower percentage of correct answers. Statistical analysis using the Kruskal-Wallis test revealed statistically significant differences between the LLMs (P<.001). The lowest correct answer percentage was provided by GPT-4-0613, at 44.1% (n=2580). Gemini Pro yielded 49.2% (n=2878) correct answers, significantly higher than GPT-4-0613 (P<.001). Claude v1 yielded 58.3% (n=2887) correct answers, statistically significantly higher than Gemini Pro (P<.001). PaLM 2 achieved 62.1% (n=3632) correct answers, significantly higher than Claude v1 (P<.001). GPT-4-1106 achieved 65.6% (n=3837) correct answers, significantly higher than PaLM 2 (P<.001). The difference between GPT-4-1106 and GPT-3.5-Turbo-1106 was not statistically significant (P=.06). Of the 39 papers analyzed, 28 (71.8%) were published before the training data cutoff date for GPT-3.5-Turbo and GPT-4-0613, while all 39 (100%) papers were published before the cutoff date for GPT-4-1106. Explicit cutoff dates for the remaining LLMs (Claude, PaLM 2, and Gemini Pro) were not publicly available and therefore could not be assessed in this study. When all LLMs are collectively considered, the 3 questions receiving the highest percentage of correct answers were question 12 (n=4025, 68.3%), question 13 (n=3695, 62.8%), and question 10 (n=3565, 60.5%). Conversely, the 3 questions with the lowest percentage of correct responses were question 8 (n=1971, 33.5%), question 15 (n=2107, 35.8%), and question 1 (n=2147, 36.5%; Table 4).

Table 4. Correct answer percentages of large language models (LLMs) for each question.

Question	Correct answers (across all LLMs), n (%)
Q1	2147 (36.5)
Q2	3061 (52)
Q3	2953 (50.2)
Q4	2713 (46.2)
Q5	3353 (57.1)
Q6	3132 (53.3)
Q7	2530 (43)
Q8	1971 (33.5)
Q9	2288 (38.9)
Q10	3565 (60.5)
Q11	3339 (56.9)
Q12	4025 (68.3)
Q13	3695 (62.8)
Q14	2578 (43.8)
Q15	2107 (35.8)

The percentages of correct answers given by all LLMs for each question are depicted in Figure 3. The median values for questions 7, 8, 9, 10, and 14 were similar across all LLMs, indicating a general consistency in performance for these specific areas of comprehension. However, significant differences were observed in the performance of different LLMs for other questions. The statistical tests used in this analysis were the Kruskal-Wallis test for comparing the medians of multiple groups and the chi-square test for comparing categorical data. For question 1, the fewest correct answers were provided by Claude (n=124, 24.8%) and Gemini Pro (n=197, 39.5%), while the most correct answers were provided by PaLM 2 (n=301, 60.3%; *P*=.01). In question 2, Claude v1 (n=366, 73.3%) achieved the highest median correct answer count (10.0, IQR 5.0-10.0), while Gemini Pro provided the fewest correct answers (n=237, 47.4%; *P*=.03). For question 3, GPT-3.5 (n=425, 85.1%) and PaLM 2 (n=434, 86.8%) had the highest median correct answer counts, while GPT-4-0613 (n=164, 32.8%) and Gemini Pro (n=189, 37.9%) had the lowest (*P*<.001). In the fourth question, PaLM 2 (n=369, 73.8%), GPT-3.5 (n=293, 58.7%), and GPT-4-1106 (n=336, 67.2%) performed best, while GPT-4-0613 (n=187, 37.4%) showed the lowest performance (*P*<.001). For questions 5 and 6, GPT-4-0613 (n=209, 41.8%) and Gemini Pro (n=186, 37.2%) provided fewer correct answers compared to the other LLMs (*P*<.001 and *P*=.001, respectively). In question 11, GPT-4-1106 (n=406, 81.2%), Claude (n=347, 69.4%), and PaLM 2 (n=406, 81.2%) performed well, while Gemini Pro (n=264, 52.8%) had the fewest correct answers (*P*=.001). For questions 12 and 13, all LLMs, except GPT-4-0613, performed well in these areas (*P*<.001). In question 15, GPT-3.5 (n=368, 73.6%) showed the highest number of correct answers (*P*<.001; Multimedia Appendix 1).

Figure 3. Comparative analysis of correct responses by large language models across 10 iterations for each question.

Questions	1	2	3	4	5	6	*P value*
Title and abstract (Q1)							
Q1. Does the paper indicate the study's design with a commonly used term in the title or the abstract?	48.5% 4.0 (3.0-8.0)	46.9% 5.0 (0.0-10.0)	37.7% 2.0 (0.0-7.0)	24.8% 1.0 (0.0-4.0)	60.3% 7.5 (1.8-10.0)	39.5% 2.0 (0.0-9.0)	**.01**
Methods (Q2-8)							
Q2. What is the observational study type: cohort, case-control, or cross-sectional studies?	62.8% 8.0 (2.0-10.0)	53.8% 6.0 (0.0-10.0)	70% 9.0 (4.0-10.0)	73.3% 10.0 (5.0-10.0)	68.5% 8.0 (4.8-10.0)	47.4% 4.0 (0.0-10.0)	**.03**
Q3. Were settings or locations mentioned in the method?	85.1% 10.0 (8.0-10.0)	32.8% 2.0 (0.0-6.0)	71% 8.0 (5.0-10.0)	48.8% 5.0 (2.0-7.5)	86.8% 10.0 (9.0-10.0)	37.9% 3.0 (0.0-7.0)	**<.001**
Q4. Were relevant dates mentioned in the method?	58.7% 7.0 (3.0-10.0)	37.4% 3.0 (0.0-7.0)	67.2% 7.0 (4.0-10.0)	50.3% 5.0 (0.0-10.0)	73.8% 9.0 (5.8-10.0)	46.2% 4.0 (0.0-8.0)	**<.001**
Q5. Were eligibility criteria for selecting participants mentioned in the method?	84.7% 10.0 (9.0-10.0)	41.8% 4.0 (0.0-8.0)	86.4% 10.0 (8.0-10.0)	80.9% 10.0 (8.0-10.0)	83.2% 10.0 (9.5-10.0)	37.2% 3.0 (0.0-7.0)	**<.001**
Q6. Were sources and methods of selection of participants mentioned in the method?	70% 10.0 (3.0-10.0)	49.5% 7.0 (0.0-9.0)	74.1% 10.0 (5.0-10.0)	72.1% 10.0 (2.5-10.0)	69.1% 10.0 (0.0-10.0)	50.3% 6.0 (0.0-9.0)	**.001**
Q7. Were any efforts to address potential sources of bias described in the method or discussion?	47.7% 4.0 (0.0-10.0)	46.7% 4.0 (1.0-9.0)	46.7% 5.0 (0.0-10.0)	59.7% 8.0 (0.0-10.0)	60.9% 10.0 (0.0-10.0)	49.7% 5.0 (0.0-10.0)	.55
Q8. Which program was used for statistical analysis?	46.7% 4.0 (0.0-9.0)	32.3% 3.0 (0.0-5.0)	41% 4.0 (1.0-7.0)	31.8% 3.0 (0.0-5.5)	50.6% 5.0 (0.0-10.0)	38.5% 3.0 (0.0-8.0)	.35
Results (Q9-12)							
Q9. Were report numbers of individuals at each stage of the study (eg, potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed) mentioned in the results?	41% 0.0 (0.0-10.0)	46.7% 4.0 (0.0-10.0)	50.5% 5.0 (3.0-7.0)	32.4% 0.0 (0.0-10.0)	63.8% 8.5 (2.8-10.0)	44.1% 0.0 (0.0-10.0)	.05
Q10. Was a flowchart used to show the reported numbers of individuals at each stage of the study?	72.3% 9.0 (4.0-10.0)	74.1% 10.0 (5.0-10.0)	78.2% 10.0 (4.0-10.0)	72.1% 10.0 (2.0-10.0)	66.2% 9.0 (3.0-10.0)	71% 10.0 (4.0-10.0)	.63
Q11. Were the study participants' demographic characteristics (eg, age and sex) given in the results?	72.8% 8.0 (5.0-10.0)	53.3% 6.0 (1.0-10.0)	81.3% 10.0 (7.0-10.0)	69.4% 10.0 (1.0-10.0)	81.2% 10.0 (8.5-10.0)	52.8% 4.0 (2.0-10.0)	**<.001**
Discussion (Q12-14)							
Q12: Does the discussion part summarize key results concerning study objectives?	80% 10.0 (6.0-10.0)	56.4% 6.0 (3.0-10.0)	93.6% 10.0 (10.0-10.0)	89.1% 10.0 (10.0-10.0)	88.8% 10.0 (10.0-10.0)	84.9% 10.0 (8.0-10.0)	**<.001**
Q13. Are the limitations of the study discussed in the paper?	98.5% 10.0 (10.0-10.0)	22.8% 1.0 (0.0-4.0)	96.9% 10.0 (10.0-10.0)	81.2% 10.0 (7.0-10.0)	86.5% 10.0 (9.8-10.0)	67.4% 10.0 (3.0-10.0)	**<.001**
Q14. Is the generalizability of the study discussed in the discussion part?	61.8% 8.0 (2.0-10.0)	39.5% 0.0 (0.0-10.0)	49.5% 4.0 (2.0-10.0)	55.8% 6.0 (2.5-10.0)	62.1% 9.0 (1.5-10.0)	47.7% 4.0 (1.0-8.0)	.15
Funding (Q15)							
Q15. Is the funding of the study mentioned in the paper?	73.6% 9.0 (4.0-10.0)	27.4% 0.0 (0.0-4.0)	39.7% 3.0 (0.0-8.0)	33% 3.0 (0.0-6.0)	60.9% 7.0 (3.0-10.0)	23.3% 0.0 (0.0-3.0)	**<.001**

1=GPT-3.5-Turbo-1106, **2**=GPT-4-0613 **3**=GPT-4-1106 **4**=Claude v1 **5**=PaLM 2-chat-bison, **6**=Gemini Pro
Each cell in the table displays the percentage of accurate answers, the minimum and maximum correct answer counts, and the median correct answer counts out of 10 trials. The analysis was conducted using the Kruskal-Wallis test.

0%-25% **25.1%-50%** **50.1%-75%** **75.1%-100%**

Discussion

Principal Findings

AI can improve the data analysis and publication process in scientific research while also being used to generate medical papers [16]. Although these fraudulent papers may appear well-crafted, their semantic inaccuracies and errors can be detected by expert readers upon closer examination [11,17]. The impact of LLMs on health care is often discussed in terms of their ability to replace health professionals, but their significant impact on medical and research writing applications and limitations is often overlooked. Therefore, physicians involved in research need to be cautious and verify information when using LLMs. As their reliance can lead to ethical concerns and inaccuracies, the scientific community should be vigilant in ensuring the accuracy and reliability of AI tools by using them as aids rather than replacements, understanding their limitations and biases [10,18]. With millions of papers published annually, AI could generate summaries or recommendations, simplifying the process of gathering evidence and enabling researchers to grasp important aspects of scientific results more efficiently [18]. Moreover, there is limited research focused on assessing the comprehension of academic papers.

This study aimed to evaluate the ability of 6 different LLMs to understand medical research papers using the STROBE checklist. We used a novel benchmark pipeline that processed 39 PubMed papers, posing 15 questions derived from the STROBE checklist to each model. The benchmark was established using the answers provided by an experienced medical professor and validated by an epidemiologist, serving as a reference standard against which the LLMs' responses were compared. To mitigate the problem of "artificial hallucinations" inherent to LLMs, our study implemented the RAG method, which involves using a web application to dissect PDF-format medical papers into text chunks and present them to the LLMs.

Our findings reveal significant variation in the performance of different LLMs, suggesting that LLMs are capable of understanding medical papers to varying degrees. While newer models like GPT-3.5-Turbo and GPT-4-1106 generally demonstrated better comprehension, GPT-3.5-Turbo outperformed even the more recent GPT-4-0613 in certain areas. This unexpected finding highlights the complexity of LLM performance, indicating that simple assumptions about newer models consistently outperforming older ones may not always hold true. The impact of training data cutoffs on LLM performance is a critical consideration in evaluating their ability to understand medical research [19]. While we were able to obtain explicitly stated cutoff dates for GPT-3.5-Turbo, GPT-4-1106, and GPT-4-0613, this information was not readily available for the remaining models. This lack of transparency regarding training data limits our ability to definitively assess the impact of knowledge cutoffs on model performance. The observation that all 39 papers were published before the cutoff date for GPT-4-1106, while only 28 papers were published

before the cutoff date for GPT-3.5-Turbo and GPT-4-0613, suggests that the knowledge cutoff may play a role in the observed performance differences. GPT-4-1106, with a more recent knowledge cutoff, has access to a larger data set, potentially including information from more recently published research. This could contribute to its generally better performance compared to GPT-3.5-Turbo. However, it is important to note that GPT-3.5-Turbo still outperformed GPT-4-0613 in specific areas, even with a similar knowledge cutoff. This suggests that factors beyond training data (eg, the number of layers, the type of attention mechanism, or the use of transformers) and compression techniques (eg, quantization, pruning, or knowledge distillation) may also play a significant role in LLM performance. Future research should prioritize transparency regarding training data cutoffs and aim to standardize how LLMs communicate these crucial details to users.

This study evaluated the performance of various LLMs in accurately answering specific questions related to different sections of a scholarly paper: title and abstract, methods, results, discussion, and funding. The results shed light on which LLMs excel in specific areas of comprehension and information retrieval from academic texts. PaLM 2 (n=219, 60.3%) showed superior performance in question 1, identifying the study design from the title or abstract, suggesting enhanced capability in understanding and identifying specific terminologies. Claude (n=82, 24.8%) and Gemini Pro (n=154, 39.5%), however, lagged, indicating a potential area for improvement in terminology recognition and interpretation. Claude v1 (n=242, 73.3%) and PaLM 2 (n=295, 86.8%) exhibited strong capabilities in identifying methodological details, such as observational study types and settings or locations (questions 2-8). This suggests a robust understanding of complex methodological descriptions and the ability to distinguish between different study frameworks. For questions regarding the results section (questions 9-11), it is evident that models like GPT-4-1106 (n=317, 81.3%), Claude (n=229, 69.4%), and PaLM 2 (n=276, 81.2%) showed superior performance in providing correct answers related to the study participants' demographic characteristics and the use of flowcharts. All LLMs except for GPT4-0613 (n=89, 22.8%) exhibited remarkable competence in summarizing key results, discussing limitations, and addressing the generalizability of the study (questions 12-14), which are critical aspects of the discussion section. GPT-3.5 (n=287, 73.6%) particularly excelled in identifying the mention of funding (question 15), indicating a nuanced understanding of acknowledgments and funding disclosures often nuanced and embedded toward the end of papers. Across the array of tested questions, both GPT-3.5 and PaLM 2 exhibit remarkable strengths in understanding and analyzing scholarly papers, with PaLM 2 generally showing a slight edge in versatility, especially in interpreting methodological details and study design. GPT-3.5, while strong in discussing study limitations, generalized findings, and funding details, indicates that improvements can be made in extracting complex methodological information. We observed that different models excelled in different areas, indicating that no single LLM currently demonstrates universal dominance in medical paper understanding. This suggests that factors like training data, model architecture, and question complexity influence performance, and further research is needed to understand the specific contributions of each factor.

Comparison to Prior Work

LLMs can be directly questioned and can generate answers from their own memory [11]. This has been extensively studied in many medical papers. According to a study, ChatGPT, an LLM, was evaluated on the United States Medical Licensing Examination. The results showed that GPT performed at or near the passing threshold for examinations without any specialized training, demonstrating a high level of concordance and insight in its explanations. These findings suggest that LLMs have the potential to aid in medical education and potentially assist with clinical decision-making [5,20]. Another study aimed to evaluate the knowledge level of GPT in medical education by assessing its performance in a multiple-choice question examination and its potential impact on the medical examination system. The results indicated that GPT achieved a satisfactory score in both basic and clinical medical sciences, highlighting its potential as an educational tool for medical students and faculties [21]. Furthermore, GPT offers information and aids health care professionals in diagnosing patients by analyzing symptoms and suggesting appropriate tests or treatments. However, advancements are required to ensure AI's interpretability and practical implementation in clinical settings [8]. The study conducted in October 2023 explored the diagnostic capabilities of GPT-4V, an AI model, in complex clinical scenarios involving medical imaging and textual patient data. Results showed that GPT-4V had the highest diagnostic accuracy when provided with multimodal inputs, aligning with confirmed diagnoses in 80.6% of cases [22]. In another study, GPT-4 was instructed to address the case with multiple-choice questions followed by an unedited clinical case report that evaluated the effectiveness of the newly developed AI model GPT-4 in solving complex medical case challenges. GPT-4 correctly diagnosed 57% of the cases, outperforming 99.98% of human readers who were also tasked with the same challenge [23]. These studies highlight the potential of multimodal AI models like GPT-4 in clinical diagnostics, but further investigation is needed to uncover biases and limitations due to the model's proprietary training data and architecture.

There are few studies in which LLMs are directly questioned, and their capacities to produce answers from their own memories are compared with each other and expert clinicians. In a study, GPT-3.5 and GPT-4 were compared to orthopedic residents in their performance on the American Board of Orthopaedic Surgery written examination, with residents scoring higher overall, and a subgroup analysis revealed that GPT-3.5 and GPT-4 outperformed residents in answering text-only questions, while residents scored higher in image interpretation questions. GPT-4 scored higher than GPT-3.5 [24]. A study aimed to evaluate and compare the recommendations provided by GPT-3 and GPT-4 with those of primary care physicians for the management of depressive episodes. The results showed that both GPT-3.5 and GPT-4 largely aligned with accepted guidelines for treating mild and severe depression while demonstrating a lack of gender or socioeconomic biases observed among primary care physicians. However, further

research is needed to refine the AI recommendations for severe cases and address potential ethical concerns and risks associated with their use in clinical decision-making [25]. Another study assessed the accuracy and comprehensiveness of health information regarding urinary incontinence generated by various LLMs. By inputting selected questions into GPT-3.5, GPT-4, and Gemini, the researchers found that GPT-4 performed the best in terms of accuracy and comprehensiveness, surpassing GPT-3.5 and Gemini [26]. According to a study that evaluates the performance of 2 GPT models (GPT-3.5 and GPT-4) and human professionals in answering ophthalmology questions from the StatPearls question bank, GPT-4 outperformed both GPT-3.5 and human professionals on most ophthalmology questions, showing significant performance improvements and emphasizing the potential of advanced AI technology in the field of ophthalmology [27]. Some studies showed that GPT-4 is more proficient, as evidenced by scoring higher than GPT-3.5 in both multiple-choice dermatology examinations and non–multiple-choice cardiology heart failure questions from various sources and outperforming GPT-3.5 and Flan-PaLM 540B on medical competency assessments and benchmark data sets [28-30]. In a study conducted on the proficiency of various open-source and proprietary LLMs in the context of nephrology multiple-choice test-taking ability, it was found that their performance on 858 nephSAP questions ranged from 17.1% to 30.6%, with Claude 2 at 54.4% accuracy and GPT-4 at 73.3%, highlighting the potential for adaptation in medical training and patient care scenarios [31]. To our knowledge, this is the first study to assess the performance of evaluating medical papers and understanding the capabilities of different LLMs. The findings reveal that the performance of LLMs varies across different questions, with some LLMs showing superior understanding and answer accuracy in certain areas. Comparative analysis across different LLMs showcases a gradient of capabilities. The results revealed a hierarchical performance ranking as follows: GPT-4-1106 equals GPT-3.5-Turbo, which is superior to PaLM 2, followed by Claude v1, then Gemini Pro, and finally, GPT-4-0613. Similar to the literature review, GPT-4-1106 and GPT-3.5 showed improved accuracy and understanding compared to other LLMs. This mirrors wider literature trends, indicating LLMs' rapid evolution and increasing sophistication in handling complex medical queries. Notably, GPT-3.5-Turbo showed better performance than GPT-4-0613, which may be counterintuitive, considering the tendency to assume newer iterations naturally perform better. This anomaly in performance between newer and older versions can be attributed to the application of compression techniques in developing new models to reduce computational costs. While these advancements make deploying LLMs more cost-effective and thus accessible, they can inadvertently compromise the performance of LLMs. The notable absence of responses from PaLM in certain instances, actually stemming from Google's policy to restrict the use of its medical information, presents an intriguing case within the scope of our discussion. Despite these constraints, PaLM's demonstrated high performance in other areas is both surprising and promising. This suggests that even when faced with limitations on accessing a vast repository of medical knowledge, PaLM's underlying architecture and algorithms enable it to make effective use of the information it can access, showcasing the robust potential of LLMs in medical settings even under restricted conditions.

Strengths and Limitations

While LLMs can be directly questioned and generate answers from their own memory, as demonstrated in numerous studies above, this approach can lead to inaccuracies known as hallucinations. Hallucinations in LLMs have diverse origins, encompassing the entire spectrum of the capability acquisition process, with hallucinations primarily categorized into 3 aspects: training, inference, and data. Architecture flaws, exposure bias, and misalignment issues in both pretraining and alignment phases induce hallucinations. To address this challenge, our study used the RAG method, ensuring that the LLMs' responses were grounded in factual information retrieved from the target paper. The RAG method intuitively addresses the knowledge gap by conditioning language models on relevant documents retrieved from an external knowledge source [12,32]. RAG provides the LLM with relevant text chunks extracted from the specific paper being analyzed. This ensures that the LLM's responses are directly supported by the provided information, reducing the risk of hallucination. While a few studies have explored the use of RAG to compare LLMs, like the one demonstrating GPT-4's improved accuracy with RAG for interpreting oncology guidelines [33], our study is the first to evaluate LLM comprehension of medical research papers using this method. This method conditions LLMs on relevant documents retrieved from an external knowledge source, ensuring their answers are grounded in factual information. The design of system prompts is crucial for LLMs, as it provides context, instructions, and formatting guidelines to ensure the desired output [34]. In this study, it is empirically determined that a foundational system and set of system prompts universally enhanced the response quality across all language models tested. This approach was designed to optimize the comprehension and summarization capabilities of each generative AI tool when processing medical research papers. The specific configuration of system settings and query structures we identified significantly contributed to improving the accuracy and relevance of the models' answers. These optimized parameters were crucial in achieving a more standardized and reliable evaluation of each model's ability to understand complex medical texts. While further research is needed to fully understand the effectiveness of RAG across different medical scenarios, our findings demonstrate its potential to enhance the reliability and accuracy of LLMs in medical research comprehension.

This study, while offering valuable insights, is subject to several limitations. The selection of 50 papers focused on obesity, and the use of a specific set of 15 STROBE-derived questions might not fully capture the breadth of medical research. Additionally, the reliance on binary and multiple-choice questions restricts the evaluation of LLMs' ability to provide nuanced answers. The rapid evolution of LLMs means that the findings might not be applicable to future versions, and potential biases within the training data have not been systematically assessed. Furthermore, the study's reliance on a single highly experienced medical professor as the benchmark, while evaluating, might

limit the generalizability of the findings. A larger panel of experts with diverse areas of specialization might provide a more comprehensive reference standard for evaluating LLM performance. Further investigation with a wider scope and more advanced methodologies is needed to fully understand the potential of LLMs in medical research.

Future Directions

In conclusion, LLMs show promise for transforming medical research, potentially enhancing research efficiency and evidence-based decision-making. This study demonstrates that LLMs exhibit varying capabilities in understanding medical research papers. While newer models generally demonstrate better comprehension, no single LLM currently excels in all areas. This highlights the need for further research to understand the complex interplay of factors influencing LLM performance. Continued research is crucial to address these limitations and ensure the safe and effective integration of LLMs in health care, maximizing their benefits while mitigating risks.

Acknowledgments

The authors gratefully acknowledge Dr Hilal Duzel for her invaluable assistance in validating the reference standard used in this study. Dr Duzel's expertise in epidemiology and statistical analysis ensured the accuracy and robustness of the benchmark against which the LLMs were evaluated. We would also like to thank Ahmet Hamza Dogan, a promising future engineer, for his contributions to the LLM analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Percentages of correct answers by large language models for each question.
[PNG File , 1049 KB - medinform_v12i1e59258_app1.png]

References

1. Lv Z. Generative artificial intelligence in the metaverse era. Cogn Robot 2023;3:208-217. [doi: 10.1016/j.cogr.2023.06.001]
2. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
3. Orrù G, Piarulli A, Conversano C, Gemignani A. Human-like problem-solving abilities in large language models using ChatGPT. Front Artif Intell 2023;6:1199350 [FREE Full text] [doi: 10.3389/frai.2023.1199350] [Medline: 37293238]
4. Chenais G, Gil-Jardiné C, Touchais H, Avalos Fernandez M, Contrand B, Tellier E, et al. Deep learning transformer models for building a comprehensive and real-time trauma observatory: development and validation study. JMIR AI 2023 Jan 12;2:e40843 [FREE Full text] [doi: 10.2196/40843] [Medline: 38875539]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inform Manage 2023 Aug;71:102642. [doi: 10.1016/j.ijinfomgt.2023.102642]
7. Akyon SH, Akyon FC, Yılmaz TE. Artificial intelligence-supported web application design and development for reducing polypharmacy side effects and supporting rational drug use in geriatric patients. Front Med (Lausanne) 2023;10:1029198 [FREE Full text] [doi: 10.3389/fmed.2023.1029198] [Medline: 36968816]
8. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122-1131.e9 [FREE Full text] [doi: 10.1016/j.cell.2018.02.010] [Medline: 29474911]
9. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform 2024 May 10;12:e53787 [FREE Full text] [doi: 10.2196/53787] [Medline: 38728687]
10. Kumar M, Mani UA, Tripathi P, Saalim M, Roy S. Artificial hallucinations by Google Bard: think before you leap. Cureus 2023 Aug;15(8):e43313 [FREE Full text] [doi: 10.7759/cureus.43313] [Medline: 37700993]
11. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: pandora's box has been opened. J Med Internet Res 2023 May 31;25:e46924 [FREE Full text] [doi: 10.2196/46924] [Medline: 37256685]
12. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. ArXiv. Preprint posted online on April 15, 2021 [FREE Full text] [doi: 10.18653/v1/2021.findings-emnlp.320]
13. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019 Apr;13(Suppl 1):S31-S34 [FREE Full text] [doi: 10.4103/sja.SJA_543_18] [Medline: 30930717]

14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007 Oct 20;335(7624):806-808 [FREE Full text] [doi: 10.1136/bmj.39335.541782.AD] [Medline: 17947786]
15. STROBE Checklist: cohort, case-control, and cross-sectional studies (combined). URL: https://www.strobe-statement.org/download/strobe-checklist-cohort-case-control-and-cross-sectional-studies-combined [accessed 2023-12-28]
16. Chen T. ChatGPT and other artificial intelligence applications speed up scientific writing. J Chin Med Assoc 2023 Apr 01;86(4):351-353. [doi: 10.1097/JCMA.0000000000000900] [Medline: 36791246]
17. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. Radiology 2023 Apr;307(2):e230171. [doi: 10.1148/radiol.230171] [Medline: 36728749]
18. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11:1166120 [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
19. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res 2023 Dec 28;25:e51580 [FREE Full text] [doi: 10.2196/51580] [Medline: 38009003]
20. Wong RS, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. JMIR Med Educ 2023 Nov 21;9:e47274 [FREE Full text] [doi: 10.2196/47274] [Medline: 37988149]
21. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare (Basel) 2023 Jul 17;11(14):2046 [FREE Full text] [doi: 10.3390/healthcare11142046] [Medline: 37510487]
22. Schubert MC, Lasotta M, Sahm F, Wick W, Venkataramani V. Evaluating the multimodal capabilities of generative AI in complex clinical diagnostics. Medrxiv 2023. [doi: 10.1101/2023.11.01.23297938]
23. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. NEJM AI 2023 Dec 11;1(1):AIp2300031. [doi: 10.1056/aip2300031]
24. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg 2023 Dec 01;31(23):1173-1179 [FREE Full text] [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]
25. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health 2023 Sep;11(4):e002391 [FREE Full text] [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
26. Coşkun B, Bayrak O, Ocakoglu G, Acar HM, Kaygisiz O. Assessing the accuracy of AI language models in providing information on urinary incontinence: a comparative study. Eur J Public Health 2023;3(3):61-70 [FREE Full text] [doi: 10.29228/ejhh.71797]
27. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial intelligence in ophthalmology: a comparative analysis of GPT-3.5, GPT-4, and human expertise in answering StatPearls questions. Cureus 2023 Jun;15(6):e40822 [FREE Full text] [doi: 10.7759/cureus.40822] [Medline: 37485215]
28. King RC, Samaan JS, Yeo YH, Mody B, Lombardo DM, Ghashghaei R. Appropriateness of ChatGPT in answering heart failure related questions. Heart Lung Circ 2024 May 30:1-5 [FREE Full text] [doi: 10.1016/j.hlc.2024.03.005] [Medline: 38821760]
29. Passby L, Jenko N, Wernham A. Performance of ChatGPT on Specialty Certificate Examination in Dermatology multiple-choice questions. Clin Exp Dermatol 2024 Jun 25;49(7):722-727. [doi: 10.1093/ced/llad197] [Medline: 37264670]
30. Nori H, King N, McKinney S, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. ArXiv. Preprint posted online on April 12, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.13375]
31. Wu S, Koo M, Blum L, Black A, Kao L, Fei Z, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. NEJM AI 2024 Jan 25;1(2):AIdbp2300092. [doi: 10.1056/aidbp2300092]
32. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. ArXiv. Preprint posted online on April 12, 2021 [FREE Full text] [doi: 10.48550/arXiv.2005.11401]
33. Ferber D, Wiest IC, Wölflein G, Ebert MP, Beutel G, Eckardt J, et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. NEJM AI 2024 May 23;1(6):AIcs2300235. [doi: 10.1056/aics2300235]
34. Chen Q, Sun H, Liu H, Jiang Y, Ran T, Jin X, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. Bioinformatics 2023 Sep 02;39(9):btad557 [FREE Full text] [doi: 10.1093/bioinformatics/btad557] [Medline: 37682111]

Abbreviations

AI: artificial intelligence
LLM: large language model
RAG: retrieval augmented generation

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by A Castonguay; submitted 07.04.24; peer-reviewed by C Wang, S Mao, W Cui; comments to author 04.06.24; revised version received 16.06.24; accepted 05.07.24; published 04.09.24.

Please cite as:
Akyon SH, Akyon FC, Camyar AS, Hızlı F, Sari T, Hızlı Ş
Evaluating the Capabilities of Generative AI Tools in Understanding Medical Papers: Qualitative Study
JMIR Med Inform 2024;12:e59258
URL: https://medinform.jmir.org/2024/1/e59258
doi:10.2196/59258
PMID:39230947

Original Paper

Comparative Study to Evaluate the Accuracy of Differential Diagnosis Lists Generated by Gemini Advanced, Gemini, and Bard for a Case Report Series Analysis: Cross-Sectional Study

Takanobu Hirosawa[1], MD, PhD; Yukinori Harada[1], MD, PhD; Kazuki Tokumasu[2], MD, PhD; Takahiro Ito[3], MD; Tomoharu Suzuki[4], MD; Taro Shimizu[1], MD, MSc, MPH, MBA, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Shimotsuga, Japan

[2]Department of General Medicine, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan

[3]Satsuki Home Clinic, Tochigi, Japan

[4]Department of Hospital Medicine, Urasoe General Hospital, Okinawa, Japan

Corresponding Author:
Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi, Mibu-cho
Shimotsuga, 321-0293
Japan
Phone: 81 282861111
Email: hirosawa@dokkyomed.ac.jp

Abstract

Background: Generative artificial intelligence (GAI) systems by Google have recently been updated from Bard to Gemini and Gemini Advanced as of December 2023. Gemini is a basic, free-to-use model after a user's login, while Gemini Advanced operates on a more advanced model requiring a fee-based subscription. These systems have the potential to enhance medical diagnostics. However, the impact of these updates on comprehensive diagnostic accuracy remains unknown.

Objective: This study aimed to compare the accuracy of the differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard across comprehensive medical fields using case report series.

Methods: We identified a case report series with relevant final diagnoses published in the *American Journal Case Reports* from January 2022 to March 2023. After excluding nondiagnostic cases and patients aged 10 years and younger, we included the remaining case reports. After refining the case parts as case descriptions, we input the same case descriptions into Gemini Advanced, Gemini, and Bard to generate the top 10 differential diagnosis lists. In total, 2 expert physicians independently evaluated whether the final diagnosis was included in the lists and its ranking. Any discrepancies were resolved by another expert physician. Bonferroni correction was applied to adjust the *P* values for the number of comparisons among 3 GAI systems, setting the corrected significance level at *P* value <.02.

Results: In total, 392 case reports were included. The inclusion rates of the final diagnosis within the top 10 differential diagnosis lists were 73% (286/392) for Gemini Advanced, 76.5% (300/392) for Gemini, and 68.6% (269/392) for Bard. The top diagnoses matched the final diagnoses in 31.6% (124/392) for Gemini Advanced, 42.6% (167/392) for Gemini, and 31.4% (123/392) for Bard. Gemini demonstrated higher diagnostic accuracy than Bard both within the top 10 differential diagnosis lists (*P*=.02) and as the top diagnosis (*P*=.001). In addition, Gemini Advanced achieved significantly lower accuracy than Gemini in identifying the most probable diagnosis (*P*=.002).

Conclusions: The results of this study suggest that Gemini outperformed Bard in diagnostic accuracy following the model update. However, Gemini Advanced requires further refinement to optimize its performance for future artificial intelligence–enhanced diagnostics. These findings should be interpreted cautiously and considered primarily for research purposes, as these GAI systems have not been adjusted for medical diagnostics nor approved for clinical use.

(JMIR Med Inform 2024;12:e63010) doi:10.2196/63010

KEYWORDS

artificial intelligence; clinical decision support; diagnostic excellence; generative artificial intelligence; large language models; natural language processing

Introduction

Diagnostic Team to Reduce Misdiagnoses

Diagnosis is a crucial step in clinical medicine, where a significant proportion of medical errors and harms are related to diagnostic errors [1]. The formation of a diagnostic team has been proposed as an effective strategy to mitigate the risks associated with misdiagnosis [2,3]. This team should promote collaboration among medical professionals, patients, and their families, and the integration of digital tools to enhance diagnostic accuracy [4]. Several research, including systematic reviews, have shown that the implementation of clinical decision support systems (CDSSs) in clinical settings has significantly improved diagnostic accuracy, patient care, and health care process [5-7].

Digital Tool for Medical Diagnosis

Various digital tools, particularly diagnostic CDSSs, have emerged for medical diagnostics. These systems are designed to provide diagnostic suggestions based on clinical data, aiding medical professionals in clinical decision-making [8]. Traditionally, diagnostic CDSSs, such as symptom checkers and differential diagnosis generators, have relied on fixed algorithms and rule-based systems derived from medical databases and expert input [9-11]. Unfortunately, these systems often experience poor accuracy and inadequate integration into clinical workflows, limiting their practical use in real-world medical settings [4]. In this context, artificial intelligence (AI), especially generative AI (GAI), has introduced a new category of CDSS [12]. This advancement suggests a future shift in how digital tools can support diagnostic processes.

GAI in Medical Diagnosis

GAI systems have shown rapid development and are increasingly influencing various fields, including medicine. This advancement is partly due to the development of machine learning techniques, such as neural networks and natural language processing. GAI represents a shift from rule-based systems to models that can autonomously generate and evaluate new data patterns. Overcoming many limitations faced by traditional CDSSs, GAI systems could significantly enhance future diagnostic processes. Notable examples include ChatGPT developed by Open AI, and Gemini and Gemini Advanced from Google [13]. These systems use advanced large language models (LLMs), which are complex neural networks trained on vast data sets through natural language processing [14]. Recent studies, including one evaluating dermoscopy image descriptions with chatbot responses, have demonstrated promising results in accuracy and diagnostic completeness by ChatGPT [15]. The language model for dialogue applications (LaMDA) developed by Google AI is one such LLM. Their ability to process and generate outputs is particularly promising for future applications in medical diagnostics, where they will analyze complex clinical information and collaborate as part of a diagnostic team.

From Bard to Gemini and Gemini Advanced

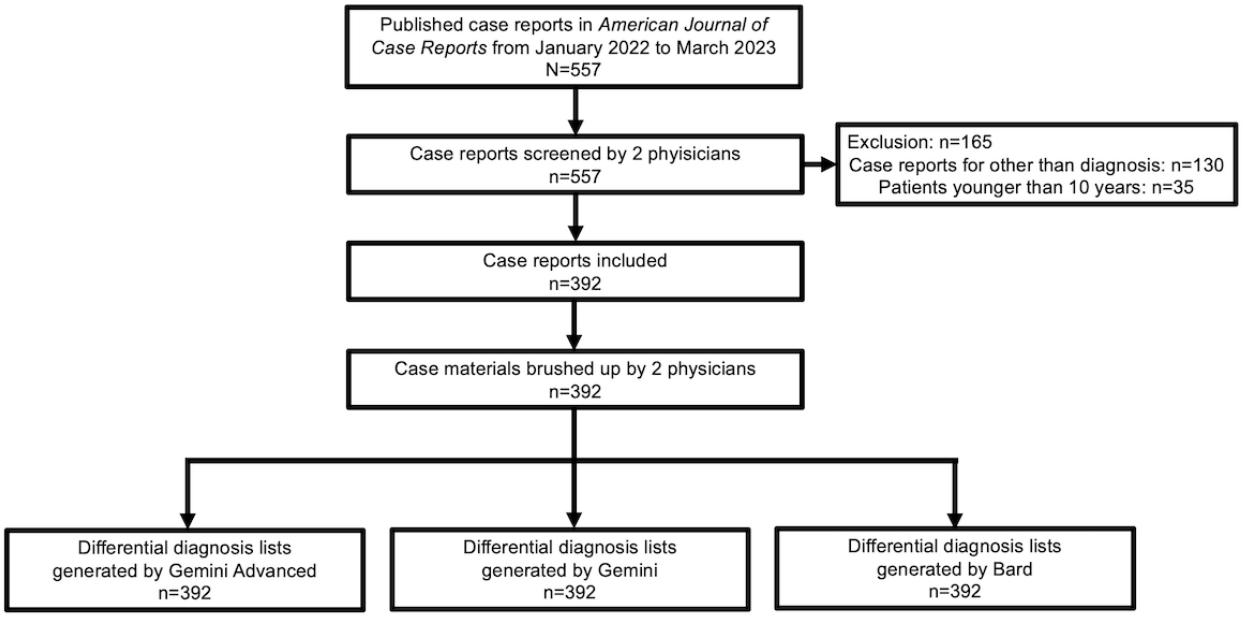
Originally, Bard was developed using the LaMDA model primarily for text generation and conversational AI and later transitioned to the Pathways Language Model (Palm 2). Subsequent developments led to the release of Gemini and Gemini Advanced in December 2023. Gemini Advanced, an upgraded version of Gemini, leverages Ultra 1.0, Google's most advanced model, offered as a fee-based service [16,17]. These developments reflect the rapid pace at which GAI technology is advancing. Recent updates have transformed Bard into Gemini and Gemini Advanced, enhancing their functionalities and applications in various fields. Previous research, including our own, has demonstrated that Bard showed promising results in medicine [18-21]. Moreover, a recent study has shown that several GAI systems, including Gemini Advanced, could achieve notable diagnostic accuracy for multiple-choice questions about clinical vignettes [22]. These findings suggest that even without specific training or reinforcement for diagnostics, GAI systems show potential for reliable use in diagnostics. Despite these advancements, the comparative diagnostic accuracy of differential diagnosis lists by these GAI systems across comprehensive medical fields remains to be fully explored. This study aims to fill that gap by evaluating the diagnostic accuracy of the differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard for case report series across various medical disciplines.

Methods

Overview

An experimental study was conducted to assess the diagnostic accuracy of Gemini Advanced, Gemini, and Bard for a comprehensive case report series. This study was conducted at the Department of Diagnostic and Generalist Medicine (General Internal Medicine) at Dokkyo Medical University, Japan. This study consisted of preparing case materials, generating differential diagnosis lists, evaluating the lists, and analyzing the diagnostic accuracy. Figure 1 shows the study flow, including the inclusion of case materials and the generation of differential diagnosis lists.

Figure 1. Study flow of inclusion case materials and generation of differential diagnosis.



Preparing Case Materials

We focused on a comprehensive series of case reports from the *American Journal of Case Reports*, covering a broad range of medical fields. The structured format of the journal facilitated easy identification of sections containing the case reports and the final diagnoses. Initially, the inclusion criteria were the case reports published in the *American Journal of Case Reports* from January 1, 2022, to March 1, 2023. A PubMed search identified 557 consecutive case reports. After excluding 130 nondiagnostic case reports and 35 pediatric case reports (patients aged 10 years and younger), 392 case reports remained. The exclusion criteria were based on previous research for CDSS [23]. We refined the case reports to prepare the case materials, which typically included the initial case report part to the definitive tests for final diagnosis. The relevant final diagnoses were typically described by the authors. We used only textual data exclusively, omitting image data. Specifically, the title, background, final diagnosis, clinical course following diagnosis, discussion, conclusion, figures, tables, and supplemental materials were excluded from the case materials. The main investigator (TH) conducted this process with validation from another investigator (YH). The PubMed search keywords are shown in Multimedia Appendix 1. For example, in a case report titled "Herpes Zoster Following COVID-19 Vaccine Booster," the final diagnosis was herpes zoster [24]. We extracted the case report part from "An 82-year-old.." to "Vesicular breath sounds were heard equally on both lung fields."

Generating Differential Diagnosis Lists

We used Gemini Advanced, Gemini, and Bard as GAI systems for this research. This was because these systems are popular AI platforms available to the public. These GAI systems were not specifically enhanced for medical diagnosis. Details about the GAI systems used in this study are provided in Table 1. To generate the top 10 differential diagnosis lists from GAI systems, the main investigator typically copied and pasted the case materials into the AI systems with the prompt, that is "Tell me the top 10 suspected illnesses for the following case: (case materials)." This prompt, developed through preliminary research, aimed to generate the top 10 differential diagnosis lists. The first list produced by the GAI systems was used as the differential diagnosis list. The data control setting was adjusted to "Not saving activity," to avoid the influence from the previous conversations. In addition, before starting a new session, the main investigator refreshed the previous session to prevent any influence from previous conversations.

Table 1. The details of generative artificial intelligence systems used in this study.

Gemini Advanced	Gemini	Bard
AI[a] model		
Ultra 1.0	Pro	Pathways Language Model (Palm 2)-based
Availability		
Fee-based subscription	Free with user login	Discontinued
The setting of the app activity		
Not saving activity	Not saving activity	Not saving activity
Access date		
April 4-9, 2024	March 12-28, 2024	July 1, 2023-August 8 2023
Prompt		
"Tell me the top 10 suspected illnesses for the following case: (case materials)."	"Tell me the top 10 suspected illnesses for the following case: (case materials)."	"Tell me the top 10 suspected illnesses for the following case: (case materials)."

Evaluating the Differential Diagnosis Lists

A total of 2 expert researchers (TI and T Suzuki) independently evaluated the differential diagnosis lists from GAI systems. A score of "1" was assigned if the differential accurately and specifically identified the final diagnosis or was sufficiently close to the final diagnosis. Conversely, a score of "0" was assigned if it diverged significantly from the final diagnosis [25]. When a GAI system could not output the differential diagnosis list, a score of "0" was labeled. When the score was "1," the evaluator assessed its ranking within the list. Any discrepancies were resolved by another expert researcher (KT). All evaluators were blinded to which GAI systems produced the differential diagnosis lists.

Analyzing the Diagnostic Accuracy

In this study, we defined diagnostic accuracy as the inclusion of the final diagnoses in the differential diagnosis lists.

Outcome

In terms of the outcomes, the primary outcome was the total score for correctly identifying the final diagnosis in the top 10 differential diagnosis lists generated by Gemini Advanced, Gemini, and Bard. The total number of included case reports was used as the denominator. The numerator was the number of case reports that correctly identified the final diagnosis in the top 10 differential diagnosis lists. The secondary outcomes were the total score for correctly identifying the final diagnosis in the top 5 differential diagnosis lists and as the top diagnosis generated from Gemini Advanced, Gemini, and Bard.

In addition, we evaluated the top 10 rankings of the most frequently named differential diagnoses across generated differential diagnosis lists by a GAI system to find the underlying patterns. We also assessed whether the items in the differential diagnosis lists corresponded to the names of existing diseases.

Moreover, we analyzed how Gemini Advanced, Gemini, and Bard rank the correct diagnosis on average when it appears in the differential diagnosis lists. This metric helps evaluate not only whether the correct diagnosis is included but also its relative priority among other suggested diagnoses. For cases where the correct diagnosis was missing, we assigned a penalty rank; specifically, we used 11 as the penalty rank.

Statistical Analysis

A chi-square test was used for the categorical or binary variables. The Mann-Whitney *U* test was applied to analyze the average rankings. For multiple comparisons, the Bonferroni correction was applied [26]. The Bonferroni correction adjusts the *P* value by dividing it by the number of comparisons being made, namely 3 GAI systems, that consist of Gemini Advanced, Gemini, and Bard, setting a more stringent threshold for individual tests to be considered significant. The Bonferroni-corrected significance level was defined as a *P* value of <.02. Note that all statistical tests were performed using the R (version 4.2.2; The R Foundation for Statistical Computing) for macOS X.

Ethical Considerations

Given the study's method of using published case reports, approval from an ethical committee was deemed not applicable.

Results

Overview of Outputs From GAI Systems

In total, 392 case reports were included. The distribution of diagnoses in these reports highlighted the prevalence of complex medical conditions, with 22.7% (89/392) of cases being diagnosed as malignancy, 22.4% (88/392) as various infections, and 17.1% (67/392) as vascular diseases.

Bard generated differential diagnosis lists for all 392 cases. In contrast, both Gemini Advanced and Gemini encountered challenges with 3 cases, failing to produce outputs. The cases where outputs were not generated involved complex clinical scenarios that were directly described in the reports as involving overdose, psychological conditions, and vaccination side effects. Representative outputs from Gemini Advanced, Gemini, and Bard are shown in Table 2. The data set of differential diagnoses generated by AI systems in this study, alongside the final diagnosis, are shown in Multimedia Appendix 2.

Table 2. Representative outputs from Gemini Advanced, Gemini, and Bard for final diagnoses.

Gemini Advanced	Gemini	Bard
Herpes Zoster		
• Shingles (Herpes Zoster) • Myocardial ischemia • Costochondritis • Pericarditis • Pleuritis • Pulmonary embolism • Pneumonia • Metastatic malignancy • Aortic dissection • Esophageal spasm	• Shingles (Herpes Zoster) • Myocardial ischemia (Angina) • Costochondritis • Musculoskeletal strain • Pleurisy • Intercostal neuritis • Pulmonary embolism • Pericarditis • Gastritis • Esophagitis	• Herpes zoster (Shingles) • Myocardial infarction (heart attack) • Aortic dissection • Pulmonary embolism • Pneumonia • Thoracic radiculopathy • Costochondritis • Intercostal neuralgia • Costoclavicular syndrome • Tietze syndrome
Diffuse alveolar hemorrhage		
• Pulmonary alveolar hemorrhage • Nocardia pneumonia • Pulmonary embolism • Acute eosinophilic pneumonia • Drug-induced pneumonitis • Legionnaires' disease • Hypersensitivity pneumonitis • Idiopathic pulmonary fibrosis • Bronchiolitis obliterans organizing pneumonia • Acute interstitial pneumonitis	• Pulmonary alveolar hemorrhage • Trastuzumab-induced pneumonitis • Aspirin-exacerbated respiratory disease • Acute heart failure • Pulmonary edema • Nocardia pneumonia • Pulmonary hypertension • Acute interstitial lung disease • Acute chronic kidney disease • Occult gastrointestinal bleed	• Acute radiation pneumonitis • Acute hypersensitivity pneumonitis • Pulmonary edema • Pulmonary embolism • Aspiration pneumonia • Bronchogenic carcinoma • Sarcoidosis • Lymphoma • Wegener’s granulomatosis • Goodpasture syndrome
Glomus tumor		
• Meralgia paresthetica • Soft tissue sarcoma • Fibroma • Myositis ossificans • Neuroma • Hemangioma • Liposarcoma • Lymphoma • Abscess • Malignant peripheral nerve sheath tumor	• Meralgia paresthetica • Lumbar radiculopathy • Femoral neuropathy • Piriformis syndrome • Muscle strain • Myofascial pain syndrome • Glomus tumor • Neuroma • Soft tissue sarcoma (unlikely) • Abscess (unlikely)	• Lipoma • Fibromatosis • Myositis • Neoplasm • Infection • Inflammation • Trauma • Hematoma • Varicose vein • Cellulitis

Diagnostic Accuracy

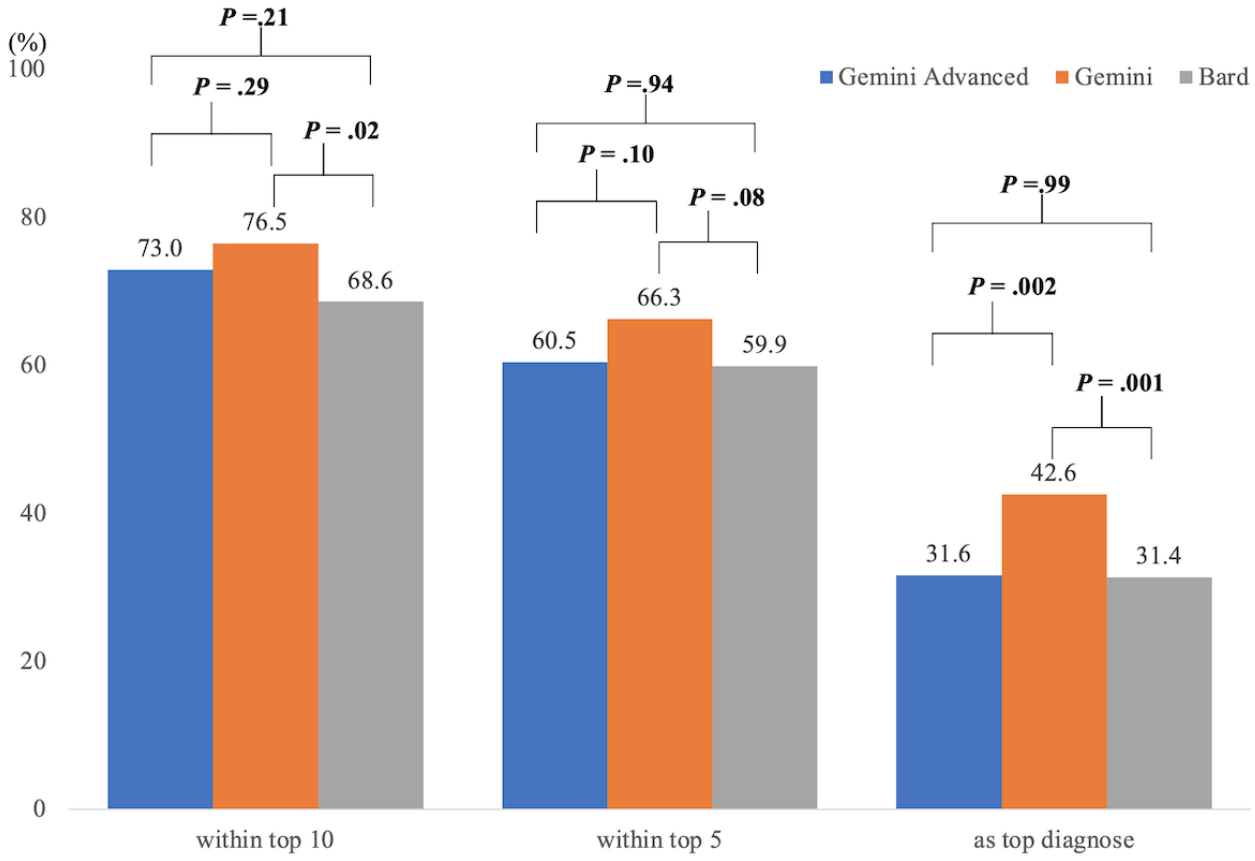
In terms of diagnostic accuracy, the inclusion rates of the final diagnoses within the top 10 differential diagnosis lists were 73.0% (286/392) for Gemini Advanced, 76.5% (300/392) for Gemini, and 68.6% (269/392) for Bard. For the top 5 differential diagnoses, the rates were 60.5% (237/392) for Gemini Advanced, 66.3% (260/392) for Gemini, and 59.9% (235/392) for Bard. The top diagnoses matched the final diagnoses in 31.6% (124/392) for Gemini Advanced, 42.6% (167/392) for Gemini, and 31.4% (123/392) for Bard. Gemini demonstrated higher diagnostic accuracy than Bard both within the top 10 differential diagnosis lists (P=.02) and as the top diagnosis (P=.001). In addition, Gemini Advanced achieved lower accuracy in identifying the most probable diagnosis, compared with Gemini with this result being statistically significant (P=.002). Other comparisons were statistically insignificant. Table 3 and Figure 2 show the diagnostic accuracy by Gemini Advanced, Gemini, and Bard.

Table 3. Diagnostic accuracy of Gemini Advanced, Gemini, and Bard.

Variable	Gemini Advanced (N=392), n (%)	Gemini (N=392), n (%)	Bard (N=392), n (%)	P value[a]		
				Gemini Advanced versus Gemini	Gemini Advanced versus Bard	Gemini versus Bard
Within the top 10	286 (73.0)	300 (76.5)	269 (68.6)	.29	.21	.02
Within the top 5	237 (60.5)	260 (66.3)	235 (59.9)	.10	.94	.08
Top diagnosis	124 (31.6)	167 (42.6)	123 (31.4)	.002	.99	.001

[a]Chi-square test. The Bonferroni-corrected significance level at a P value <.02.

Figure 2. Diagnostic accuracy of Gemini Advanced, Gemini, and Bard. *P* values were derived from the chi-square test. The Bonferroni-corrected significance level at a *P* value <.02.



Most Frequently Named Differential Diagnoses

Regarding the top 10 most frequently named differential diagnoses, all rankings included sepsis, pneumonia, pulmonary embolism, lymphoma, and meningitis. Notably, the top 3 most frequently named differential diagnoses by Gemini Advanced and Gemini were the same. Table 4 shows the top 10 most frequently named differential diagnoses generated by Gemini Advanced, Gemini, and Bard.

Table 4. The top 10 most frequently named differential diagnoses were generated by Gemini Advanced, Gemini, and Bard.

The raking in the top 10 most frequently named differentials, (N)	Gemini Advanced (n)	Gemini (n)	Bard (n)
1	Sepsis (43)	Sepsis (42)	Sarcoidosis (51)
2	Pneumonia (34)	Pneumonia (28)	Sepsis (42)
3	Pulmonary embolism (33)	Pulmonary embolism (20)	Pneumonia (41)
4	Acute kidney injury (28)	Sarcoidosis (15)	Lymphoma (40)
5	Lymphoma (25)	Pericarditis (14)	Pulmonary embolism (39)
6	Urinary tract infection (24)	Meningitis (14)	Meningitis (31)
7	Heart failure (23)	Lymphoma (13)	Inflammatory bowel disease (29)
8	Meningitis (22)	Myocarditis (13)	Tuberculosis (26)
9	Myocardial infarction (20)	Acute kidney injury (12)	Encephalitis (25)
10	Pericarditis (18)	Systemic lupus erythematosus (12)	Myocarditis (24)

Inappropriate Diseases Names

From all differential diagnosis lists output by generative AIs, we identified inappropriate disease names: 11 items from Gemini Advanced, 9 items from Gemini, and 5 items from Bard. Notably, Gemini Advanced and Gemini both erroneously listed "Wegner's granulomatosis," a misspelling of the previous correct term, "Wegener's granulomatosis," which has now been updated to "Granulomatosis with Polyangiitis" [27]. Another error by Gemini Advanced involved "Microcytic colitis," likely a confusion between "microcystic anemia" and "microscopic colitis." Table 5 lists the inappropriate disease names generated by Gemini Advanced, Gemini, and Bard.

Table 5. Inappropriate disease name generated by Gemini Advanced, Gemini, and Bard.

Correct disease name, [cell number in Multimedia Appendix 2]	Inappropriate disease name by Gemini Advanced	Inappropriate disease name by Gemini	Inappropriate disease name by Bard
Drug reaction (Dasatinib)	Drug reation (Dasatinib) [O15]	—[a]	—
Nonketotic hyperglycemic hyperosmolar coma	—	Nonketotic hyperglycemia hyperosmolar coma [X29]	—
Lipedema	—	Lipoderma [U34]	—
Small bowel angiodysplasia	Small bowel angiodisplasia [M40]		—
Granulomatosis with polyangiitis	Granulomatous with polyangiitis (L68), Wegner's granulomatosis [N111]	Wegner's granulomatosis [U186]	—
Costochondritis	Costochondritisa [P86]	—	—
Maxillary sinus carcinoma	—	Maxillary sinus cycinoma [L106]	—
Constrictive pericarditis	Conrictive pericarditis [L110]	—	—
Scleroderma-related interstitial lung disease	—	Scleroderma-related interstitial lung disease [S117]	—
Pericoronitis	—	—	Pericoronatitis [AA133]
Osteitis	—	—	Osteoitis [AC133]
Microscopic colitis	Microcytic colitis [L152]	—	—
Pneumocystis jirovecii	—	—	Pneumocystis jerovecii [AG156]
Leukoencephalopathy	—	Leukoencephalomyopathy [X195]	—
Strumal carcinoid	—	Struma carcinoid [W208]	—
Restrictive ventilatory impairment	Restricted ventilatory impairment [J360]	—	—
Moebius syndrome	—	Mobius syndrome [Z369]	—
Endometriosis	—	—	Endometrios [AE385]
Cryptococcus neoformans	—	—	Chryptococcus neoformans [AJ389]
Unknown	Ytzinger hernia [M197]	Y-type appendicitis [V197]	—
Unknown	(There was partly Arabic language) [L292]	—	—
Unknown (Transaminase elevation is also not disease name)	Transaminitis elevation (N354)	—	—

[a]Not applicable.

Average Ranking

In terms of average ranking, the scores were 5.25 (SD 4.16) for Gemini Advanced, 4.54 (SD 4.21) for Gemini, and 5.33 (SD 4.29) for Bard. The differences in average rankings were not statistically significant between Gemini Advanced and Gemini (*P*=.99), between Gemini Advanced and Bard (*P*=.17), and between Gemini and Bard (*P*=.99).

Discussion

Principal Findings

In the following, we discuss our principal findings. Our findings indicate that Gemini demonstrated superior diagnostic accuracy compared with Bard, not only within the top 10 differential diagnosis lists but also in identifying the most likely diagnosis. Specifically, Gemini's diagnostic accuracy for the top 10 lists was 76.5% (300/392), compared to Bard's 68.6% (269/392), with a statistically significant difference (*P*=.02). Moreover, as the top diagnosis, Gemini's diagnostic accuracy was 42.6% (167/392) versus Bard's 31.4% (123/392), also significant (*P*=.001). This enhancement in Gemini's diagnostic performance may be attributed to its advanced algorithmic framework, which likely incorporates more nuanced medical data and learns from recent case inputs, leading to more refined diagnostic predictions.

However, the performance of Gemini did not statistically outperform in the top 5 differential diagnosis lists. This outcome may suggest that while Gemini's algorithm is effective in a broader exploratory context, its precision may falter when

constrained to a narrower list of top diagnoses. This indicates that a balance between breadth of exploration and depth of focus is crucial for optimizing diagnostic accuracy in such AI systems.

Conversely, our analysis showed that Gemini Advanced did not perform as well as expected when compared with Gemini. Despite expectations that the advanced model would provide enhanced diagnostic capabilities, it achieved lower accuracy in identifying the most probable diagnosis with 31.6% (124/392) compared to Gemini's 42.6% (167/392), with this result being statistically significant (P=.002). This outcome suggests that the additional features or complexity added in Gemini Advanced may not necessarily translate into improved diagnostic performance. These findings underscore the need for further refinement and optimization of Gemini Advanced to harness its potential for future AI-enhanced diagnostics.

In addition, our analysis identified issues with inappropriate disease naming in the outputs from GAI systems, with Gemini Advanced and Gemini producing outdated or misspelled terms for vasculitis, instead of using the updated name. These inaccuracies highlight the challenges in ensuring up-to-date and precise medical terminology in AI outputs, which is crucial for maintaining trust and reliability in AI-assisted diagnostics. Furthermore, these misspellings are often found in published medical articles, suggesting that GAIs may have learned these errors from these sources. The fact that both Gemini Advanced and Gemini exhibited the same mistakes indicates potential similarities in their underlying models or training data.

Regarding average rankings, there were no statistically significant differences among generative AI systems. This indicates a level of parity in how each model ranks diagnoses when they include the correct diagnosis, suggesting that while there are differences in overall accuracy, the ranking mechanisms of each model are relatively similar.

Given the current performance metrics, our analysis supports prioritizing the adjustment and enhancement of Gemini for future applications in medical diagnostics, rather than Gemini Advanced. Despite the theoretically superior capabilities of Gemini Advanced [17], Gemini's framework appears more aligned with practical diagnostic needs and shows greater promise in real-world applications. However, it is essential to verify this trend across a variety of sources to ensure that these findings are not specific to the data sets used in this study. Further investigations involving diverse clinical environments and different types of medical data are crucial to confirm the consistency and reliability of Gemini's superior performance.

Finally, the comparative analysis of the differentials by Gemini Advanced, Gemini, and Bard revealed consistent inclusion of sepsis, pneumonia, pulmonary embolism, lymphoma, and meningitis among their top 10 differentials. This underscores not only a shared prioritization of these conditions but also the effectiveness of systems in recognizing critical and prevalent diseases. The consistent identification of sepsis, particularly its second-place ranking by Bard, underscores the potential of these AI systems to enhance diagnostic accuracy and reduce errors in the identification of life-threatening conditions [28]. Importantly, the top 3 differentials by Gemini Advanced and Gemini—sepsis, pneumonia, and pulmonary embolism—are among the most harmful diseases where reducing diagnostic errors is crucial [1]. This suggests a potential for GAI systems to alert medical professionals about the inclusion of these important diseases during diagnosis. Such an understanding could facilitate more effective use of these GAI systems in future diagnostics processes.

Strengths

This study had several strengths. First, the strengths of this study lie in its direct comparison of 3 cutting-edge AI systems and its demonstration of the dynamic improvements in their diagnostic accuracy. Unlike some CDSSs like symptom checkers, whose performance has plateaued [29], these GAI systems evaluated in this research show considerable enhancements with each iteration. Second, we evaluated the diagnostic accuracy of GAI systems using a series of case reports. These case reports often describe rare diseases and atypical presentations, as opposed to common diseases and typical presentations [30]. This showcases the system's diagnostic capabilities under challenging conditions. Third, the comprehensive range of medical conditions covered by the differential diagnosis lists generated by the AI systems represents a significant strength of this study. This extensive coverage demonstrates the systems' capacity to handle a broad spectrum of medical knowledge and its applicability to various clinical scenarios.

Limitations

Several limitations should be discussed. First, the use of case report series might not fully reflect real-world clinical scenarios. This limitation arises because case reports typically focus on novel or rare aspects of diseases rather than typical presentations and common diseases [30]. Second, the exclusive use of a single case report journal could introduce selection bias. Third, there was no well-established method for evaluating AI diagnostics. In our study, we used binary evaluation methods. In contrast, other research on CDSSs used several rating methods [31,32] and the ranking averages in the differential diagnosis lists [33]. Fourth, we used only text data; excluding image data could influence the diagnostic performance. These factors limit the generalizability of these findings.

Concerning the GAI systems, all platforms used in this study were not designed for clinical use and have not received approval for medical diagnostics. These systems were not specifically reinforced or enhanced for medical diagnostic purposes. According to a preprint, Med-Gemini, a specialized model in medicine, was developed [34] but is not available to the public. In addition, we could not include all currently available GAI systems; thus, these findings cannot be generalized to other systems or different clinical scenarios. There was also a risk that these GAI systems may have learned from the published case reports used in this study.

Moreover, the use of user data to refine models, as seen in Gemini Advanced and Gemini, highlights significant privacy concerns [35]. Future research should address the development of locally deployable LLM solutions tailored specifically for CDSS. Although our data set is sourced from an open journal, careful consideration must be given to the ethical deployment of these models within health care settings. Finally, given the

rapid pace of GAI technology development, such as the evolution from Bard to Gemini and from ChatGPT-3 to ChatGPT-4 and ChatGPT-4o, our findings may have a limited shelf-life.

Future Direction

Future research will aim to explore the diagnostic accuracy of GAI systems following medical enhancements and adjustments. Once approved for medical use, it will also be essential to investigate the performance of GAI systems across various populations and settings, including remote medical consultations, to ensure their effectiveness in real-world diagnostics. Moreover, assessing the impact of AI-enhanced diagnostics on the decision-making process of medical professionals will be crucial.

In addition, future studies should focus on integrating GAI systems with existing electronic health record systems to understand how AI can augment data accessibility and analysis. This integration will be essential to evaluate how GAI can improve clinical workflows, reduce the cognitive burden and the time to diagnosis, and enhance patient outcomes.

Finally, the development of ethical guidelines and governance frameworks for the use of GAI in diagnostics is imperative [36]. As AI technologies become more prevalent in health care, it is crucial to establish clear protocols that safeguard patient privacy, ensure data security, and maintain transparency in AI decision-making processes.

Comparison With Previous Work

Our research builds on previous findings. We revealed that the diagnostic accuracy of ChatGPT-4 was 86.7% (340/392) for the final diagnoses included in the top 10 differential diagnosis lists, and 54.6% (214/392) for the top diagnosis [37]. ChatGPT-4's performances were still higher than that of Gemini in the lists (76.5% vs 86.7%) and as a top diagnosis (42.6% vs 54.6%); it was similar to Gemini Advanced in the lists (73.0% vs 86.7%) and as a top diagnosis (31.6% vs 54.6%).

Expanding our findings, another study showed that Isabel Pro, a successful CDSS developed by Isabel Healthcare, Ltd [38], correctly identified diagnoses in 87.1% (175/201) of cases, compared with 82.1% (165/201) for ChatGPT-4 in a series of clinical cases [33]. These findings are partly attributed to the earlier launch of Isabel Pro and the ChatGPT series, allowing them to receive more user feedback and undergo updates to improve performance.

In addition, another research focused on multiple choice questions on clinical vignettes revealed that ChatGPT-4 achieved a high accuracy rate of 73.3% for Clinical Challenges from the *Journal of the American Medical Association* (*JAMA*) and 88.7% for Image Challenges from the *New England Journal of Medicine* (*NEJM*). In contrast, Gemini, referred to as Gemini Pro in that study, achieved 63.6% for Clinical Challenges from *JAMA* and 68.7% for Image Challenges from the *NEJM* [22]. While these previous findings and current results revealed certain diagnostic performances of generative AI systems, comparing these results poses significant challenges due to methodological differences. Variations stem from differences in data set preparation, the types of clinical vignettes used, and the specific challenges or images included, which may influence performance outcomes. In addition, the evaluation criteria used to assess accuracy might differ significantly, affecting the comparability. For instance, the scoring systems or the definitions of a "correct" answer could vary, necessitating caution when drawing direct comparisons between these findings and those of this study.

In contrast to the serial evaluation approach of a symptom checker [29], which demonstrated an accuracy of 44.3% (97/219) in the first year and 47.7% (43/90) in the third year without significant difference, the performance of generative AI systems presents a different dynamic. Specifically, the serial evaluation of generative AI indicated that Gemini outperformed Bard over a relatively short period. This superiority can be attributed in part to the adaptability of generative AI systems to incorporate additional data. However, it is crucial to note that this adaptability does not consistently translate into improved diagnostic accuracy, as evidenced by the current comparison between Gemini Advanced and Bard. This observation highlights the nuanced interplay between technological advancement and clinical efficacy, underscoring the need for continued research and validation in integrating these systems into medical practice effectively.

Conclusions

The results of this study suggest that Gemini outperformed Bard in diagnostic accuracy following the model update. However, Gemini Advanced requires further refinement to optimize its performance for future AI-enhanced diagnostics. These findings should be interpreted cautiously and considered primarily for research purposes, as these GAI systems have not been adjusted for medical diagnostics nor approved for clinical use. The potential and limitations highlighted by this study underscore the need for ongoing development and evaluation of GAI systems within medical diagnostics.

Acknowledgments

This research was funded by JSPS KAKENHI (grant 22K10421). This study was conducted using resources from the Department of Diagnostics and Generalist Medicine at Dokkyo Medical University.

Authors' Contributions

TH, YH, KT, TI, T Suzuki, and T Shimizu contributed to the study of concept and design. TH performed the statistical analyses. TH contributed to the drafting of the manuscript. YH, KT, TI, T Suzuki, and T Shimizu contributed to the critical revision of the manuscript for relevant intellectual content. All the authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The PubMed search keywords.
[DOCX File , 19 KB - medinform_v12i1e63010_app1.docx]

Multimedia Appendix 2
The data set of differential diagnosis generated by artificial intelligence systems in this study, alongside the final diagnosis.
[XLSX File (Microsoft Excel File), 185 KB - medinform_v12i1e63010_app2.xlsx]

References

1. Newman-Toker DE, Nassery N, Schaffer AC, Yu-Moe CW, Clemens GD, Wang Z, et al. Burden of serious harms from diagnostic error in the USA. BMJ Qual Saf 2024;33(2):109-120. [doi: 10.1136/bmjqs-2021-014130] [Medline: 37460118]
2. Singh H, Connor DM, Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. BMJ 2022;376:e068044. [doi: 10.1136/bmj-2021-068044] [Medline: 35172968]
3. Balogh EP, Miller BT, Ball JR. Improving Diagnosis in Health Care. Washington DC: National Academies Press; 2015.
4. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
5. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
6. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
7. Sibbald M, Monteiro S, Sherbino J, LoGiudice A, Friedman C, Norman G. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of isabel. BMJ Qual Saf 2022;31(6):426-433 [FREE Full text] [doi: 10.1136/bmjqs-2021-013493] [Medline: 34611040]
8. van Baalen S, Boon M, Verhoef P. From clinical decision support to clinical reasoning support systems. J Eval Clin Pract 2021;27(3):520-528 [FREE Full text] [doi: 10.1111/jep.13541] [Medline: 33554432]
9. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. PLoS One 2016;11(3):e0148991 [FREE Full text] [doi: 10.1371/journal.pone.0148991] [Medline: 26954234]
10. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res 2022;24(5):e31810 [FREE Full text] [doi: 10.2196/31810] [Medline: 35536633]
11. Castaneda C, Nalley K, Mannion C, Bhattacharyya P, Blake P, Pecora A, et al. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 2015;5(1):4 [FREE Full text] [doi: 10.1186/s13336-015-0019-3] [Medline: 25834725]
12. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. JMIR Med Inform 2024;12:e52073 [FREE Full text] [doi: 10.2196/52073] [Medline: 38506918]
13. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJPC. Generative AI for transformative healthcare: a comprehensive study of emerging models, applications, case studies, and limitations. IEEE Access 2024;12:31078-31106. [doi: 10.1109/access.2024.3367715]
14. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. Computer Science Computation and Language 2023 [FREE Full text]
15. Karampinis E, Toli O, Georgopoulou KE, Kampra E, Spyridonidou C, Roussaki Schulze AV, et al. Can artificial intelligence "Hold" a dermoscope?-The evaluation of an artificial intelligence chatbot to translate the dermoscopic language. Diagnostics (Basel) 2024;14(11):1165 [FREE Full text] [doi: 10.3390/diagnostics14111165] [Medline: 38893694]
16. Pichai S, Hassabis D. Introducing Gemini: our largest and most capable AI model. Google. 2023. URL: https://blog.google/technology/ai/google-gemini-ai/ [accessed 2023-12-06]

17. Team G, Anil R, Borgeaud S, Wu Y, Alayrac J, Yu J, et al. Gemini: a family of highly capable multimodal models. Computer Science Computation and Language 2023 [FREE Full text]
18. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023;93(5):1090-1098. [doi: 10.1227/neu.0000000000002551]
19. Doshi RH, Amin K, Khosla P, Bajaj S, Chheang S, Forman H. Utilizing large language models to simplify radiology reports: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, Google bard, and microsoft bing. medRxiv 2023:23290786. [doi: 10.1101/2023.06.04.23290786]
20. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the efficacy of large language models in health literacy: a comprehensive cross-sectional study. Yale J Biol Med 2024;97(1):17-27 [FREE Full text] [doi: 10.59249/ZTOZ1966] [Medline: 38559461]
21. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative evaluation of diagnostic accuracy between Google bard and physicians. Am J Med 2023;136(11):1119-1123.e18. [doi: 10.1016/j.amjmed.2023.08.003] [Medline: 37643659]
22. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative analysis of multimodal large language model performance on clinical vignette questions. JAMA 2024;331(15):1320-1321. [doi: 10.1001/jama.2023.27861] [Medline: 38497956]
23. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008;23 Suppl 1(Suppl 1):37-40 [FREE Full text] [doi: 10.1007/s11606-007-0271-8] [Medline: 18095042]
24. Shahrudin MS, Mohamed-Yassin MS, Nik Mohd Nasir NM. Herpes zoster following COVID-19 vaccine booster. Am J Case Rep 2023;24:e938667 [FREE Full text] [doi: 10.12659/AJCR.938667] [Medline: 36650730]
25. Krupat E, Wormwood J, Schwartzstein RM, Richards JB. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. Med Educ 2017;51(11):1127-1137. [doi: 10.1111/medu.13382] [Medline: 28857266]
26. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. New York: John Wiley & Sons; 2003.
27. Falk RJ, Gross WL, Guillevin L, Hoffman G, Jayne DRW, Jennette JC, et al. Granulomatosis with polyangiitis (Wegener's): an alternative name for Wegener's granulomatosis. Ann Rheum Dis 2011;70(4):704. [doi: 10.1136/ard.2011.150714] [Medline: 21372195]
28. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, CDC Prevention Epicenter Program. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. JAMA 2017;318(13):1241-1249 [FREE Full text] [doi: 10.1001/jama.2017.13836] [Medline: 28903154]
29. Harada Y, Sakamoto T, Sugimoto S, Shimizu T. Longitudinal changes in diagnostic accuracy of a differential diagnosis list developed by an AI-Based symptom checker: retrospective observational study. JMIR Form Res 2024;8:e53985 [FREE Full text] [doi: 10.2196/53985] [Medline: 38758588]
30. Riley DS, Barber MS, Kienle GS, Aronson JK, von Schoen-Angerer T, Tugwell P, et al. CARE guidelines for case reports: explanation and elaboration document. J Clin Epidemiol 2017;89:218-235. [doi: 10.1016/j.jclinepi.2017.04.026] [Medline: 28529185]
31. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330(1):78-80 [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
32. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med 2012;27(2):213-219 [FREE Full text] [doi: 10.1007/s11606-011-1804-8] [Medline: 21789717]
33. Bridges JM. Computerized diagnostic decision support systems - a comparative performance study of isabel pro vs. ChatGPT4. Diagnosis (Berl) 2024;11(3):250-258 [FREE Full text] [doi: 10.1515/dx-2024-0033] [Medline: 38709491]
34. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of gemini models in medicine. Computer Science Artificial Intelligence 2024 [FREE Full text]
35. Gemini apps privacy notice. URL: https://support.google.com/gemini/answer/13594961#privacy_notice [accessed 2024-05-29]
36. Newman-Toker DE, Sharfstein JM. The role for policy in AI-Assisted medical diagnosis. JAMA Health Forum 2024;5(4):e241339 [FREE Full text] [doi: 10.1001/jamahealthforum.2024.1339] [Medline: 38635262]
37. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Diagnostic performance of generative artificial intelligences for a series of complex case reports. DIGITAL HEALTH 2024;10. [doi: 10.1177/20552076241265215] [Medline: 39229463]
38. Ren LY. Isabel pro. J Can Health Libr Assoc 2019;40(2):63-69. [doi: 10.29173/jchla29418]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
GAI: generative artificial intelligence
JAMA: Journal of the American Medical Association
LaMDA: language model for dialogue applications

LLM: large language model
NEJM: New England Journal of Medicine

Edited by C Lovis; submitted 07.06.24; peer-reviewed by E Karampinis, X Ai; comments to author 22.07.24; revised version received 29.07.24; accepted 06.08.24; published 02.10.24.

Please cite as:
Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T
Comparative Study to Evaluate the Accuracy of Differential Diagnosis Lists Generated by Gemini Advanced, Gemini, and Bard for a Case Report Series Analysis: Cross-Sectional Study
JMIR Med Inform 2024;12:e63010
URL: https://medinform.jmir.org/2024/1/e63010
doi:10.2196/63010
PMID:

Review

Health Care Language Models and Their Fine-Tuning for Information Extraction: Scoping Review

Miguel Nunes[1], BSc; Joao Bone[2], MSc; Joao C Ferreira[1,3,4], PhD; Luis B Elvas[1,3,4], MSc

[1]ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal

[2]Select Data, Anaheim, CA, United States

[3]Department of Logistics, Molde, University College, Molde, Norway

[4]INOV Inesc Inovação, Instituto de Novas Tecnologias, Lisbon, Portugal

Corresponding Author:
Luis B Elvas, MSc
Department of Logistics, Molde, University College
Britvegen 2, Noruega
Molde, 6410
Norway
Phone: 47 969152334
Email: luis.m.elvas@himolde.no

Abstract

Background: In response to the intricate language, specialized terminology outside everyday life, and the frequent presence of abbreviations and acronyms inherent in health care text data, domain adaptation techniques have emerged as crucial to transformer-based models. This refinement in the knowledge of the language models (LMs) allows for a better understanding of the medical textual data, which results in an improvement in medical downstream tasks, such as information extraction (IE). We have identified a gap in the literature regarding health care LMs. Therefore, this study presents a scoping literature review investigating domain adaptation methods for transformers in health care, differentiating between English and non-English languages, focusing on Portuguese. Most specifically, we investigated the development of health care LMs, with the aim of comparing Portuguese with other more developed languages to guide the path of a non–English-language with fewer resources.

Objective: This study aimed to research health care IE models, regardless of language, to understand the efficacy of transformers and what are the medical entities most commonly extracted.

Methods: This scoping review was conducted using the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) methodology on Scopus and Web of Science Core Collection databases. Only studies that mentioned the creation of health care LMs or health care IE models were included, while large language models (LLMs) were excluded. The latest were not included since we wanted to research LMs and not LLMs, which are architecturally different and have distinct purposes.

Results: Our search query retrieved 137 studies, 60 of which met the inclusion criteria, and none of them were systematic literature reviews. English and Chinese are the languages with the most health care LMs developed. These languages already have disease-specific LMs, while others only have general–health care LMs. European Portuguese does not have any public health care LM and should take examples from other languages to develop, first, general-health care LMs and then, in an advanced phase, disease-specific LMs. Regarding IE models, transformers were the most commonly used method, and named entity recognition was the most popular topic, with only a few studies mentioning Assertion Status or addressing medical lexical problems. The most extracted entities were diagnosis, posology, and symptoms.

Conclusions: The findings indicate that domain adaptation is beneficial, achieving better results in downstream tasks. Our analysis allowed us to understand that the use of transformers is more developed for the English and Chinese languages. European Portuguese lacks relevant studies and should draw examples from other non-English languages to develop these models and drive progress in AI. Health care professionals could benefit from highlighting medically relevant information and optimizing the reading of the textual data, or this information could be used to create patient medical timelines, allowing for profiling.

(JMIR Med Inform 2024;12:e60164) doi:10.2196/60164

KEYWORDS

language model; information extraction; healthcare; PRISMA-ScR; scoping literature review; transformers; natural language processing; European Portuguese

Introduction

The health care sector generates a vast amount of structured and unstructured data, including images from medical exams, text written in electronic medical records (EMRs) or Electronic Health Records (EHRs), and structured data from relational databases that store patient and admission information, as well as all the data collected during a patient's hospitalization [1]. Approximately 30% of the world's data volume is generated by the health care sector, and projections indicate that by 2025, the compound annual growth rate of data for health care will reach 36% [2].

Medical texts present several challenges due to the use of unfamiliar context-specific terminologies that differ from everyday language. In addition, physicians often use abbreviations and acronyms to save time and space. However, the same abbreviation can have different meanings, adding an additional layer of complexity when trying to understand the content of medical texts [3]. All these characteristics pose challenges when attempting to apply artificial intelligence (AI) techniques to interpret the text.

In the field of natural language processing (NLP), the introduction of transformers [4] has revolutionized the field, achieving state-of-the-art performance for numerous NLP tasks [5]. Their general architecture comprises an encoder, which receives the input and builds a representation of it, and a decoder that uses the encoder's representation along with other inputs to generate a target sequence. The introduction of the self-attention mechanism further revolutionized NLP by allowing the model to weigh the importance of different words in a sentence regardless of their position. This enables better handling of long-range dependencies compared with traditional deep learning (DL) architectures like recurrent neural networks (RNNs) and long short-term memory Networks [6]. In the context of medical text, transformers excel in interpreting and extracting medically relevant information by effectively handling context and meaning, even in complex and specialized language.

Transformers can be trained as language models (LMs) on raw text in a self-supervised manner, enabling them to develop a statistical understanding of the text they were trained on [7]. However, the benefits of this approach are only fully realized when fine-tuning a downstream task.

Another important concept is called domain adaptation, which stands for the process of adapting or adjusting something to be suitable within a different domain or context. In the field of machine learning (ML), domain adaptation is used to align the disparity between domains so that the trained model can generalize into the domain of interest [8]. For transformers, domain adaptation involves continuing the pretraining of an LM with text data from a different domain than the one it was originally trained on [9]. This approach allows for leveraging the learning capabilities of general-scope LMs and refining them for specific contexts. For example, if we consider a general-scope LM, one that was trained using textual data from various domains, and continue its pretraining with health care–specific textual data, it will help the LM to refine its understanding of the health care data, leading to improvements when fine-tuning the LM for downstream tasks related to health care. To explore this further, we can take a health care LM who was trained using EMRs from a hospital and continue its pretraining using only text from patients with a specific disease. It will allow the LM to adjust its weights and become more precise when interpreting texts related to that particular disease.

An example of domain adaptation is the BioBERT model [10], which resulted from the continuation of the pretraining of the Bidirectional Encoder Representations from Transformers (BERT) [10] model on biomedical text. The BioBERT model outperformed its predecessor in biomedical named entity recognition (NER), relation extraction, and question-answering tasks. Alzheimer's Disease-BERT [11] and CancerBERT [12] are 2 examples of applying domain adaptation to a more restricted domain. Both models outperformed their respective baselines on downstream tasks related to their respective diseases. Summing up, performing domain adaptation for the health care sector appears inevitable to improve results, for example, for information extraction (IE) models, where a better understanding of medical terminologies and lexicon would make it easier to identify and extract information [13].

The European Portuguese (PT-PT) language does not generate the same amount of data as the English language, resulting in limitations in the literature and the published models. A study published in 2023 by the Ernst & Young Audit highlights the following areas where AI can play a relevant role in Portugal's health care; disease diagnoses, precision medicine, remote monitoring and prevention, data management and hospital efficiency, and health policies [14]. Recently, a project was launched in Portugal, funded by the European Union, with the aim of creating PT-PT NLP solutions for the health care sector. Under this scope, the objective is to create PT-PT medical LM and IE models to automatically identify medically relevant entities.

Therefore, in this study, we aim to present a scoping literature review (SLR), in which we will begin by exploring the creation of health care LMs through domain adaptation and analyze their results. In addition, we aim to focus on the geographical domain to understand the current state-of-the-art for the Portuguese language and compare it to other, potentially more developed, languages to identify further steps. We also want to explore IE models in the health care sector, regardless of their data language, to understand the most commonly extracted medical entities and the methods used in doing so. Despite the literature being rich in studies focused on health care large language models (LLMs), there is a lack of studies that evaluate the current state-of-the-art of health care LMs not only in English but also in other less-resourced languages. This will enable us

to grasp how the community is using the capabilities of transformers and whether the advantages of using them are indeed present in the health care domain. In addition, researchers will have 1 study about health care LMs that could guide their path and help them understand how the literature has developed in their respective languages. Finally, we will present the corresponding discussion and the conclusion drawn from the SLR.

Methods

Overview

To complete our goal, we have conducted an SLR to gain a better understanding of the research conducted in the application of health care–domain LMs and the development of IE models within the health care domain. In the first stage, our study encompasses health care–domain LMs in various languages, with a focus on the Portuguese language. In the second stage, we searched for studies related to IE models to evaluate the methods most frequently used. In terms of methodology, we followed the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) [15] to ensure a systematic and transparent approach in conducting and reporting our scoping review.

Search Strategy and Inclusion Criteria

This SLR was conducted in November 2023 and focused exclusively on studies and reviews published in journals within the last 5 years (2019-2023) that were written in English or Portuguese. It was not an arbitrary data range since limiting the search to the last 5 years ensures that the review includes the most recent and relevant studies, reflecting the latest advancements, technologies, and methodologies in the field. Since BERT [10], one of the most popular transformer architectures, and LM were launched at the end of 2018, we searched for studies from 2019 onward. The primary databases used for this review were Scopus [16] and the Web of Science Core Collection (WOSCC) [17] since both databases are renowned for indexing a wide array of peer-reviewed journals across multidisciplinary fields [18,19]. While acknowledging that additional databases might offer further insights, the significant overlap with these resources ensures that relevant studies are unlikely to be missed.

The criteria were defined to include studies focused on continuing the pretraining of LMs to achieve health care–domain LMs or studies focused on creating IE models within the health care field. Therefore, we formulated a query that includes the training or fine-tuning LMs or IE Models within the context of health care or similar, using EMRs or EHRs as data.

Since there is a significant semantic similarity between LMs and LLMs, we decided to exclude the second from the search query because it has a different purpose from the aim of our study. LLMs are typically composed of more than 7B parameters and are suited for text generation. LMs are models that are not, by themselves, suited to perform any downstream NLP task, needing to be readjusted or fine-tuned with labeled data to be able to perform downstream tasks.

Our final query is as follows: "("Language Model" OR "Masked Language Model" OR "Information Extraction" OR "Content Extraction") AND ("EHR" OR "EMR" OR "Electronic Health Record" OR "Electronic Medical Record") AND ("Fine-Tuned" OR "Fine-tuning" OR "Training" OR "Trained") AND ("Healthcare" OR "Health Care" OR "Clinical" OR "Medical") AND NOT ("Large Language Model" OR "LLM")."

According to our objectives, a study was considered valid if it documented a continuation of the pretraining of an LM to create a health care LM or if it focused on the creation of health care IE models.

Study Selection

To minimize the risk of bias in the study selection, the process was conducted independently by 3 researchers. A total of 2 researchers were responsible for reading and judging the studies according to the inclusion criteria, while the third researcher was involved in cases of disagreement.

Data Charting and Synthesis

A data-charting form was jointly developed by two reviewers to extract relevant information from the selected studies systematically. The form included variables such as study title, year of publication, language focus (English or non-English), domain adaptation techniques for Transformer-based models, healthcare-specific information extraction tasks, evaluation metrics used, and the specific health-related entities being extracted. Both reviewers independently charted the data to ensure comprehensive coverage of healthcare language models in English and non-English languages, with particular attention to languages other than English (referred to as non-English). Discrepancies in the extracted data were discussed and resolved through consensus. As the review progressed, the data-charting form was iteratively updated to capture emerging themes, especially regarding the disparity between language resources and technological development for healthcare information extraction across different languages.

Results

The query retrieved 137 papers, with the vast majority of these studies being retrieved from Scopus, adding up to 90 when compared with the 47 studies WOSCC has yielded. The PRISMA-ScR methodology was then followed, as seen in Figure 1. Since we included studies from sources beyond the 2 selected databases, we adhered to the updated PRISMA-ScR guideline [20]. In the following subsections, we explained the decision to include studies by other methods.

The first step was to identify and remove duplicated papers, resulting in 101 studies. Following a screening of titles and abstracts, 10 records were deemed out of scope, and 1 could not be retrieved, leaving us with 90 fully reviewed studies.

After screening all the papers that matched our criteria, we realized that 30 of them did not meet our inclusion criteria. Some studies referred to the fine-tuning of pretrained LMs for tasks unrelated to IE, or they lacked relevant information to contribute to this study, or even though we excluded them from our search query, they mentioned the use of LLMs.

As we were focused on the Portuguese language, our study also emphasized the geographical domain, with an aim to comprehend the medical data language used in health care LMs. Table 1 resumes the studies focused on the pretraining of LMs separated by the language of their data.

From the reading of Table 1, we can understand that English is the main language, which can be explained by the much higher availability of English data and the overwhelming presence and applicability of this language throughout the world. However, we notice that in the Chinese language, there are studies attempting to fill the gap of being non-English, creating in-domain LMs aware of their benefits. We also found studies in Brazilian Portuguese (PT-BR), Spanish, and PT-PT, and we acknowledge that there might be other studies in different languages, even though they did not match our search query criteria.

Changing the view for the health care IE studies, Figure 2 resumes the distribution of studies by topic.

From the reading of Figure 2, NER appears as the main topic on the IE, with only 2 studies performing Assertion Status and 3 studies focused on solving medical lexical problems.

To provide a more in-depth review of each study, we present the subsequent 3 subsections where we differentiate between non-Portuguese health care LMs, Portuguese health care LMs, and health care IE models.

Figure 1. PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) workflow diagram. WOSCC: Web of Science Core Collection.

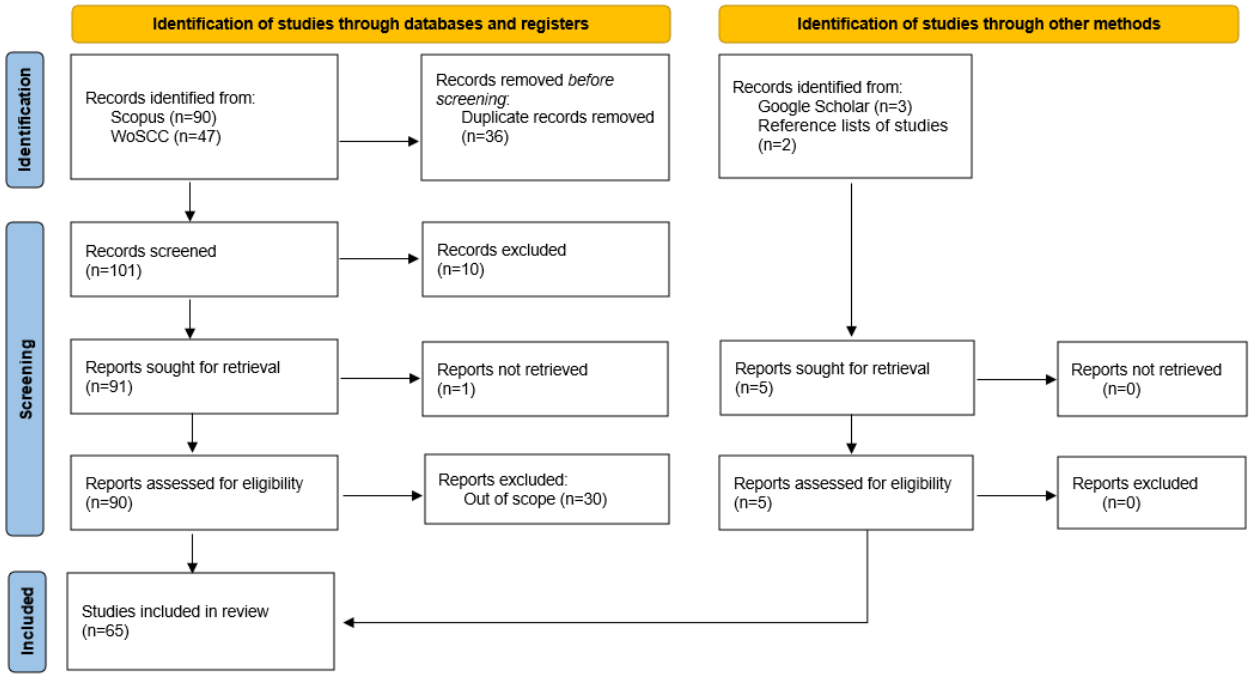
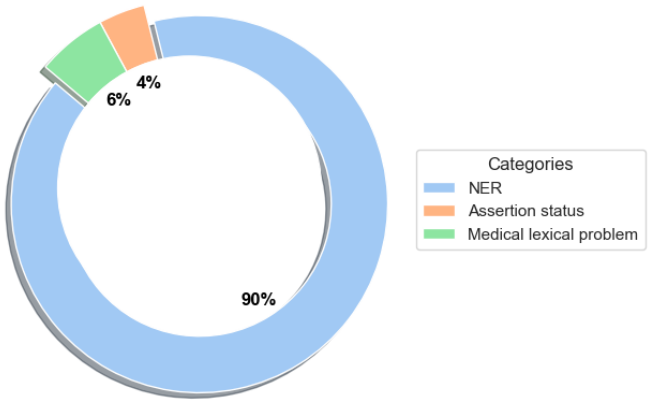


Table 1. Studies for pretraining language models (LMs) were reviewed by their data language.

Medical data language	Reference	Studies, n
English	[11,12,21-25]	7
Chinese	[26-29]	4
Brazilian Portuguese	[30,31]	2
Spanish	[32]	1
European Portuguese	[33]	1

Figure 2. Distribution of health care information extraction (IE) studies by topic. NER: named entity recognition.



Non-Portuguese Health Care Language Models

A study by Zhou et al [12] introduces CancerBERT, a domain-specific LM, that resulted from continuing the pretraining of the BlueBERT model [34] with a cancer corpus, resulting in various checkpoints of CancerBERT. The evaluation was performed for the cancer phenotyping NER task, with the results showing that the CancerBERT model pretrained with the cancer corpus outperformed the checkpoint using the original BERT [10] vocabulary.

A similar approach was conducted in a study by Mao et al [11], where the objective was to predict the risk of disease progression from Mild Cognitive Impairment to Alzheimer disease. A BERT model specifically tailored for Alzheimer disease (ie, AD-BERT) was pretrained with clinical notes, and its comparison with other models in experiments showed the benefits of domain adaptation.

Within the same scope, the identification of fall incidents from EHRs is discussed in the study by Fu et al [21]. A context-aware LM, BERT-based, was trained and integrated into a hybrid architecture along with post hoc heuristic rules. The performance of the BERT-based model was compared with DL methods, and the conclusions highlighted that the BERT model achieved superior results in identifying fall events.

In a study by Wang et al [26], a Chinese medical text corpus was used to pretrain BERT and obtain the Chinese BERT model. The results were aligned with previous studies, and domain adaptation demonstrated better results than traditional DL models and other pretrained LMs.

Studies by Roitero et al [22] and Agnikula Kshatriya et al [23] once again mention the pretraining of BERT models on a medical corpus, achieving comparable or better performance than state-of-the-art models. In a study by Zhang et al [24], an unsupervised adversarial domain adaptation framework with a pretrained LM for clinical event sequences is presented. Another example can be found in a study by Chen et al [25], where a contextual LM is used in combination with rule-based preprocessing methods to develop a model for *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) multilabel classification. The results demonstrate superiority over state-of-the-art models. Studies by Wen G et al and Wen C et al [27,28] refer to the training of a domain-specific pretrained LM on unlabeled medical data, with the evaluation being made through NER.

In the process of reviewing all the papers, we observed references to papers that aligned with our requirements despite not being explicitly included within our search query criteria. It is the case of studies by Zhang et al [29] and Carrino et al [32]. Carrino et al [32] present a large-scale biomedical Spanish LM, where the models were pretrained from scratch, using a RoBERTa [35] base model, and then fine-tuned on 3 clinical NER tasks. The comparison between general-domain and other available Spanish clinical models revealed the superiority of the models presented in the paper. Zhang et al [29] share a similar scenario with BERT being pretrained on Chinese biomedical corpora, and MC-BERT, an in-domain LM, was developed. The results are consistent with previous studies, with MC-BERT outperforming BERT-based models in all evaluated tasks.

Portuguese Health Care Language Models

Our search query did not retrieve any studies for the Portuguese language. To address this scarcity of studies and since it is one of the objectives of this research, we carried out a broader search on Google Scholar [36] to include studies that mentioned the creation of Portuguese health care LMs.

The PT-BR language has already presented several studies, with BioBERTpt [30] being one example. The authors used clinical notes and biomedical abstracts to pretrain 3 BERT-based checkpoints that were fine-tuned for the NER task to assess their performance. The results align with others, showing that the in-domain models achieved better performance. Another example is the study by Schneider et al [31], where several clinical BERT-based checkpoints were developed resulting from the continuation of the pretraining of BERTimbau [35], mBERT [11], and all 3 BioBERTpt checkpoints on 150,000 clinical narratives from cardiology ambulatory. The results of fine-tuning for NER align with previous studies, demonstrating that the in-domain models outperformed general LMs.

For the PT-PT language, we found the literature to be scarcer, with only 1 study being found that mentioned the continuation of the pretraining of an LM to achieve health care–domain LM. Coutinho and Martins [33] propose a BERT-based model for assigning *ICD-10* codes to causes of death by using BERTimbau and continuing its pretraining with death certificates. The

evaluation was made through NER, with all the checkpoints involved being fine-tuned for the classification task, and the findings indicated that transformer models could produce promising outcomes for health care tasks involving the analysis of relatively short documents.

Health Care IE Models

To better organize this section, we decided to categorize the studies by topic. Therefore, the first subsection presents NER studies where the authors attempted to automatically identify and extract medical information. The second subsection contains Assertion Status models, where entities are classified according to their status (present or absent), and finally, the third subsection presents studies that attempt to solve medical lexical problems.

NER

Zhou et al [37] evaluated the performance of CancerBERT along with ML models for the breast cancer phenotype extraction task, with the results proving that CancerBERT has superior learning ability and generalizability for this task. Rahman et al [38] refer to the use of BERT to identify the presence of a diagnosis in EHRs. With BERT's ability to understand the context of text and based on conditions presented in EHRs, a pipeline was successfully designed to identify EHRs with the presence of a diagnosis, reducing the manual note review load. Crema et al [39], use an Italian biomedical BERT model, fine-tuning it for the NER task with the entities of interest, including diagnoses, symptoms, drugs, and medical assessments, achieving an F_1-score of nearly 0.85 values. Entity-BERT was introduced in the study by Lu et al [40], a DL-based model for entity IE that is capable of recognizing entities such as medical terminologies, disease names, or drug information. Zhang et al [41] propose the combination of data augmentation and domain information using the Adapter Transformer Encoder Model for Clinical Event Detection. It uses the BioBERT model to generate word-level features, addressing the issue of many obscure professional terms in EMRs leading to poor recognition performance of the model. The results were reported to be superior to those of other existing models. A multilingual transformer was fine-tuned in a study by Kim et al [42], where researchers successfully extracted alcohol-related information from unstructured clinical notes with an extraction accuracy of 0.84 as measured by the macro F_1-score. Kormilitzin et al [43], initially pretrained a model on the task of predicting the next word and subsequently fine-tuned it for the NER task, extracting various categories of drugs and achieving performance with an F_1-score above 0.95 values. Solarte-Pabón et al [44] evaluate the fine-tuning of several pretrained LMs for the NER task, aiming to identify breast cancer concepts in the Spanish language. The results show that BERT-based and RoBERTa-based LMs exhibit competitive performance on this task. Liu et al [45] propose the use of BERT-BiLSTM-CRF for the NER task of rheumatoid arthritis vocabulary and then MC-BERT for the entity extraction task, with results showing F_1-scores above 90%. Wang et al [46] compare the use of 4 pretrained transformer-based LMs fine-tuned for the NER task with a baseline regular expression model in order to extract ophthalmic examination components, demonstrating that transformers achieve superior results. In the study by Singh et al [47], a pretrained transformer-based LM was fine-tuned with cardiac magnetic resonance imaging annotations to effectively extract measurements from clinical reports, and it achieved high extraction performance without requiring heuristics or expert annotations.

Several studies focus on extracting information about family history, such as studies by Kim et al [48], He et al [49], Silva et al [50], Dai et al [51], and Zhan et al [52]. They use ML methods, incorporating rule-based approaches, multitask-based artificial neural networks (ANN), attention-based neural networks, and even combinations such as convolutional neural networks (CNNs) BiLSTM and BERT. The goal was to automatically extract entities such as people's names, residence, birth date, or death date, and in some cases, there is an additional subtask related to relation extraction, which involves identifying relationships between family members. Overall, the results have proven to be satisfactory, particularly in the NER task.

CNNs are highly popular methods in the scientific community for extracting clinical information and studies by Yang et al [53], Santus et al [54], Mahajan and Rana [55], and Landlosi et al [56] primarily used them, often supplemented with rule-based approaches or feature optimization in some cases. The use of these methods lies in extracting clinical information from EHRs, tasks that could be time-consuming if done manually. Within the broader category of neural networks, RNNs are also a method used for IE in which the authors of studies [57-66] all use RNNs, with BiLSTM-CRF (Bidirectional Long Short-Term Memory - Conditional Random Field) being a very popular network among these studies. The main topics extracted include terms related to specific diseases, drug names with associated attributes (dosage, frequency, duration, route, and condition), adverse drug events, the presence of a diagnosis, or even important information in medical image reports, with the results globally proven to be promising.

Studies [67-69] use ML methods, with the first focusing on automatically classifying the outcomes of specific tasks related to the clinical conditions of stroke survivors, the second aiming to extract useful information in abdominopelvic radiology reports, and the third one focused on extracting travel history mentions from clinical documents. In Malmasi et al [70], the use of different methods to extract low-prevalence concepts is discussed, specifically in the case of insulin rejection by patients with attempts at both sentence-level and token-level approaches using ML and DL methods, but the results showed that it is challenging to automatically identify low-prevalence concepts. Similar proposals have been presented in studies [71-79] using spaCy's [80] pipeline for IE, contextual embeddings such as embeddings from language models (ELMo) [81] and BERT, position-attention mechanisms, knowledge graph embeddings, word segmentation models, or even NLP models developed using Java for extracting medical information, for example, extracting details related to drugs, drug attributes, or diagnoses.

In Lee and Uppal [82], a web-based summarization and visualization tool is introduced for extracting salient information from clinical and biomedical text, featuring sentence ranking by relevance and facilitating early medical risk detection in

clinical settings. Chen et al [83] aimed to create a model to extract concept embeddings from EHRs for disease pattern retrieval and subsequent classification tasks.

Assertion Status

Sykes et al [84] address the issue of negation and non-negation of clinical terms in EHRs. It is an Assertion Status case, in which the text can be characterized by cases where diseases are stated to be absent or only hypothesized. In this study, they propose various methods to address this issue, including rule-based, ML, or DL approaches, and all proposals yielded good results in a test set, achieving an F_1-score of more than 0.95. In Chaturvedi et al [85], a corpus annotated with mentions of pain was developed, considering the presence or absence of pain. It is another example of an Assertion Status problem aimed at facilitating further studies using the corpus to better understand how pain is mentioned in clinical notes.

Medical Lexical Problems

From a different perspective, there have been studies focusing on medical lexical problems. Newman-Griffis et al [86] discuss the presence of ambiguous words and attempts to normalize medical concepts to standardize vocabularies, while the study by Jaber et al [87] addresses the problem of the frequent use of abbreviations by proposing a method, by fine-tuning a pretrained LM, to successfully disambiguate clinical abbreviations. Lee et al [88] propose a typographical error correction model that considers context, based on a masked LM, to address the issue of typographical errors in real-world medical data. They conclude that typographical errors in unstructured text negatively impact the performance of NLP tasks, and their method is robust and applicable in real-world environments.

Discussion

Principal Findings

Continuing the pretraining of LMs to develop health care LMs has proven beneficial. The most common method to evaluate this approach is by fine-tuning both the baseline and the in-domain LM on downstream NLP tasks and comparing the results.

In IE models, NER is the most popular topic aimed at automatically identifying and extracting medically relevant information. Transformers are the preferred technology for this purpose, with fine-tuning of medical LMs consistently achieving superior results.

To conclude our SLR, we engaged in a deeper discussion divided into health care LMs and health care IE models.

Health Care Language Models

On a global scale, we have identified numerous studies that continued the pretraining of LMs to develop domain-specific LM, specifically medical LMs. In general, the findings across almost all of these studies substantiate the advantages of in-domain training before undertaking any other downstream tasks. The favorite evaluation task is NER, with almost every study mentioning the fine-tuning of LMs for the NER task.

As shown in Table 1, English and Chinese are the languages with the most studies and published models due to the available resources in terms of data and hardware power. The level of domain adaptation for these languages is more advanced, with dedicated health care LMs developed for specific diseases such as Alzheimer Disease-BERT [11] and CancerBERT [12], which represent very focused domains. These studies offer advantages by achieving better performance in extracting specific concepts from textual data related to these diseases compared with general health care LMs.

For non-English languages, the process is not so developed, which can be considered as expected since they have their known limitations, such as the scarcity of data and resources available. Nevertheless, there have been concerted efforts to create general health care LMs, underscoring the community's recognition of the use of these models. The Portuguese language fits this context, and despite initial strides that have already been taken, there exists ample room for improvement, particularly in the context of PT-PT where the only published study is [33], yet, to the best of our knowledge, the model is not publicly available.

Non-English languages, particularly Portuguese, should draw inspiration from advancements and results in medical domain adaptation studies. Despite limited resources and available data, efforts should first focus on creating general medical LMs. In a subsequent phase, efforts should be directed toward narrowing down to specific diseases while performing domain adaptation. This approach ensures that knowledge previously acquired by the LMs is refined within the medical domain and then adapted to smaller medical domains without losing the previously acquired knowledge completely. This initiative aims to foster the development of AI technologies in Portuguese, thereby promoting health care equality and access in languages with fewer resources. These models can be further fine-tuned for medical NLP tasks, such as IE, aimed at automatically identifying or highlighting specific information or structuring medical information extracted from textual data for ML analysis to aid health care professionals.

Health Care IE Models

Several methods have been used to create health care IE models. The most common method is the use of transformers, followed by the application of other DL and ML methods (Table 2). As previously discussed, the most popular topic was NER, where authors attempted to identify and extract medically relevant information.

The results indicate that the most successful approach involves using pretrained LMs fine-tuned for IE tasks, benefiting from the contextual understanding of the text to achieve better results. The most commonly identified entities were diagnoses or diseases and drugs, along with specific phenotypes related to certain diseases.

Table 2. Number of studies used per method.

Methods	Articles
Transformers	16
Other DL[a]	15
RNN[b]	10
Other ML[c]	9
CNN[d]	5
Rule-based	4

[a]DL: deep learning.

[b]RNN: recurrent neural network.

[c]ML: machine learning.

[d]CNN: convolutional neural network.

It is also relevant to mention that in our query, 2 studies were focused on Assertion Status. This task involves classification at the sentence level aimed at assessing an entity based on its presence or absence in the text. Examples of absence include negation or hypothesizing medical information. From another perspective, to address the problems presented by medical text, we also found 2 studies that propose solutions to disambiguate the multitude of abbreviations present in medical text and 1 study that presents a typo correction model. Both solutions aimed to improve text quality and seek to correct issues in the text that are considered inevitable by health care professionals. These 5 studies could also be seen as an improvement to NER results. The ones focused on correcting the text could be viewed as a preprocessing step that would enhance the understanding of the medical text, while the Assertion Status studies could help ascertain whether an identified entity is present or absent in a patient's condition. When compared with the distribution of NER, these 2 topics lack development, as together, they account for only 10% of the health care IE studies found. The community would benefit from more studies using different technologies and identifying new challenges to be solved.

Conclusions

Our SLR highlights the benefits of in-domain training for health care LMs and the effectiveness of transformers in IE tasks, addressing a research gap regarding the lack of studies on health care LMs. Transformers excel in NER, identifying diagnoses, diseases, drugs, and phenotypes. English and Chinese lead in research and LM development, while non-English languages such as Portuguese show promise but need exploration. Challenges include Assertion Status and text disambiguation, necessitating diverse methodologies and research in health care IE.

We have identified several health care–domain LMs, but there is a clear gap for non-English languages where the data and resources available are low. There is much to improve in those languages, with Portuguese being an example. The benefits of creating a medical-domain LM are already proven, and the health care sector could benefit greatly from a symbiosis with AI. Therefore, non-English languages should be motivated by the scarce studies already published and try to replicate them for their own language in order to fill this existing gap.

From another point of view, the use of transformers appears to be the better technique to automatically identify medical information. Despite the annotation process for any supervised learning task being very time-consuming, transformers achieve better results on fewer annotations, making their usage on new tasks relatively easier. This task also benefits from an in-domain medical LM. The entities most commonly extracted are diagnosis or disease, posology-related entities, symptoms, and phenotypes related to specific diseases.

Despite our belief that this was the right choice, we highlight the 2 databases that we searched, and we acknowledge that, despite our best efforts, there is always a possibility that not all relevant papers will be found when formulating a query. These are the limitations of our study. The chosen timeframe may also limit the availability of relevant studies, even though we believe it is the right timeframe to include studies that establish the current state-of-the-art with new technologies. While we focused on the Portuguese language, we acknowledge that our conclusions cannot be generalized to all non-English languages. However, other languages with similar characteristics in terms of available data and resources can certainly gain insights from this SLR.

Globally, the development and research in these topics for the English language are very advanced compared with non-English languages. In English, several studies have been presented that perform domain adaptation for smaller domains, such as specific-disease LMs, which have improved results in extracting medical information related to these diseases. The next steps should involve continuing the pretraining for different medical areas or diseases to ensure the most comprehensive coverage with LMs. In addition, fine-tuning the already available models to meet the specific requirements of health care professionals is essential.

Non-English languages are still performing domain adaptation for general domains, such as medical or biomedical fields, and should be motivated by these studies to overcome the barriers inherent in their respective language. In the next step, they should focus on performing domain adaptation, aiming to narrow down to specific medical areas or diseases. They should strive to replicate studies on Assertion Status or even those focused on resolving the frequent presence of abbreviations and typos

in the text. In non-English languages where there is a scarcity of available data, it would be beneficial to have open corpora, even if distributed under licenses that protect data privacy, to enable more researchers to develop models.

These types of studies are important to assess and guide the development of non-English languages trying to bridge the gaps and capitalize on the opportunities provided by these technologies to promote equity and improve access to health care all over the world. The differences in the available data and resources are almost impossible to correct but at least should be minimized.

This effort aims to harness AI to enhance health care by developing advanced LMs tailored for non-English languages, thereby supporting health care professionals with decision-making tools that alleviate their workload and improve patient care indirectly.

Acknowledgments

MN's work was supported by the project Blockchain.PT, (PRR RE-C05-i01.02: AGENDAS/ALIANÇAS VERDES PARA A INOVAÇÃO EMPRESARIAL).

Authors' Contributions

MN, JB, and LBE performed conceptualization. MN and LBE conducted the investigation. JB, JCF, and LBE performed supervision. JB, LBE, and JCF conducted validation. MN wrote the original draft. JB, JCF, and LBE performed review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 509 KB - medinform_v12i1e60164_app1.pdf]

References

1. HealthTech. How to navigate structured and unstructured data as a healthcare organization. URL: https://healthtechmagazine.net/article/2023/05/structured-vs-unstructured-data-in-healthcare-perfcon [accessed 2024-06-14]
2. RBC Capital Markets. Navigating the changing face of healthcare episode. URL: https://www.rbccm.com/en/gib/healthcare/story.page [accessed 2023-10-30]
3. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Computational Stats 2021;13(6):e1549. [doi: 10.1002/wics.1549]
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Computer Science Computation and Language 2023. [doi: 10.5860/choice.189890]
5. Chernyavskiy A, Ilvovsky D, Nakov P. Transformers: "The End of History" for NLP? Computer Science Computation and Language 2021 [FREE Full text]
6. Rokon OF. RNN vs. LSTM vs. Transformers: unraveling the secrets of sequential data processing. Medium. 2023. URL: https://tinyurl.com/432k5mn5 [accessed 2024-05-20]
7. How do Transformers work? - Hugging Face NLP Course. URL: https://huggingface.co/learn/nlp-course/chapter1/4 [accessed 2023-12-06]
8. Farahani A, Voghoei S, Rasheed K, Arabnia HR. A brief review of domain adaptation. 2020. URL: http://arxiv.org/abs/2010.03978 [accessed 2024-04-21]
9. Guo X, Yu H. On the domain adaptation and generalization of pretrained language models: a survey. 2022. URL: http://arxiv.org/abs/2211.03154 [accessed 2024-04-21]
10. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ACL Anthology 2019. [doi: 10.5260/chara.21.2.8]
11. Mao C, Xu J, Rasmussen L, Li Y, Adekkanattu P, Pacheco J, et al. AD-BERT: using pre-trained language model to predict the progression from mild cognitive impairment to alzheimer's disease. J Biomed Inform 2023;144:104442 [FREE Full text] [doi: 10.1016/j.jbi.2023.104442] [Medline: 37429512]
12. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc 2022;29(7):1208-1216 [FREE Full text] [doi: 10.1093/jamia/ocac040] [Medline: 35333345]
13. Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z, et al. Pre-trained language models in biomedical domain: a systematic survey. ACM Comput Surv 2023;56(3):1-52. [doi: 10.1145/3611651]
14. A Inteligência Artificial na Saúde, uma Breve Perspectiva. URL: https://www.ey.com/pt_pt/health/a-inteligencia-artificial-na-saude-uma-breve-perspectiva [accessed 2024-05-18]

15. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336-341 [FREE Full text] [doi: 10.1016/j.ijsu.2010.02.007] [Medline: 20171303]
16. Scopus - Document search. URL: https://www.scopus.com/search/form.uri?display=basic#basic [accessed 2023-05-15]
17. Web of science core collection. Clarivate. URL: https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/web-of-science-core-collection/ [accessed 2024-01-26]
18. Pranckutė R. Web of science (WoS) and scopus: The titans of bibliographic information in today's academic world. Publications 2021;9(1):12. [doi: 10.3390/publications9010012]
19. Zhu J, Liu W. A tale of two databases: the use of web of science and scopus in academic papers. Scientometrics 2020;123(1):321-335. [doi: 10.1007/s11192-020-03387-8]
20. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
21. Fu S, Thorsteinsdottir B, Zhang X, Lopes GS, Pagali SR, LeBrasseur NK, et al. A hybrid model to identify fall occurrence from electronic health records. Int J Med Inf 2022;162:104736 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104736] [Medline: 35316697]
22. Roitero K, Portelli B, Popescu MH, Mea VD. DiLBERT: cheap embeddings for disease related medical NLP. IEEE Access 2021;9:159714-159723. [doi: 10.1109/access.2021.3131386]
23. Agnikula Kshatriya BS, Sagheb E, Wi CI, Yoon J, Seol HY, Juhn Y, et al. Identification of asthma control factor in clinical notes using a hybrid deep learning model. BMC Med Inform Decis Mak 2021;21(Suppl 7):272 [FREE Full text] [doi: 10.1186/s12911-021-01633-4] [Medline: 34753481]
24. Zhang T, Chen M, Bui AAT. AdaDiag: adversarial domain adaptation of diagnostic prediction with clinical event sequences. J Biomed Inform 2022;134:104168 [FREE Full text] [doi: 10.1016/j.jbi.2022.104168] [Medline: 35987449]
25. Chen PF, Chen KC, Liao WC, Lai F, He TL, Lin SC, et al. Automatic international classification of diseases coding system: deep contextualized language model with rule-based approaches. JMIR Med Inform 2022;10(6):e37557 [FREE Full text] [doi: 10.2196/37557] [Medline: 35767353]
26. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J Cloud Comp 2021;10(1). [doi: 10.1186/s13677-020-00218-2]
27. Wen G, Chen H, Li H, Hu Y, Li Y, Wang C. Cross domains adversarial learning for Chinese named entity recognition for online medical consultation. J Biomed Inform 2020;112:103608 [FREE Full text] [doi: 10.1016/j.jbi.2020.103608] [Medline: 33132138]
28. Wen C, Chen T, Jia X, Zhu J. Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary. Data Intell 2021;3(3):402-417. [doi: 10.1162/dint_a_00105]
29. Zhang N, Jia Q, Yin K, Dong L, Gao F, Hua N. Conceptualized representation learning for Chinese biomedical text mining. Computer Science Computation and Language. 2020. URL: http://arxiv.org/abs/2008.10813 [accessed 2023-10-19]
30. Schneider ETR, de Souza JVA, Knafou J, Oliveira LES, Copara J, Gumiel YB, et al. BioBERTpt - a portuguese neural language model for clinical named entity recognition. : Association for Computational Linguistics; 2020 Presented at: Proceedings of the 3rd Clinical Natural Language Processing Workshop; August 27, 2024; USA p. 65-72. [doi: 10.18653/v1/2020.clinicalnlp-1.7]
31. Schneider ETR, Gumiel YB, De Souza JVA, Mie ML, Emanuel SEOL, De SRM, et al. CardioBERTpt: transformer-based models for cardiology language representation in portuguese. 2023 Presented at: Proceedings of the 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS); June 22-24, 2023; L'Aquila, Italy p. 378-381. [doi: 10.1109/CBMS58004.2023.00247]
32. Carrino CP, Llop J, Pàmies M, Gutiérrez-Fandiño A, Armengol-Estapé J, Silveira-Ocampo J, et al. Pre-trained biomedical language models for clinical NLP in Spanish. ACL Anthology 2022:193-199. [doi: 10.18653/v1/2022.bionlp-1.19]
33. Coutinho I, Martins B. Transformer-based models for ICD-10 coding of death certificates with portuguese text. J Biomed Inform 2022;136:104232 [FREE Full text] [doi: 10.1016/j.jbi.2022.104232] [Medline: 36307020]
34. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural Language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. Computer Science Computation and Language 2019. [doi: 10.18653/v1/w19-5006]
35. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. Annual Reviews 2019;21(2):8-10. [doi: 10.5260/chara.21.2.8]
36. Google Académico. URL: https://scholar.google.com/ [accessed 2023-11-03]
37. Zhou S, Wang N, Wang L, Sun J, Blaes A, Liu H, et al. A cross-institutional evaluation on breast cancer phenotyping NLP algorithms on electronic health records. Comput Struct Biotechnol J 2023;22:32-40 [FREE Full text] [doi: 10.1016/j.csbj.2023.08.018] [Medline: 37680211]
38. Rahman P, Ye C, Mittendorf KF, Lenoue-Newton M, Micheel C, Wolber J, et al. Accelerated curation of checkpoint inhibitor-induced colitis cases from electronic health records. JAMIA Open 2023;6(1):ooad017 [FREE Full text] [doi: 10.1093/jamiaopen/ooad017] [Medline: 37012912]

39. Crema C, Buonocore TM, Fostinelli S, Parimbelli E, Verde F, Fundarò C, et al. Advancing italian biomedical information extraction with transformers-based models: methodological insights and multicenter practical application. J Biomed Inform 2023;148:104557 [FREE Full text] [doi: 10.1016/j.jbi.2023.104557] [Medline: 38012982]
40. Lu W, Jiang J, Shi Y, Zhong X, Gu J, Huangfu L, et al. Application of entity-BERT model based on neuroscience and brain-like cognition in electronic medical record entity recognition. Front Neurosci 2023;17:1259652 [FREE Full text] [doi: 10.3389/fnins.2023.1259652] [Medline: 37799340]
41. Zhang Z, Liu D, Zhang M, Qin X. Combining data augmentation and domain information with TENER model for clinical event detection. BMC Med Inform Decis Mak 2021;21(Suppl 9):261 [FREE Full text] [doi: 10.1186/s12911-021-01618-3] [Medline: 34789246]
42. Kim HK, Park Y, Park Y, Choi E, Kim S, You H, et al. Identifying alcohol-related information from unstructured bilingual clinical notes with multilingual transformers. IEEE Access 2023;11:16066-16075. [doi: 10.1109/access.2023.3245523]
43. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: a transferable clinical natural language processing model for electronic health records. Artif Intell Med 2021;118:102086. [doi: 10.1016/j.artmed.2021.102086] [Medline: 34412834]
44. Solarte-Pabón O, Montenegro O, García-Barragán A, Torrente M, Provencio M, Menasalvas E, et al. Transformers for extracting breast cancer information from Spanish clinical narratives. Artif Intell Med 2023;143:102625 [FREE Full text] [doi: 10.1016/j.artmed.2023.102625] [Medline: 37673566]
45. Liu F, Liu M, Li M, Xin Y, Gao D, Wu J, et al. Automatic knowledge extraction from chinese electronic medical records and rheumatoid arthritis knowledge graph construction. Quant Imaging Med Surg 2023;13(6):3873-3890 [FREE Full text] [doi: 10.21037/qims-22-1158] [Medline: 37284084]
46. Wang SY, Huang J, Hwang H, Hu W, Tao S, Hernandez-Boussard T. Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. Int J Med Inform 2022;167:104864 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104864] [Medline: 36179600]
47. Singh P, Haimovich J, Reeder C, Khurshid S, Lau ES, Cunningham JW, et al. One clinician is all you need-cardiac magnetic resonance imaging measurement extraction: deep learning algorithm development. JMIR Med Inform 2022;10(9):e38178 [FREE Full text] [doi: 10.2196/38178] [Medline: 35960155]
48. Kim Y, Heider PM, Lally IRH, Meystre SM. A hybrid model for family history information identification and relation extraction: development and evaluation of an end-to-end information extraction system. JMIR Med Inform 2021;9(4):e22797 [FREE Full text] [doi: 10.2196/22797] [Medline: 33885370]
49. He K, Yao L, Zhang J, Li Y, Li C. Construction of genealogical knowledge graphs from obituaries: multitask neural network extraction system. J Med Internet Res 2021;23(8):e25670 [FREE Full text] [doi: 10.2196/25670] [Medline: 34346903]
50. Silva JF, Almeida JR, Matos S. Extraction of family history information from clinical notes: deep learning and heuristics approach. JMIR Med Inform 2020;8(12):e22898 [FREE Full text] [doi: 10.2196/22898] [Medline: 33372893]
51. Dai HJ, Lee YQ, Nekkantti C, Jonnagaddala J. Family history information extraction with neural attention and an enhanced relation-side scheme: algorithm development and validation. JMIR Med Inform 2020;8(12):e21750. [doi: 10.2196/21750] [Medline: 33258777]
52. Zhan K, Peng W, Xiong Y, Fu H, Chen Q, Wang X, et al. Novel graph-based model with biaffine attention for family history extraction from clinical text: modeling study. JMIR Med Inform 2021;9(4):e23587 [FREE Full text] [doi: 10.2196/23587] [Medline: 33881405]
53. Yang Z, Pou-Prom C, Jones A, Banning M, Dai D, Mamdani M, et al. Assessment of natural language processing methods for ascertaining the expanded disability status scale score from the electronic health records of patients with multiple sclerosis: algorithm development and validation study. JMIR Med Inform 2022;10(1):e25157 [FREE Full text] [doi: 10.2196/25157] [Medline: 35019849]
54. Santus E, Li C, Yala A, Peck D, Soomro R, Faridi N, et al. Do neural information extraction algorithms generalize across institutions? JCO Clin Cancer Inform 2019;3:1-8 [FREE Full text] [doi: 10.1200/CCI.18.00160] [Medline: 31310566]
55. Mahajan P, Rana D. Feature optimization in CNN using MROA for disease classification. IDT 2023;17(2):301-315. [doi: 10.3233/idt-220097]
56. Landolsi MY, Hlaoua L, Romdhane LB. Hybrid method to automatically extract medical document tree structure. Eng Appl Artif Intell 2023;120:105922. [doi: 10.1016/j.engappai.2023.105922]
57. Wunnava S, Qin X, Kakar T, Sen C, Rundensteiner EA, Kong X. Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. Drug Saf 2019;42(1):113-122. [doi: 10.1007/s40264-018-0765-9] [Medline: 30649736]
58. Viani N, Miller TA, Napolitano C, Priori SG, Savova GK, Bellazzi R, et al. Supervised methods to extract clinical events from cardiology reports in Italian. J Biomed Inform 2019;95:103219 [FREE Full text] [doi: 10.1016/j.jbi.2019.103219] [Medline: 31150777]
59. Wang S, Pang M, Pan C, Yuan J, Xu B, Du M, et al. Information extraction for intestinal cancer electronic medical records. IEEE Access 2020;8:125923-125934. [doi: 10.1109/access.2020.3005684]
60. Ju M, Short AD, Thompson P, Bakerly ND, Gkoutos GV, Tsaprouni L, et al. Annotating and detecting phenotypic information for chronic obstructive pulmonary disease. JAMIA Open 2019;2(2):261-271 [FREE Full text] [doi: 10.1093/jamiaopen/ooz009] [Medline: 31984360]

61. Lopes F, Teixeira C, Gonçalo Oliveira H. Comparing different methods for named entity recognition in portuguese neurology text. J Med Syst 2020;44(4):77. [doi: 10.1007/s10916-020-1542-8] [Medline: 32112285]
62. Alfattni G, Belousov M, Peek N, Nenadic G. Extracting drug names and associated attributes from discharge summaries: text mining study. JMIR Med Inform 2021;9(5):e24678 [FREE Full text] [doi: 10.2196/24678] [Medline: 33949962]
63. Chen T, Dredze M, Weiner JP, Hernandez L, Kimura J, Kharrazi H. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. JMIR Med Inform 2019;7(1):e13039 [FREE Full text] [doi: 10.2196/13039] [Medline: 30862607]
64. Li Z, Ren J. Fine-tuning ERNIE for chest abnormal imaging signs extraction. J Biomed Inform 2020;108:103492 [FREE Full text] [doi: 10.1016/j.jbi.2020.103492] [Medline: 32645382]
65. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A. Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. JMIR Med Inform 2021;9(3):e17934. [doi: 10.2196/17934]
66. Li C, Ma K. Entity recognition of Chinese medical text based on multi-head self-attention combined with BILSTM-CRF. Math Biosci Eng 2022;19(3):2206-2218 [FREE Full text] [doi: 10.3934/mbe.2022103] [Medline: 35240782]
67. Zanotto BS, Beck da Silva Etges AP, Dal Bosco A, Cortes EG, Ruschel R, De Souza AC, et al. Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and nonneural classifiers. JMIR Med Inform 2021;9(11):e29120 [FREE Full text] [doi: 10.2196/29120] [Medline: 34723829]
68. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS. Toward complete structured information extraction from radiology reports using machine learning. J Digit Imaging 2019;32(4):554-564 [FREE Full text] [doi: 10.1007/s10278-019-00234-y] [Medline: 31218554]
69. Peterson KS, Lewis J, Patterson OV, Chapman AB, Denhalter DW, Lye PA, et al. Automated travel history extraction from clinical notes for informing the detection of emergent infectious disease events: algorithm development and validation. JMIR Public Health Surveill 2021;7(3):e26719 [FREE Full text] [doi: 10.2196/26719] [Medline: 33759790]
70. Malmasi S, Ge W, Hosomura N, Turchin A. Comparing information extraction techniques for low-prevalence concepts: The case of insulin rejection by patients. J Biomed Inform 2019;99:103306 [FREE Full text] [doi: 10.1016/j.jbi.2019.103306] [Medline: 31618679]
71. Chen Y, Hao L, Zou VZ, Hollander Z, Ng RT, Isaac KV. Automated medical chart review for breast cancer outcomes research: a novel natural language processing extraction system. BMC Med Res Methodol 2022;22(1):136 [FREE Full text] [doi: 10.1186/s12874-022-01583-z] [Medline: 35549854]
72. Sterckx L, Vandewiele G, Dehaene I, Janssens O, Ongenae F, De Backere F, et al. Clinical information extraction for preterm birth risk prediction. J Biomed Inform 2020;110:103544 [FREE Full text] [doi: 10.1016/j.jbi.2020.103544] [Medline: 32858168]
73. Cen X, Yuan J, Pan C, Tang Q, Ma Q. Contextual embedding bootstrapped neural network for medical information extraction of coronary artery disease records. Med Biol Eng Comput 2021;59(5):1111-1121. [doi: 10.1007/s11517-021-02359-1] [Medline: 33893606]
74. Dandala B, Joopudi V, Tsou CH, Liang JJ, Suryanarayanan P. Extraction of information related to drug safety surveillance from electronic health record notes: joint modeling of entities and relations using knowledge-aware neural attentive models. JMIR Med Inform 2020;8(7):e18417 [FREE Full text] [doi: 10.2196/18417] [Medline: 32459650]
75. Zhou J, Guo X, Duan L, Yao Y, Shang Y, Wang Y, et al. Moving toward a standardized diagnostic statement of pituitary adenoma using an information extraction model: a real-world study based on electronic medical records. BMC Med Inform Decis Mak 2022;22(1):319 [FREE Full text] [doi: 10.1186/s12911-022-02031-0] [Medline: 36476365]
76. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. Artif Intell Med 2021;117:102083. [doi: 10.1016/j.artmed.2021.102083] [Medline: 34127232]
77. Dewaswala N, Chen D, Bhopalwala H, Kaggal VC, Murphy SP, Bos JM, et al. Natural language processing for identification of hypertrophic cardiomyopathy patients from cardiac magnetic resonance reports. BMC Med Inform Decis Mak 2022;22(1):272 [FREE Full text] [doi: 10.1186/s12911-022-02017-y] [Medline: 36258218]
78. Chen L, Song L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from chinese electronic medical records. Int J Med Inform 2019;124:6-12. [doi: 10.1016/j.ijmedinf.2019.01.004] [Medline: 30784428]
79. Zhang Q, Wu M, Lv P, Zhang M, Yang H. Research on named entity recognition of chinese electronic medical records based on multi-head attention mechanism and character-word information fusion. IFS 2022;42(4):4105-4116. [doi: 10.3233/jifs-212495]
80. spaCy. Language processing pipelines. URL: https://spacy.io/usage/processing-pipelines/ [accessed 2023-11-01]
81. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Walker M, Ji H, Stent A, editors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018:2227-2237.

82. Lee EK, Uppal K. CERC: an interactive content extraction, recognition, and construction tool for clinical and biomedical text. BMC Med Inform Decis Mak 2020;20(Suppl 14):306 [FREE Full text] [doi: 10.1186/s12911-020-01330-8] [Medline: 33323109]
83. Chen YP, Lo YH, Lai F, Huang CH. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: algorithm development and validation study. J Med Internet Res 2021;23(1):e25113 [FREE Full text] [doi: 10.2196/25113] [Medline: 33502324]
84. Sykes D, Grivas A, Grover C, Tobin R, Sudlow C, Whiteley W, et al. Comparison of rule-based and neural network models for negation detection in radiology reports. Nat Lang Eng 2021;27(2):203-224. [doi: 10.1017/s1351324920000509]
85. Chaturvedi J, Chance N, Mirza L, Vernugopan V, Velupillai S, Stewart R, et al. Development of a corpus annotated with mentions of pain in mental health records: natural language processing approach. JMIR Form Res 2023;7:e45849 [FREE Full text] [doi: 10.2196/45849] [Medline: 37358897]
86. Newman-Griffis D, Divita G, Desmet B, Zirikly A, Rosé CP, Fosler-Lussier E. Ambiguity in medical concept normalization: an analysis of types and coverage in electronic health record datasets. J Am Med Inform Assoc 2021;28(3):516-532 [FREE Full text] [doi: 10.1093/jamia/ocaa269] [Medline: 33319905]
87. Jaber A, Martínez P. Disambiguating clinical abbreviations using a one-fits-all classifier based on deep learning techniques. Methods Inf Med 2022;61(S 01):e28-e34 [FREE Full text] [doi: 10.1055/s-0042-1742388] [Medline: 35104909]
88. Lee EB, Heo GE, Choi CM, Song M. MLM-based typographical error correction of unstructured medical texts for named entity recognition. BMC Bioinformatics 2022;23(1):486 [FREE Full text] [doi: 10.1186/s12859-022-05035-9] [Medline: 36384464]

Abbreviations

AI: artificial intelligence
ANN: artificial neural network
BERT: Bidirectional Encoder Representations from Transformers
BiLSTM-CRF: Bidirectional Long Short-Term Memory - Conditional Random Field
CNN: convolutional neural network
DL: deep learning
EHR: electronic health record
EMR: electronic medical record
ELMo: embeddings from language models
ICD-10: International Statistical Classification of Diseases, Tenth Revision
IE: information extraction
LLM: large language model
LM: language model
ML: machine learning
NER: named entity recognition
NLP: natural language processing
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
PT-BR: Brazilian Portuguese
PT-PT: European Portuguese
RNN: recurrent neural network
SLR: systematic literature review
WOSCC: Web of Science Core Collection

Edited by A Castonguay; submitted 03.05.24; peer-reviewed by A Kocian, DR Carvalho; comments to author 13.06.24; revised version received 14.07.24; accepted 06.08.24; published 21.10.24.

Please cite as:
Nunes M, Bone J, Ferreira JC, Elvas LB
Health Care Language Models and Their Fine-Tuning for Information Extraction: Scoping Review
JMIR Med Inform 2024;12:e60164
URL: https://medinform.jmir.org/2024/1/e60164
doi:10.2196/60164
PMID:39432345

Original Paper

Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation

Xingyuan Li[1*], MS; Ke Liu[1*], MS; Yanlin Lang[1*], MS; Zhonglin Chai[2], Prof Dr; Fang Liu[1], Prof Dr Med

[1]Department of Nephrology, West China Hospital, Sichuan University, Chengdu, China

[2]Department of Diabetes, Central Clinical School, Monash University, Melbourne, Australia

[*]these authors contributed equally

Corresponding Author:
Fang Liu, Prof Dr Med
Department of Nephrology
West China Hospital
Sichuan University
37 Guoxue Alley
Wuhou District
Chengdu, 610041
China
Phone: 86 13890790651
Email: liufangfh@163.com

Abstract

Background: Artificial intelligence (AI) has shown great promise in assisting medical diagnosis, but its application in renal pathology remains limited.

Objective: We evaluated the performance of an advanced AI language model, Claude 3 Opus (Anthropic), in generating diagnostic descriptions for renal pathological images.

Methods: We carefully curated a dataset of 100 representative renal pathological images from the *Diagnostic Atlas of Renal Pathology* (3rd edition). The image selection aimed to cover a wide spectrum of common renal diseases, ensuring a balanced and comprehensive dataset. Claude 3 Opus generated diagnostic descriptions for each image, which were scored by 2 pathologists on clinical relevance, accuracy, fluency, completeness, and overall value.

Results: Claude 3 Opus achieved a high mean score in language fluency (3.86) but lower scores in clinical relevance (1.75), accuracy (1.55), completeness (2.01), and overall value (1.75). Performance varied across disease types. Interrater agreement was substantial for relevance (κ=0.627) and overall value (κ=0.589) and moderate for accuracy (κ=0.485) and completeness (κ=0.458).

Conclusions: Claude 3 Opus shows potential in generating fluent renal pathology descriptions but needs improvement in accuracy and clinical value. The AI's performance varied across disease types. Addressing the limitations of single-source data and incorporating comparative analyses with other AI approaches are essential steps for future research. Further optimization and validation are needed for clinical applications.

(JMIR Med Inform 2024;12:e65033) doi:10.2196/65033

KEYWORDS

artificial intelligence; Claude 3 Opus; renal pathology; diagnostic performance; large language model; LLM; performance evaluation; medical diagnosis; AI language model; diagnosis; pathology images; pathologist; clinical relevance; accuracy; language fluency; pathological diagnosis

Introduction

Artificial intelligence (AI) has demonstrated remarkable capabilities in analyzing complex medical data and assisting clinical decision-making across various fields [1]. In particular, AI's potential for interpreting histopathological images has been increasingly recognized, offering novel insights into disease pathogenesis and diagnosis [2]. However, the application of AI in renal pathology, a field characterized by high complexity and variability, remains relatively unexplored. Recent advancements

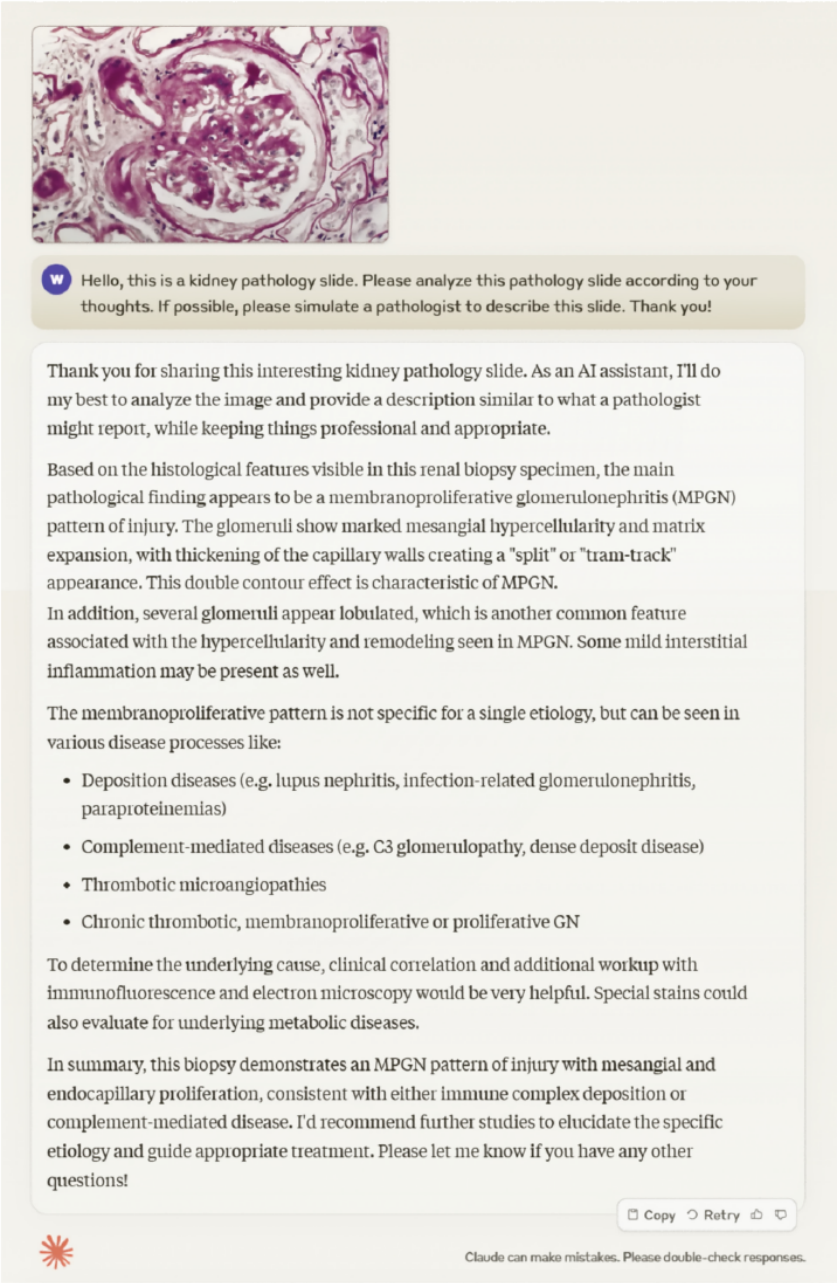
in natural language processing, such as the development of large language models (LLMs) like GPT-3 (OpenAI) and Claude 3 Opus (Anthropic), have opened up new possibilities for AI-assisted pathological diagnosis [3,4]. These models can capture semantic and contextual information from textual data and generate coherent, human-like responses. Despite the proven utility of AI in other pathology domains, such as oncology and dermatology, its performance and feasibility in renal pathology have not been systematically evaluated. To address this gap, we conducted a pioneering study to investigate the potential of using Claude 3 Opus, a state-of-the-art AI language model, for renal pathological diagnosis. By assessing the model's ability to generate accurate and clinically relevant diagnostic descriptions for a wide range of renal pathologies, we aimed to provide initial evidence and insights into the strengths and limitations of AI in this challenging field.

Methods

We carefully curated a dataset of 100 representative renal pathological images from the *Diagnostic Atlas of Renal Pathology* (3rd edition) [5]. The image selection process aimed to cover a wide spectrum of common renal diseases, ensuring a balanced and comprehensive dataset. The number of images per disease type ranged from 2 to 9 (eg, immunoglobulin A nephropathy: n=5; acute tubular injury: n=4; diabetic nephropathy: n=9), proportional to the relative prevalence and morphological diversity of each condition. Multimedia Appendix 1 provides the complete list of images and their corresponding disease labels.

Claude 3 Opus, an advanced AI language model, was used to generate diagnostic descriptions for each image. For each image, the model was given the following prompt: "Describe the key morphological features and provide a diagnostic impression for this renal biopsy image." No additional disease-specific information or background knowledge was provided, allowing us to assess Claude 3 Opus's standalone performance in renal pathology interpretation. The generated descriptions were evaluated by 2 experienced renal pathologists using a comprehensive 5-point scale across 5 key dimensions: clinical relevance, accuracy, language fluency, detail completeness, and overall clinical value. These evaluation dimensions were selected based on established frameworks for assessing the quality and utility of pathology reports [6,7]. They collectively cover the essential aspects of an effective pathology description, from clinical applicability and correctness to clarity and comprehensiveness. Descriptive statistics (means and SDs) of the evaluation scores were calculated for each pathological type. The interrater agreement between the 2 pathologists' scores was assessed using the Cohen κ statistic. Figure 1 shows an example of a prompt to Claude 3 Opus and its response.

Figure 1. Example of a prompt to Claude 3 Opus and its response.



Results

The performance evaluation results for Claude 3 Opus in generating renal pathological descriptions are presented in Table 1. The AI model achieved a high overall score in language fluency (mean score 3.86, SD 0.68), indicating its ability to produce grammatically correct and easily readable reports. However, the model's performance in other key aspects was suboptimal, with lower scores for clinical relevance (mean score 1.75, SD 0.77), accuracy (mean score 1.55, SD 0.66), detail completeness (mean score 2.01, SD 0.84), and overall clinical value (mean score 1.75, SD 0.74). Notably, the AI model's performance varied across different renal pathological types. Higher scores (>4) were observed for certain diseases, such as membranoproliferative glomerulonephritis and subacute bacterial endocarditis–associated glomerulonephritis, suggesting the model's potential in assisting the diagnosis of these specific conditions. Conversely, the model's performance was subpar for several other types, including acute interstitial nephritis (mean score 1.00, SD 0.00) and collapsing glomerulopathy (mean score 1.17, SD 0.24), indicating limitations in capturing their key diagnostic features. The inter-rater agreement analysis revealed substantial agreement between the 2 pathologists on clinical relevance (κ=0.627) and overall clinical value (κ=0.589), as well as moderate agreement on accuracy (κ=0.485) and detail completeness (κ=0.458).

Table 1. Performance of Claude 3 Opus in generating descriptions for 100 renal pathological images across 27 disease types.

Standard pathological diagnosis	Clinical relevance, mean score	Accuracy of description, mean score	Language fluency, mean score	Detail completeness, mean score	Overall clinical applicability, mean score
Immunoglobin A nephropathy (n=5)	2.10	1.10	4.30	2.90	1.90
Immunoglobin G4–related tubulointerstitial nephritis (n=3)	1.67	1.17	4.33	2.50	1.67
Proliferative glomerulonephritis with monoclonal deposits (n=3)	2.33	1.83	4.17	2.83	2.50
Autosomal dominant polycystic kidney disease (n=2)	1.25	1.50	3.75	2.50	1.75
Henoch-Schönlein purpura nephritis (n=5)	1.10	1.10	3.90	2.00	1.40
Acute postinfectious glomerulonephritis (n=4)	1.63	1.38	3.88	2.38	1.63
Acute interstitial nephritis (n=3)	1.17	1.00	3.67	1.50	1.00
Acute tubular injury (n=4)	1.25	1.38	3.50	1.88	1.38
Acute pyelonephritis (n=4)	1.38	1.38	3.88	2.38	1.63
Focal segmental glomerulosclerosis (n=4)	1.63	1.75	3.75	2.25	1.88
Anti–glomerular basement membrane antibody–mediated glomerulonephritis (n=4)	2.38	2.38	4.00	2.75	2.63
Chronic interstitial fibrosis and tubular atrophy (n=2)	1.50	1.00	4.00	2.00	2.00
Chronic pyelonephritis (n=2)	2.00	1.50	4.00	2.00	2.00
Diffuse mesangial sclerosis (n=3)	1.83	1.50	3.50	1.50	1.83
Membranoproliferative glomerulonephritis (n=3)	4.17	4.00	3.83	3.67	4.00
Arterionephrosclerosis (n=4)	1.13	1.00	4.00	1.00	1.13
Medullary cystic disease (n=3)	1.33	1.17	3.83	1.67	1.33
Collapsing glomerulopathy (n=3)	1.17	1.17	3.50	1.17	1.17
Diabetic nephropathy (n=9)	1.50	1.44	3.61	1.50	1.44
Preeclampsia (n=8)	1.44	1.56	3.75	1.94	1.69
Fibrillary glomerulonephritis (n=2)	1.50	1.25	3.75	1.50	1.25
Microscopic polyangiitis (n=3)	1.83	1.67	3.83	1.67	1.83
Hemoglobinuric acute renal failure (n=2)	2.25	1.75	4.00	2.75	2.25
Thrombotic microangiopathy (n=3)	1.50	1.50	4.00	1.67	1.67
Subacute bacterial endocarditis–associated glomerulonephritis (n=3)	3.17	3.00	4.00	2.67	2.83
Scleroderma (n=4)	2.00	1.88	4.00	1.75	1.88
Kidney biopsies from healthy people (n=5)	1.50	1.10	3.50	1.40	1.30
Overall	1.75	1.55	3.86	2.01	1.75

Discussion

Principal Findings

This study provides initial evidence for the potential of advanced AI language models, such as Claude 3 Opus, in assisting renal pathological diagnosis. The model demonstrated promise in generating fluent and readable pathological descriptions, which could streamline the reporting process and alleviate pathologists' workloads. However, the suboptimal performance in accuracy, clinical relevance, and overall value highlights the need for further improvement before clinical implementation. The interrater agreement analysis revealed substantial agreement for clinical relevance and overall value, but only moderate agreement for accuracy and completeness. This discrepancy might stem from the inherent subjectivity in evaluating granular aspects of pathology descriptions. Pathologists' individual expertise, expectations, and interpretive styles could influence their assessments of accuracy and completeness. Developing standardized scoring rubrics and involving larger, multicenter

expert panels in future studies could help mitigate this variability and improve evaluation reliability [8,9].

The AI model's performance varied notably across different renal pathological types. The higher scores for conditions like membranoproliferative glomerulonephritis and infection-related glomerulonephritis could be attributed to their distinct morphological features, such as characteristic immune-complex deposits or structural alterations [10]. These overt patterns might be more readily discernible by the AI algorithms. Conversely, the lower performance for diseases like acute interstitial nephritis and collapsing glomerulopathy might reflect their subtler or more heterogeneous histological manifestations [11], posing challenges for automated interpretation.

While Claude 3 Opus exhibited potential in generating fluent descriptions, the limited accuracy and clinical relevance underscore the challenges of applying LLMs to complex medical image data. As highlighted by recent studies [12,13], LLMs excel at processing textual information but may struggle with the intricacies of specialized visual tasks like histopathology interpretation. Continued research on architectures and training strategies tailored for medical vision applications is crucial for realizing the full potential of AI in this domain.

Our study's reliance on images from a single atlas dataset may have introduced some biases and limits the generalizability to real-world clinical scenarios. Although the *Diagnostic Atlas of Renal Pathology* is widely recognized as a high-quality reference, external validation using diverse, multicenter biopsy datasets is essential to assess the robustness and transferability of our findings [14]. Future studies should prioritize prospective validation on independent clinical cohorts to establish the real-world performance of AI models like Claude 3 Opus.

Comparing the performance of Claude 3 Opus with other AI models or traditional diagnostic methods could offer valuable insights into its relative strengths and areas for improvement. While direct comparisons were beyond the scope of this initial study, recent work has reported promising results with deep learning–based approaches for renal pathology classification. Shen et al [15] developed a convolutional neural network (CNN) model that achieved an accuracy of 87.5% in classifying 6 common glomerular diseases. Similarly, Hermsen et al [16] demonstrated the effectiveness of a multiclass CNN for diagnosing various renal pathologies, with an area under the receiver operating characteristic curve ranging from 0.88 to 0.99. These studies highlight the potential of specialized AI architectures for renal pathology diagnosis, serving as benchmarks for evaluating the performance of LLMs like Claude 3 Opus. Future research should aim to conduct head-to-head comparisons and explore synergistic integrations of LLMs with image-based AI models to leverage their complementary strengths.

To fully realize AI's potential in renal pathology, future research should focus on optimizing the model's training process with comprehensive and balanced datasets, incorporating expert feedback into the learning process, and integrating AI with advanced digital pathology tools for more accurate and objective diagnoses. The scarcity of large-scale annotated datasets and concerns about AI interpretability remain major challenges to be addressed. Amidst these challenges, the collaboration between AI researchers, pathologists, and clinicians is crucial for developing reliable and clinically applicable AI models. Standardized evaluation frameworks and best practices for the responsible use of AI in pathology are also needed. As AI continues to advance, we anticipate its increasing role in enhancing diagnostic accuracy and efficiency, ultimately benefiting patient care.

Conclusion

In conclusion, our study provides an initial assessment of Claude 3 Opus's potential for AI-assisted renal pathology diagnosis. While the model showed promise in generating fluent descriptions, improvements in accuracy and clinical relevance are necessary for practical implementation. The observed performance variations across disease types highlight the need for targeted model optimizations. Addressing the limitations of single-source data and incorporating comparative analyses with other AI approaches are essential steps for future research. As AI continues to advance, close collaboration between pathologists, AI researchers, and clinicians will be instrumental in developing reliable, integrated diagnostic solutions that enhance patient care in renal pathology.

Acknowledgments

This study was supported by Health Commission of Sichuan Province Program (21ZD001)

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete list of images and their corresponding disease labels.
[PDF File (Adobe PDF File), 213 KB - medinform_v12i1e65033_app1.pdf]

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
2. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019 May;20(5):e253-e261 [FREE Full text] [doi: 10.1016/S1470-2045(19)30154-8] [Medline: 31044723]
3. Claude: A breakthrough AI assistant based on constitutional AI. Anthropic. URL: https://www.anthropic.com/claude.html [accessed 2023-05-15]
4. Brown T, Mann B, Ryder N. Language models are few-shot learners. arXiv Preprint posted online May 28, 2020. [doi: 10.48550/arXiv.2005.14165]
5. Fogo A, Kashgarian M. Diagnostic Atlas of Renal Pathology, 3rd edition. Amsterdam, Netherlands: Elsevier; 2016.
6. Hewitt KJ, Wiest IC, Carrero ZI, Bejan L, Millner TO, Brandner S, et al. Large language models as a diagnostic support tool in neuropathology. J Pathol Clin Res 2024 Nov;10(6):e70009 [FREE Full text] [doi: 10.1002/2056-4538.70009] [Medline: 39505569]
7. Loor-Torres R, Wu Y, Esteban Cabezas, Borras-Osorio M, Toro-Tobon D, Duran M, et al. Use of natural language processing to extract and classify papillary thyroid cancer features from surgical pathology reports. Endocr Pract 2024 Nov;30(11):1051-1058. [doi: 10.1016/j.eprac.2024.08.008] [Medline: 39197747]
8. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016 Jan;69:245-247. [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
9. Liu X, Duan C, Kim M, Zhang L, Jee E, Maharjan B, et al. Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. JMIR Med Inform 2024 Aug 06;12:e59273 [FREE Full text] [doi: 10.2196/59273] [Medline: 39106482]
10. Sethi S, Haas M, Markowitz G, D'Agati VD, Rennke HG, Jennette JC, et al. Mayo Clinic/Renal Pathology Society consensus report on pathologic classification, diagnosis, and reporting of GN. J Am Soc Nephrol 2016 May;27(5):1278-1287 [FREE Full text] [doi: 10.1681/ASN.2015060612] [Medline: 26567243]
11. Mubarak M. Collapsing focal segmental glomerulosclerosis: increasing the awareness. J Nephropathol 2012 Jul;1(2):77-80 [FREE Full text] [doi: 10.5812/nephropathol.7474] [Medline: 24475392]
12. Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. Indian J Radiol Imaging 2024 Apr;34(2):269-275 [FREE Full text] [doi: 10.1055/s-0043-1777289] [Medline: 38549881]
13. Nowak S, Schneider H, Layer YC, Theis M, Biesner D, Block W, et al. Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers. Eur Radiol 2024 May;34(5):2895-2904 [FREE Full text] [doi: 10.1007/s00330-023-10373-0] [Medline: 37934243]
14. Subramanian S, Viswanathan VK, Ramani S, Rajendiran P. Generalizability in the age of deep learning: The case of renal histology. Kidney Int Rep 2022;7(8):1623-1625 [FREE Full text] [doi: 10.1016/j.ekir.2022.05.011]
15. Shen L, Sun W, Zhang Q, Wei M, Xu H, Luo X, et al. Deep learning-based model significantly improves diagnostic performance for assessing renal histopathology in lupus glomerulonephritis. Kidney Dis (Basel) 2022 Jul;8(4):347-356 [FREE Full text] [doi: 10.1159/000524880] [Medline: 36157261]
16. Hermsen M, de Bel T, den Boer M, Steenbergen EJ, Kers J, Florquin S, et al. Deep learning-based histopathologic assessment of kidney tissue. J Am Soc Nephrol 2019 Oct;30(10):1968-1979 [FREE Full text] [doi: 10.1681/ASN.2019020144] [Medline: 31488607]

Abbreviations

AI: artificial intelligence
CNN: convolutional neural network
LLM: large language model

Edited by C Lovis; submitted 02.08.24; peer-reviewed by X Liu, PK Sarangi, GK Gupta; comments to author 22.09.24; revised version received 28.09.24; accepted 19.10.24; published 15.11.24.

Please cite as:
Li X, Liu K, Lang Y, Chai Z, Liu F
Exploring the Potential of Claude 3 Opus in Renal Pathological Diagnosis: Performance Evaluation
JMIR Med Inform 2024;12:e65033
URL: https://medinform.jmir.org/2024/1/e65033
doi:10.2196/65033
PMID:

Review

Task-Specific Transformer-Based Language Models in Health Care: Scoping Review

Ha Na Cho[1*], PhD; Tae Joon Jun[2*], PhD; Young-Hak Kim[3*], MD, PhD; Heejun Kang[4], MSc; Imjin Ahn[1], MSc; Hansle Gwon[1], MSc; Yunha Kim[5], BSc; Jiahn Seo[5], BSc; Heejung Choi[5], BSc; Minkyoung Kim[5], BSc; Jiye Han[5], MSc; Gaeun Kee[1], MSc; Seohyun Park[1], BSc; Soyoung Ko[1], BSc

[1]Department of Information Medicine, Asan Medical Center, Seoul, Republic of Korea

[2]Big Data Research Center, Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

[3]Division of Cardiology, Department of Information Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[4]Division of Cardiology, Asan Medical Center, Seoul, Republic of Korea

[5]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Tae Joon Jun, PhD
Big Data Research Center
Asan Institute for Life Sciences
Asan Medical Center
88, Olympicro 43gil
Songpagu
Seoul, 05505
Republic of Korea
Phone: 82 10 2956 6101
Email: saigram89@gmail.com

Abstract

Background: Transformer-based language models have shown great potential to revolutionize health care by advancing clinical decision support, patient interaction, and disease prediction. However, despite their rapid development, the implementation of transformer-based language models in health care settings remains limited. This is partly due to the lack of a comprehensive review, which hinders a systematic understanding of their applications and limitations. Without clear guidelines and consolidated information, both researchers and physicians face difficulties in using these models effectively, resulting in inefficient research efforts and slow integration into clinical workflows.

Objective: This scoping review addresses this gap by examining studies on medical transformer-based language models and categorizing them into 6 tasks: dialogue generation, question answering, summarization, text classification, sentiment analysis, and named entity recognition.

Methods: We conducted a scoping review following the Cochrane scoping review protocol. A comprehensive literature search was performed across databases, including Google Scholar and PubMed, covering publications from January 2017 to September 2024. Studies involving transformer-derived models in medical tasks were included. Data were categorized into 6 key tasks.

Results: Our key findings revealed both advancements and critical challenges in applying transformer-based models to health care tasks. For example, models like MedPIR involving dialogue generation show promise but face privacy and ethical concerns, while question-answering models like BioBERT improve accuracy but struggle with the complexity of medical terminology. The BioBERTSum summarization model aids clinicians by condensing medical texts but needs better handling of long sequences.

Conclusions: This review attempted to provide a consolidated understanding of the role of transformer-based language models in health care and to guide future research directions. By addressing current challenges and exploring the potential for real-world applications, we envision significant improvements in health care informatics. Addressing the identified challenges and implementing proposed solutions can enable transformer-based language models to significantly improve health care delivery

and patient outcomes. Our review provides valuable insights for future research and practical applications, setting the stage for transformative advancements in medical informatics.

(JMIR Med Inform 2024;12:e49724) doi:10.2196/49724

KEYWORDS

transformer-based language models; medicine; health care; medical language model

Introduction

Background

Transformer models have revolutionized natural language processing (NLP) with their exceptional state-of-the-art performance in various applications such as conversation, translation, text classification, and text generation. A transformer model is a type of deep learning model designed to process and generate sequences of data, such as text. The key innovation of transformer models is the self-attention mechanism, which allows the model to weigh the importance of different words in the input sequence, regardless of their position. Self-attention allows the model to focus on different parts of an input sequence simultaneously, rather than processing the sequence in a fixed order. This mechanism enables the model to capture complex patterns and relationships within the context more effectively than previous models, which is particularly useful for understanding and generating natural language. These models hold significant promise for the health care sector, addressing clinical challenges and unlocking new opportunities in medical informatics (eg, disease prediction, clinical decision support, and patient interaction).

Since the introduction of the transformer model by Google [1] in 2017, it has become the foundation for various pretrained language models (PLMs). PLMs are transformer models that have been initially trained on a large text corpus before being fine-tuned for specific tasks. This pretraining allows the models to leverage vast amounts of unstructured data to improve their performance in various NLP tasks. Two of the most widely used PLM architectures in medical research are Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT). GPT is designed to generate coherent text based on a given input, making it useful for tasks like dialogue generation [2]. BERT, on the other hand, is designed to understand the context of words in a sentence from both directions, making it highly effective for tasks like question answering and text classification [3]. Transformer-based language models have revolutionized the field of NLP and continued to advance the state-of-the-art in NLP with their impressive performance.

Despite the success of transformer-based language models in many domains, there is a significant gap in comprehensive reviews specifically focused on their application in the health care domain. In health care, transformer-based language models have been used for crucial tasks such as disease prediction, decision-making, and image analysis [4]. With the abundance of free text sources, such as medical documentation in free text, including social media, electronic medical records (EMRs), physician-patient conversations, and online encyclopedias, more significant challenges to language models are needed. The application of NLP in health care is not without controversy, particularly concerning data privacy, ethical implications, and the integration of artificial intelligence (AI) systems into clinical practices. Debates continue about the extent to which AI can replace human judgment, the transparency of AI decision-making processes, and the potential biases in AI models trained on unbalanced datasets. By addressing these concerns, our paper contributes to the timely and critical discourse on the responsible deployment of transformer-based language models in health care, emphasizing the need for transparency, fairness, and ethical considerations in AI development.

Objective

The objective of this paper is to provide a comprehensive scoping review of task-specific transformer-based language models in health care. By focusing on models pretrained on medical corpora, we aim to address the gap in existing literature where detailed surveys specifically tailored to health care applications are lacking. We seek to highlight the strengths, limitations, and potential of these models, offering valuable insights for future research and practical applications in medical informatics.

Related Work

While many review studies of NLP have been conducted in the medical field [5-13], on transformer-based language models [14-20], and in health-related domains [21-25], comprehensive surveys and broader and up-to-date transformer-based language models in health care are lacking, leaving a gap in understanding their full potential and limitations. Pandey et al [5] introduced RedBERT, a model focusing on topic discovery and deep sentiment classification of COVID-19 online discussions, demonstrating the application of NLP in understanding public health concerns. Iroju and Olaleke [6] conducted a systematic review of NLP applications, identifying key areas where NLP can enhance clinical decision-making and patient care. Similarly, Locke et al [7] provided a comprehensive overview of NLP in medicine, emphasizing the potential of NLP technologies in transforming medical practice. Adyashreem et al [8] surveyed various NLP techniques in the biomedical field, shedding light on how these techniques can be applied to biomedical text for improved information extraction and analysis. Wang et al [9] reviewed the application of NLP in clinical medicine, highlighting the advancements and challenges in integrating NLP with clinical workflows.

Khanbhai et al [11] applied NLP and machine learning techniques to patient experience feedback, providing insights into patient satisfaction and areas for improvement in health care services. Casey et al [12] focused on NLP applications in radiology reports, identifying how NLP can streamline the interpretation and reporting of radiological findings. Zhou et al

[13] discussed the broader applications of NLP for smart health care, envisioning a future where NLP-driven systems enhance patient care and operational efficiency.

In the realm of transformer-based language models, Zhang et al [14] surveyed their applications in bioinformatics, highlighting how these models have advanced the analysis of biological data. Yang [15] and Lin et al [16] explored the progress and applications of transformer models in Korean and general NLP tasks, respectively, highlighting their growing importance and versatility. Chitty-Venkata et al [17] reviewed neural architecture search for transformers, underscoring the potential of these models in optimizing NLP tasks. Gillioz et al [18] provided an overview of transformer-based models for various NLP tasks, illustrating their adaptability and efficiency. Han et al [19] focused on multimodal pretrained models, emphasizing their capability to handle diverse data types, including text, image, and audio. Greco et al [20] and Albalawi et al [21] discussed transformer models' applications in mental health and Arabic social media, respectively, highlighting their potential in understanding and addressing specific health-related issues. Kalyan et al [22] and Shamshad et al [23] provided comprehensive surveys on biomedical PLMs and their applications in medical imaging, respectively, showcasing the transformative impact of transformers in these fields.

Our review categorizes these models into 6 key tasks: dialogue generation, question answering, summarization, text classification, sentiment analysis, and named entity recognition (NER). Ultimately, advancements in transformer-based language models hold the promise of significantly transforming health care delivery and improving patient outcomes. By enabling more accurate disease prediction, enhancing clinical decision support, and facilitating better patient-provider communication, these models can lead to more efficient, effective, and personalized health care. Our review underscores the broader implications of these technologies, advocating for continued research and development to harness their full potential in revolutionizing medical informatics and patient care.

Methods

Information Source and Search Strategy

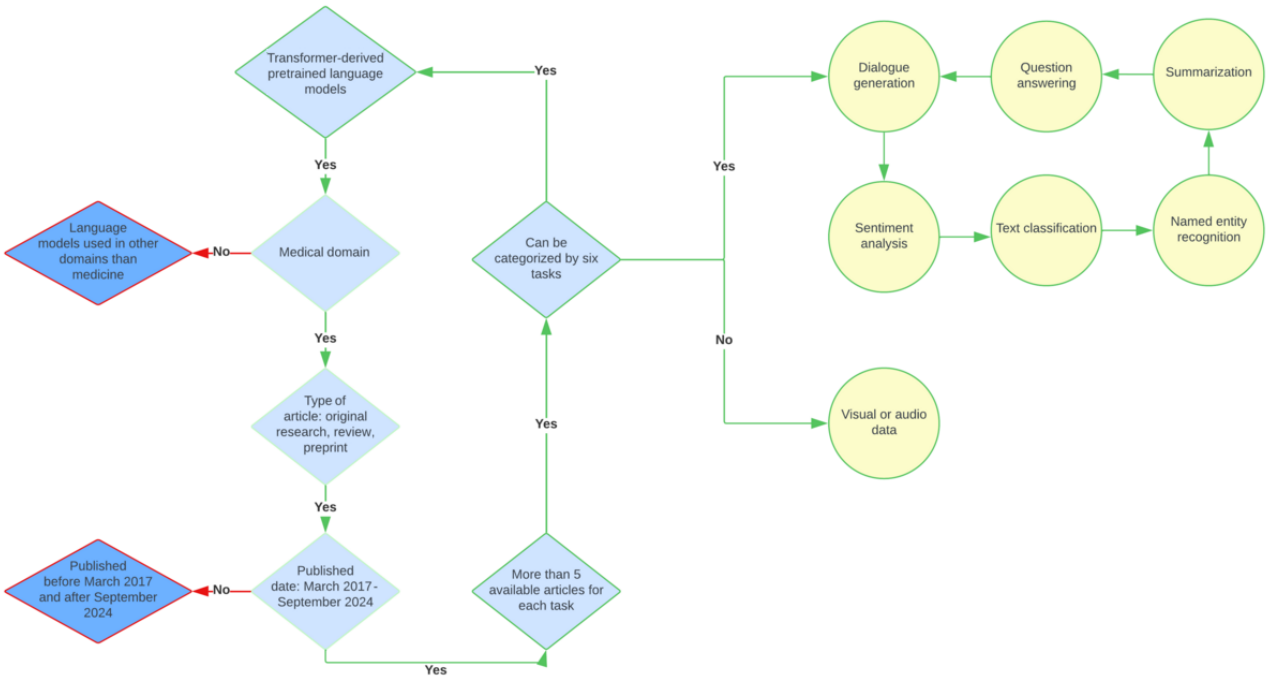
We followed the Cochrane scoping review protocol to conduct and map the available literature in an efficient and systematic approach. This method involves defining the research question, identifying relevant studies, selecting studies based on predefined criteria, charting data, and summarizing the results to clarify key concepts and identify research gaps [24].

Our research team (mainly HNC and TJJ) conducted a comprehensive literature review for identifying studies in the field that met the inclusion and exclusion criteria. The screening and selection of papers were conducted by 2 independent reviewers (HNC and TJJ). Initially, titles and abstracts were screened to identify relevant studies. Full texts of potentially eligible studies were then reviewed to ensure they met the inclusion criteria. Disagreements between reviewers were resolved through discussion and consensus, with a third reviewer (YHK) consulted if necessary. Our literature search was conducted across several scientific databases, including Google Scholar and PubMed, which were selected for their comprehensive coverage of relevant journals and peer-reviewed studies in the medical and academic fields. We covered publications from January 01, 2017, to September 30, 2024, and used specific combinations of keywords and Boolean operators, such as "transformer-based AND language models AND medical domain," "health care AND language models," "NLP AND medicine AND survey," and "GPT AND BERT AND health care." Data extraction involved summarizing key findings, model names, and training datasets. The extracted data were cross-verified by both reviewers to ensure accuracy and consistency. Any discrepancies were resolved through discussion.

We included studies that involved transformer-derived models applied to medical tasks, were published in peer-reviewed journals, and were written in English. The exclusion criteria involved studies focusing solely on non-text data (eg, audio, image, and video) or those not meeting the inclusion requirements. The selection of tasks (dialogue generation, question answering, summarization, text classification, sentiment analysis, and NER) was based on their critical role in advancing health care applications of transformer models. The specific process is illustrated in Figure 1, with details of each stage of filtering from the initial identification of articles to the final selection. The inclusion criteria were rigorously applied at each step, beginning with the screening of titles and abstracts, followed by a full-text review, and culminating in the inclusion of studies that met all predefined criteria. This methodical approach allowed us to compile a comprehensive and focused set of articles for our scoping review, ensuring that our findings are both robust and reliable.

These tasks cover a wide range of functionalities essential for improving clinical workflows, enhancing patient interactions, and facilitating efficient information retrieval and analysis, making them vital for the advancement of transformer-based language models in the medical domain. Languages and model types were chosen to represent a diverse range of medical contexts and applications.

Figure 1. Article filtering process with inclusion criteria.



In this section, studies that have used language models in health care applications were examined. Based on the literature review, Table 1 provides a comprehensive list of transformer-based models applied in the medical domain, comparing each task based on the authors, model name, training dataset, PLM model, key metric, score, and purpose or findings of the study. These English-written PLMs in the health care domain were categorized into 6 distinct tasks, namely dialogue generation, question answering, summarization, text classification, sentiment analysis, and NER. The articles within each task are listed in no sequential order. In Figure 2, the evolution timeline of transformer-based language models provides an overview of significant models that have been developed for use in medicine. It illustrates key milestones and the deployment criteria used to guide the inclusion of studies in our review. This historical context provides a foundation for understanding the methodological choices made in our scoping review. This visual representation highlights the emergence of models over time and their increasing significance in health care applications. We provide insights into the progress made in this field and anticipate future advancements by tracking the development of these models.

Table 1. Summary of the applications of pretrained language models subdivided into tasks.

Task and author (year)	Model name	Training dataset	PLM[a] model	Key metric	Score (%)	Key findings
Conversation						
Varshney et al [26], 2023	Medical Entity Prediction (MEP)	UMLS	BERT	Accuracy	85	Integrated triples from knowledge graphs to enhance medical predictions using a large pretrained model.
Yuan et al [27], 2022	BioBART	PubMed	BART	Rouge-2	65	Adapted and improved biomedical context understanding through advanced generative techniques.
Zhao et al [28], 2022	MedPIR	MedDG, MedDialog	BERT, GPT	F1	82	Used a knowledge-aware dialogue graph encoder (KDGE) and recall-enhanced generator (REG) to improve clinical responses.
Chen et al [29], 2023	OPAL	Wikipedia, WOZ, CamRest676	BART	BLEU	21.5	Tailored for task-oriented medical dialogues by incorporating domain-specific ontologies.
Liang et al [30], 2021	MKA-BERT-GPT	MedDG, MedDialog-CN	BERT, GPT	Relevance improvement	15	First scalable model to integrate a medical knowledge graph into a large pretrained model, enhancing biomedical understanding.
Compton et al [31], 2021	MEDCOD	KB, doctor edits	GPT-3	Emotive accuracy	90	Generated diverse, emotive, and empathetic sentences for health care interactions.
Li et al [32], 2023	ChatDoctor	5000 doctor-patient conversations	LlaMA	Precision, recall, F1	83.7, 84.5, 84.1	Fine-tuned LLaMa model using tailored doctor-patient dialogues for medical NLP[b] tasks.
Tang et al [33], 2023	-w terms+AL	MedDialog	BART	Annotation accuracy	87	Automated large-scale medical conversation text annotation with terminology extraction.
Zeng et al [34], 2020	Transformer-DST	MultiWOZ	BERT	DST accuracy	54.6	Proposed a transformer-based framework using a flat encoder-decoder architecture for dialogue state tracking in medical contexts.
Suri et al [35], 2021	MeDia-BERT	MeDiaQA	BERT	Accuracy	64.3	Employed a hierarchical approach to medical dialogue analysis, including multiple-choice question answering.
Phan et al [36], 2021	SciFive	PubMed	T5	Accuracy	86.6	A medical T5 text-to-text model effective for various clinical downstream tasks.
Wu et al [37], 2023	PMC-LLaMA	PubMed, 30K Medical Books	LlaMA	Accuracy	64.43	Transitioned a general-purpose model to a high-performing medical language model via comprehensive fine-tuning, achieving state-of-the-art performance in medical question answering.
Zhang et al [38], 2023	HuatuoGPT	Huatuo26M	GPT	BLEU, ROUGE, distinct	25.6, 27.76, 93	Chinese health care LLM: Tailored for the Chinese medical domain, providing state-of-the-art results in medical consultation tasks.
Question answering						
Lee et al [39], 2019	BioBERT	PubMed, EHR[c], clinical notes, patents	BERT	MRR improvement	12.24	First domain-specific BERT-based model for biomedical text mining, outperforming standard BERT in medical tasks.
Luo et al [40], 2023	BioGPT	PubMed	GPT	Accuracy	78.2	Pretrained on a 15M PubMed corpus, this model outperforms GPT-2 in biomedical text generation.
Shin et al [41], 2020	BioMegatron	Wikipedia, news, OpenWebtext	Megatron-LM	Bias	40	Enhanced the representation of biomedical entities across a large corpus for better entity understanding.
Rasmy et al [42], 2020	MED-BERT	Cerner Health Facts, Truven	BERT	AUC[d]	20 boosts	First proof-of-concept BERT model for integrating electronic health records.
Yasunaga et al [43], 2022	LinkBERT	Wikipedia	BERT	Improvement	5	Effective in multi-hot reasoning and few-shot question answering by linking documents.
Michalopoulos et al [44], 2020	UmlsBERT	MIMIC-III	BERT	F1	86	Learned the association of clinical terms within the UMLS metathesaurus.

Task and author (year)	Model name	Training dataset	PLM[a] model	Key metric	Score (%)	Key findings
Zhang et al [45], 2021	SMedBERT	ChineseBLUE	BERT	Accuracy	78	Introduced a mention-neighbor hybrid attention model for heterogeneous medical entity information.
Yang et al [46], 2022	ExKidney-BERT	EMR[e]	BERT	Accuracy	95.8	A specialized model focused on renal transplant-pathology integration.
Mitchell et al [47], 2021	CaBERTnet	Pathology reports	BERT	Accuracy	85	An automatic system for extracting tumor sites and histology information.
Trieu et al [48], 2021	BioVAE	PubMed	SciBERT, GPT2	VAE	72.9	First large-scale pretrained language model using the OPTIMUS framework in the biomedical domain.
Khare et al [49], 2021	MMBERT	COntext (ROCO)	BERT	Accuracy	72	Proposed masked language modeling for radiology text representations.
Yasunaga et al [43], 2022	BioLinkBERT	Wikipedia, Book Corpus	BERT	BLURB	84	A novel linking method for predicting document relations in pretraining models.
Nguyen et al [50], 2022	SPBERTQA	ViHealthQA	SBERT	Mean average precision	69.5	A 2-step question answering system addressing linguistic disparities with BM25 and Sentence BERT.
Luo et al [51], 2023	BioMEDGPT	PubMed	GPT	Accuracy	76.1	First multimodal GPT capable of aligning biological modalities with human language for medical text analysis.
Toma et al [52], 2023	Clinical Camel	PubMed, USMLE, MedMCQA	LlaMA-2	Five-shot accuracy	74.3, 54.3, 47.0	A model that outperforms GPT-3.5 by using efficient fine-tuning techniques.
Han et al [53], 2023	MedAlpaca	Medical flash cards, Wikidoc	Alpaca	Accuracy	21.1-24.1	Highlighted privacy protection in medical artificial intelligence and demonstrated significant performance enhancements in medical certification exams through fine-tuning.
Singhal et al [54], 2023	MedPaLM-2	MedMCQA, MedQA, PubMedQA, MMLU	PaLM	Accuracy	67.6	Instruction prompt tuning undergoes rigorous human evaluation to assess harm avoidance, comprehension, and factual accuracy.
Chen et al [55], 2023	MEDITRON	PubMED	LlaMA-2	Accuracy	79.8	Achieved 6% improvement over the best public baseline and 3% gain over fine-tuned Llama-2 models.
Summarization						
Yan et al [56], 2022	RadBERT	Open-I chest radiograph report	BERT	Accuracy, F1	97.5, 95	Adapted a bidirectional encoder representation for radiology text.
Du et al [57], 2020	BioBERTSum	PubMed	BERT	ROUGE-L	68	Introduced the first transformer-based model for extractive summarization in the biomedical domain.
Li et al [58], 2022	Clinical-Longformer & Clinical-Big Bird	MIMIC-III	Longformer, Big Bird	F1	97	Reduced memory usage through sparse attention in a long-sequence transformer.
Moro et al [59], 2022	DAMEN	MS2	BERT, BART	Accuracy	75	Developed a multi-document summarization method using token probability.
Chen et al [60], 2020	AlphaBERT	HER (NTUH-iMD)	BERT	Accuracy	69.3	Designed a diagnoses summarization model based on character-level tokens.
Alsentzer et al [61], 2019	Bio+Clinical BERT	MIMIC-III	BERT	F1	11 improvements	Released the first BERT-based model specifically for clinical text.
Cai et al [62], 2021	ChestXRay-BERT	MIMIC	BERT	Accuracy	73	Automatically generates abstractive summarization of radiology reports.
Yalunin et al [63], 2022	LF2BERT	UMLS, EHR	BERT	ROUGE-1 F1, ROUGE-2 F1, ROUGE-L F1	67, 56.4, 64.5	Developed a neural abstractive model for summarizing long medical texts.
Balde et al [64], 2024	MEDVOC	PubMed, BioASQ, EBM	GPT	ROUGE	51.49, 47.54, 19.51	Efficiently reduced fine-tuning time and improved vocabulary adaptation for medical texts.

Task and author (year)	Model name	Training dataset	PLM[a] model	Key metric	Score (%)	Key findings
Text classification						
Yang et al [65], 2022	GatorTron	Clinical notes, UF Health clinical corpus MIMIC-III, PubMed, Wikipedia	GPT	Pearson correlation	89	Outperformed previous biomedical and clinical domain models.
Gu et al [66], 2021	PubMED-BERT	PubMed	BERT	BLURB	81.2	Established a leaderboard for biomedical NLP, with robustness against noisy and incomplete biomedical text.
Huang et al [67], 2020	Clinical-BERT	EHR (clinical notes)	BERT	Accuracy, precision, recall, AUROC[f], AUPRC[g]	72.7, 37.6, 54.2, 74.2, 42.0	Introduced "catastrophic forgetting prevention" and generated visualized interpretable embeddings.
Gupta et al [68], 2022	MatSciB-ERT	Wikipedia, clinical database, Book Corpus	BERT	F1	81.5	Effective transformer model for scientific text analysis.
Fang et al [69], 2023	Bioformer	PubMed, PMC	BERT	Performance, speed	60 reduced model size, 2-3× speed	Reduced model size by 60% for biomedical text mining.
Gururangan et al [70], 2020	BioMed-RoBERTa	CHEMPROT, PubMed	RoBER-Ta	F1	83.4	Proposed domain and task-adaptive pretraining with a data selection strategy.
Liao et al [71], 2023	Mask-BERT	PubMed, NICTA-PIBOSO, symptoms	BERT	Accuracy, F1, PR-AUC[h]	91.8, 89.6, 93.1	Improved a BERT-based model for multiple tasks with masked input text.
He et al [72], 2022	KG-MTT-BERT	EHR	BERT	Accuracy	82	Developed a model for multi-type medical tests using a knowledge graph.
Yang et al [73], 2023	Trans-formEHR	EHR	BERT	AUROC, AUPRC	81.95, 78.64	Set a new standard in clinical disease prediction using longitudinal EHRs.
Pedersen et al [74], 2023	MeDa-BERT	EMR	BERT	Accuracy	86.7-97.1	Tailored embeddings for Danish medical text processing.
Hong et al [75], 2023	SCHOLAR-BERT	Public resource	BERT	F1	85.49	Leveraged a public resource-driven dataset for scientific NLP.
Abu Tareq Rony et al [76], 2024	MediGPT	Illness dataset	GPT	Accuracy, F1	90.0, 88.7	Improved medical text classification tasks showing performance gains up to 22.3% compared to traditional methods.
Sentiment analysis						
Ji et al [77], 2021	Mental-BERT/Mental-RoBERTa	Reddit	BERT, RoBER-Ta	F1, recall	81.76, 81.82	A pretrained masked model designed for mental health detection.
Taghizadeh et al [78], 2021	SINA-BERT	Self-gathered collection of texts from online sources	BERT	Precision, recall, macro F1, accuracy	94.91, 94.63, 94.77, 96.14	Developed a pretrained language model for the Persian medical domain.
AlBadani et al [79], 2022	SGTN	SemEval, SST2, IMDB, Yelp	BERT	Accuracy	80	Proposed the first sentiment analysis model using a transformer-based graph algorithm.
Pandey et al [80], 2021	RedBERT	Reddit	BERT	Accuracy	86.05	Introduced a sentiment classification method from web-scraped data.
Palani et al [81], 2021	T-BERT	Twitter	BERT	Accuracy	90.81	Designed a sentiment classification method for microblogging platforms.

Task and author (year)	Model name	Training dataset	PLM[a] model	Key metric	Score (%)	Key findings
Mao et al [82], 2022	AKI-BERT	MIMIC-III	BERT	AUC, precision, recall/sensitivity, F1, specificity, negative predictive value	74.7, 35.6, 61.9, 45.2, 76.8, 90.7	Created a BERT model for predicting acute kidney injury.
Chaudhary et al [83], 2020	TopicBERT	Ohsumed	BERT	Cost optimization	70	Improved computational efficiency by combining topic and language models for fine-tuning.
Qudar et al [84], 2020	TweetBERT	NCBI, BC5CDR, BIOSSES, MedNLI, Chemprot, GAD, JNLPBA	BERT	F1	87.1	Achieved state-of-the-art performance on biomedical datasets using Twitter data for pretraining.
Wouts et al [85], 2021	BelabBERT	DBRD	RoBERT	Accuracy	95.9	Developed a Dutch language model for psychiatric disease classification.
Named entity recognition						
Li et al [86], 2020	BEHRT	EHR	BERT	Accuracy	81	Interpretable model for multi-heterogeneous medical concepts.
Shang et al [87], 2019	G-BERT	EHR	BERT	Jaccard, PR-AUC, F1	45.7, 69.6, 61.5	The first pretraining method for medication recommendation in the medical domain.
Lentzen et al [88], 2022	BioGottBERT	Wikipedia, drug leaflets from AMIce, LIVIVO	RoBERTa, GottBERT	Accuracy	78	Introduced the first transformer model for German medical texts.
Davari et al [89], 2020	TIMBERT	PubMed	BERT	Precision, recall, F1	90.5, 91.2, 90.9	Developed a BERT-based model for automated toponym identification.
Peng et al [90], 2019	BlueBERT	PubMed, MIMIC-III	BERT	Masked token score	77.3	Demonstrated strong generalization ability across biomedical texts and cross-lingual tasks.
Miolo et al [91], 2021	ELECTRAMed	NCBI	BERT	Precision, recall, F1	85.9, 89.3, 87.5	The first ELECTRA-based model for the biomedical domain.
Khan et al [92], 2020	MT-BioNER	BC2GM, BC5CDR, NCBI-Disease	BERT	Precision, recall, F1	88.4, 90.52, 89.5	A multi-task transformer model for slot tagging in the biomedical domain.
Naseem et al [93], 2020	BioALBERT	PubMed, PMC	BERT	Precision, recall, F1	97.4, 94.4, 95.9	Trained on large biomedical corpora using ALBERT for biomedical text mining.
Yang et al [94], 2021	BIBC	Textbooks, research papers, clinical guidelines	BERT	Accuracy	78	Designed a new architecture for processing long text inputs in diabetes literature.
Martin et al [95], 2020	CamemBERT	Wikipedia	RoBERTa	Accuracy	85.7	Developed the first monolingual RoBERTa model for French medical text.
Kraljevic et al [96], 2021	MedGPT	EHR	GPT	Precision	64	Efficiently handled noise in EHR data using NER and MedCAT.
Li et al [97], 2019	EhrBERT	EHR	BERT	F1	93.8	Proposed an entity normalization technique for 1.5 million EHR notes.
Gwon et al [98], 2024	HeartBERT	EMR	BERT	Accuracy	74	Emphasized the importance of department-specific language models, with a focus on cardiology.
Mannion et al [99], 2023	UMLS-KGI-BERT	UMLS	BERT	Precision	85.05	Introduced a graph-based learning method with masked-language pretraining for clinical text extraction.
Schneider et al [100], 2023	CardioBERTpt	EHR	BERT	FL-score	83	Specialized in extracting Portuguese cardiology terms, demonstrating that data volume and representation improve NER performance.

Task and author (year)	Model name	Training dataset	PLM[a] model	Key metric	Score (%)	Key findings
Saleh et al [101], 2024	TocBERT	MIMIC-III	BERT	F1	84.6	Outperformed a rule-based solution in differentiating titles and subtitles for a discharge summary dataset.

[a]PLM: pretrained language model.

[b]NLP: natural language processing.

[c]EHR: electronic health record.

[d]AUC: area under the curve.

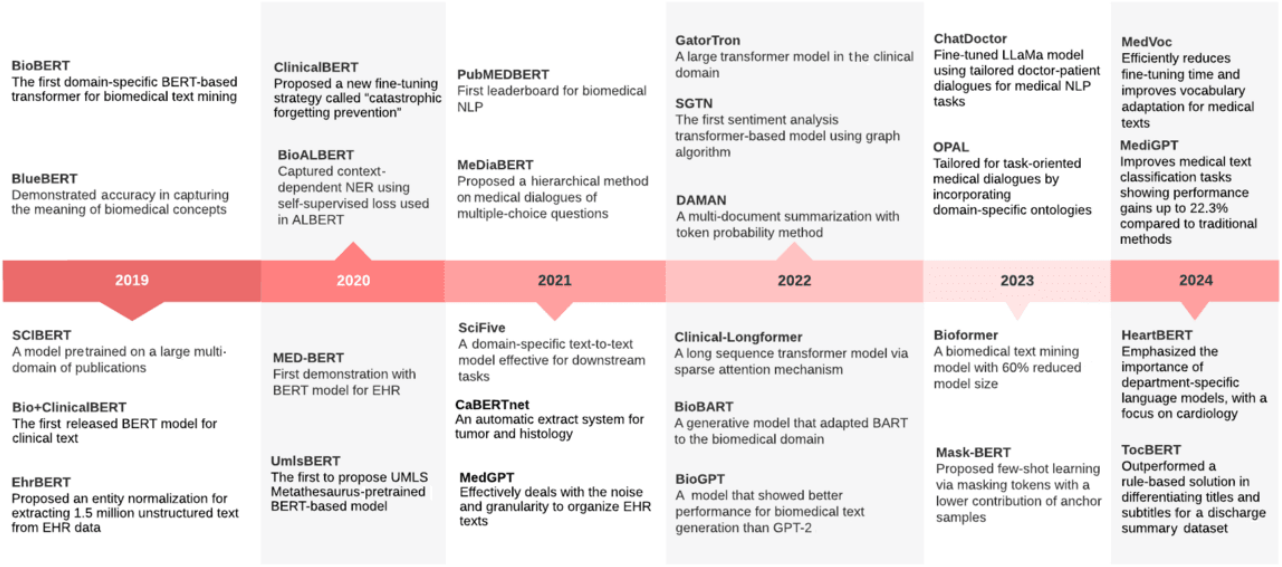
[e]EMR: electronic medical record.

[f]AUROC: area under the receiver operating characteristic curve.

[g]AUPRC: area under the precision-recall curve.

[h]PR-AUC: precision-area under curve.

Figure 2. Timeline of significant transformer-based models in health care. EHR: electronic health record; NER: named entity recognition; NLP: natural language processing.

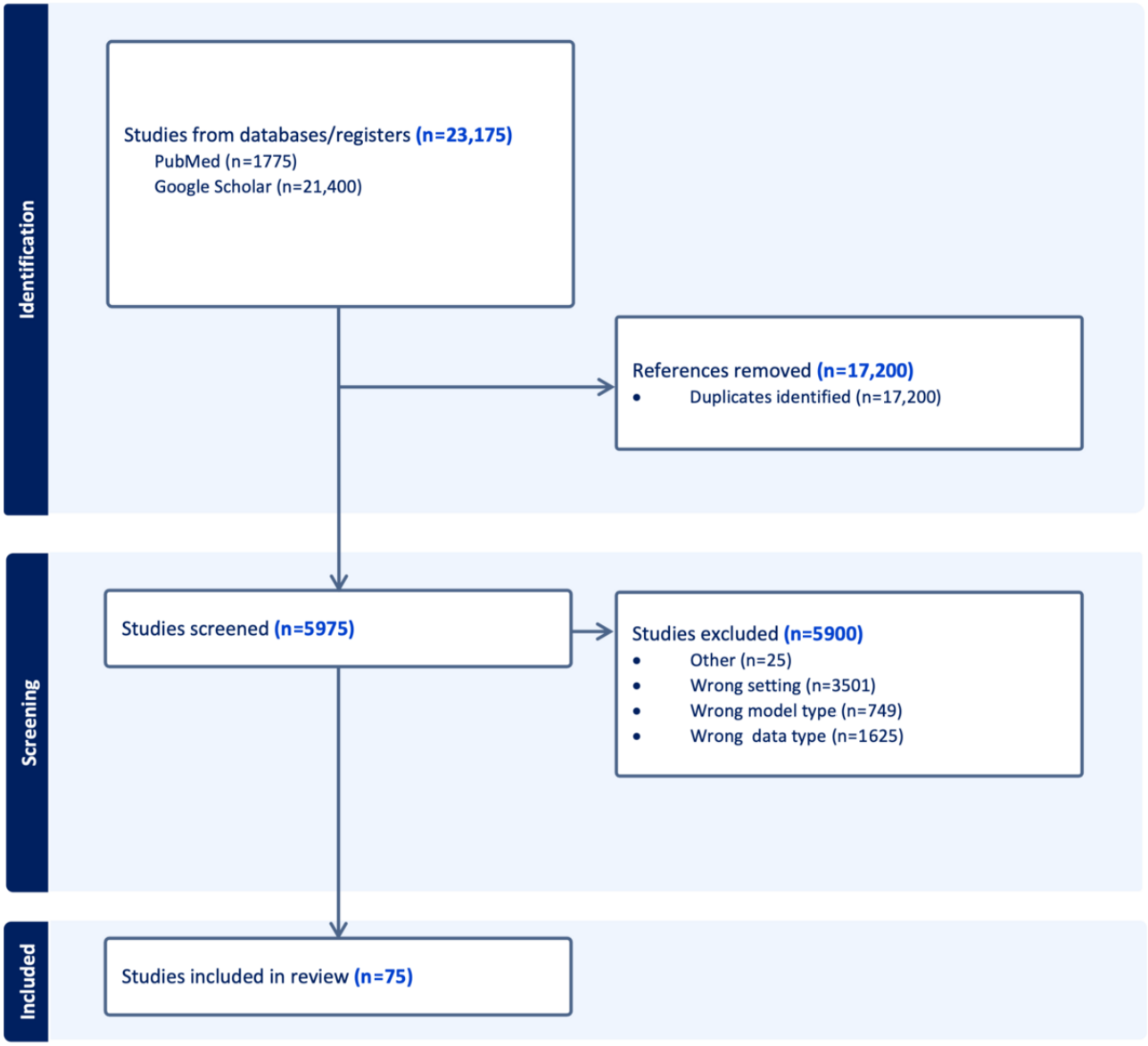


Results

Selected Studies

A total of 75 models were identified through our comprehensive review. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart is presented in Figure 3. The PRISMA checklist is presented in Multimedia Appendix 1. These papers encompass various research areas related to transformer-based models and their applications in the medical domain. The selection of these papers was based on predefined inclusion criteria, ensuring the relevance of each study to the scope of our review.

Figure 3. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the review process.



Applications of Language Models in Health Care: Task-Specific

Dialogue Generation

Conversation generation generates responses to a given dialogue. GPT models, including DialoGPT and DialogBERT, can effectively generate human-like dialogues based on large corpora and contextualized representations of text [25,102-105]. In the medical domain, conversation generation focuses on developing conversations related to medical information [106]. Chatbots in health care can be classified into 6 types: screening and diagnosis, treatment, monitoring, support, workflow efficiency, and health promotion. These tasks involve aiding patient consultation, acting as a physician's decision support system, collaborating with interdisciplinary research, and providing care instructions and medical education [107,108].

The key models are MEP, BioBART, MedPIR, MEDCOD, Transformer-DST, MeDiaBERT, ChatDoctor, and SciFive.

Research efforts on conversation generation in medicine have also incorporated knowledge graphs. MKA-BERT-GPT was the first scalable work to integrate a medical knowledge graph mechanism into a large pretrained model. Meanwhile, MedPIR proposed a recall-enhanced generator framework by using a knowledge-aware dialogue graph encoder to strengthen the relationship between the user input and the response via past conversation information [28,30]. They achieved an F1 score of 82% and a bilingual evaluation understudy (BLEU) score of 21.5. Varshney et al [26] proposed the Masked Entity Dialogue (MED) model to train smaller corpora texts, addressing a 10% improvement in entity prediction accuracy for the problem of local embeddings in entity embeddings by incorporating conversation history into triples in the graphs, resulting in an automatic prediction of the medical entities model.

On the other hand, MEDCOD [31] used the GPT pretrained model to integrate emotive and empathetic aspects into the output sentences, which further imitates a human physician–like feature to better communicate with patients. The Transformer-DST [34] model addresses dialogue state tracking,

optimizing state operation prediction, and value generation with high accuracy by suggesting to ask the DST model to consider the whole dialogue and the previous state. Moreover, MeDiaBERT [35], using a hierarchical approach, achieves 63.8% accuracy in multiple choice medical queries by building a transformer encoder contained within another in a hierarchical manner.

BioBART [27], a BART-based model, used patient descriptions and conversation histories as input for the model to autoregressively generate replies to user inputs. The model outperformed the BART model by 1.71 on Rouge-2 with a BLEU score of 4.45, and pretraining on PubMed abstracts supported the model's performance. OPAL [29] and -w terms+AL models also used BART for pretraining. OPAL's proposed method involves 2 phases: pretraining on large-scale contextual texts with structured information extracted using an information-extracting tool and fine-tuning the pretrained model on task-oriented dialogues. The results showed a significant performance boost, overcoming the problems created by annotated data with large structured dialogue data. Recently, the -w terms+AL model proposed a framework for improving dialogue generation by incorporating domain-specific terminology through an automatic terminology annotation framework using a self-attention mechanism [33].

While other models are based on BERT, GPT, or BART, ChatDoctor [32], SciFive [36], and PMC-LLaMA [37] use LLaMA or T5 PLMs. To improve accuracy and provide informed advice in medical consultations, ChatDoctor used Meta's open-source LLaMA [109], which was fine-tuned using real-world patient-physician conversations and autonomous knowledge retrieval capabilities, achieving 91.25% accuracy. SciFive, a Text-To-Text Transfer Transformer–based model, was pretrained on large biomedical corpora, indicating its significant potential for learning large and extended outputs. The SciFive model was trained using a maximum likelihood objective with "teacher forcing" [110] for multi-task learning by leveraging task-specific tokens in the input sequence. Both models outperformed previous baseline methods.

More recently, the HuatuoGPT model [38], specifically tailored for the Chinese medical domain, provided state-of-the-art results in medical consultation tasks.

Question Answering

The question-answering task involves answering questions posed by users based on the texts in documents. It aims to generate an accurate response that directly answers the question input, contributing to clinical decision-making, medical education, and patient communication. Allowing physicians and researchers to obtain valuable answers quickly from electronic health records (EHRs) and various medical literature will effectively reduce the time and effort required when the procedure is done manually. While the dialogue generation and question-answering tasks both involve providing answers, the former focuses on generating responses within a conversation, whereas the latter focuses on developing specific answers to user questions.

The key models are BioBERT, BioGPT, BioMegatron, Med-BERT, UmlsBERT, SMedBERT, and BioVAE.

BERT-based language models have become increasingly popular in biomedical text mining as they can understand the context and generate accurate predictions. BioBERT [39], the first domain-specific BERT-derived transformer language model for biomedical text mining applications, achieved 89% accuracy on the MedQA dataset and outperformed BERT in medical text applications. A BioMegatron model [41], based on Megatron-LM [111], was also experimented on a question-answering task, and it was found that the domain and task-specific language model affected the overall performance rather than the model size. Shin et al [41] found that model size is not closely related to the performance rate, but rather the domain and task-specific language model affects the overall performance. Med-BERT [42], another BERT-inspired model, improved the prediction accuracy by 20% in disease prediction studies by pretraining on EHR datasets.

More recently, researchers have built BERT-based models for specific domains and tasks [112]. UmlsBERT [44] first built a semantic embedding linking concepts with the words in the UMLS Metathesaurus and proposed multi-label loss function–masked modeling. SMedBERT [45] also presented a similar approach with the knowledge semantic representation but structured the neighboring entities to learn heterogeneous information. UmlsBERT and SMedBERT enhanced performance, with F1 scores of 84% and 86%, respectively. Similarly, LinkBERT and BioLinkBERT [43] incorporated ontological knowledge to better understand a linking system between entities in the corpus. LinkBERT used a multi-task learning framework on several related tasks simultaneously to extract relations between entities in the corpus more effectively. ExKidneyBERT, CaBERTnet, and MMBERT extracted more precise answers from individual departmental reports [46,47,49].

On the other hand, BioVAE [48] used the OPTIMUS framework pretrained with SciBERT [113] and GPT-2 [114,115] and outperformed the baseline models on biomedical text mining. To address the issues on linguistic disparity, SPBERTQA [50] proposed a 2-stage multilingual language model pretrained on the SBERT model [116] to reply to user questions using multiple negative ranking losses with Bert multilingual 25.

However, previous studies using the BERT structure are a better fit for understanding the context, rather than generating texts. To this end, BioMedLM, a GPT architecture model, was built mainly for biomedical question-answering tasks [117] in recent studies of question-answering benchmarks and achieved 50% accuracy on summarizations of the patient's quest even in real situations with fewer data. BioGPT [40] applied a 2-step fine-tuning method to remove the noise in data and achieved 6.0% improved results compared with BioLinkBERT in the medical domain for question-answering tasks.

Recent studies have introduced significant advancements, such as BioMEDGPT [51], the first multimodal GPT for aligning biological data with human language, achieving 76.1% accuracy. Clinical Camel [52], using LLaMA-2, demonstrated superior performance with 5-shot accuracy ranging from 47.0% to 74.3%, outperforming GPT-3.5. MedAlpaca [53] focused on privacy

and medical certifications, attaining 21.1%-24.1% accuracy. MedPaLM-2 [54] reached 67.6% accuracy through instruction prompt tuning, and MEDITRON [55] achieved 79.8% accuracy, marking a 6% improvement upon existing models and setting a new benchmark.

Summarization

For many years, the medical field has suffered from the challenge of finding efficient and rapid access to understanding the fast-growing and immensely increasing amount of data formation. The key to timely and efficient clinical workflow is providing automatic summarization in clinical text. Summarization in health care is an important technique in NLP as it automatically summarizes the medical contexts into a concise summary of text. Summarization can be applied to medical records, literature, clinical trial reports, and other types of medical texts that aim to provide clinical providers with quick access to relevant information, without the need to skim through lengthy documents. Overall summarization can aid clinicians with decision-making through effective and prompt communication during the physician-patient meeting, as well as knowledge discovery for medical research [57].

The key models are BioBERTSum, AlphaBERT, ClinicalBertSum, ChestXRayBERT, RadBERT, LF2BERT, and DAMEN.

To alleviate the problems of biomedical literature summarization, which can have difficulties in learning sentence and document-level features, Du et al [57] proposed the first PLM for medical extractive summarization application called BioBERTSum. BioBERTSum captures a domain-aware token and sentence-level context by using a sentence position embedding mechanism that inserts structural information into a vector representation. It achieved a ROUGE-L score of 0.68, outperforming standard BERT models. AlphaBERT [60] proposed a diagnostic summary extractive model using a character-level token to reduce the model size and achieved a ROUGE-L score of 0.693, reducing the burden of physicians in the emergency department regarding reading complex discharge notes of patients.

To better use clinical notes, ClinicalBertSum [118] used the ClinicalBERT, SciBERT, and BertSum models during the fine-tuning and summarization process to automatically extract summaries from clinical abstracts. Similarly, ChestXRayBERT used BERT to perform an automatic abstractive summarization on radiology reports [62], with ROUGE-1 scores of 0.70 and 0.73, respectively. RadBERT [56], which was fine-tuned for radiology report summarization, achieved 10% fewer annotated sentences during the training, demonstrating the benefit of domain-specific pretraining to increase the overall performance.

LF2BERT [63] applied a Longformer neural network and BERT in an encoder-decoder framework to process longer sequence inputs and performed better than human summarization, according to doctors' evaluations. DAMEN [59] used BERT together with BART to discriminate important topic-related sentences in summarization, outperforming previous methods to summarize multiple medical literature via the token probability distribution method. The proposed probabilistic method selected only related significant chunks of information and then provided the probabilities of the tokens within the chunk, rather than the sentence level, to effectively reduce redundancy. Moreover, to overcome the long sequence issue, Li et al [58] comparably proposed Clinical-Longformer and Clinical-Big Bird pretrained on the Longformer [58] and Big Bird [119] models, respectively. Both proposed models used sparse attention mechanisms and linear level sequence lengths to mitigate memory consumption, thus increasing long-term dependency to train extensive clinical notes.

Recent development has also introduced MEDVOC [64], which uses GPT architecture to improve the adaptation of vocabulary in medical texts. By efficiently reducing fine-tuning time, MEDVOC achieves competitive performance, with ROUGE scores of 51.49, 47.54, and 19.51 across different datasets, such as PubMed, BioASQ, and EBM, respectively.

Text Classification

The medical text classification task categorizes medical text datasets into predefined categories based on the content and context within the text [120]. Disease classification, medical image classification, drug classification, and sentiment analysis are some of the standard text classification applications in health care.

The key models are jpCR+jpW, BioMed-RoBERTa, ClinicalBERT, Mask-BERT, KG-MTT-BERT, EduDistilBERT, PathologyBERT, KD distilledBERT, MatSciBERT, and Bioformer.

Wada et al [120] proposed a BERT model, jpCR+jpW, that uses a classification method. The method pretrains the medical BERT model once following the up-sampling step of domain-specific word amplification. This is done to achieve better performance on a smaller medical corpus. Similarly, BioMed-RoBERTa [70] used the RoBERTA model and applied a domain and task-adaptive pretraining strategy with a simple data selection approach for domain-specific classification. By pretraining on domain-specific and unlabeled data, the model achieved 87% accuracy. ClinicalBERT [67] represents clinical notes effectively, with a word similarity accuracy of 90% to generate visualized and interpretable embeddings for capturing semantic associations between clinical texts.

Yogarajan et al [121] suggested applying multi-labels (eg, using more than 300 labels for longer documents) to enhance the performance of medical classification tasks. Furthermore, to solve the imbalance class problems, Rodrawangpai et al [122] demonstrated a framework of adding normalization layers and dropout to BERT-based models, which improved the classification performance by 4% on data that included imbalance target labels. Similar efforts have been made by Nguyen [50] to address label-abandoning problems in medical abstract classification. The author proposed that a BERT model with label attention in the fine-tuning process raised the F1 score by 0.3 and supported the explainability of the prediction results. Learning is difficult with insufficient labeled data in a low-resource experiment setting. To alleviate this problem, Mask-BERT [71] proposed a framework for few-shot learning, where the mask is applied to the input text and enables the

gathering of more definitive tokens. Masked learning leads to filtered results on anchored samples from the data being used for representation, increasing the robustness of the output features.

Several language-specific models have been developed, including RuBioBERT and RuBioRoBERTa [123] for Russian text, BERTurk [124] for Turkish text, BioGottBERT for German text [88], and a Spanish text model [125], demonstrating the applicability of BERT-based models beyond English. Moreover, various disease-specific classification models have been developed. For example, PathologyBERT [126] is a pretrained masked language model used for classifying the severity of breast cancer diagnoses, raising the importance of applying domain-specific tokenization. KD distilled_BERT [127] is a response-embedded knowledge distillation framework that used pretrained BERT for depression classification and achieved a high accuracy of 97%. MatSciBERT [68] presents a biomedical domain-specific classification model on abstracts of the literature with binary classification application. The model extracted the context of the embeddings alongside the topic and had 2.75% higher accuracy than SciBERT.

To overcome the over-fitting and dimensionality problems for extracting numerous features in the text classification task, AFKF [128] proposed a fusion block with Kalman filters onto features of EMRs. This led to a 20% increase in accuracy compared with previous models. Likewise, to classify the features in EMRs, BERT-MSA [129] showed that a multilayered self-attention mechanism improved accuracy in obtaining relevant features. EduDistilBERT [130] demonstrated that adapting a smaller BERT model with limited parameter usage increases overall performance by 95% while reducing the computation cost. Another BERT-based fusion approach by Al-Garadi et al [131] explored architecture to fuse BERT, ALBERT, and RoBERTa model probabilities using a naive Bayes classifier, achieving an F1 score of 0.67 in classifying medication abuse texts.

Recently, the Bioformer [69] model demonstrated a 60% reduced model size and a 2- to 3-fold increase in performance speed. The model used a whole-word masking approach with 15% masking, which provided contextual information. However, KG-MTT-BERT [72] raised a question on limitations for multi-type clinical text classification. Concatenating numerous texts may be more efficient in developing relevant contextual information, and using only BERT may misplace crucial details. Therefore, the model extended the BERT model with a knowledge graph during fine-tuning, demonstrating effective handling in classifying patients into diagnosis-related groups.

Although the models were pretrained on the BERT model, Gao et al [132] showed that BERT-structured models did not gain better accuracy on clinical classification tasks, such as classifying discharge summaries or pathology reports, compared with nontransformer language models. Gao et al asserted that, in addition to the knowledge obtained through the entities, grammar patterns should also play a role in the model's mechanism. Furthermore, beyond the applications mentioned above, text classification can be used with other tasks. For example, Wang et al [133] applied a question-answering task along with the classification task by using the BERT model to classify texts in question inputs from patient inquiries regarding their symptoms.

Recent advancements in text classification models include TransformEHR [73], which uses BERT and longitudinal EHR data for clinical disease prediction, achieving area under the receiver operating characteristic curve and area under the precision-recall curve scores of 81.95 and 78.64, respectively. MeDa-BERT [74] tailored embeddings for Danish medical text, with accuracy ranging from 86.7% to 97.1%. SCHOLARBERT [75] leveraged public resource-driven datasets for scientific NLP, obtaining an F1 score of 85.49%. MediGPT [76] improved medical text classification tasks, with accuracy and F1 scores of 90.0% and 88.7%, respectively, showing a 22.3% performance gain over traditional methods.

Sentiment Analysis

The sentiment analysis task captures and identifies expressions and opinions [134] in medical contexts, including clinical notes, social media posts related to medicine, or patient feedback. For instance, sentiment analysis can capture the perception of people expressed in social media during the COVID-19 outbreak [135-137]. Emotions, such as positive, neutral, and negative sentiments, expressed by the public dominated during the pandemic [138]. Additionally, multi-label sentiment classification proved that the BERT model provided better performance compared with the LSTM model [139]. Moreover, the opinions of patients and physicians can be used to describe the symptoms and diagnosis to facilitate the decision-making process and support the decisions in clinical patterns [77]. The primary goal of sentiment analysis in health care is to provide insights into patient experiences, such as attitudes toward health care services and overall medical experience satisfaction. It not only assists patients but also supports clinicians to identify any underlying issues in patient care.

The key models are MentalBERT, MeentalRoBERTa, SINA-BERT, SGTN, RedBERT, T-BERT, AKI-BERT, TopicBERT, TweetBERT, and BelabBERT.

In mental health, patients' written texts have become a valuable source for supporting hypotheses and providing insights into the emotions expressed by patients [85,140]. While more research using PLMs needs to be conducted in this field, MentalBERT, MentalRoBERTa [77], PsychBERT [140], and belabBERT [85] applied mental health texts and achieved sentiment classification accuracies of 75%, 86%, and 90%, respectively. Additionally, transformer language models have been studied for sentiment analysis in languages other than English. Some language models are being developed to accommodate the unique structure and characteristics of different languages.

To achieve an effective model adaptation, researchers have explored HeBERT and HebEMO [141] for Hebrew, AraBERT and MARBERT [142] for Arabic, SINA-BERT [78] for Persian, and Fine-tuned BERT [143] for Chinese. However, few studies have conducted disease-specific sentiment analysis. RedBERT [80] involved a sentiment model for COVID-19, where BERT

was used for classifying sentiments of Reddit comments to grasp insights into the pandemic. Mao et al [82] proposed AKI-BERT, where the model was developed to support the early prediction of acute kidney injury.

Social media data are often used as a source for medical sentiment analysis as they are more informal and conversational, making them useful for modeling the nuances of language models. The COVID-TWITTER-BERT model [144], which was pretrained on Twitter messages regarding COVID-19, showed improved performance on COVID-19–related datasets. In particular, TweetBERT [84] exhibited improved performance on COVID-19–related and biomedical datasets. TwitterBERT was evaluated on 12 different biomedical datasets and outperformed previous BERT models, such as SciBERT [113] and BERT [145].

Comparably, TopicBERT [83], a memory-efficient BERT model, fine-tuned and enhanced sentimental analysis performance, and a complementary topic framework was applied to improve its performance. Beyond the proposed frameworks presented above, AlBadani et al [79] proposed a graph transformer model, SGTN, which used BERT to pretrain node embeddings and aggregated neighboring information to efficiently learn sentiments. It showed 5% improvement over baseline models.

Named Entity Recognition

The NER task identifies the named entities in unstructured text data. In health care, NER is used to automatically extract and define relevant medical entities, including diseases, medications, procedures, and other clinical concepts, from medical texts in research papers or EMRs [146,147]. The common applications of NER in medicine are as follows: (1) identify and analyze medical entities and relationships in medical literature to support biomedical findings [148]; (2) extract patient data, such as diagnosis, medication, laboratory results, and physical measurements, from EMRs to improve the decision-making for clinicians and the overall care [149]; and (3) extract and categorize data from medical claims and hospital admission and discharge data to improve health care management and resource allocation [150,151].

The key models are Bio+ClinicalBERT, Med-BERT, G-BERT, BioALBERT, GatorTron, ELECTRAMed, CamemBERT, BioGottBERT, Ra-RC, and RG-FLAT-CRF.

Numerous language models have been developed to implement EMR or EHR data for NER tasks in the clinical field. Bio+Clinical BERT [61] achieved superior results in clinical texts, with an F1 score of 83%. While the ClinicalBERT and Clinical BioBERT models were trained on EHRs, the Bio+Clinical BERT model did not perform well on deidentification text. Other models, such as G-BERT, also used EHR data to propose a language model that combined graph neural networks and BERT for representing medication information and predicting drug recommendations [87]. MedGPT [96] effectively processed noises by organizing medical text in multi-step procedures. In the first stage of the proposed model, unstructured data were converted into a standardized ontology using NER+L. Then, GPT was used for forecasting diagnosis events.

Moreover, studies attempted to tackle problems regarding representing and learning long medical entities [94,152]. Liu et al [152] proposed Med-BERT, using a Span-FLAT method for longer medical entities, and it achieved an F1 score of 84%. By contrast, the BIBC model built by Yang et al [94] captured both local and global sequence features to efficiently solve long text input issues. Additionally, models were trained on an extensive collection of biomedical texts to overcome the limited amount of training data. For example, BioALBERT and GatorTron attempted to develop a large medical language model [65,93]. BioALBERT used vocabulary specifically tailored to the biomedical domain and applied the ALBERT structure. On the other hand, GatorTron used the byte pair encoding algorithm and was pretrained on the GPT model to scale up the language model up to 8.9 billion parameters, showing 9.6% accuracy improvement. Furthermore, the datasets in the medical domain face the challenge of not only limited training data but also low-quality labeled training data. Therefore, multi-task learning was presented by Khan et al [92], and the slot tagging problem was approached with MT-BioNER, a multi-task transformer-based model that enhanced memory performance and time efficiency in slot tagging, with 10% better performance than single-task models.

While recent studies have heavily relied on BERT-based structures, transformer models used other PLMs for improving NER tasks. ELECTRAMed [91] proposed an ELECTRA-based model for the biomedical domain, which reduced the sequence length and training phases. Additionally, many models have focused on multilingualism, including the CamemBERT, BioGottBERT, Ra-RC, and RG-FLAT-CRF models [95,153,154], focusing on efficiently learning features in languages other than English, such as French, German, and Chinese.

More recent studies include HeartBERT [98], which emphasizes department-specific models, focusing on cardiology and achieving 74% accuracy. UMLS-KGI-BERT [99] introduced graph-based learning for clinical text extraction, with a precision of 85.05%. CardioBERTpt [100], which is specialized in Portuguese cardiology terms, improved NER performance, with an FL-score of 83%. Finally, TocBERT [101], which is fine-tuned on the MIMIC-III dataset, outperformed rule-based methods for segmenting discharge summaries, achieving an F1 score of 84.6%.

Discussion

Principal Findings

This study examined previous studies on transformer-based language models in the medical domain. We reviewed a total of 75 recently studied models that aligned with our inclusion criteria. The initial step of the method involves categorizing the models based on the tasks they perform, such as dialogue generation, question answering, summarization, text classification, sentiment analysis, and NER. Then, each study is analyzed based on the key findings, frameworks, pretraining

models used, and model names. Finally, the limitations of each task application are discussed. The use of transformer-derived language models in medicine has shown numerous advantages, such as high accuracy, language comprehension, automated diagnosis, adaptability, and efficiency. However, these models also face several challenges, including the lack of standardization, the need for domain-specific knowledge, limited annotated training data, safety concerns, interoperability and interpretability issues, integration complexities, ethical considerations, and evaluation issues. In this discussion section, we explore key limitations and future directions of models in terms of each task and the generalizability of the explored models across different health care settings and population considerations. Model challenges, their potential solutions, and the future of natural language models in the medical domain will be discussed.

Specific Task-Based Challenges and Future Directions

Dialogue Generation

Dialogue generation models like DialoGPT and ChatDoctor face challenges such as handling the complexity and specificity of medical terminology, ensuring data privacy, and providing accurate, contextually relevant, and empathetic responses. Privacy and security issues are critical since these models deal with sensitive patient information. The risk of privacy, chances of errors, ethical constraints, and security issues remain to be addressed. The challenge comes from the specificity and complexity of medical terminology, although the medical dialogue system certainly should provide only accurate and informative knowledge tailored to the level of expertise of the end user. Therefore, human experts need to conduct regular risk and security audits.

The following suggestions are made to further improve medical dialogue research. First, continuous learning and training of the dialogue system are necessary to incorporate up-to-date knowledge for users. Additionally, language translation could be integrated into the dialogues to enable universal access to data and promote a more profound exchange of insights without language barriers. Moreover, chatbots [107] should be integrated in a real medical setting to reduce medical costs and physician burdens. Proper and accurate usage of the dialogue system may assist patients in navigating through the vast amount of freely available online data, finding correct information, and avoiding falsified or unsolved answers. Lastly, automated data augmentation techniques can be used to create unbiased dialogues. These suggestions can lead to further advancements in medical dialogue research, leading to more efficient communication between patients and medical professionals.

Question Answering

Medical question-answering systems like BioBERT and UmlsBERT struggle with the complexity of medical terminology for nonexperts in the medical field. Patients who are experiencing an illness may find it difficult to filter and search for relevant information. These models need to handle diverse linguistic data and adapt regional variations in medical practices. One approach to addressing these limitations is to integrate multilingual models to handle questions in various languages. Another approach is the incorporation of region-specific medical data to improve model generalizability and accuracy. Further, enhancing the ability to integrate summarization tasks on top of the question-answering system may provide comprehensive responses. However, such a multi-task system requires several human experts to evaluate the provided answers in order to judge the task performance accurately.

Summarization

Medical summarization is a crucial application in language model tasks to facilitate the hospital's process and significantly reduce the workload and burnout of clinicians. However, challenges emerge due to the complexity of health care terminologies and the need for expert knowledge to comprehend them. The ability to achieve concise and faithful summaries is critical for avoiding physician burnout and patient dissatisfaction. Models like BioBERTSum and ClinicalBERTSum face challenges in learning sentence and document-level features, handling complex medical terminologies, and ensuring summaries are concise and accurate. The risk of physician burnout due to extensive documentation can be mitigated by effective summarization. Future work can focus on developing a system of human expert assessments to validate the summarization quality. Additionally, combining extractive methods and abstractive summarization methods is suggested. A fine-tuned summarization model for a particular task should consider tense information and personal information. We recommend building an ensemble method to improve pretraining and fine-tuning datasets for summarization effectiveness. Medical summarization is a crucial application in language model tasks to facilitate the hospital's process and significantly reduce the workload and burnout of clinicians.

Text Classification

Improving the accuracy and effectiveness of classification tasks poses several challenges and limitations that need to be addressed. Models, such as BioMed-RoBERTa and ClinicalBERT, need to address issues related to class imbalance, limited annotated data, and the complexity of medical terminologies. Limited training data, for instance, can be addressed by collaborating with different institutions to gather various information options in vocabulary usage and text structure, and high-quality annotated data can thus be developed. Ambiguity, variation, concept drift, data privacy, language complexity, and class imbalance can be addressed by employing domain-specific approaches, and pretraining language models can be leveraged on similar datasets. Domain-specific approaches can resolve ambiguity issues and achieve active learning to reduce the reliance on large volumes of labeled data. These strategies will facilitate better model performance results.

Sentiment Analysis

Despite previous research results, medical sentiment analysis remains a challenging task due to personalized information required to accurately measure meaning and interpret emotions in context. MentalBERT and RedBERT, for instance, need to accurately interpret emotions in medical contexts, handle personalized information, and manage the complexity of evaluating representations in the biomedical domain [155].

Organizing emotions in context requires sentiments, including sarcasm, emojis, and misspelled words, which create subjectivity, as noted by Brezulianu et al [156].

These limitations can be overcome by defining emotional polarity for annotations and integrating cultural, economic, and medical contexts into the model. Future research should consider using a domain-specific sentiment dataset, adapting the specific medical source (despite the lack of an available dataset, mostly from a single source), creating highly effective and defined labels in data, performing analysis based on the context, building both the explicit and implicit sentiment lexicon, and addressing the lack of a mental health–related sentiment lexicon. By addressing these challenges, future work can develop more accurate and effective medical sentiment analysis language models.

Named Entity Recognition

NER downstream tasks are imperative to address its limitations. The limited annotated data in medical text datasets is a major challenge due to the high cost and time involved in labeling, resulting in limited labeled data for model training. The Bio+ClinicalBERT and Med-BERT models face challenges in handling limited annotated data, normalizing various terminologies, and ensuring accurate entity extraction across different medical texts.

Therefore, we suggest collaborative effort among health care providers, biomedical researchers, and computer engineering experts to develop effective and robust NER models. Improving the annotation algorithms and creating extensive and accurately labeled medical text datasets can significantly enhance the performance. Moreover, standardized clinical entities can prevent ambiguity arising from abbreviations and context. The use of transfer learning techniques and domain PLMs can be beneficial in addressing the limited annotated data issue. Developing domain-specific dictionaries and ontologies can aid in improving the model performance.

Generalizability Challenges

Health care systems vary widely in their practices, protocols, and terminologies. For instance, a model trained on data from the United States may not perform optimally in a health care setting in Asia or Europe due to differences in clinical infrastructure and settings. The availability of resources, such as EHRs and technological infrastructure, can also differ between urban and rural settings, and between developed and underdeveloped countries. This variability can significantly affect the implementation and performance of the models.

Moreover, patients from different ethnic and cultural backgrounds may present symptoms differently and may have varying health behaviors, and models need to account for these varying characteristics to avoid biases and ensure equitable health care delivery. Moreover, multilingual populations pose a challenge for language models trained predominantly on English language data. The inclusion of diverse linguistic data during model training can mitigate this issue with a language-specific pretraining stage followed by a shared fine-tuning stage to improve the model's applicability across different regions.

By incorporating diverse datasets during training, language models can support personalized medicine initiatives. This involves tailoring medical treatments to individual patient characteristics, leading to more effective and efficient care. Developing adaptable models that can be fine-tuned with local data ensures scalability across different health care settings, to address regional variations in medical practices and patient demographics.

Standardizing Medical Data for Improved Model Performance

The quality and consistency of medical data may vary across health care settings. Models trained on high-quality standardized data may not perform as well when applied to settings with less structured and lower quality data. The lack of standardized terminologies in medical texts, which encompass a vast array of terminologies from disease-specific to domain-specific language, is a notable challenge. Currently available datasets often have a restricted range of medical entities, posing difficulties in accurately extracting relevant entities.

To address this, we suggest creating standardized clinical entities. This would enable the normalization of different names or abbreviations to accurately normalize entities for standard medical terminology, thereby improving data consistency and model performance. Building standardized forms that are widely adopted and available in multiple languages will facilitate standardized medical learning. Additionally, developing domain-specific models has proven effective in enhancing model performance. For instance, an open-source package for detecting clinical entities from medical texts, which can recognize risk factors, medications, and diagnoses, can be developed to support this initiative.

The fuel of building and training language models is data. Collecting accurate information and precisely fabricating the data design during the preprocessing step is crucial. The challenges in creating such quality data for medical language models include a lack of key annotation and limited training data. Annotating medical text is time-consuming and costly, resulting in limited labeled data for training models. First, using diagnosis codes on weak supervision for training labels is suggested. Second, the pipeline should support the automated retrieval of datasets and multiple types of clinical entities to enable the preservation of annotation relationships across different languages. The automated retrieval of datasets and the development of speedy and supportive algorithms can aid in data integration and preprocessing. In addition to technical solutions, emphasizing the importance of multidisciplinary collaboration can significantly enhance the development and implementation of these models. By integrating expertise from various fields, we can overcome challenges, develop innovative solutions, and further advance the field of AI in health care. Collaborative efforts among data scientists, clinicians, bioinformaticians, and ethicists are crucial for building robust, reliable, and ethically sound models.

Ethical Considerations

Interoperability and cybersecurity pose significant challenges in medicine. EMRs and clinical decision support systems often

have difficulty interacting with each other, leading to inefficiencies in patient care. To overcome these challenges, it is important to develop strategies focused on informed consent, safety, transparency, and algorithmic fairness for bias prevention. Ensuring that patients provide informed consent for the use of their data is critical. This involves informing patients about the use of their data, the benefits and risks, and their rights to withdraw consent at any time. This process upholds patient autonomy and enhances trust in AI systems.

Safety and transparency are fundamental to the ethical deployment of AI models in health care. Models handling sensitive patient data and providing clinical recommendations, such as BioBERTSum, must be rigorously validated and continuously monitored to detect and rectify errors promptly. Transparency can be achieved by making algorithms and decision-making processes understandable to users. This includes documenting how models are trained, the types of data used, and the underlying mechanisms of the algorithms.

Ensuring algorithmic fairness is crucial to prevent biases in AI models, which could lead to unequal treatment of patients. AI models trained on biased datasets can perpetuate existing disparities in health care. For example, models must include diverse and representative data to avoid underrepresentation of certain populations, ensuring fairness and accuracy across different groups such as ethnicity, cultural background, gender, and age. By addressing these ethical considerations within mitigated guidelines, we can ensure the reliability of transformer language models in medicine to improve overall health care while preserving fairness.

Evaluation Metrics

Furthermore, the rapid evolution of medical knowledge poses a challenge for language models to adapt and remain up to date with innovative discoveries. Ensuring the interpretability of language models is also crucial to address the trust issue and support the decision-making process. To evaluate medical language models, multiple metrics, including the F1 score, Biomedical Language Understanding Evaluation, Biomedical Language Understanding & Reasoning Benchmark (BLURB), and Chinese Biomedical Language Understanding Evaluation [157], should be used to overcome unbalanced performance issues.

Conclusion

We presented a comprehensive survey of task-specific transformer-derived models employed for diverse medical tasks, demonstrating their significant potential in the medical domain. Numerous studies have highlighted their capabilities in improving health outcomes, extending beyond disease prediction and medical classification studies. Our work clearly delineates the applications of transformer-based language models in various medical tasks such as dialogue generation, question answering, summarization, text classification, sentiment analysis, and NER. We identified innovative models and their unique contributions to the field. These findings distinguish our work from existing literature by providing a detailed, task-specific analysis of transformer-based models in health care.

Despite the promising advancements, several challenges must be addressed to develop effective models. These include standardization, limited annotated data, interoperability, and ethical considerations. To overcome these challenges, it is crucial to emphasize multidisciplinary collaboration. Future research should investigate transformer models that incorporate visual or audio data sources to provide a more comprehensive understanding of medical contexts.

Developing models that support patients' experiences and assist health care practitioners in focusing solely on critical tasks by providing evidence-based recommendations and identifying potential diagnostic and treatment options can remarkably improve patient care. AI-driven tools rationalize administrative tasks, reduce paperwork, and improve workflow efficiency, eventually saving time for health care providers. Further, policymakers can leverage insights from transformer-based models to inform health care policies and allocate resources more effectively, ensuring equitable health care delivery.

This review solely focused on transformer language models that used text data. While the findings are promising, the applicability of these models may vary across different medical settings and populations. Our findings highlight the transformative potential of transformer-based language models in the medical field. By addressing the identified challenges and focusing on innovative research directions, the health care domain can advance significantly. We encourage researchers to build upon our work, address these challenges, and explore new frontiers in medical AI to improve patient care and clinical decision-making.

Acknowledgments

This work was supported by the Korea Medical Device Development Fund grant funded by the Korean government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, Republic of Korea, and the Ministry of Food and Drug Safety) (project number: 202012B06) and by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR21C0198). The funders of the study had no role in the study design, data collection, data analysis, data interpretation, or manuscript writing. All authors had full access to the data in the study and accepted the responsibility to submit it for publication.

Data Availability

Supporting data are available from the corresponding author upon reasonable request.

Authors' Contributions

HNC designed the study, conducted the literature review, created the figures and tables, and wrote the manuscript. IA, HG, HJK, YK, JS, HC, MK, JH, GK, and SP reviewed all records. TJJ and YK created the search strategy and supervised the overall project. All authors reviewed and approved the submission of the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 881 KB - medinform_v12i1e49724_app1.pdf]

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention Is All You Need. arXiv. 2017. URL: https://arxiv.org/abs/1706.03762 [accessed 2024-10-26]
2. GPT-4 Technical Report. OpenAI. 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf [accessed 2024-10-26]
3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics; June 3-5, 2019; Minneapolis, Minnesota URL: https://aclanthology.org/N19-1423/
4. Ji Y, Zhou Z, Liu H, Davuluri R. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 2021 Aug 09;37(15):2112-2120 [FREE Full text] [doi: 10.1093/bioinformatics/btab083] [Medline: 33538820]
5. Pandey C. redBERT: A Topic Discovery and Deep Sentiment Classification Model on COVID-19 Online Discussions Using BERT NLP Model. medRxiv. 2021. URL: https://www.medrxiv.org/content/10.1101/2021.03.02.21252747v1.full.pdf [accessed 2024-10-26]
6. Iroju OG, Olaleke JO. A systematic review of natural language processing in healthcare. IJITCS 2015 Jul 08;7(8):44-50. [doi: 10.5815/ijitcs.2015.08.07]
7. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen G. Natural language processing in medicine: A review. Trends in Anaesthesia and Critical Care 2021 Jun;38:4-9 [FREE Full text] [doi: 10.1016/j.tacc.2021.02.007]
8. Adyashreem M, Rajatha K, Rakesh K. Survey on NLP Techniques in Biomedical field. International Journal of Scientific Research in Computer Science Applications and Management Studies 2018:7.
9. Wang S, Ren F, Lu H. A review of the application of natural language processing in clinical medicine. 2018 Presented at: 13th IEEE Conference on Industrial Electronics and Applications (ICIEA); May 31-June 02, 2018; Wuhan, China. [doi: 10.1109/ICIEA.2018.8398172]
10. Liu Z, He M, Jiang Z, Wu Z, Dai H, Zhang L, et al. Survey on natural language processing in medical image analysis. Zhong Nan Da Xue Xue Bao Yi Xue Ban 2022 Aug 28;47(8):981-993. [doi: 10.11817/j.issn.1672-7347.2022.220376] [Medline: 36097765]
11. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform 2021 Mar 02;28(1):e100262 [FREE Full text] [doi: 10.1136/bmjhci-2020-100262] [Medline: 33653690]
12. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021 Jun 03;21(1):179 [FREE Full text] [doi: 10.1186/s12911-021-01533-7] [Medline: 34082729]
13. Zhou B, Yang G, Shi Z, Ma S. Natural language processing for smart healthcare. IEEE Rev Biomed Eng 2024;17:4-18. [doi: 10.1109/RBME.2022.3210270] [Medline: 36170385]
14. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bioinform Adv 2023;3(1):vbad001 [FREE Full text] [doi: 10.1093/bioadv/vbad001] [Medline: 36845200]
15. Yang K. Transformer-based Korean Pretrained Language Models: A Survey on Three Years of Progress. arXiv. 2021. URL: https://arxiv.org/abs/2112.03014 [accessed 2024-10-26]
16. Lin T, Wang Y, Liu X, Qiu X. A Survey of Transformers. arXiv. 2021. URL: https://arxiv.org/abs/2106.04554 [accessed 2024-10-26]
17. Chitty-Venkata KT, Emani M, Vishwanath V, Somani AK. Neural architecture search for transformers: A survey. IEEE Access 2022;10:108374-108412. [doi: 10.1109/access.2022.3212767]
18. Gillioz A, Casas J, Mugellini E, Khaled O. Overview of the Transformer-based Models for NLP Tasks. In: Proceedings of the 2020 Federated Conference on Computer Science and Information Systems. 2020 Presented at: 2020 Federated Conference on Computer Science and Information Systems; September 6-9, 2020; Sofia, Bulgaria. [doi: 10.15439/2020F20]

19. Han X, Wang Y, Feng J, Deng C, Chen Z, Huang Y, et al. A survey of transformer-based multimodal pre-trained modals. Neurocomputing 2023 Jan;515:89-106 [FREE Full text] [doi: 10.1016/j.neucom.2022.09.136]
20. Greco C, Simeri A, Tagarelli A, Zumpano E. Transformer-based language models for mental health issues: A survey. Pattern Recognition Letters 2023 Mar;167:204-211 [FREE Full text] [doi: 10.1016/j.patrec.2023.02.016]
21. Albalawi Y, Nikolov NS, Buckley J. Pretrained transformer language models versus pretrained word embeddings for the detection of accurate health information on Arabic social media: Comparative study. JMIR Form Res 2022 Jun 29;6(6):e34834 [FREE Full text] [doi: 10.2196/34834] [Medline: 35767322]
22. Kalyan K, Rajasekharan A, Sangeetha S. AMMU: A survey of transformer-based biomedical pretrained language models. J Biomed Inform 2022 Feb;126:103982 [FREE Full text] [doi: 10.1016/j.jbi.2021.103982] [Medline: 34974190]
23. Shamshad F, Khan S, Zamir S, Khan M, Hayat M, Khan F, et al. Transformers in Medical Imaging: A Survey. arXiv. 2022. URL: https://arxiv.org/abs/2201.09873 [accessed 2024-10-26]
24. Scoping reviews: what they are and how you can do them. Cochrane Training. URL: https://training.cochrane.org/resource/scoping-reviews-what-they-are-and-how-you-can-do-them [accessed 2024-10-26]
25. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. OpenAI. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2024-10-26]
26. Varshney D, Zafar A, Behera N, Ekbal A. Knowledge grounded medical dialogue generation using augmented graphs. Sci Rep 2023 Feb 27;13(1):3310 [FREE Full text] [doi: 10.1038/s41598-023-29213-8] [Medline: 36849466]
27. Yuan H, Yuan Z, Gan R, Zhang J, Xie Y, Yu S. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. In: Proceedings of the 21st Workshop on Biomedical Language Processing. 2022 Presented at: 21st Workshop on Biomedical Language Processing; May 26, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.bionlp-1.9]
28. Zhao Y, Li Y, Wu Y, Hu B, Chen Q, Wang X, et al. Medical Dialogue Response Generation with Pivotal Information Recalling. In: KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022 Presented at: 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; August 14-18, 2022; Washington, DC, USA. [doi: 10.1145/3534678.3542674]
29. Chen Z, Liu Y, Chen L, Zhu S, Wu M, Yu K. OPAL: Ontology-aware pretrained language model for end-to-end task-oriented dialogue. Transactions of the Association for Computational Linguistics 2023;11:68-84 [FREE Full text] [doi: 10.1162/tacl_a_00534]
30. Liang K, Wu S, Gu J. MKA: A scalable medical knowledge-assisted mechanism for generative models on medical conversation tasks. Comput Math Methods Med 2021;2021:5294627 [FREE Full text] [doi: 10.1155/2021/5294627] [Medline: 34976109]
31. Compton R, Valmianski I, Deng L, Huang C, Katariya N, Amatriain X, et al. MEDCOD: A medically-accurate, emotive, diverse, and controllable dialog system. Proceedings of Machine Learning Research 2021;158:110-128 [FREE Full text]
32. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. arXiv. 2023. URL: https://arxiv.org/abs/2303.14070 [accessed 2024-10-26]
33. Tang C, Zhang H, Loakman T, Lin C, Guerin F. Terminology-aware Medical Dialogue Generation. arXiv. 2022. URL: https://arxiv.org/abs/2210.15551 [accessed 2024-10-26]
34. Zeng Y, Nie JY. Multi-Domain Dialogue State Tracking based on State Graph. arXiv. 2020. URL: https://arxiv.org/abs/2010.11137 [accessed 2024-10-26]
35. Suri H, Zhang Q, Huo W, Liu Y, Guan C. MeDiaQA: A Question Answering Dataset on Medical Dialogues. arXiv. 2021. URL: https://arxiv.org/abs/2108.08074 [accessed 2024-10-26]
36. Phan L, Anibal J, Tran H, Chanana S, Bahadroglu E, Peltekian A, et al. SciFive: a text-to-text transformer model for biomedical literature. arXiv. 2021. URL: https://arxiv.org/abs/2106.03598 [accessed 2024-10-26]
37. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inform Assoc 2024 Sep 01;31(9):1833-1843 [FREE Full text] [doi: 10.1093/jamia/ocae045] [Medline: 38613821]
38. Zhang H, Chen J, Jiang F, Yu F, Chen Z, Chen G, et al. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023:10859-10885.
39. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
40. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022 Nov 19;23(6):bbac409 [FREE Full text] [doi: 10.1093/bib/bbac409] [Medline: 36156661]
41. Shin HC, Zhang Y, Bakhturina E, Puri R, Patwary M, Shoeybi M, et al. BioMegatron: Larger Biomedical Domain Language Model. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020

Presented at: 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); November 16-20, 2020; Online. [doi: 10.18653/v1/2020.emnlp-main.379]
42. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021 May 20;4(1):86 [FREE Full text] [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
43. Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022 Presented at: 60th Annual Meeting of the Association for Computational Linguistics; May 22-27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.551]
44. Michalopoulos G, Wang Y, Kaka H, Chen H, Wong A. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021 Presented at: 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics; June 6-11, 2021; Online. [doi: 10.18653/v1/2021.naacl-main.139]
45. Zhang T, Cai Z, Wang C, Qiu M, Yang B, He X. SMedBERT: A Knowledge-Enhanced Pre-trained Language Model with Structured Semantics for Medical Text Mining. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021 Presented at: 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; August 2021; Online. [doi: 10.18653/v1/2021.acl-long.457]
46. Yang T, Sucholutsky I, Jen K, Schonlau M. exKidneyBERT: a language model for kidney transplant pathology reports and the crucial role of extended vocabularies. PeerJ Comput Sci 2024;10:e1888. [doi: 10.7717/peerj-cs.1888] [Medline: 38435545]
47. Mitchell JR, Szepietowski P, Howard R, Reisman P, Jones JD, Lewis P, et al. A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT Network): Development study. J Med Internet Res 2022 Mar 23;24(3):e27210 [FREE Full text] [doi: 10.2196/27210] [Medline: 35319481]
48. Trieu H, Miwa M, Ananiadou S. BioVAE: a pre-trained latent variable language model for biomedical text mining. Bioinformatics 2022 Jan 12;38(3):872-874 [FREE Full text] [doi: 10.1093/bioinformatics/btab702] [Medline: 34636886]
49. Khare Y, Bagal V, Mathew M, Devi A, Priyakumar UD, Jawahar CV. MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. arXiv. URL: https://arxiv.org/abs/2104.01394 [accessed 2024-10-26]
50. Nguyen N, Ha P, Nguyen L, Van Nguyen K, Nguyen N. SPBERTQA: A Two-Stage Question Answering System Based on Sentence Transformers for Medical Texts. In: Memmi G, Yang B, Kong L, Zhang T, Qiu M, editors. Knowledge Science, Engineering and Management. KSEM 2022. Lecture Notes in Computer Science, vol 13369. Cham: Springer; 2022:371-382.
51. Luo Y, Zhang J, Fan S, Yang K, Wu Y, Qiao M, et al. BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. arXiv. 2023. URL: https://arxiv.org/abs/2308.09442 [accessed 2024-10-26]
52. Toma A, Lawler P, Ba J, Krishnan R, Rubin B, Wang B. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. arXiv. 2023. URL: https://arxiv.org/abs/2305.12031 [accessed 2024-10-26]
53. Han T, Adams L, Papaioannou JM, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca - An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv. 2023. URL: https://arxiv.org/abs/2304.08247 [accessed 2024-10-26]
54. Singhal K, Azizi S, Tu T, Mahdavi S, Wei J, Chung H, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
55. Chen Z, Cano A, Romanou A, Matoba K, Salvi F, Pagliardini M, et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv. URL: https://arxiv.org/abs/2311.16079 [accessed 2024-10-26]
56. Yan A, McAuley J, Lu X, Du J, Chang E, Gentili A, et al. RadBERT: Adapting transformer-based language models to radiology. Radiol Artif Intell 2022 Jul;4(4):e210258 [FREE Full text] [doi: 10.1148/ryai.210258] [Medline: 35923376]
57. Du Y, Li Q, Wang L, He Y. Biomedical-domain pre-trained language model for extractive summarization. Knowledge-Based Systems 2020 Jul;199:105964 [FREE Full text] [doi: 10.1016/j.knosys.2020.105964]
58. Li Y, Wehbe R, Ahmad F, Wang H, Luo Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. arXiv. 2022. URL: https://arxiv.org/abs/2201.11838 [accessed 2024-10-26]
59. Moro G, Ragazzi L, Valgimigli L, Freddi D. Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022 Presented at: 60th Annual Meeting of the Association for Computational Linguistics; May 22-27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.15]
60. Chen Y, Chen Y, Lin J, Huang C, Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): Development and performance evaluation. JMIR Med Inform 2020 Apr 29;8(4):e17787 [FREE Full text] [doi: 10.2196/17787] [Medline: 32347806]
61. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. arXiv. 2019. URL: https://arxiv.org/abs/1904.03323 [accessed 2024-10-26]

62. Cai X, Liu S, Han J, Yang L, Liu Z, Liu T. ChestXRayBERT: A pretrained language model for chest radiology report summarization. IEEE Trans Multimedia 2023;25:845-855. [doi: 10.1109/tmm.2021.3132724]
63. Yalunin A, Umerenkov D, Kokh V. Abstractive summarization of hospitalisation histories with transformer networks. arXiv. 2022. URL: https://arxiv.org/abs/2204.02208 [accessed 2024-10-26]
64. Balde G, Roy S, Mondal M, Ganguly N. MEDVOC: Vocabulary Adaptation for Fine-tuning Pre-trained Language Models on Medical Text Summarization. arXiv. 2024. URL: https://arxiv.org/abs/2405.04163 [accessed 2024-10-26]
65. Yang X, PourNejatian N, Shin H, Smith K, Parisien C, Compas C, et al. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. medRxiv. URL: https://www.medrxiv.org/content/10.1101/2022.02.27.22271257v1 [accessed 2024-10-26]
66. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
67. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv. 2019. URL: https://arxiv.org/abs/1904.05342 [accessed 2024-10-26]
68. Gupta T, Zaki M, Krishnan N. MatSciBERT: A materials domain language model for text mining and information extraction. npj Comput Mater 2022 May 03;8(1):A [FREE Full text] [doi: 10.1038/s41524-022-00784-w]
69. Fang L, Chen Q, Wei CH, Lu Z, Wang K. Bioformer: an efficient transformer language model for biomedical text mining. arXiv. 2023. URL: https://arxiv.org/abs/2302.01588 [accessed 2024-10-26]
70. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.740]
71. Liao W, Liu Z, Dai H, Wu Z, Zhang Y, Huang X, et al. Mask-guided BERT for Few Shot Text Classification. arXiv. 2023. URL: https://arxiv.org/abs/2302.10447 [accessed 2024-10-26]
72. He Y, Wang C, Zhang S, Li N, Li Z, Zeng Z. KG-MTT-BERT: Knowledge Graph Enhanced BERT for Multi-Type Medical Text Classification. arXiv. 2022. URL: https://arxiv.org/abs/2210.03970 [accessed 2024-10-26]
73. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nat Commun 2023 Nov 29;14(1):7857 [FREE Full text] [doi: 10.1038/s41467-023-43715-z] [Medline: 38030638]
74. Pedersen J, Laursen M, Vinholt P, Savarimuthu T. MeDa-BERT: A medical Danish pretrained transformer model. In: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). 2023 Presented at: 24th Nordic Conference on Computational Linguistics (NoDaLiDa); May 22-24, 2023; Tórshavn, Faroe Islands.
75. Hong Z, Ajith A, Pauloski J, Duede E, Chard K, Foster I. The Diminishing Returns of Masked Language Models to Science. In: Findings of the Association for Computational Linguistics: ACL 2023. 2023 Presented at: 61st Annual Meeting of the Association for Computational Linguistics: Industry Track; July 9-14, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.findings-acl.82]
76. Abu Tareq Rony M, Shariful Islam M, Sultan T, Alshathri S, El-Shafai W. MediGPT: Exploring potentials of conventional and large language models on medical data. IEEE Access 2024;12:103473-103487 [FREE Full text] [doi: 10.1109/ACCESS.2024.3428918]
77. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. arXiv. 2021. URL: https://arxiv.org/abs/2110.15621 [accessed 2024-10-26]
78. Taghizadeh N, Doostmohammadi E, Seifossadat E, Rabiee H, Tahaei M. SINA-BERT: A pre-trained Language Model for Analysis of Medical Texts in Persian. arXiv. 2021. URL: https://arxiv.org/abs/2104.07613 [accessed 2024-10-26]
79. AlBadani B, Shi R, Dong J, Al-Sabri R, Moctard O. Transformer-based graph convolutional network for sentiment analysis. Applied Sciences 2022 Jan 26;12(3):1316 [FREE Full text] [doi: 10.3390/app12031316]
80. Pandey C. redBERT: A topic discovery and deep sentiment classification model on COVID-19 online discussions using BERT NLP model. International Journal of Open Source Software and Processes 2021;12(3):32-47. [doi: 10.4018/IJOSSP.2021070103]
81. Palani S, Rajagopal P, Pancholi S. T-BERT - Model for Sentiment Analysis of Micro-blogs Integrating Topic Model and BERT. arXiv. 2021. URL: https://arxiv.org/abs/2106.01097 [accessed 2024-10-26]
82. Mao C, Yao L, Luo Y. A Pre-trained Clinical Language Model for Acute Kidney Injury. 2020 Presented at: 2020 IEEE International Conference on Healthcare Informatics (ICHI); November 30-December 03, 2020; Oldenburg, Germany. [doi: 10.1109/ichi48887.2020.9374312]
83. Chaudhary Y, Gupta P, Saxena K, Kulkarni V, Runkler T, Schütze H. TopicBERT for Energy Efficient Document Classification. In: Findings of the Association for Computational Linguistics: EMNLP 2020. 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing; November 16-20, 2020; Online. [doi: 10.18653/v1/2020.findings-emnlp.152]
84. Qudar M, Mago V. TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis. arXiv. 2020. URL: https://arxiv.org/abs/2010.11091 [accessed 2024-10-26]

85. Wouts J, de Boer J, Voppel A, Brederoo S, van Splunter S, Sommer I. belabBERT: a Dutch RoBERTa-based language model applied to psychiatric classification. arXiv. 2021. URL: https://arxiv.org/abs/2106.01091 [accessed 2024-10-26]
86. Li Y, Rao S, Solares J, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: Transformer for electronic health records. Sci Rep 2020 Apr 28;10(1):7155 [FREE Full text] [doi: 10.1038/s41598-020-62922-y] [Medline: 32346050]
87. Shang J, Ma T, Xiao C, Sun J. Pre-training of Graph Augmented Transformers for Medication Recommendation. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence AI for Improving Human Well-being. 2019 Presented at: Twenty-Eighth International Joint Conference on Artificial Intelligence AI for Improving Human Well-being; August 10-16, 2019; Macao. [doi: 10.24963/ijcai.2019/825]
88. Lentzen M, Madan S, Lage-Rupprecht V, Kühnel L, Fluck J, Jacobs M, et al. Critical assessment of transformer-based AI models for German clinical notes. JAMIA Open 2022 Dec;5(4):ooac087 [FREE Full text] [doi: 10.1093/jamiaopen/ooac087] [Medline: 36380848]
89. Davari MR, Kosseim L, Bui T. TIMBERT: Toponym Identifier For The Medical Domain Based on BERT. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020 Presented at: 28th International Conference on Computational Linguistics; December 8-13, 2020; Barcelona, Spain (Online).
90. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. arXiv. 2019. URL: https://arxiv.org/abs/1906.05474 [accessed 2024-10-26]
91. Miolo G, Mantoan G, Orsenigo C. ELECTRAMed: a new pre-trained language representation model for biomedical NLP. arXiv. 2021. URL: https://arxiv.org/abs/2104.09585 [accessed 2024-10-26]
92. Khan M, Ziyadi M, AbdelHady M. MT-BioNER: Multi-task Learning for Biomedical Named Entity Recognition using Deep Bidirectional Transformers. arXiv. URL: https://arxiv.org/abs/2001.08904 [accessed 2024-10-26]
93. Naseem U, Khushi M, Reddy V, Rajendran S, Razzak I, Kim J. BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition. 2021 Presented at: 2021 International Joint Conference on Neural Networks (IJCNN); July 18-22, 2021; Shenzhen, China. [doi: 10.1109/ijcnn52387.2021.9533884]
94. Yang L, Fu Y, Dai Y. BIBC: A Chinese named entity recognition model for diabetes research. Applied Sciences 2021 Oct 16;11(20):9653 [FREE Full text] [doi: 10.3390/app11209653]
95. Martin L, Muller B, Suárez P, Dupont Y, Romary L, de la Clergerie É, et al. CamemBERT: a Tasty French Language Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.645]
96. Kraljevic Z, Shek A, Bean D, Bendayan R, Teo J, Dobson R. MedGPT: Medical Concept Prediction from Clinical Narratives. arXiv. 2021. URL: https://arxiv.org/abs/2107.03134 [accessed 2024-10-26]
97. Li F, Jin Y, Liu W, Rawat B, Cai P, Yu H. Fine-tuning Bidirectional Encoder Representations From Transformers (BERT)-based models on large-scale electronic health record notes: An empirical study. JMIR Med Inform 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
98. Gwon H, Seo H, Park S, Kim Y, Jun T. HeartBERT : A language model pre-trained on anopen source dataset for cardiac text mining. Research Square. URL: https://www.researchsquare.com/article/rs-4137702/v1 [accessed 2024-10-26]
99. Mannion A, Schwab D, Goeuriot L. UMLS-KGI-BERT: Data-Centric Knowledge Integration in Transformers for Biomedical Entity Recognition. In: Proceedings of the 5th Clinical Natural Language Processing Workshop. 2023 Presented at: 5th Clinical Natural Language Processing Workshop; July 14, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.clinicalnlp-1.35]
100. Schneider E, Gumiel Y, de Souza J, Mukai L, Silva e Oliveira L, de Sa Rebelo M. CardioBERTpt: Transformer-based Models for Cardiology Language Representation in Portuguese. 2023 Presented at: 36th International Symposium on Computer-Based Medical Systems (CBMS); June 22-24, 2023; L'Aquila, Italy. [doi: 10.1109/cbms58004.2023.00247]
101. Saleh M, Baghdadi S, Paquelet S. TocBERT: Medical Document Structure Extraction Using Bidirectional Transformers. arXiv. 2024. URL: https://doi.org/10.48550/arXiv.2406.19526 [accessed 2024-10-26]
102. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. Papers With Code. URL: https://paperswithcode.com/paper/language-models-are-unsupervised-multitask [accessed 2024-10-26]
103. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing System. 2020 Presented at: 34th International Conference on Neural Information Processing System; December 6-12, 2020; Vancouver, BC, Canada. [doi: 10.5555/3495724.3495883]
104. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv. 2020. URL: https://arxiv.org/pdf/2005.14165 [accessed 2024-10-26]
105. Zhang Y, Sun S, Galley M, Chen YC, Brockett C, Gao X, et al. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-demos.30]
106. Krishna K, Pavel A, Schloss B, Bigham J, Lipton Z. Extracting Structured Data from Physician-Patient Conversations By Predicting Noteworthy Utterances. arXiv. 2020. URL: https://arxiv.org/abs/2007.07151 [accessed 2024-10-26]

107. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. JMIR Cancer 2021 Nov 29;7(4):e27850 [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
108. Parmar P, Ryu J, Pandya S, Sedoc J, Agarwal S. Health-focused conversational agents in person-centered care: a review of apps. NPJ Digit Med 2022 Feb 17;5(1):21 [FREE Full text] [doi: 10.1038/s41746-022-00560-6] [Medline: 35177772]
109. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv. 2023. URL: https://arxiv.org/abs/2302.13971 [accessed 2024-10-26]
110. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 2020;21:1-67 [FREE Full text]
111. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv. 2019. URL: https://arxiv.org/abs/1909.08053 [accessed 2024-10-26]
112. Cai Z, Zhang T, Wang C, He X. EMBERT: A Pre-trained Language Model for Chinese Medical Text Mining. In: Spaniol M, Sakurai Y, Chen J, editors. Web and Big Data. APWeb-WAIM 2021. Lecture Notes in Computer Science, vol 12858. Cham: Springer; 2021:242-257.
113. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; November 3-7, 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1371]
114. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. OpenAI. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [accessed 2024-10-26]
115. Zhou S, Zhang Y. DATLMedQA: A data augmentation and transfer learning based solution for medical question answering. Applied Sciences 2021 Nov 26;11(23):11251 [FREE Full text] [doi: 10.3390/app112311251]
116. Henderson M, Al-Rfou R, Strope B, Sung Y, Lukacs L, Guo R, et al. Efficient Natural Language Response Suggestion for Smart Reply. arXiv. 2017. URL: https://arxiv.org/abs/1705.00652 [accessed 2024-10-26]
117. Bolton E, Hall D, Yasunaga M, Lee T, Manning C, Liang P. BioMedLM. Stanford Center for Research on Foundation Models. URL: https://crfm.stanford.edu/2022/12/15/biomedlm.html [accessed 2024-10-26]
118. Lu M, Jin X, Wang Z. ClinicalBertSum: RCT Summarization by Using Clinical BERT Embeddings. Stanford CS224N Custom Project. URL: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1204/reports/custom/report29.pdf [accessed 2024-10-26]
119. Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, et al. Big Bird: Transformers for Longer Sequences. 2020 Presented at: 34th Conference on Neural Information Processing Systems; December 6-12, 2020; Online.
120. Wada S, Takeda T, Manabe S, Konishi S, Kamohara J, Matsumura Y. Pre-training technique to localize medical BERT and enhance biomedical BERT. arXiv. URL: https://arxiv.org/abs/2005.07202 [accessed 2024-10-26]
121. Yogarajan V, Montiel J, Smith T, Pfahringer B. Transformers for Multi-label Classification of Medical Text: An Empirical Comparison. In: Tucker A, Henriques Abreu P, Cardoso J, Pereira Rodrigues P, Riaño D, editors. Artificial Intelligence in Medicine. AIME 2021. Lecture Notes in Computer Science, vol 12721. Cham: Springer; 2021:114-123.
122. Rodrawangpai B, Daungjaiboon W. Improving text classification with transformers and layer normalization. Machine Learning with Applications 2022 Dec;10:100403. [doi: 10.1016/j.mlwa.2022.100403]
123. Yalunin A, Nesterov A, Umerenkov D. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. arXiv. 2022. URL: https://arxiv.org/abs/2204.03951 [accessed 2024-10-26]
124. Çelıkten A, Bulut H. Turkish Medical Text Classification Using BERT. 2021 Presented at: 29th Signal Processing and Communications Applications Conference (SIU); June 09-11, 2021; Istanbul, Turkey. [doi: 10.1109/SIU53274.2021.9477847]
125. Blanco A, Perez A, Casillas A. Exploiting ICD hierarchy for classification of EHRs in Spanish through multi-task transformers. IEEE J Biomed Health Inform 2022 Mar;26(3):1374-1383. [doi: 10.1109/jbhi.2021.3112130]
126. Santos T, Tariq A, Das S, Vayalpati K, Smith G, Trivedi H, et al. PathologyBERT - Pre-trained Vs. A New Transformer Language Model for Pathology Domain. arXiv. 2022. URL: https://arxiv.org/abs/2205.06885 [accessed 2024-10-26]
127. Zeberga K, Attique M, Shah B, Ali F, Jembre Y, Chung T. A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. Comput Intell Neurosci 2022;2022:7893775 [FREE Full text] [doi: 10.1155/2022/7893775] [Medline: 35281185]
128. Li J, Huang Q, Ren S, Jiang L, Deng B, Qin Y. A novel medical text classification model with Kalman filter for clinical decision making. Biomedical Signal Processing and Control 2023 Apr;82:104503 [FREE Full text] [doi: 10.1016/j.bspc.2022.104503]
129. Zhang X, Song X, Feng A, Gao Z. Multi-self-attention for aspect category detection and biomedical multilabel text classification with BERT. Mathematical Problems in Engineering 2021 Nov 30;2021:1-6 [FREE Full text] [doi: 10.1155/2021/6658520]
130. Clavie B, Gal K. EduBERT: Pretrained Deep Language Models for Learning Analytics. arXiv. 2019. URL: https://arxiv.org/abs/1912.00690 [accessed 2024-10-26]

131. Al-Garadi M, Yang Y, Cai H, Ruan Y, O'Connor K, Graciela G, et al. Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC Med Inform Decis Mak 2021 Jan 26;21(1):27 [FREE Full text] [doi: 10.1186/s12911-021-01394-0] [Medline: 33499852]
132. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon HJ, et al. Limitations of transformers on clinical text classification. IEEE J Biomed Health Inform 2021 Sep;25(9):3596-3607. [doi: 10.1109/jbhi.2021.3062322]
133. Wang X, Tao M, Wang R, Zhang L. Reduce the medical burden: An automatic medical triage system using text classification BERT based on Transformer structure. 2021 Presented at: 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE); September 24-26, 2021; Zhuhai, China. [doi: 10.1109/ICBASE53849.2021.00133]
134. Naseem U, Razzak I, Khushi M, Eklund PW, Kim J. COVIDSenti: A large-scale benchmark twitter data set for COVID-19 sentiment analysis. IEEE Trans Comput Soc Syst 2021 Aug;8(4):1003-1015. [doi: 10.1109/tcss.2021.3051189]
135. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, et al. Social network analysis of COVID-19 sentiments: Application of artificial intelligence. J Med Internet Res 2020 Aug 18;22(8):e22590 [FREE Full text] [doi: 10.2196/22590] [Medline: 32750001]
136. Chandra R, Krishna A. COVID-19 sentiment analysis via deep learning during the rise of novel cases. PLoS One 2021;16(8):e0255615 [FREE Full text] [doi: 10.1371/journal.pone.0255615] [Medline: 34411112]
137. Yang J, Xiao L, Li K. Modelling clinical experience data as an evidence for patient-oriented decision support. BMC Med Inform Decis Mak 2020 Jul 09;20(Suppl 3):138 [FREE Full text] [doi: 10.1186/s12911-020-1121-4] [Medline: 32646414]
138. Jabreel M, Maaroof N, Valls A, Moreno A. Introducing sentiment analysis of textual reviews in a multi-criteria decision aid system. Applied Sciences 2020 Dec 28;11(1):216. [doi: 10.3390/app11010216]
139. Rajput A. Chapter 3 - Natural Language Processing, Sentiment Analysis, and Clinical Analytics. In: Lytras M, Sarirete A, editors. Innovation in Health Informatics. Cambridge, MA: Academic Press; 2020:79-97.
140. Vajre V, Naylor M, Kamath U, Shehu A. PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. 2021 Presented at: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 9-12, 2021; Houston, TX. [doi: 10.1109/BIBM52615.2021.9669469]
141. Chriqui A, Yahav I. HeBERT and HebEMO: A Hebrew BERT model and a tool for polarity analysis and emotion recognition. INFORMS Journal on Data Science 2022 Apr;1(1):81-95 [FREE Full text] [doi: 10.1287/ijds.2022.0016]
142. Alturayeif N, Luqman H. Fine-grained sentiment analysis of Arabic COVID-19 tweets using BERT-based transformers and dynamically weighted loss function. Applied Sciences 2021 Nov 12;11(22):10694 [FREE Full text] [doi: 10.3390/app112210694]
143. Wang T, Lu K, Chow KP, Zhu Q. COVID-19 sensing: Negative sentiment analysis on social media in China via BERT model. IEEE Access 2020;8:138162-138169. [doi: 10.1109/access.2020.3012595]
144. Müller M, Salathé M, Kummervold PE. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. Front Artif Intell 2023 Mar 14;6:1023281 [FREE Full text] [doi: 10.3389/frai.2023.1023281] [Medline: 36998290]
145. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT 2019. 2019 Presented at: NAACL-HLT 2019; June 2-7, 2019; Minneapolis, Minnesota.
146. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]
147. Zhang K, Zhang C, Ye Y, Zan H, Liu X. Named Entity Recognition in Electronic Medical Records Based on Transfer Learning. In: ICIMH '22: Proceedings of the 2022 International Conference on Intelligent Medicine and Health. 2022 Presented at: 4th International Conference on Intelligent Medicine and Health; August 19-21, 2022; Xiamen, China. [doi: 10.1145/3560071.3560086]
148. Bose P, Srinivasan S, Sleeman W, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Applied Sciences 2021 Sep 08;11(18):8319 [FREE Full text] [doi: 10.3390/app11188319]
149. Thiyagu TM, Manjula D, Shridhar S. Named entity recognition in biomedical domain: A survey. International Journal of Computer Applications 2019 Feb 15;181(41):30-37. [doi: 10.5120/ijca2019918469]
150. Raza S, Reji D, Shajan F, Bashir S. Large-scale application of named entity recognition to biomedicine and epidemiology. PLOS Digit Health 2022 Dec;1(12):e0000152 [FREE Full text] [doi: 10.1371/journal.pdig.0000152] [Medline: 36812589]
151. Syed M, Al-Shukri S, Syed S, Sexton K, Greer M, Zozus M, et al. DeIDNER corpus: Annotation of clinical discharge summary notes for named entity recognition using BRAT tool. Stud Health Technol Inform 2021 May 27;281:432-436 [FREE Full text] [doi: 10.3233/SHTI210195] [Medline: 34042780]
152. Liu N, Hu Q, Xu H, Xu X, Chen M. Med-BERT: A pretraining framework for medical records named entity recognition. IEEE Trans Ind Inf 2022 Aug;18(8):5600-5608. [doi: 10.1109/tii.2021.3131180]
153. Wu Y, Huang J, Xu C, Zheng H, Zhang L, Wan J. Research on named entity recognition of electronic medical records based on RoBERTa and radical‐level feature. Wireless Communications and Mobile Computing 2021 Jun 28;2021(1):1-10. [doi: 10.1155/2021/2489754]

154. Li J, Liu R, Chen C, Zhou S, Shang X, Wang Y. An RG-FLAT-CRF model for named entity recognition of Chinese electronic clinical records. Electronics 2022 Apr 18;11(8):1282. [doi: 10.3390/electronics11081282]
155. Tawfik N, Spruit M. Evaluating sentence representations for biomedical text: Methods and experimental results. J Biomed Inform 2020 May;104:103396 [FREE Full text] [doi: 10.1016/j.jbi.2020.103396] [Medline: 32147441]
156. Brezulianu A, Burlacu A, Popa IV, Arif M, Geman O. "Not by our feeling, but by other's seeing": Sentiment analysis technique in cardiology-An exploratory review. Front Public Health 2022 Apr 11;10:880207 [FREE Full text] [doi: 10.3389/fpubh.2022.880207] [Medline: 35480589]
157. Zhang N, Chen M, Bi Z, Liang X, Li L, Shang X. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022 Presented at: 60th Annual Meeting of the Association for Computational Linguistics; May 22–27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.544]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
EHR: electronic health record
EMR: electronic medical record
GPT: Generative Pre-trained Transformer
NER: named entity recognition
NLP: natural language processing
PLM: pretrained language model

Edited by A Castonguay; submitted 07.06.23; peer-reviewed by D Chrimes, I Gabashvili; comments to author 06.07.23; revised version received 10.07.23; accepted 21.10.24; published 18.11.24.

Please cite as:
Cho HN, Jun TJ, Kim YH, Kang H, Ahn I, Gwon H, Kim Y, Seo J, Choi H, Kim M, Han J, Kee G, Park S, Ko S
Task-Specific Transformer-Based Language Models in Health Care: Scoping Review
JMIR Med Inform 2024;12:e49724
URL: https://medinform.jmir.org/2024/1/e49724
doi:10.2196/49724
PMID:

CHDmap: One Step Further Toward Integrating Medicine-Based Evidence Into Practice

Jef Van den Eynde, MD

Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

Corresponding Author:
Jef Van den Eynde, MD

Related Article:

https://medinform.jmir.org/2024/1/e49138

Abstract

Evidence-based medicine, rooted in randomized controlled trials, offers treatment estimates for the average patient but struggles to guide individualized care. This challenge is amplified in complex conditions like congenital heart disease due to disease variability and limited trial applicability. To address this, medicine-based evidence was proposed to synthesize information for personalized care. A recent article introduced a patient similarity network, CHDmap, which represents a promising technical rendition of the medicine-based evidence concept. Leveraging comprehensive clinical and echocardiographic data, CHDmap creates an interactive patient map representing individuals with similar attributes. Using a k-nearest neighbor algorithm, CHDmap interactively identifies closely resembling patient groups based on specific characteristics. These approximate matches form the foundation for predictive analyses, including outcomes like hospital length of stay and complications. A key finding is the tool's dual capacity: not only did it corroborate clinical intuition in many scenarios, but in specific instances, it prompted a reevaluation of cases, culminating in an enhancement of overall performance across various classification tasks. While an important first step, future versions of CHDmap may aim to expand mapping complexity, increase data granularity, consider long-term outcomes, allow for treatment comparisons, and implement artificial intelligence–driven weighting of various input variables. Successful implementation of CHDmap and similar tools will require training for practitioners, robust data infrastructure, and interdisciplinary collaboration. Patient similarity networks may become valuable in multidisciplinary discussions, complementing clinicians' expertise. The symbiotic approach bridges evidence, experience, and real-life care, enabling iterative learning for future physicians.

(JMIR Med Inform 2024;12:e52343) doi:10.2196/52343

KEYWORDS

artificial intelligence; clinical practice; congenital heart disease; decision-making; evidence-based medicine; machine learning; medicine-based evidence; patient similarity networks; precision medicine; randomized controlled trials

Evidence-based medicine (EBM), built on the foundations of randomized controlled trials (RCTs), is good at providing average estimates for treatments or outcomes in the average patient. While EBM has resulted in important clinical guidelines, it does not solve the real clinical quandaries: patients appear for care individually, each may differ in important ways from an RCT cohort, and the physician will wonder each time if following EBM will provide best guidance for this unique patient. This is particularly the case for complex and heterogeneous populations, such as those with congenital heart disease (CHD). Indeed, in congenital cardiology, RCTs are both difficult to conduct and commonly not definitive. The complexity of disease, clinical heterogeneity within lesions, and the small number of patients with specific forms of CHD severely degrade the precision and value of estimates of average treatment effects in the average patient provided by RCTs.

In response to mounting concern about the value of EBM for decision-making, we have previously proposed medicine-based evidence (MBE) as a means of synthesizing all available information and applying it to the individual patient [1]. Briefly, we proposed that whenever a physician needs to decide a patient's treatment plan, a library of patient profiles would be interrogated. A nearest neighbor algorithm would then find "approximate matches," a group of patients who share the greatest similarity with the index case. Some of these matches would and others would not have received a certain treatment or developed a certain outcome, such that specific analyses tailored to the clinical question could be performed within this pool of approximate matches. We envisioned that this approach would represent a major step toward true personalized medicine, as individualization of treatment would shift from today's intrinsically subjective human-driven assessment toward a more objective, data- and model-driven process that is more descriptive, integrative, and predictive.

In their recent article, Li et al [2] introduced CHDmap, an innovative patient similarity network (PSN) designed to prognosticate outcomes among patients with CHD. By leveraging comprehensive clinical and echocardiographic data sets from 4774 surgical cases, the PSN manifests as an interactive, zoomable electronic cartography, wherein each node symbolizes an individual patient, and internode distances delineate their similarity. This user-centric software empowers practitioners to delineate specific patient attributes—such as age, gender, CHD classification, and echocardiographic metrics—tailoring the analysis to the case at hand. The program subsequently uses a k-nearest neighbor algorithm to identify a cohort of closely resembling peers according to the top-k parameter or similarity threshold. This assemblage of approximate matches serves as the foundation for diverse predictive analyses, encompassing variables like hospital length of stay, complications, and survival. This way, CHDmap allows for conducting real-time clinical trials that are specifically tailored to the individual patient, based on historical cases with a similar clinical profile. A key finding from the study by Li et al [2] was the tool's dual capacity: not only did it corroborate clinical intuition in many scenarios, but in specific instances, it prompted a reevaluation of cases, culminating in an enhancement of overall performance across various classification tasks.

The tool has been made publicly available [3] and represents a promising technical rendition of the MBE concept. According to the authors, future generations of the software will be uploaded in time, further expanding the possibilities of CHDmap, including the following: (1) Labeling and visualization of increasingly complex and rare CHD types—currently, only some major subtypes (atrial septal defect, patent foramen ovale, ventricular septal defect, patent ductus arteriosus) are depicted in the map overview; as the underlying data set expands, patients with more complex anatomy may be visualized as well. (2) More granularity in data—in a similar manner, the width of the underlying data set (ie, number of cases) and its depth (ie, number of variables) will likely increase, allowing for more precise matching and examination of more aspects of decision-making. (3) Long-term outcomes—currently, only in-hospital outcomes can be considered within CHDmap, but future generations of the software may allow for long-term outcomes to be analyzed. (4) Comparisons of specific treatment options—once in-depth data on various treatments become available, the optimal treatment for an individual patient may be examined through real-time clinical trials within CHDmap, where outcomes after initiation of various treatments are compared among a group of approximate matches. (5) Artificial intelligence–driven weighting of indicators—the default setting in CHDmap allocates to each indicator the same weighting, whereas physicians can modify these weights based on their prior knowledge; the latter option allows accounting for the fact that weights are likely to differ depending on the clinical setting and the question at hand. With future generations of CHDmap, the authors may implement an artificial intelligence model to dynamically allocate weights to each of the indicators.

CHDmap undeniably signifies a significant stride toward the actualization of MBE. Just as with any statistical methodology, the principles of implementation science will play a pivotal role in optimizing the widespread integration of this tool into clinical practice [4]. Medical practitioners will need to be trained to use these tools correctly and to ensure they are aware of the perks and pitfalls of the PSN (eg, knowing that there is a trade-off between increasing similarity and increasing statistical power or being able to correctly interpret the certainty associated with a specific prediction). Data infrastructure will need to be in place, and continued efforts should be made to establish multicenter clinical registries with in-depth and up-to-date information collection. Furthermore, collaboration between health care professionals and experts in data science will be required to ensure these novel technologies can benefit our patients, taking into account issues regarding data quality and privacy.

Finally, at some point in the future, tools like CHDmap may become routinely used to support team discussions. Rather than replacing the clinician, they should be embraced as assistive technology enhancing overall clinical efficacy. This symbiotic approach serves to harmonize real-life patient care with prior experience and established evidence. This way, we can truly start to achieve the incremental benefits of future generations of physicians learning from previous ones.

Conflicts of Interest

None declared.

References

1. Van den Eynde J, Manlhiot C, Van De Bruaene A, et al. Medicine-based evidence in congenital heart disease: how artificial intelligence can guide treatment decisions for individual patients. Front Cardiovasc Med 2021 Dec;8:798215. [doi: 10.3389/fcvm.2021.798215] [Medline: 34926630]
2. Li H, Zhou M, Sun Y, et al. A patient similarity network (CHDmap) to predict outcomes after congenital heart surgery: development and validation study. JMIR Med Inform 2024 Jan 19;12:e49138. [doi: 10.2196/49138] [Medline: 38297829]
3. CHDmap online interactive tool. Clinical Genetic Test Report Platform. URL: http://chdmap.nbscn.org/ [accessed 2024-04-02]
4. Manlhiot C, Van den Eynde J, Kutty S, Ross HJ. A primer on the present state and future prospects for machine learning and artificial intelligence applications in cardiology. Can J Cardiol 2022 Feb;38(2):169-184. [doi: 10.1016/j.cjca.2021.11.009] [Medline: 34838700]

Abbreviations

CHD: congenital heart disease
EBM: evidence-based medicine
MBE: medicine-based evidence
PSN: patient similarity network
RCT: randomized controlled trial

Edited by C Lovis, S Gardezi; submitted 31.08.23; peer-reviewed by H Li; accepted 10.03.24; published 19.04.24.

Please cite as:
Van den Eynde J
CHDmap: One Step Further Toward Integrating Medicine-Based Evidence Into Practice
JMIR Med Inform 2024;12:e52343
URL: https://medinform.jmir.org/2024/1/e52343
doi:10.2196/52343

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org